REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
03-04-2022	Final Report	14-Sep-2012 - 13-Mar-2017

4. TITLE AND SUBTITLE
Final Report: Will the revolution be tweeted? Weak Ties, Social Media, and Flash Movements

5a. CONTRACT NUMBER
W911NF-12-1-0556

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Northeastern University
360 Huntington Avenue
490 RP
Boston, MA 02115 -5005

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
62550-NS.16

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			David Lazer
UU	UU	UU	UU		19b. TELEPHONE NUMBER 617-877-5873

RPPR Final Report
as of 11-Jan-2023

Agency Code: 21XD

Proposal Number: 62550NS **Agreement Number: W911NF-12-1-0556**
INVESTIGATOR(S):

Name: Alan E. Mislove
Email: A.Mislove@northeastern.edu
Phone Number: 6173737069
Principal: N

Name: David Lazer
Email: d.lazer@northeastern.edu
Phone Number: 6178775873
Principal: Y

Organization: **Northeastern University**
Address: 360 Huntington Avenue, Boston, MA 021155005
Country: USA
DUNS Number: 001423631 EIN: 104167998

Report Date: 13-Jun-2017 Date Received: 03-Apr-2022
Final Report for Period Beginning 14-Sep-2012 and Ending 13-Mar-2017
Title: Will the revolution be tweeted? Weak Ties, Social Media, and Flash Movements
Begin Performance Period: 14-Sep-2012 **End Performance Period:** 13-Mar-2017
Report Term: 0-Other

Submitted By: David Lazer Email: d.lazer@northeastern.edu
Phone: (617) 877-5873

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 4 **STEM Participants:**

Major Goals: The major goal of the project focused on understanding how to use social media (and, especially, Twitter) to understand mass-- especially political and social information.

Accomplishments: This project provided the support for a series of papers examining online collective behavior. This grant has yielded an impressive corpus of computational social science papers that have already been cited many thousands of times. For example: We identified flaws in "Google flu trends" in a paper in Science, which has become required reading on the issues in the use of "big data." We developed a method to link administrative data to twitter handles to produce valid estimates of the prevalence of "fake news" on Twitter. (These methods and the resulting data have resulted in an additional stream of papers post grant.) We developed methods to study the dynamics of collective human attention, and to identify from contextual traces linkages among political actors and to detect political "spin" in textual content.

Training Opportunities: This project provided for the professional development of a number of junior scholars, including:

Oren Tsur (postdoc)
Lisa Friedland (postdoc)
Stefan Wojcik (postdoc)
Jason Radford (postdoc)
Aniko Hannak (student)
Yabing Liu (student)
Waiofeng Yang (student)
Arash Molavi Kakhi (student)

Results Dissemination: This grant supported the publication of 6 articles (including two in Science), 6 terminal conference publications, and 1 Annual Review chapter.

RPPR Final Report
as of 11-Jan-2023

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Jason Radford
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Lisa Friedland
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Stefan Wojcik
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: PD/PI
Participant: David Lazer
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co PD/PI
Participant: Alan Mislove
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Aniko Hannak
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Yabing Liu

RPPR Final Report
as of 11-Jan-2023

Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Arash Molavi Kakhki
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Xiaofeng Yao
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: PLoS ONE
Publication Identifier Type: Publication Identifier:
Volume: 0 Issue: 5 First Page #: 0
Date Submitted: Date Published:
Publication Location:
Article Title: Rising Tides or Rising Stars?: Dynamics of Shared Attention on Twitter During Media Events
Authors:
Keywords: social media
Abstract: “Media events” generate conditions of shared attention as many users simultaneously tune in with the dual screens of broadcast and social media to view and participate. We examine how collective patterns of user behavior under conditions of shared attention are distinct from other “bursts” of activity like breaking news events. Using 290 million tweets from a panel of 193,532 politically active Twitter users, we compare features of their behavior during eight major events during the 2012 U.S. presidential election to examine how patterns of social media use change during these media events compared to “typical” time and whether these changes are attributable to shifts in the behavior of the population as a whole or shifts from particular segments such as elites. Compared to baseline time periods, our findings reveal that media events not only generate large volumes of tweets, but they are also associated with (1) substantial declines in interpersonal communication, (2) more highly conc
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support:

RPPR Final Report
as of 11-Jan-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: PLoS ONE
Publication Identifier Type: Publication Identifier:
Volume: 1.0E+001 Issue: 4.0E+000 First Page #: 0
Date Submitted: Date Published:
Publication Location:
Article Title: Timescales of Massive Human Entrainment
Authors:
Keywords: human dynamics, social media
Abstract: The past two decades have seen an upsurge of interest in the collective behaviors of complex systems composed of many agents entrained to each other and to external events. In this paper, we extend concepts of entrainment to the dynamics of human collective attention. We conducted a detailed investigation of the unfolding of human entrainment—as expressed by the content and patterns of hundreds of thousands of messages on Twitter—during the 2012 US presidential debates. By time locking these data sources, we quantify the impact of the unfolding debate on human attention. We show that collective social behavior covaries second-by-second to the interactional dynamics of the debates: A candidate speaking induces rapid increases in mentions of his name on social media and decreases in mentions of the other candidate. Moreover, interruptions by an interlocutor increase the attention received. We also highlight a distinct time scale for the impact of salient moments in the debate: Mentions i
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support:

Publication Type: Journal Article Peer Reviewed: N **Publication Status:** 1-Published
Journal: Journal of the Association for Information Science and Technology
Publication Identifier Type: Publication Identifier:
Volume: 0 Issue: 0 First Page #: 0
Date Submitted: Date Published:
Publication Location:
Article Title: Uncovering social semantics from textual traces: A theory?driven approach and evidence from public statements of US Members of Congress
Authors:
Keywords: social networks, text analysis
Abstract: The increasing abundance of digital textual archives provides an opportunity for understanding human social systems.Yet the literature has not adequately considered the disparate social processes by which texts are produced. Drawing on communication theory, we identify three common processes by which documents might be detectably similar in their textual features—authors sharing subject matter, sharing goals, and sharing sources. We hypothesize that these processes produce distinct, detectable relationships between authors in different kinds of textual overlap. We develop a novel n-gram extraction technique to capture such signatures based on n-grams of different lengths. We test the hypothesis on a corpus where the author attributes are observable: the public statements of the members of the U.S. Congress. This article presents the first empirical finding that shows different social relationships are detectable through the structure of overlapping textual features. Our study has impor
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support:

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Proceedings of the second ACM conference on Online social networks. ACM
Publication Identifier Type: Publication Identifier:
Volume: 0 Issue: 0 First Page #: 0
Date Submitted: 11/14/16 12:00AM Date Published:
Publication Location:
Article Title: Beyond CPM and CPC: Determining the Value of Users on OSNs
Authors: Diego Saez-Trumper, Yabing Liu, Ricardo Baeza-Yates, Balachander Krishnamurthy, Alan Mislove
Keywords: social media, social networks
Abstract: Not all of the over one billion users of online social networks (OSNs) are equally valuable to the OSNs. The current business model of monetizing advertisements targeted to users does not appear to be based on any visible grouping of the users. The primary metrics remain CPM (cost per mille— i.e., thousand impressions) and CPC (cost per click) of ads that are shown to users. However, there is significant diversity in the actions of users—some users upload interesting content triggering additional views and comments leading to further cascades of action. Beyond direct impressions, a user's action can generate indirect impressions by actions induced on friends and other users. Identifying the valuable user segments requires examination of profile data, friendships, and most importantly, their activity. Here we explore an alternate approach for measuring the value of users in OSNs by proposing a framework from the viewpoint of a popular OSN. Using a real dataset on the social network and
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics
Publication Identifier Type: Publication Identifier:
Volume: 0 Issue: 0 First Page #: 0
Date Submitted: 11/13/16 12:00AM Date Published:
Publication Location:
Article Title: Automatic Discovery of Partisan campaigns, AgendaSetting and Political Spin in Press releases
Authors: Oren Tsur, Dan Calacci and David Lazer
Keywords: text analysis, agenda setting
Abstract: Press releases (PR) serve as an important tool by which political figures (and business) communicate their messages to news editors and journalists, which in turn deliver it to their audience. Indeed, much of the media coverage of political events is based on, responds to or quotes press releases submitted by politicians. With the increasing number of PR communications it is important to provide journalists with tools for easy processing of press releases at large scale. In this paper we present a system for automatic discovery of agenda setting efforts, framing strategies and political spin as evident in a large corpus of press releases. The system combines topic models, sentiment analysis and autoregressive-distributed-lag models. We automatically analyze over 130000 PR communications released by members of the House and the Senate in the years 2010-2013 and find significant differences in topic ownership and sentiment as well as significant evidence for coordinated campaigns for age
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: ICWSM
Date Received: 14-Nov-2016 Conference Date: 27-May-2014 Date Published:
Conference Location: Oxford, UK
Paper Title: The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior
Authors: Yabing Liu, Chloe Kliman-Silver, Alan Mislove
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 11-Jan-2023

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: ICWSM
Date Received: 14-Nov-2016 Conference Date: 27-May-2014 Date Published: 27-May-2014
Conference Location: Oxford, UK
Paper Title: Of Pins and Tweets: Investigating how users behave across image- and text-based social networks
Authors: Raphael Ottoni, Diego B. Las Casas, Joao Paulo Pesce, Wagner Meira Jr, Christo Wilson, Alan Mislove
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: ICWSM
Date Received: 14-Nov-2016 Conference Date: 01-Jun-2015 Date Published: 01-Jun-2016
Conference Location: Ann Arbor Michigan
Paper Title: Get back! You don't know me like that: The social mediation of fact checking interventions in Twitter conversations
Authors: Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 14th ACM Internet Measurement Conference (IMC'14)
Date Received: 14-Nov-2016 Conference Date: 05-Nov-2015 Date Published: 05-Nov-2015
Conference Location: Vancouver, Canada
Paper Title: Measuring Price Discrimination and Steering on E-commerce Web Sites
Authors: Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, Christo Wilson
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguisticsand the 7th International Joint Conference on Natural Language Processing
Date Received: 14-Nov-2016 Conference Date: 26-Jul-2015 Date Published:
Conference Location: Beijing, China
Paper Title: A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns.
Authors: Oren Tsur, Dan Calacci, and David Lazer
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Ninth International AAAI Conference on Web and Social Media
Date Received: 14-Nov-2016 Conference Date: 26-May-2015 Date Published:
Conference Location: Oxford, UK
Paper Title: Don’t Let Me Be #Misunderstood: Linguistically MotivatedAlgorithm for Predicting the Popularity of Textual Memes
Authors: Oren Tsur, Ari Rappoport.
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 11-Jan-2023

Partners

,

I certify that the information in the report is complete and accurate:
Signature: David Lazer
Signature Date: 4/3/22 5:16PM

ANNUAL REVIEWS

Annual Review of Sociology

Data ex Machina: Introduction to Big Data

David Lazer[1,2] and Jason Radford[1,3]

[1]Department of Political Science and College of Computer and Information Science, Northeastern University, Boston, Massachusetts 02115; email: d.lazer@neu.edu, j.radford@neu.edu

[2]Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts 02138

[3]Department of Sociology, University of Chicago, Chicago, Illinois 60637

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Annu. Rev. Sociol. 2017. 43:19–39

First published as a Review in Advance on May 17, 2017

The *Annual Review of Sociology* is online at soc.annualreviews.org

https://doi.org/10.1146/annurev-soc-060116-053457

Copyright © 2017 by Annual Reviews.
All rights reserved

ANNUAL REVIEWS **Further**

Click here to view this article's online features:

- Download figures as PPT slides
- Navigate linked references
- Download citations
- Explore related articles
- Search keywords

Keywords

big data, computational social science, technology, research methodology, quantitative sociology, research ethics, research design, social media, CDR data, data linkage, networks, mobility

Abstract

Social life increasingly occurs in digital environments and continues to be mediated by digital systems. Big data represents the data being generated by the digitization of social life, which we break down into three domains: digital life, digital traces, and digitalized life. We argue that there is enormous potential in using big data to study a variety of phenomena that remain difficult to observe. However, there are some recurring vulnerabilities that should be addressed. We also outline the role institutions must play in clarifying the ethical rules of the road. Finally, we conclude by pointing to a number of nascent but important trends in the use of big data.

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

INTRODUCTION

Archives of human activity go back millennia; however, the increasingly comprehensive digital archives of human behavior, combined with the exponential growth of computational power, create the potential for a transformation of fields such as sociology. Core constructs of sociology, such as interaction, collective action, expression, and diffusion of behavior lurk in these archives. It is possible to study the connectivity of entire human societies, including who communicates with whom about what, how people move through space, who says what, and who buys what, all with a temporal granularity of seconds. The coming generation will witness a transformation of sociological theory through these improvements in our ability to observe dynamic social systems (boyd & Crawford 2012, Golder & Macy 2014).

Big data does pose distinctive challenges for scholars. These digital archives are not the product of scientific design. The information captured in these archives is not what a social scientist would choose. Further, what is captured is constantly, and sometimes abruptly, changing. The signal in big data is vulnerable to manipulation, sometimes purposive, sometimes incidental. Further, relevant behaviors are split into many archives, with no practical way of conjoining them. For example, there may not be a strong conceptual or empirical distinction or boundary to be drawn between behaviors captured by different cell phone carriers, but most research on cell phone data is based on data from a single carrier.

Big data also presents enormous institutional challenges to sociology. The large majority of sociologically relevant analysis of big data is done by computer scientists, and there is relatively little reflection of the big data revolution in top sociology journals. Only 6 of 182 articles published between 2012 and 2016 in the *American Journal of Sociology* (*AJS*) and 9 of 240 in *American Sociological Review* (*ASR*) involve the use of big data.[1]

The objective of this review, therefore, is to critically assess the potential of big data for the collective intellectual endeavor that is sociology. In what follows, we provide a brief overview of what big data is and then review recent big data projects. We break these projects down into three types of big data, enumerate the promises and pitfalls common across them, and offer guidance to sociology for moving forward. We conclude by drawing together the challenges that remain across all three areas and outline the work that needs to be done to put the field on more solid ethical, methodological, and epistemic ground. Our hope is to demonstrate the value of big data research while articulating its most pressing problems and offering reasonable strategies for advancing the field.

WHAT IS BIG DATA?

The term "big data" has proliferated across society and owes its recent popularization to a report by McKinsey (Manyika et al. 2011). The issue of scale of data is relevant across the academy, from astronomy to the humanities. The manifestation of big data, of course, varies. In astronomy, big data takes the form of images that take petabytes of storage; in the humanities, it can take the form of millions of digitized books.

The dominant data paradigm for the quantitative social sciences has been tabular—variables in columns, cases in rows. Big sociological data, in contrast, comes in an infinite array of other

[1]For our review, we looked for any studies using big data sources or existing data with big data methods such as automated text analysis. We especially looked for social media data, data from online platforms like crowdfunding and online dating websites, data from smartphones including CDR data, and large scale administrative data sets. At *AJS*, these works include Burt (2012), Knigge et al. (2014a), Legewie (2016), Legewie & Schaeffer (2016), Lin & Lundquist (2013), and de Vaan et al. (2015). For *ASR*, the relevant literature includes Bail (2012), Curington et al. (2015), Diekmann et al. (2014), Goldberg et al. (2016), Hall et al. (2015), Knigge et al. (2014b), Leung (2014), van de Rijt et al. (2013), and Vasi et al. (2015).

forms, such as pictures, video, tweets, text, space, networks, streaming, etc. In early formulations, its "bigness" meant that the data could not be processed with extant software (Gartner 2011, Manovich 2012). Big data was big to the extent that new technologies had to be created by specialists to collect, store, and analyze it. The term big data thus refers to data that are so large (volume), complex (variety), and/or variable (velocity) that the tools required to understand them must first be invented (Laney 2001, Monroe 2013). Big data thus requires a computational social science—a mash-up of computer science for inventing new tools to deal with complex data, and social science, because the substantive substrate of the data is the collective behavior of humans (Lazer et al. 2009).

The big data challenges of today have analogues in history. Two generations ago, running multivariate linear models required punch cards and a computer the size of a basement. The challenge was to build tools that leveraged ever-increasing processing power. Today, the technical and theoretical challenges confronting big data research are different, stemming from the complex and heterogeneous form of contemporary big data. These challenges, too, can be met. Sociologists must learn how to cull through large amounts of unstructured text data, program mobile phones, and build data pipelines that scrape, process, store, and make available large amounts of data of wide varieties and types.

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Big Data Sources

Big data is vast and heterogeneous, encompassing everything from YouTube to digital archives of books. A comprehensive review of all big data sources is beyond the scope of this review. There are many discrete literatures around different big data sources, and even a complete list of those literatures would soon be obsolete. Instead, we offer a fuzzy typology of big data sources, based on the loci of data collection. We begin with a discussion of digital life—the capturing of digitally mediated social behaviors—that likely accounts for the majority of big data research. We then discuss digital trace data, the archival exhaust of the modern bureaucratic organization, and conclude with digitalized life data, the movement of intrinsically analog behavior into digital form.

Digital life. An increasing fraction of life is intrinsically digitally mediated. Twitter, Facebook, and Wikipedia are all platforms where behaviors are all online. These behaviors may relate to offline occurrences, but behaviors like tweeting are inherently digital. Behaviors on such platforms are typically substantially captured by platform owners, because their business models rely on inferences (such as ad targeting) that can be made from these data streams. Further, it is possible for third parties to harvest data from these platforms. Facebook allows users to download portions of their data; the entire edit history of Wikipedia can be downloaded for analysis; Google allows some access to search query volume. Twitter data are the most used by scholars because of their accessibility. Third parties can negotiate and pay for access for large samples of Twitter data or harvest sizable targeted subsamples.

Digital platform data may be viewed in two ways. The first is to view these platforms as generalizable microcosms of society (Tufekci 2014). Kossinets & Watts (2006) examine email to study the role of different social foci in emergent network structure. Barberá et al. (2015) use Twitter to study political mobilization. The dissemination of news and rumors has been studied on Wikipedia (Keegan et al. 2013, Keegan & Brubaker 2015), Twitter (Bakshy et al. 2011, Romero et al. 2011, Yang & Counts 2010), and across platforms (Goel et al. 2012, Kim et al. 2014). Online markets such as Airbnb and Kickstarter have been used to study patterns of social inequality (Edelman & Luca 2014, Greenberg & Mollick 2017). The Billion Prices Project uses price data scraped from online retailers to track inflation (Cavallo & Rigobon 2016). In this vein too is

Google Flu Trends, which sought to track cases of flu using Google search data (Ginsberg et al. 2009).

The second way to view these platforms is as distinctive realms in which much of the human experience now resides. Exemplary in this regard is the research on whether Facebook creates or accentuates an informational filter around individuals such that people only see ideologically compatible content (Bakshy et al. 2015; also see Lazer 2015). The objective of this article is not to resolve the filter bubble question more generally—the study's finding of only moderate filtering effects tells us nothing about Google or Twitter, for example. And its entire scientific relevance rests on the (correct) assumption that Facebook is, by itself, general enough to be worthy of study.

We note that not all behavior on a platform may be captured. Indeed, the entire Internet may be viewed as an enormous platform for human expression and, although there are efforts to capture records of the Internet, most notably by the Internet Archive, generally these capture only snapshots of particular moments for particular URLs.

Further, as we discuss below, the mapping between what happens on these platforms and the phenomena of interest to a sociologist may be somewhat weak. All friends are not Facebook friends, and not all Facebook friends are friends.

Digital traces. The modern complex organization creates a steady output of records that chronicle actions taken (sometimes labeled metadata). Call detail records (CDRs) from phone calls are illustrative (Onnela et al. 2007, Toole et al. 2015). CDRs from cell phones, for example, typically offer time stamps and duration of calls, identifiers for initiator and recipient of calls, and identifiers for the cell towers accessed during the call. Such information has been used to critically examine the strength of weak ties (Onnela et al. 2007), predict individual and collective level unemployment (Toole et al. 2015), and model the spread of malaria (Wesolowski et al. 2012). Governmental data, such as voter records, political contribution data (Bonica 2014), and tax data (Chetty et al. 2016) are other examples. What distinguishes trace data from digital life is that the trace is only a record of the action, not the action itself.

Digitalized life. Lastly, digitalized life represents the capture of nonintrinsically digital life (i.e., most of life) in digital form. Thus, for example, phones can be programmed to continuously identify nearby Bluetooth devices, thus capturing the proximity of individuals (Eagle et al. 2009). The constant video recording of major modern cities creates ongoing records of human interaction. Informational objects that predate computers can be easily scanned into manipulable digital form. Exemplary in this regard are Google Books and related endeavors (Michel et al. 2011), which involved scanning millions of books, as well as the use of newspaper data to study the dynamics of fame (van de Rijt et al. 2013).

Instrumentation of human behavior. Crosscutting these three types of big data is the possibility for the proactive and purposive instrumentation of human behavior. It is possible to identify and collect data on subsets of behaviors: to monitor a stream of certain types of tweets, track web browsing behaviors from particular locations, or scrape select data from select websites, such as with the Billion Prices Project (Cavallo & Rigobon 2016). Smartphones can be instrumented to collect a wide array of ambient data (e.g., Eagle et al. 2009). Specially designed hardware, so-called sociometers, can track minute details regarding face-to-face social interactions, and have been used, for example, to study the role of gender in collaboration (Onnela et al. 2014).

These examples highlight the variety of uses for these new data. These projects rely heavily on a new breed of tools and techniques, largely from computer science.

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Summary

Big data represents both the new kinds of digital data available to sociologists as well as the tools and technologies required to access these data. The explosion of applications of big data to long-standing social questions reveals a variety of important opportunities for sociologists to extend our knowledge. In the following, we outline these opportunities and provide suggestions for taking advantage of them. These applications, too, both in their successes and their failures, reveal new challenges for big data research, which must be addressed if we are to advance in a responsible, equitable, and scientifically sound way.

OPPORTUNITIES

Our review of the literature reveals a set of distinct opportunities present in a world replete with big data. Much of these data are what is generally called massive, passive data: data generated in the process of meaningful social behavior rather than data reported for research. The prevalence of such systems and digital devices means that whole systems are captured by these data. And because these systems are always running, they act as controls for experiments offline, online, and even in the lab. One final opportunity in big data comes from making big data small by finding the special populations within it (Foucault Welles 2014).

Massive, Passive: Behavioral Data at Scale

In principle, big data archives offer measures of actual behaviors, as compared with self-reports of behaviors. The literature is rife with evidence of the problems with self-reported behavior. Generally, self-reported behavior is noisy, with a variety of systemic biases. For example, people systematically lie about everything from whether they voted, to what their weight and height are. Certain types of behavior are entirely inaccessible via self-reports.

To focus on one example, social network researchers have long struggled to accurately measure social ties (Bernard et al. 1984, Marsden 1990). Respondents are biased in the ties they remember and how they remember them. Respondents provide very different networks when asked who they go to for advice, who they would ask for a loan, who their friends are, and who they spend the most time with. And the same question may be interpreted differently by different people. These differences present a range of problems for survey-based social network research, including ambiguity in defining what a tie means, difficulty interpreting concepts such as social influence, and uncertainty about omitted ties.

Behavioral data enable us to observe social networks through interactions. Eagle et al. (2009) compare self-report data to behavioral data. They handed out 94 smartphones to students for nine months. The phones were programmed with applications that "recorded and sent the researcher data about call logs, Bluetooth devices in proximity of approximately five meters, cell tower IDs, application usage, and phone status" (2009, p. 15274). The researchers constructed a behavioral measure of participants' social networks using the call log data and spatial proximity captured via Bluetooth. Participants were behaviorally related if they talked to one another or were in physical proximity. In the middle of the study, participants filled out a questionnaire about their physical and social proximity to one another. The survey constituted the self-report measure of students' networks.

Comparing reported physical proximity to behavioral measures of proximity, Eagle et al. find recency and salience biases in subjects' recall. People tended to overreport physical proximity in general. Additionally, friends were much more accurate in recalling their physical proximity to

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

one another than those who did not consider one another friends. The researchers also find that subjects who reported being friends behaved in stable and substantially different ways from those who reported not being friends. Friends were much more likely to be spatially proximate to one another at night or on the weekends even if they were at work or somewhere else. Moreover, self-reported friends in January exemplified this pattern five months later in May.

This study helps validate passive network inference by showing that behavioral measures can capture self-reported friendship. At the same time, it reveals a substantial shortcoming of self-reported measures in capturing our weak ties—the colleagues and acquaintances who we see on a regular basis but who we do not consider close friends.

The Copenhagen Network Study (Stopczynski et al. 2014a,b) extends behavioral research on social networks by comparing different modes of behavioral inference. Researchers handed out 1,000 phones to students entering the Technical University of Denmark in 2012 and 2013. Researchers used the phones to infer Bluetooth proximity, geographical proximity via GPS, and interaction via calls and text messages. They combined this with students' Facebook data, their proximity to routers, and qualitative field observations from an anthropologist. They find that nearly the entire call network is captured by Bluetooth proximity, and 80% of students' Facebook friends could be captured by Bluetooth proximity. Yet only 20% of students' Facebook friend network was captured by call logs. Stopczynski et al. conclude that passive behavioral measures are not interchangeable and different digital systems can lead to different social networks with different properties. That is, there is not a single social network for anyone, but a series of shifting networks based on the organizations and technologies individuals use to form and sustain relationships. Thus, the choice of which systems to use to collect data from subjects and how to integrate these different data sources will likely affect study results.

Nowcasting

Monitoring social phenomena is an essential part of social science research. Surveys such as the University of Michigan's Survey of Consumers or the Bureau of Labor Statistics' Current Employment Statistics survey are fielded regularly to monitor the economic health of the United States. The Centers for Disease Control and Prevention and World Health Organization use a network of testing labs, healthcare providers, and government agencies to generate regular estimates of flu prevalence and virulence.

These regularly updated, vital statistics monitor phenomena that are essential to the workings of contemporary institutions. However, these surveillance systems are very expensive to operate, time consuming to deploy, and inaccurate at high levels of temporal and geographic granularity. Through the digitization of social life, these phenomena are increasingly becoming visible in big data. Such digitization offers the potential to reduce the costs, improve the accuracy, and increase the scale of societal monitoring.

Researchers have used existing large-scale digital data such as CDR data and scraped web data as sensors for monitoring social phenomena. CDR data have been used to detect unemployment (Toole et al. 2015). Google Flu Trends sought to track cases of flu using Google search data (Ginsberg et al. 2009; although see Lazer et al. 2014). Finally, Beauchamp (2016) and Hopkins & King (2010) use Twitter data to generate estimates of public opinion. Each of these studies represents an attempt to validate the use of big data to generate estimates of socially relevant phenomena. This area of research, called nowcasting, seeks to generate descriptions of the world as they happen.

The Billion Prices Project is one such ambitious project. Using price data scraped from online retailers, Cavallo, Rigobon, and colleagues produce daily estimates of prices that are essential to

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

everything from exchange rates to inflation to the real value of wages (see Cavallo & Rigobon 2016 for an overview). Much like traditional price indexing, researchers at the Billion Prices Project scrape data for a preselected "bag" of goods from a curated set of online retailers from around the world. They compare the prices among the same goods within retailers within countries every day to compute a daily consumer price index for over 70 countries.

This approach has unearthed several new empirical insights. The standard data used to study price changes showed that price changes were normally distributed, with most changes being small. However, Cavallo (2017) finds that the distribution of price changes is actually bimodal, with large price increases and decreases, and very few small price changes. Cavallo argues that the price imputations made during the construction of standard data produce small but erroneous price changes. In essence, measurement error led to a decade's worth of poor theory on how prices change.

A second result of the Billion Prices Project has been to document violations of the law of one price (Cavallo et al. 2014). This law asserts that a good should generally have the same price no matter where it is sold, controlling for differences in the value of currency. Cavallo et al. show that the law of one price only holds among countries that share a currency.

Finally, Cavallo (2013) demonstrates the social impact of nowcasting. For example, the government of Argentina began manipulating its estimates of inflation in 2007. Cavallo attempts to identify the extent of manipulation by estimating the divergence between Argentina's official estimates and those derived from online prices. As a control, Cavallo also compares Consumer Price Indexes computed via online prices to those published by the governments of Brazil, Venezuela, Chile, and Colombia. The results show that although indexes computed using online prices track prices in the control countries, they diverge substantially in the case of Argentina. In 2015, Argentina stopped publishing inflation numbers, and the Billion Prices Project has been used to infer inflation in their place.

The Billion Prices Project demonstrates some of the implications of using big data to estimate social phenomena. First, it acts as an alternative method of estimating social phenomena, illuminating problems in traditional methods. Second, it can act as a measure of something that is otherwise unmeasured or whose measure may be disputed. Finally, nowcasting demonstrates big data's potential to generate measures of more phenomena, with better geographic and temporal granularity, much more cost-effectively than traditional methods. It is worth noting that this granularity is most useful when fused with traditional methods rather than used as a replacement for them (Lazer et al. 2014).

Data on Social Systems

Perhaps most exciting about big data is the opportunity to build a science of society, a science that would study society at scale, composed of subsystems and individuals that are dynamically connected in particular ways and locations. For example, State et al. (2015) examine networks based on interactions among millions of people on Twitter and between Yahoo! email users. Both data sets revealed strong intracultural correlations among individuals across the globe.

Analysis of the aggregate activity of Twitter reveals expected and unexpected patterns in global behavior. For example, there are strong diurnal patterns across the globe (Golder & Macy 2011). People are generally happier in the morning and earlier in the week, but these patterns vary substantially with work and season and differ systematically for so-called night owls. Dodds et al. (2011) show that big data can capture annual emotional cycles, particularly around holidays, as well as divergences during global events such as the financial crisis.

Data on systems have been used to answer long-standing questions about human mobility. Online data and data from mobile phones have been used to characterize human mobility (Brockmann

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

et al. 2006, González et al. 2008). This mobility information has been linked to geographical, cultural, and political information in order to study patterns of interaction at the national and regional levels (Blanford et al. 2015, Sevtsuk & Ratti 2010).

One such study is that of Toomet et al. (2015), who use cell phone data to examine the patterns of interaction among members of the Estonian majority and Russian-speaking minority members in the Estonian capital of Tallinn. They used CDR data from Estonia's largest mobile provider, Telia Eesti (formerly EMT), to capture individuals' locations to within several hundred meters. When individuals made calls or sent texts in a shared location at a particular time, Toomet et al. counted that as copresence. To distinguish Estonian and Russian-speaking users, they used the preferred language settings that cell phone providers link to a phone's SIM (subscriber identity module) card. Finally, Toomet and colleagues inferred individuals' home and work locations based on their most frequent geographic location during work and evening hours.

Toomet et al. analyze the ethnic dissimilarity of copresent mobile users at work, home, and during free time. They find high similarity (i.e., substantial segregation) among individuals at home and at work. However, they find substantial dissimilarity among individuals during free time periods. This mixing during free time was most prominent in the city's central district, but also existed even in the less integrated suburbs. The findings suggest that although major social institutions engender segregation, people desegregate in less structured social life. The study raises questions for the long history of research on segregation and social inequality, which almost uniformly define segregation as residential segregation (Massey & Denton 1993, Wilson 1987). As Small (2004, ch. 5) suggests, segregation is as much a process of where you live and work as it is the cultural and physical boundaries that shape who can and cannot participate in the life of a neighborhood. This study shows that high-resolution social mobility data offer a more nuanced picture of segregation bridging residence, employment, and community.

Natural and Field Experiments

Digitally mediated social systems such as Facebook and Twitter and newly digital institutions like the Internal Revenue Service (IRS) capture data irrespective of events in the environment. In addition, the administrators of these systems make innumerable changes to them over time. As such, it is plausible that all kinds of natural experiments may be hidden within large-scale data. Big data offers a milieu for studying the effect of external events on ongoing social processes. For example, Phan & Airoldi (2015) use Facebook data to examine how social networks were affected by Hurricane Ike in 2008. Ayers et al. (2011) use search traffic data to determine whether tax increases on cigarettes lead to an increased interest in tobacco cessation.

Big data can capture the effects of experiments in the field through data linkage. Exemplary in this regard is a series of studies by Chetty and collaborators who merge IRS data with data from prior research (Chetty et al. 2014a,b, 2016). Chetty et al. (2016) linked participants in the Moving to Opportunity (MTO) field experiment to their tax filings with the IRS decades later. MTO sought to investigate neighborhood effects by examining the impact of receiving vouchers and moving to wealthier and safer neighborhoods on the economic, social, and psychological well-being of low-income families. By linking IRS data to the MTO data, Chetty et al. (2016) were able to measure the effect of the experimental treatment on children's future earnings. They found that the economic impact only occurred for children who were younger when they moved, indicating that neighborhood effects require long exposure periods to affect children's future earnings.

Big data systems themselves can create natural experiments by changing user behavior through subtle and not-so-subtle changes to their policies and practices. For example, Brown et al. (2010) use changes to eBay's interface to test the effect of suppressing price information, specifically shipping costs, on purchase decisions. Researchers can use changes to policies, designs, or algorithms

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

as experimental manipulations. The very malleability of digital worlds make them powerful vessels for conducting very large experiments.

Facebook has been the setting for several large-scale field experiments using randomized manipulations of what people could see about their peers to study social influence (Bond et al. 2012, Kramer et al. 2014). For example, Bond et al. conducted a field experiment using all Facebook users in the United States who were over the age of 18 and who logged into Facebook on the date of the US election in 2012. Six hundred thousand users were put into an informational condition and received a message encouraging them to vote, providing information on where to vote, and allowing them to click an "I Voted" button. Sixty million users were randomly put into the social condition and received the same information but with a list of all of their friends who clicked the "I Voted" button. Finally, a control group of six hundred thousand users saw no message at all.

They examined the extent to which study participants actually voted, engaged with the "I Voted" button, or clicked the link to view voter information. To measure actual voting they linked a subsample of the study population to public voting records. They found a direct effect: Those in the informational condition were more likely to vote than those in the control group, and those in the social condition were more likely than those in the informational condition to vote. They also found an indirect effect: Those who had friends in the social condition were 0.255% more likely to vote per friend in the condition and 0.012% more likely to seek out information per friend in the condition. These findings are important, proving that online social networks contribute to the diffusion of offline behavior. It presents another insight beyond this. The effects of the study are miniscule but, in aggregate, still represent hundreds of thousands of voters. Big data systems offer an unprecedented degree of precision in measuring small but meaningful effects.

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Making Big Data Small

The power of big data is often the small data contained within. Although many online platforms such as Reddit, Wikipedia, and Microsoft's Xbox Live are dominated by young, white, Western men, within their millions of users, one can still find hundreds or thousands of people who are older, nonwhite, female, and non-Western. Studies that "make big data small" (Foucault Welles 2014) either use big data to observe traditionally hard to reach populations or utilize the vast array of very specific kinds of cases to generate robust estimates.

Big data provides access to data on traditionally underrepresented populations. Twitter data have been used to study people who suffer from PTSD, suicidal ideation, and depression (Coppersmith et al. 2014; De Choudhury et al. 2013, 2016). Jackson & Foucault Welles (2016) and Barberá et al. (2015) use Twitter to identify individuals at the center of emergent social movements including Black Lives Matter, the Gezi Park protest in Turkey, Occupy Wall Street, and the Spanish Indignados.

Barberá et al. use Twitter to compare the patterns of mobilization between a successful protest mobilization at Taksim Gezi Park in 2013 (one locus of the Arab Spring) and two unsuccessful mobilizations by Occupy Wall Street and the Indignados in the spring of 2012. They searched Twitter's public application program interface (API) for keywords and hashtag keywords to collect samples of tweets from the three movements. They also collected two other samples of tweets from widespread, nonprotest activity to act as a control (the 2014 Academy Awards and a year's worth of tweets related to raising the minimum wage in the United States). They constructed mobilization networks among users who posted messages containing the keywords and hashtags and users who reposted those messages (i.e., retweets).

They find peripheral members of the Gezi Park protest mobilized more people than the core Occupy and Indignados protesters did. They find no such core-periphery patterns in the nonpolitical mobilization cases. They argue that successful mobilization involves getting peripheral

members to recruit still more peripheral people to protest: The mobilized must mobilize others themselves. This adds to contemporary social movement theory, showing that resources and organizational capacity are not enough to mobilize protest. Network diffusion can separate success from failure.

As studies of online protest show, using big data for research takes advantage of the increasing digitization of social life. Social movement studies have largely depended on newspapers and organizational archives because they are the few sources regularly logging social movement activities (Earl et al. 2004). Social media have become an essential mode for people at the margins of society to connect with one another and express thoughts and sentiments that otherwise go unsaid.

A second use for the small data within big data is in the robust estimates they can generate from the many narrow samples contained within. One example of this is the use of big data to generate population estimates. Paramount in this vein is Wang et al.'s (2015) use of surveys from a panel of Xbox users to predict national election polls in the United States. Xbox users skew heavily toward young, white, Western men. However, even though men made up 93% of the panel, the poll of roughly 340,000 still contained roughly 24,000 women. Using multilevel regression and poststratification, Wang et al. generate estimates of state and national polls for "demographic cells" (Wang et al. 2015, p. 981) such as college educated women over 55 in Florida who identify as Republican and voted for McCain in 2008. They aggregate these cell estimates to predict polls at the state and national level. They find that this reweighting of highly skewed big data accurately predicts national and most state-level polls for the presidential race and Obama's eventual election victory.

In addition, large samples contain enough unusual cases to robustly estimate heterogenous effects. Small data sets are blunt tools able only to detect large average effects. However, many associations of interest in sociology are contingent on individual and contextual factors. The study of intersectionality is premised on the belief that the main forces in society differ by the particular combinations of race, class, gender, sexuality, and other identities (Collins 1998). Big data allows for robustly estimating effects at the intersection. Others are developing algorithms that seek to inductively identify groups for whom effects are especially large (Athey & Imbens 2016, Green & Kern 2012, Imai & Ratkovic 2013, Taddy et al. 2016).

As survey researchers have long known, a well-defined sample of even a small size can tell us more than millions of poorly defined cases (Squire 1988). These studies by Wang et al., Athey & Imbens, and others show that we can recover the power of well-defined, small data from big data. The difference is that big data and big populations like Xbox users already exist, making them substantially more cost efficient to use to gather data than traditional sample building methods.

VULNERABILITIES

The core issue with any data is who and what get represented. With surveys, one might ask, for example, which respondents are accessible, and what they can accurately reveal. The scale of big data sets creates the illusion that they contain all relevant information on all relevant people. However, the difference between big and everything is still infinite, and the core issues of social science research around validity and generalizability still apply. Further, certain big data can be quite brittle, vulnerable to changes in the data generation process and to attacks motivated by the fact that they are materially consequential.

Generalizability

Big data are almost always convenience samples offering a distinct set of advantages and disadvantages. However, the data now easily available are unlike most convenience samples hitherto

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

common in the social sciences. Many are often convenience censuses: a complete record of a certain set of individuals or behaviors that match certain criteria. Scale and seeming comprehensiveness of data often obscure major issues regarding inclusion and selection and therefore representativeness and generalizability.

Many big data census efforts aspire to capture all possible data but do so without a systematic sampling frame. For example, projects like EventRegistry (Leban et al. 2014) and GDELT (Leetaru & Schrodt 2013) identify global events from major media sources such as the *New York Times* and Associated Press, news aggregators such as Google News and LexisNexis, and regional news sources, which are added continuously. They crawl as many sources as possible with the goal of creating as near a census of world events as possible. Yet these projects lack principles for sampling on critical features such as geography, publication frequency, journalistic practices, and type of news. This means that different kinds of events like sporting events or weather may be covered in geographically and temporally uneven ways. Beyond this, more data cannot solve the long-standing problems of using news sources to study events (Earl et al. 2004).

"Big data hubris" (Lazer et al. 2014, p. 1203) is the belief that volume can solve all problems. With near-census projects, trying to create a census without a sampling frame causes errors associated with selection, missing data, and thin coverage to be inestimable (Japec et al. 2015). To use an analogy, these near-census projects are like sending millions of people out into the streets to count the population of the United States. You would count a large number of people, but you could not know the kinds of people who are counted twice or not counted at all, and therefore you could not know what kinds of people are over- or underrepresented and to what extent.

However, big data often is a census of a particular, conveniently accessible social world. All of Twitter is a census of Twitter. Data from Kickstarter are a census of Kickstarter. However, even census data have limits when they come from a single platform. Tufekci (2014) notes that Twitter has become to social media scholars what the fruit fly is to biologists—a model organism. Tufekci argues that relying on a single platform produces issues for generalizability unique to model organisms. For example, different social media platforms have different rules for following or friending others and posting or sending messages. Thus, the patterns of friendship and diffusion differ across platforms (Tufekci 2014, p. 507). Different model organisms behave differently and therefore can give divergent results. In addition, platforms differ in the age, gender, race, and class backgrounds of their users (Perrin 2015). Even phone call data can exclude or overrepresent certain populations (Pestre et al. 2016). Not all social processes and forces are well-represented in a given model organism.

Generalizability is always a question of reference: To what do we want to generalize? With big data, the widespread availability of convenient census and near-census data leads many to overstate the reach of their findings. As research using multiple social media platforms shows, the results from one population of users do not necessarily apply to another. And convenient near-census data may play to our big data hubris that volume will trump sampling. The solution is to use data from multiple sources to validate the findings from any one of them. This will take a commitment from scientists to create institutional structures that can provide access to these data sources. In addition, near-census projects need to take sampling more seriously in order to estimate error in the data produced (Japec et al. 2015).

Too Many Big Data

Being tied to individual platforms presents another problem: when the scientifically relevant behavior spans these platforms. For example, consider trying to measure a simple construct like "who an individual regularly talks with." People use different modes to interact with different people,

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

such as text messaging, face-to-face, or by phone. And people can use functionally equivalent tools within a mode, such as cell phones, land lines, Skype, and Google Hangouts to reach different people. Thus, the data on who someone interacts with exist in a variety of different data sets. As society continues to develop new ways for people to interact with one another, our data on social interaction will continue to fragment. We call this the problem of too many big data.

This issue is even worse for many important sociocultural constructs, such as friendship, which are arguably more cognitive and normative than behavioral. How does one observe love, affection, or deceit from cell phone data?

These issues with too many big data are potentially surmountable. They point to the need to create a fusion of emerging computational methods with existing social science methods. For example, one study has shown promise in tracking interaction across digital and physical modes. As we discussed earlier, the Copenhagen Network Study has proven we can track interaction across digital and physical modes. The results provide a rare glimpse into the ways in which these data can overlap or diverge from one another. However, this multimode data collection is expensive and intrusive.

Other work could compare use of a particular platform relative to population averages to understand the potential limits of the platform. It is also plausible that nonbehavioral constructs like friendship, love, and trust are observable in the sense that there may be certain behaviors strongly correlated with certain cognitive constructs. To identify the best signals, we need a Rosetta stone connecting behavioral big data constructs to theoretically motivated social constructs (Margolin et al. 2013).

Artifacts and Reactivity

The caveats to big data extend beyond the problems of only studying a specific data set. Big data systems are themselves susceptible to various kinds of error and misappropriation. In the next section, we discuss how social forces can manipulate these data in unexpected and difficult-to-detect ways. In this section, we talk about artifacts, the errors and anomalies that systems produce, and reactivity, the changes in data resulting from technical changes rather than underlying changes in behavior.

Platforms do not merely represent data but generate them. In some cases, it can be difficult to distinguish observations resulting from errors in the system from those representing a real change in underlying behavior. For example, in the Google Ngram project, the word "fuck" is used with startling frequency in books published through 1800, and drops to near zero during the 1800s. Upon closer inspection, it is clear that this did not reflect some dramatic shift in social mores, but rather is an artifact of contemporary optical character recognition systematically misinterpreting an archaic version of "s" as an "f."[2]

When platforms change how they operate, both behavior and the way behavior is recorded can change. Earlier we argued that these changes can act as natural experiments. Here, we show these changes can have negative consequences for science. Google Flu Trends was perhaps one of the most prominent uses of big data, estimating the prevalence of flu using Google search traffic (Ginsberg et al. 2009, Ortiz et al. 2011). However, it seems likely that Google changed its search to make it more useful for finding health-related information, leading people to perform more searches for the flu during the peak of the season. The result was that the number of

[2] **http://searchengineland.com/when-ocr-goes-bad-googles-ngram-viewer-the-f-word-59181**.

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

searches increased relative to the number of flu cases, and Google Flu Trends started dramatically overreporting the flu (Lazer et al. 2014).

Behavior on any platform is part emergent, part mechanical. The hashtag convention on Twitter was not imagined or proposed by the company but instead originated from users who found keywords searchable when they began with the hash. Only later did Twitter make these hashtags into hyperlinks. There will always be changes in a platform, and artifacts inevitably slip in. Social scientists must be aware of and wary of the technical and social history of the platforms from which their data are generated.

The Ideal User Assumption: Bots, Puppets, and Manipulation

In big data analysis, we often assume that the data have been generated by a specific type of user, most often single, unique people who express themselves honestly through their personal accounts. This ideal user assumption fails to hold under a wide variety of critical circumstances. Many accounts are not operated by human beings. Additionally, users can have multiple accounts, sometimes with the intent to conceal the user's true identity. Finally, people, organizations, and even nation states put platforms to unintended uses. Taken together, the characteristics of the ideal user need to be validated rather than assumed, and nonideal users generating big data need to be studied in their own right.

The first type of violation of the ideal user assumption is when we incorrectly assume all users are human. This is most often violated in the case of robots and organizations. It is likely that robots are present in all big data systems, whether bots on social media (Ferrara et al. 2016), nonplayer characters in online games (Lee & Ramler 2015), or even robocalls in CDR data (Gupta et al. 2015). Many organizations and institutions are present and active in these systems as well, including corporations, governments, and terrorist organizations (McCorriston et al. 2015). Robots and organizations engage on these platforms for a variety of prosocial and antisocial purposes, making them difficult to detect and theorize (Lee et al. 2011, Ferrara 2015).

The ideal-user assumption is also violated when we assume everyone is who they claim to be. Humans misrepresent themselves in a variety of ways outlined by Wang et al. (2006). Individuals can create multiple accounts for different purposes, including alters to express different identities or so-called sock puppets to fabricate support for themselves or their cause (Bu et al. 2013, Zheng et al. 2006). Individuals can also be deceptive about critical aspects of their identity, as in the case of catfishing. Finally, people can use a fraudulent identity, presenting themselves to be another particular person.

The final violation of the ideal user assumption is when users manipulate the platform in unintended ways to achieve surreptitious goals. In analyzing big data, we often assume that the data are being generated by people behaving in good faith. Yet, many users attempt to game the system or create rigged systems (Ferrara 2015). News media have reported cases in which countries and election campaigns have used robots and real people on social media to artificially inflate their own positions and influence (Durgin 2016, Matthew 2015). In 2016, the Chinese peer-to-peer lending platform Ezubao was revealed to be a Ponzi scheme in which 95% of the proposals were fake (Xinhua 2016). Finally, actors, particularly nation states, can directly control these platforms. As King et al. (2014) find, the Chinese government conducts mass censorship of online media to suppress news of any events with the potential to yield extragovernmental political mobilizations such as protests.

Social media platforms that are highly permeable and very cheap to manipulate are easy targets for robots, deception, and manipulation. However, even high cost, impermeable platforms like CDR and newswire data can be subject to this activity. The challenge is that each of these violations

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

of the ideal user assumption typically occurs under different conditions. Sock puppetry and political manipulation are probably more likely to occur in contentious arenas and topics, whereas fraud and market manipulation are probably likely to occur at the intersection of the financially vulnerable and the financially influential. In analyzing big data, sample control is critical for making inferences about human beings acting in good faith. However, it is important to remember that these forms of nonideal-user behavior are worthy of study on their own.

RESEARCH ETHICS

There are major ethical issues regarding the acquisition and use of big data for researchers, institutions, and society at large. Some issues are new, but many are new versions of long-standing issues. The problem, however, is that there is no consensus on what the rules should be, and the policies and recommendations set forth by scientific associations vary substantially, often contradicting one another. Rules will eventually become clear, but the risks to researchers, universities, and the public remain high until they do. Further, these ethical issues in turn suggest interesting researchable questions, ranging from issues around reidentification (Sweeney 2002) to the meaning and management of consent by subjects (Stopczynski et al. 2014a,b).

The National Research Council proposed to amend the definition of human subjects research to "a systematic investigation designed to develop or contribute to generalizable knowledge by obtaining data about a living individual directly through interaction or intervention, or by obtaining identifiable private information about an individual" (NRC 2014, recommendation 2.1, p. 40). However, rules regarding obtaining subject consent are typically obviated when the data have been collected by a third party and have been anonymized. Is this an acceptable standard given the prevalence of collaboration between companies and academics? Furthermore, given the ease of deidentifying many kinds of data, what standards can be set for anonymization?

Regulating informed consent, the heart of human subjects research, is only one of the central issues that have yet to be settled. Other open issues include the rights of secondary subjects, measuring harms from the loss of privacy, and regulating the status of leaked data like the Panama Papers, which revealed international tax evasion by the world's elite, or the data dump of the US adultery website Ashley Madison. The role of the university becomes critical here because it is part of the regulatory apparatus enforcing rules and protecting scholars, but also because it provides the training infrastructure that empowers compliance.

FUTURE TRENDS

Researchers have used big data to answer old questions in new ways and new questions never before answerable. Their successes and failures have helped us identify the promises and pitfalls of research in this area and the kinds of investments institutions need to make for the future of this research. In this final section, we point to six trends that will likely affect the big data landscape in hopes of helping sociology get ahead of the curve.

More Data Are Coming

Big data will continue to grow into more domains. For example, EventRegistry provides data on events, and it also acts as a repository providing real-time access to more than 100,000 news publishers (**http://www.EventRegistry.org**). Big data will also continue to reach back into the past as libraries digitize their collections, newspapers digitize their archives, and initiatives such as

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Google Books and Project Gutenberg digitize books. The question for data coverage will continue to be less about whether the data exist, but more about what can be studied with the available data.

More linkages between different big data will become more common. The work of Chetty and colleagues (2014a,b, 2016) with IRS data is only the beginning. Another opportunity is connecting online data to offline data, such as linking social media accounts to voter records and data collected by brokers such as Acxiom. One revolutionary form of big data linkage is wikification, which involves linking words and phrases in text to entities in Wikipedia (Mihalcea & Csomai 2007). Entity linkage allows researchers to use the structured and unstructured data Wikipedia maintains on entities to enhance the contextual information associated with texts.

Different Data Are Coming

The majority of data being created on big data platforms are still unusable for social scientific research. These are the images, audio, and video being created, discussed, and shared. The tools for providing meaningful structure to these data at scale have lagged behind those of other types of data, such as text. They are quickly catching up. For example, image processing using convolutional neural nets and other deep learning methods has become as easy to use in Python and R as methods for text analysis. Furthermore, the tools to analyze these data are increasingly being made available through publicly accessible interfaces like Google Cloud Vision API. With publicly accessible models, researchers upload their files to the service, which uses pretrained models to make inferences about the file and then sends these inferences back as metadata.

Models Will Become More Generic

Google Cloud Vision API is a prime example of another critical trend in machine learning: creating generic models and making them available to the public. There is a long precedent of creating and sharing models for processing unstructured data. The Linguistic Inquiry and Word Count dictionary is perhaps the best known in the social sciences (Tausczik & Pennebaker 2010). However, such models are now being published in a variety of methods. For example, Jozefowicz et al. (2016) trained a deep learning model on the One Billion Word Benchmark and are publishing the model itself as an alternative to the model released in 2014 in Stanford's GloVe (Pennington et al. 2014). Like the Vision API, researchers can use these models on their own texts to generate word embeddings.

Generic models allow researchers to use pretrained machine learning models on their own data, rather than having to deal with the issues of data processing and model specification. These out-of-the-box machine learning projects aspire to use big data to create the most effective models and then make those models the standard for processing unstructured data. However, generic models are not necessarily better at applied tasks than specialized models. And, in the absence of social theory, such generic models may miss obvious social patterns in data, potentially reinforcing long-standing social biases (Caliskan et al. 2017).

Data from Multiple Platforms Will Become Standard

As big data systems proliferate and multiple systems offer similar services, it will become increasingly possible and easier for researchers to perform studies on different platforms. The CrowdBerkeley project is offering data for multiple crowdfunding websites such as Kickstarter, Indiegogo, and Kiva. Multiple forms of big data will also be used to create parallel measures. For example, new models of political partisanship are being generated by Federal Election

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Commission data, Twitter, press releases, and floor speeches (Barberá 2015, Bonica 2014, Gentzkow et al. 2016, Tsur et al. 2015). Interestingly, each of these measures provides different narratives about the emergence of partisanship and demonstrates the importance of using different data to approach the same phenomena.

Qualitative Approaches to Big Data

While big data is typically viewed as quantitative social science on steroids, we anticipate innovative approaches weaving together qualitative methods and computational approaches to large-scale data. There is a long history of archival research in the social sciences. Digital archives present the challenge of vast amounts of information beyond the capacity of armies of grad students to read. Searching and sorting of archives becomes essential to qualitative understanding. At the simplest level, this might simply require keyword searches, but certainly more complex, computationally enabled approaches will emerge. For example, consider hypothetical research examining the Internet Archive's version of **http://www.congress.gov**. The snapshots of the members' home pages present too large a data set to comprehensively read, but it is feasible to query the data for policy statements on health care by every member for targeted reading and hand-coding.

Methodological Integration

The prior point highlights a more general lesson: Big data will increasingly be integrated with existing research methods in sociology. Big data offers strengths and weaknesses that are quite different than existing data sources (Lazer et al. 2014). The most compelling sociological research in the twenty-first century will not be big data but a fusion of data sources related to important questions. Survey data will be linked to a tiny portion of archival data, providing inferential power to the entire archive that it otherwise would not have. Interesting or typical cases in big data can be identified for qualitative exploration. The scientific payoff should, in turn, be insight into phenomena that heretofore have been neglected, related to the connectivity and dynamics of entire societies.

The future of big data is as bright and fraught as its past. While sociology has generally lagged behind in using big data, there are many opportunities for the field to take advantage of and many challenges and debates to confront. Further, the increasing presence of digitally mediated social activity and the increasingly digital social life mean that the need to integrate big data approaches into sociology will increase for the foreseeable future, with the corresponding need for sociologists to contribute to our understanding of an increasingly digital and digitalized world.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF0920053 (the ARL Network Science CTA) and in part by a grant from the US Army Research Office W911NF1210556. Any opinions expressed are the authors' alone.

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

LITERATURE CITED

Athey S, Imbens G. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS* 113(27):53–60

Ayers JW, Ribisl K, Brownstein JS. 2011. Using search query surveillance to monitor tax avoidance and smoking cessation following the United States' 2009 "SCHIP" cigarette tax increase. *PLOS ONE* 6(3):e16777

Bail CA. 2012. The fringe effect. *Am. Sociol. Rev.* 77(6):55–79

Bakshy E, Hofman JM, Mason WA, Watts DJ. 2011. Everyone's an influencer: quantifying influence on twitter. *Proc. 4th ACM Conf. Web Search Data Mining*, pp. 65–74. New York: ACM

Bakshy E, Messing S, Adamic LA. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239):1130–32

Barberá P. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Polit. Anal.* 23(1):76–91

Barberá P, Wang N, Bonneau R, Jost JT, Nagler J, et al. 2015. The critical periphery in the growth of social protests. *PLOS ONE* 10(11):1–15

Beauchamp N. 2016. Predicting and interpolating state-level polls using Twitter textual data. *Am. J. Political Sci.* 61:490–503

Bernard HR, Killworth P, Kronenfeld D, Sailer L. 1984. The problem of informant accuracy: the validity of retrospective data. *Annu. Rev. Anthropol.* 13:495–517

Blanford JI, Huang Z, Savelyev A, MacEachren AM. 2015. Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. *PLOS ONE* 10(6):e0129202

Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, et al. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–98

Bonica A. 2014. Mapping the ideological marketplace. *Am. J. Polit. Sci.* 58(2):367–86

boyd d, Crawford K. 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* 15(5):662–79

Brockmann D, Hufnagel L, Geisel T. 2006. The scaling laws of human travel. *Nature* 439(7075):462–65

Brown J, Hossain T, Morgan J. 2010. Shrouded attributes and information suppression: evidence from the field. *Q. J. Econ.* 125(2):859–76

Bu Z, Xia Z, Wang J. 2013. A sock puppet detection algorithm on virtual spaces. *Knowl.-Based Syst.* 37:366–77

Burt RS. 2012. Network-related personality and the agency question: multirole evidence from a virtual world. *Am. J. Sociol.* 118(3):543–91

Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–86

Cavallo A. 2013. Online and official price indexes: measuring Argentina's inflation. *J. Monet. Econ.* 60(2):152–165

Cavallo A. 2017. Scraped data and sticky prices. *Rev. Econ. Stat.* In press. **http://dx.doi.org/10.1162/REST_a_00652**

Cavallo A, Neiman B, Rigobon R. 2014. Currency unions, product introductions, and the real exchange rate. *Q. J. Econ.* 129(2):529–95

Cavallo A, Rigobon R. 2016. The Billion Prices Project: using online prices for measurement and research. *J. Econ. Perspect.* 30(2):151–78

Chetty R, Friedman JN, Rockoff JE. 2014a. Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* 104(9):2593–632

Chetty R, Friedman JN, Rockoff JE. 2014b. Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. *Am. Econ. Rev.* 104(9):2633–79

Chetty R, Hendren N, Katz LF. 2016. The effects of exposure to better neighborhoods on children: new evidence from the Moving to Opportunity experiment. *Am. Econ. Rev.* 106(4):855–902

Collins PH. 1998. It's all in the family: intersections of gender, race, and nation. *Hypatia* 13(3):62–82

Coppersmith G, Harman C, Dredze M. 2014. Measuring post traumatic stress disorder in Twitter. *Proc. 8th Int. AAAI Conf. Weblogs Soc. Media*, pp. 579–82. **http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8079**

Curington CV, Lin K-H, Lundquist JH. 2015. Positioning multiraciality in cyberspace. *Am. Sociol. Rev.* 80(4):764–88

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

De Choudhury M, Gamon M, Counts S, Horvitz E. 2013. Predicting depression via social media. *Proc. 7th Int. AAAI Conf. Weblogs Soc. Media*, pp. 128–37. **http://course.duruofei.com/wp-content/uploads/2015/05/Choudhury_Predicting-Depression-via-Social-Media_ICWSM13.pdf**

De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Proc. 2016 CHI Conf. Hum. Factors Comput. Syst.*, pp. 2098–2110. New York: ACM Press

De Vaan M, Vedres B, Stark D. 2015. Game changer: the topology of creativity. *Am. J. Sociol.* 120(4):1144–94

Diekmann A, Jann B, Przepiorka W, Wehrli S. 2014. Reputation formation and the evolution of cooperation in anonymous online markets. *Am. Sociol. Rev.* 79(1):65–85

Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PLOS ONE* 6(12):e26752

Durgin C. 2016. Inside Donald Trump's Potemkin Twitter army. *National Review*, Apr. 8. **http://www.nationalreview.com/article/433870/donald-trumps-twitter-supporters-might-be-fake**

Eagle N, Pentland AS, Lazer D. 2009. Inferring friendship network structure by using mobile phone data. *PNAS* 106(36):15274–78

Earl J, Martin A, McCarthy JD, Soule SA. 2004. The use of newspaper data in the study of collective action. *Annu. Rev. Sociol.* 30(1):65–80

Edelman BG, Luca M. 2014. *Digital discrimination: the case of Airbnb.com*. Working Pap. 14–054, NOM Unit, Harvard Bus. Sch.

Ferrara E. 2015. "Manipulation and abuse on social media" by Emilio Ferrara with Ching-man Au Yeung as coordinator. *ACM SIGWEB Newsletter*, *Spring* 4:1–9

Ferrara E, Varol O, Davis C, Menczer F, Flammini A. 2016. The rise of social bots. *Commun. ACM.* 59(7):96–104

Foucault Welles B. 2014. On minorities and outliers: The case for making big data small. *Big Data Soc.* 1(1):1–2

Gartner. 2011. *Gartner says solving "big data" challenge involves more than just managing volumes of data*. News Release, June 27. **http://www.gartner.com/newsroom/id/1731916**

Gentzkow M, Shapiro JM, Taddy M. 2016. Measuring polarization in high-dimensional data: method and application to congressional speech. NBER Work. Pap. 22423, Natl. Bur. Econ. Res., Cambridge, MA

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–14

Goel S, Watts DJ, Goldstein DG. 2012. The structure of online diffusion networks. *Proc. 13th ACM Conf. Electron. Commer.*, pp. 623–38. New York: ACM

Goldberg A, Srivastava SB, Manian VG, Monroe W, Potts C. 2016. Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *Am. Sociol. Rev.* 81(6):1190–222

Golder SA, Macy MW. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–81

Golder SA, Macy MW. 2014. Digital footprints: opportunities and challenges for online social research. *Annu. Rev. Sociol.* 40:129–52

González MC, Hidalgo CA, Barabási A-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–82

Green DP, Kern HL. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* 76(3):491–511

Greenberg J, Mollick E. 2017. Activist choice homophily and the crowdfunding of female founders. *Adm. Sci. Q.* 62:341–74

Gupta P, Srinivasan B, Balasubramaniyan V, Ahamad M. 2015. *Phoneypot: data-driven understanding of telephony threats*. Brief. Pap., NDSS Symp. 2015, San Diego, CA

Hall M, Crowder K, Spring A. 2015. Neighborhood foreclosures, racial/ethnic transitions, and residential segregation. *Am. Sociol. Rev.* 80(3):526–49

Hopkins DJ, King G. 2010. A method of automated nonparametric content analysis for social science. *Am. J. Political Sci.* 54(1):229–47

Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7(1):443–70

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Jackson SJ, Foucault Welles B. 2016. #Ferguson is everywhere: initiators in emerging counterpublic networks. *Inf. Commun. Soc.* 19(3):397–418

Japec L, Kreuter F, Berg M, Biemer P, Decker P, et al. 2015. *AAPOR report on big data*. Am. Assoc. Public Opin. Res., Oakbrook Terrace, IL

Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y. 2016. Exploring the limits of language modeling. arXiv:1602.02410 [cs.CL]

Keegan BC, Brubaker JR. 2015. "Is" to "was": coordination and commemoration in posthumous activity on Wikipedia biographies. *Proc. 18th ACM Conf. Comput. Support. Coop. Work Soc. Comput.*, pp. 533–46. New York: ACM

Keegan BC, Gergle D, Contractor N. 2013. Hot off the wiki: structures and dynamics of Wikipedia's coverage of breaking news events. *Am. Behav. Sci.* 57(5):595–622

Kim M, Newth D, Christen P. 2014. Trends of news diffusion in social media based on crowd phenomena. *Proc. 23rd Int. Conf. World Wide Web*, pp. 753–58. New York: ACM

King G, Pan J, Roberts ME. 2014. Reverse-engineering censorship in China: randomized experimentation and participant observation. *Science* 345(6199):1–10

Knigge A, Maas I, van Leeuwen MHD. 2014a. Sources of sibling (dis)similarity: total family impact on status variation in the Netherlands in the nineteenth century. *Am. J. Sociol.* 120(3):908–48

Knigge A, Maas I, van Leeuwen MHD, Mandemakers K. 2014b. Status attainment of siblings during modernization. *Am. Sociol. Rev.* 79(3):549–74

Kossinets G, Watts DJ. 2006. Empirical analysis of an evolving social network. *Science* 311(5757):88–90

Kramer ADI, Guillory JE, Hancock JT. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* 111(29):8788–90

Laney D. 2001. *3D data management: controlling data volume, velocity and variety*. Res. Note, META Group, Stamford, CT. **https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf**

Lazer D. 2015. The rise of the social algorithm. *Science* 348(6239):1090–91

Lazer D, Kennedy R, King G, Vespignani A. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343(6176):1203–5

Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, et al. 2009. Life in the network: the coming age of computational social science. *Science* 323(5915):721

Leban G, Fortuna B, Brank J, Grobelnik M. 2014. Event Registry: learning about world events from news. *Proc. 23rd Int. Conf. World Wide Web*, pp. 107–10. New York: ACM

Lee C-S, Ramler I. 2015. Rise of the bots: bot prevalence and its impact on match outcomes in League of Legends. *Int. Worksh. Netw. Syst. Support Games (NetGames), Zagreb, Dec. 3–4*, pp. 1–6

Lee K, Eoff BD, Caverlee J. 2011. Seven months with the devils: a long-term study of content polluters on Twitter. *5th Int. AAAI Conf. Weblogs Soc. Media.* **https://pdfs.semanticscholar.org/1dd5/355e62b9fc37a355e135d5909ed28128d653.pdf**

Leetaru K, Schrodt PA. 2013. GDELT: global data on events, location, and tone, 1979–2012. *Int. Stud. Assoc. Annu. Conf., San Diego*. **http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.686.6605&rep=rep1&type=pdf**

Legewie J. 2016. Racial profiling and use of force in police stops: how local events trigger periods of increased discrimination. *Am. J. Sociol.* 122(2):379–424

Legewie J, Schaeffer M. 2016. Contested boundaries: explaining where ethnoracial diversity provokes neighborhood conflict. *Am. J. Sociol.* 122(1):125–61

Leung MD. 2014. Dilettante or Renaissance person? How the order of job experiences affects hiring in an external labor market. *Am. Sociol. Rev.* 79(1):136–58

Lin K-H, Lundquist J. 2013. Mate selection in cyberspace: the intersection of race, gender, and education. *Am. J. Sociol.* 119(1):183–215

Manovich L. 2012. Trending: the promises and the challenges of big social data. In *Debates in the Digital Humanities*, Vol. 2, ed. MK Gold, pp. 460–75. Minneapolis, MN: Univ. Minn. Press

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, et al. 2011. *Big data: the next frontier for innovation, competition, and productivity*. Rep., McKinsey Global Inst. **http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation**

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Margolin D, Lin Y-R, Brewer D, Lazer D. 2013. Matching data and interpretation: towards a Rosetta stone joining behavioral and survey data. *7th Int. AAAI Conf. Weblogs Soc. Media*, pp. 9–10. **http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6267**

Marsden PV. 1990. Network data and measurement. *Annu. Rev. Sociol.* 16:435–63

Massey DS, Denton NA. 1993. *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA: Harvard Univ. Press

Matthew S. 2015. Revealed: how Russia's "troll factory" runs thousands of fake Twitter and Facebook accounts to flood social media with pro-Putin propaganda. *The Daily Mail*, March 28

McCorriston J, Jurgens D, Ruths D. 2015. Organizations are users too: characterizing and detecting the presence of organizations on Twitter. *9th Int. AAAI Conf. Web Soc. Media*. **http://www-cs.stanford.edu/~jurgens/docs/mccorriston-jurgens-ruths_icwsm-2015.pdf**

Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–82

Mihalcea R, Csomai A. 2007. Wikify!: Linking documents to encyclopedic knowledge. *Proc. 16th ACM Conf. Inf. Knowl. Manag.*, pp. 233–42. New York: ACM

Monroe BL. 2013. The five Vs of big data political science: introduction to the Virtual Issue on Big Data in Political Science. *Polit. Anal.* 19(5):66–86

NRC (Natl. Res. Counc.). 2014. *Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences*. Washington, DC: Natl. Acad. Press

Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, et al. 2007. Structure and tie strengths in mobile communication networks. *PNAS* 104(18):7332–36

Onnela J-P, Waber BN, Pentland A, Schnorf S, Lazer D. 2014. Using sociometers to quantify social interaction patterns. *Sci. Rep.* 4:5604

Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CH. 2011. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. *PLOS ONE* 6(4):1–9

Pennington J, Socher R, Manning CD. 2014. GloVe: global vectors for word representation. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, pp. 1532–43

Perrin A. 2015. *Social networking usage: 2005–2015*. Rep., Pew Res. Cent., Washington, DC

Pestre G, Letouzé E, Zagheni E. 2016. *The ABCDE of big data: assessing biases in call-detail records for development estimates*. Presented at Annu. Bank Conf. Dev. Econ., June 20–21, Washington, DC. **http://pubdocs.worldbank.org/pubdocs/publicdoc/2016/6/551311466182785065/Pestre-Letouze-Zagheni-ABCDE-May-2016.pdf**

Phan TQ, Airoldi EM. 2015. A natural experiment of social network formation and dynamics. *PNAS* 112(21):6595–600

Romero DM, Meeder B, Kleinberg J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. *Proc. 20th Int. Conf. World Wide Web*, pp. 695–704. New York: ACM

Sevtsuk A, Ratti C. 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *J. Urban Technol.* 17(1):41–60

Small ML. 2004. *Villa Victoria: The Transformation of Social Capital in a Boston Barrio*. Chicago: Univ. Chicago Press

Squire P. 1988. Why the 1936 *Literary Digest* poll failed. *Public Opin. Q.* 52(1):125–33

State B, Park P, Weber I, Macy M. 2015. The mesh of civilizations in the global network of digital communication. *PLOS ONE* 10(5):e0122543

Stopczynski A, Pietri R, Pentland A, Lazer D, Lehmann S. 2014a. Privacy in sensor-driven human data collection: a guide for practitioners. arXiv:1403.5299 [cs.CY]

Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, et al. 2014b. Measuring large-scale social networks with high resolution. *PLOS ONE* 9(4):e95978

Sweeney L. 2002. K-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 10(05):557–70

Taddy M, Gardner M, Chen L, Draper D. 2016. A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. *J. Bus. Econ. Stat.* 34(4):661–72

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Tausczik YR, Pennebaker JW. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29(1):24–54

Toole JL, Lin Y-R, Muehlegger E, Shoag D, González MC, Lazer D. 2015. Tracking employment shocks using mobile phone data. *J. R. Soc. Interface* 12(107):20150185

Toomet O, Silm S, Saluveer E, Ahas R, Tammaru T. 2015. Where do ethno-linguistic groups meet? How copresence during free-time is related to copresence at home and at work. *PLOS ONE*. 10(5):e0126093

Tsur O, Calacci D, Lazer D. 2015. A frame of mind: using statistical models for detection of framing and agenda setting campaigns. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Joint Conf. Nat. Lang. Process., Beijing, July 26–31*, pp. 1629–38. **https://pdfs.semanticscholar.org/f5c8/dbeea0112227486b7fc3bd20a73726ffea88.pdf**

Tufekci Z. 2014. Big questions for social media big data: representativeness, validity and other methodological pitfalls. arXiv:1403.7400 [cs.SI]

van de Rijt A, Shor E, Ward C, Skiena S. 2013. Only 15 minutes? The social stratification of fame in printed media. *Am. Sociol. Rev.* 78(2):266–89

Vasi IB, Walker ET, Johnson JS, Tan HF. 2015. "No fracking way!" Documentary film, discursive opportunity, and local opposition against hydraulic fracturing in the United States, 2010 to 2013. *Am. Sociol. Rev.* 80(5):934–59

Wang GA, Chen H, Xu JJ, Atabakhsh H. 2006. Automatically detecting criminal identity deception: an adaptive detection algorithm. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* 36(5):988–99

Wang W, Rothschild D, Goel S, Gelman A. 2015. Forecasting elections with non-representative polls. *Int. J. Forecast.* 31(3):980–91

Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, et al. 2012. Quantifying the impact of human mobility on malaria. *Science*. 338(6104):267–70

Wilson WJ. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago: Univ. Chicago Press

Xinhua. 2016. Online P2P lender suspected of $US 7.6 billion fraud. *Xinhua*, Feb. 1. **http://news.xinhuanet.com/english/2016-02/01/c_135065022.htm**

Yang J, Counts S. 2010. Predicting the speed, scale, and range of information diffusion in Twitter. *ICWSM* 10:355–58

Zheng R, Li J, Chen H, Huang Z. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol.* 57(3):378–93

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Annual Review
of Sociology
Volume 43, 2017

Contents

Prefatory Article

A Life in Sociology
Robert M. Hauser 1

Theory and Methods

Data ex Machina: Introduction to Big Data
David Lazer and Jason Radford 19

Field Experiments Across the Social Sciences
Delia Baldassarri and Maria Abascal 41

Genealogical Microdata and Their Significance for Social Science
Xi Song and Cameron D. Campbell 75

Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling
Douglas D. Heckathorn and Christopher J. Cameron 101

New Developments in Survey Data Collection
Mick P. Couper 121

Replication in Social Science
Jeremy Freese and David Peterson 147

Studying the Digital: Directions and Challenges for Digital Methods
Keith N. Hampton 167

Theorizing in Sociological Research: A New Perspective, a New Departure?
Richard Swedberg 189

Social Processes

Decision-Making Processes in Social Contexts
Elizabeth Bruch and Fred Feinberg 207

Social Networks and Macrosocial Change
Emily Erikson and Nicholas Occhiuto 229

Toward a Sociology of Privacy
Denise Anthony, Celeste Campos-Castillo, and Christine Horne 249

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Formal Organizations

The Social Bases of Philanthropy
Emily Barman 271

Political and Economic Sociology

The Demand Side of Hiring: Employers in the Labor Market
David B. Bills, Valentina Di Stasio, and Klarita Gërxhani 291

Differentiation and Stratification

Categorical Inequality: Schools as Sorting Machines
Thurston Domina, Andrew Penner, and Emily Penner 311

Gender Quotas for Legislatures and Corporate Boards
Melanie M. Hughes, Pamela Paxton, and Mona Lena Krook 331

Graduate Education and Social Stratification
Julie R. Posselt and Eric Grodsky 353

Wealth Inequality and Accumulation
Alexandra Killewald, Fabian T. Pfeffer, and Jared N. Schachner 379

Individual and Society

Skin Color and Colorism: Global Research, Concepts, and Measurement
Angela R. Dixon and Edward E. Telles 405

The Development of Transgender Studies in Sociology
Kristen Schilt and Danya Lagos 425

Demography

Social Structure, Adversity, Toxic Stress, and Intergenerational Poverty: An Early Childhood Model
Craig A. McEwen and Bruce S. McEwen 445

The Second Demographic Transition Theory: A Review and Appraisal
Batool Zaidi and S. Philip Morgan 473

Urban and Rural Community Sociology

Ethnographies of Race, Crime, and Justice: Toward a Sociological Double-Consciousness
Victor M. Rios, Nikita Carney, and Jasmine Kelekay 493

Explicating Divided Approaches to Gentrification and Growing Income Inequality
Japonica Brown-Saracino 515

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

Policy

The Social Safety Net After Welfare Reform: Recent Developments and Consequences for Household Dynamics
Laura Tach and Kathryn Edin 541

Indexes

Cumulative Index of Contributing Authors, Volumes 34–43 563

Cumulative Index of Article Titles, Volumes 34–43 567

Errata

An online log of corrections to *Annual Review of Sociology* articles may be found at http://www.annualreviews.org/errata/soc

Annu. Rev. Sociol. 2017.43:19-39. Downloaded from www.annualreviews.org
Access provided by 50.233.106.125 on 08/31/17. For personal use only.

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

Big Data Hubris

"Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9*–*11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (*13*). This should have been a warning that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the nonseasonal 2009 influenza A–H1N1 pandemic (*2*, *14*). In short, the initial version of GFT was part flu detector, part winter detector. GFT engineers updated the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (*16*–*19*), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near–real-time health data (*2*, *20*). For example, by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone (see the chart). This is no substitute for ongoing evaluation and improvement, but, by incorporating this information, GFT could have largely healed itself and would have likely remained out of the headlines.

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy.

*Corresponding author. E-mail: d.lazer@neu.edu.

CREDIT: ADAPTED FROM AXEL KORES/DESIGN & ART DIRECTION/ISTOCKPHOTO.COM

Algorithm Dynamics

All empirical research stands on a foundation of measurement. Is the instrumentation actually capturing the theoretical construct of interest? Is measurement stable and comparable across cases and over time? Are measurement errors systematic? At a minimum, it is quite likely that GFT was an unstable reflection of the prevalence of the flu because of algorithm dynamics affecting Google's search algorithm. Algorithm dynamics are the changes made by engineers to improve the commercial service and by consumers in using that service. Several changes in Google's search algorithm and user behavior likely affected GFT's tracking. The most common explanation for GFT's error is a media-stoked panic last flu season (*1*, *15*). Although this may have been a factor, it cannot explain why GFT has been missing high by wide margins for more than 2 years. The 2009 version of GFT has weathered other media panics related to the flu, including the 2005–2006 influenza A/H5N1 ("bird flu") outbreak and the 2009 A/H1N1 ("swine flu") pandemic. A more likely culprit is changes made by Google's search algorithm itself.

The Google search algorithm is not a static entity—the company is constantly testing and improving search. For example, the official Google search blog reported 86 changes in June and July 2012 alone (SM). Search patterns are the result of thousands of decisions made by the company's programmers in various subunits and by millions of consumers worldwide.

There are multiple challenges to replicating GFT's original algorithm. GFT has never documented the 45 search terms used, and the examples that have been released appear misleading (*14*) (SM). Google does provide a service, Google Correlate, which allows the user to identify search data that correlate with a given time series; however, it is limited to national level data, whereas GFT was developed using correlations at the regional level (*13*). The service also fails to return any of the sample search terms reported in GFT-related publications (*13*, *14*).

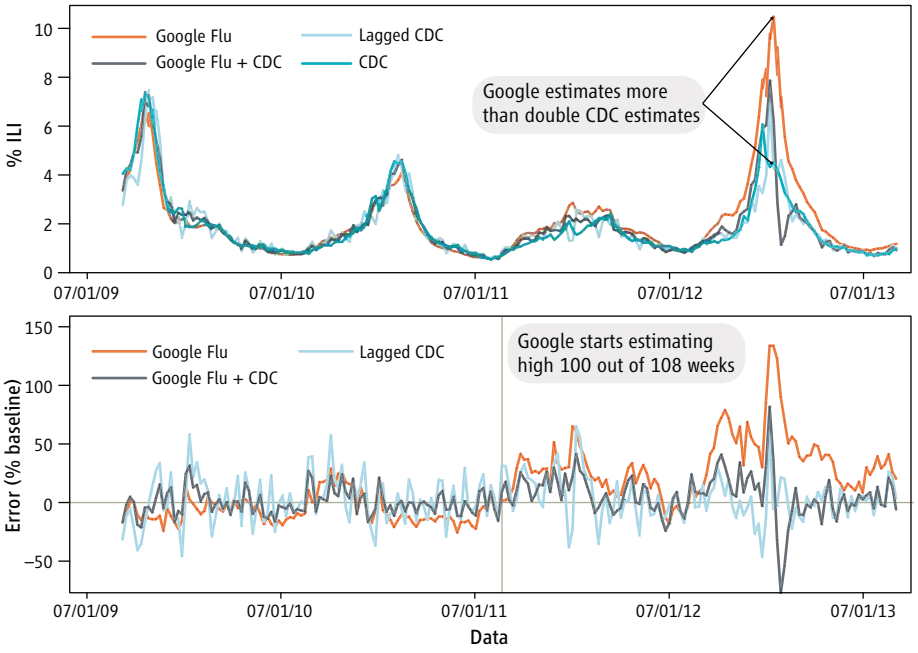
Nonetheless, using Google Correlate to compare correlated search terms for the GFT time series to those returned by the CDC's data revealed some interesting differences. In particular, searches for treatments for the flu and searches for information on differentiating the cold from the flu track closely with GFT's errors (SM). This points to the possibility that the explanation for changes in relative search behavior is "blue team" dynamics—where the algorithm producing the data (and thus user utilization) has been modified by the service provider in accordance with their business model. Google reported in June 2011 that it had modified its search results to provide suggested additional search terms and reported again in February 2012 that it was now returning potential diagnoses for searches including physical symptoms like "fever" and "cough" (*21*, *22*). The former recommends searching for treatments of the flu in response to general flu inquiries, and the latter may explain the increase in some searches to distinguish the flu from the common cold. We document several other changes that may have affected GFT (SM).

In improving its service to customers, Google is also changing the data-generating process. Modifications to the search algorithm are presumably implemented so as to support Google's business model—for example, in part, by providing users useful information quickly and, in part, to promote more advertising revenue. Recommended searches, usually based on what others have searched, will increase the relative magnitude of certain searches. Because GFT uses the relative prevalence of search terms in its model, improvements in the search algorithm can adversely affect GFT's estimates. Oddly, GFT bakes in an assumption that relative search volume for certain terms is statically related to external events, but search behavior is not just exogenously determined, it is also endogenously cultivated by the service provider.

Blue team issues are not limited to Google. Platforms such as Twitter and Facebook are always being re-engineered, and whether studies conducted even a year ago on data collected from these platforms can be replicated in later or earlier periods is an open question.

Although it does not appear to be an issue in GFT, scholars should also be aware of the potential for "red team" attacks on the systems we monitor. Red team dynamics occur when research subjects (in this case Web searchers) attempt to manipulate the data-generating process to meet their own goals, such as economic or political gain. Twitter polling is a clear example of these tactics. Campaigns and companies, aware that news media are monitoring Twitter, have used numerous tactics to make sure their candidate or product is trending (*23*, *24*).

Similar use has been made of Twitter and Facebook to spread rumors about stock prices and markets. Ironically, the more successful we become at monitoring the behavior of people using these open sources of information, the more tempting it will be to manipulate those signals.



GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage {[Non-CDC estmate) − (CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

Transparency, Granularity, and All-Data

The GFT parable is important as a case study where we can learn critical lessons as we move forward in the age of big data analysis.

Transparency and Replicability. Replication is a growing concern across the academy. The supporting materials for the GFT-related papers did not meet emerging community standards. Neither were core search terms identified nor larger search corpus provided. It is impossible for Google to make its full arsenal of data available to outsiders, nor would it be ethically acceptable, given privacy issues. However, there is no such constraint regarding the derivative, aggregated data. Even if one had access to all of Google's data, it would be impossible to replicate the analyses of the original paper from the information provided regarding the analysis. Although it is laudable that Google developed Google Correlate ostensibly from the concept used for GFT, the public technology cannot be utilized to replicate their findings. Clicking the link titled "match the pattern of actual flu activity (this is how we built Google Flu Trends!)" will not, ironically, produce a replication of the GFT search terms (*14*). Oddly, the few search terms offered in the papers (*14*) do not seem to be strongly related with either GFT or the CDC data (SM)—we surmise that the authors felt an unarticulated need to cloak the actual search terms identified.

What is at stake is twofold. First, science is a cumulative endeavor, and to stand on the shoulders of giants requires that scientists be able to continually assess work on which they are building (*25*). Second, accumulation of knowledge requires fuel in the form of data. There is a network of researchers waiting to improve the value of big data projects and to squeeze more actionable information out of these types of data. The initial vision regarding GFT—that producing a more accurate picture of the current prevalence of contagious diseases might allow for life-saving interventions—is fundamentally correct, and all analyses suggest that there is indeed valuable signal to be extracted.

Google is a business, but it also holds in trust data on the desires, thoughts, and the connections of humanity. Making money "without doing evil" (paraphrasing Google's motto) is not enough when it is feasible to do so much good. It is also incumbent upon academia to build institutional models to facilitate collaborations with such big data projects—something that is too often missing now in universities (*26*).

Use Big Data to Understand the Unknown. Because a simple lagged model for flu prevalence will perform so well, there is little room for improvement on the CDC data for model projections [this does not apply to other methods to directly measure flu prevalence, e.g., (*20*, *27*, *28*)]. If you are 90% of the way there, at most, you can gain that last 10%. What is more valuable is to understand the prevalence of flu at very local levels, which is not practical for the CDC to widely produce, but which, in principle, more finely granular measures of GFT could provide. Such a finely granular view, in turn, would provide powerful input into generative models of flu propagation and more accurate prediction of the flu months ahead of time (*29*–*33*).

Study the Algorithm. Twitter, Facebook, Google, and the Internet more generally are constantly changing because of the actions of millions of engineers and consumers. Researchers need a better understanding of how these changes occur over time. Scientists need to replicate findings using these data sources across time and using other data sources to ensure that they are observing robust patterns and not evanescent trends. For example, it is eminently feasible to do controlled experiments with Google, e.g., looking at how Google search results will differ based on location and past searches (*34*). More generally, studying the evolution of socio-technical systems embedded in our societies is intrinsically important and worthy of study. The algorithms underlying Google, Twitter, and Facebook help determine what we find out about our health, politics, and friends.

It's Not Just About Size of the Data. There is a tendency for big data research and more traditional applied statistics to live in two different realms—aware of each other's existence but generally not very trusting of each other. Big data offer enormous possibilities for understanding human interactions at a societal scale, with rich spatial and temporal dynamics, and for detecting complex interactions and nonlinearities among variables. We contend that these are the most exciting frontiers in studying human behavior. However, traditional "small data" often offer information that is not contained (or containable) in big data, and the very factors that have enabled big data are enabling more traditional data collection. The Internet has opened the way for improving standard surveys, experiments, and health reporting (*35*). Instead of focusing on a "big data revolution," perhaps it is time we were focused on an "all data revolution," where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world.

References and Notes

1. D. Butler, *Nature* **494**, 155 (2013).
2. D. R. Olson *et al.*, *PLOS Comput. Biol.* **9**, e1003256 (2013).
3. A. McAfee, E. Brynjolfsson, *Harv. Bus. Rev.* **90**, 60 (2012).
4. S. Goel *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17486 (2010).
5. A. Tumasjan *et al.*, in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Atlanta, Georgia, 11 to 15 July 2010 (Association for Advancement of Artificial Intelligence, 2010), pp. 178.
6. J. Bollen *et al.*, *J. Comput. Sci.* **2**, 1 (2011).
7. F. Ciulla *et al.*, *EPJ Data Sci.* **1**, 8 (2012).
8. P. T. Metaxas *et al.*, in *Proceedings of PASSAT—IEEE Third International Conference on Social Computing*, Boston, MA, 9 to 11 October 2011 (IEEE, 2011), pp. 165; doi:10.1109/PASSAT/SocialCom.2011.98.
9. D. Lazer *et al.*, *Science* **323**, 721 (2009).
10. A. Vespignani, *Science* **325**, 425 (2009).
11. G. King, *Science* **331**, 719 (2011).
12. D. Boyd, K. Crawford *Inform. Commun. Soc.* **15**, 662 (2012).
13. J. Ginsberg *et al.*, *Nature* **457**, 1012 (2009).
14. S. Cook *et al.*, *PLOS ONE* **6**, e23610 (2011).
15. P. Copeland *et al.*, *Int. Soc. Negl. Trop. Dis.* **2013**, 3 (2013).
16. C. Viboud *et al.*, *Am. J. Epidemiol.* **158**, 996 (2003).
17. W. W. Thompson *et al.*, *J. Infect. Dis.* **194** (Suppl. 2), S82–S91 (2006).
18. I. M. Hall *et al.*, *Epidemiol. Infect.* **135**, 372 (2007).
19. J. B. S. Ong *et al.*, *PLOS ONE* **5**, e10036 (2010).
20. J. R. Ortiz *et al.*, *PLOS ONE* **6**, e18687 (2011).
21. Organizing lists of related searches, Google; http://insidesearch.blogspot.com/2011/06/organizing-lists-of-related-searches_16.html.
22. Improving health searches, because your health matters, Google; http://insidesearch.blogspot.com/2012/02/improving-health-searches-because-your.html.
23. E. Mustafaraj, P. Metaxas, in *Proceedings of the Web-Sci10*, Raleigh, NC, 26 and 27 April 2010 (Web Science Trust, 2010); http://journal.webscience.org/317/.
24. J. Ratkiewicz *et al.*, in *Proceedings of 5th International AAAI Conference on Weblogs and Social Media*, San Francisco, CA, 7 to 11 August 2011 (AAAI, 2011), p. 297–304.
25. G. King, *PS Polit. Sci. Polit.* **28**, 443 (1995).
26. P. Voosen, *Chronicle of Higher Education*, 13 September 2013; http://chronicle.com/article/Researchers-Struggle-to-Secure/141591/.
27. R. Lazarus *et al.*, *BMC Public Health* **1**, 9 (2001).
28. R. Chunara *et al.*, *Online J. Public Health Inform.* **5**, e133 (2013).
29. D. Balcan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21484 (2009).
30. D. L. Chao *et al.*, *PLOS Comput. Biol.* **6**, e1000656 (2010).
31. J. Shaman, A. Karspeck, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 20425 (2012).
32. J. Shaman *et al.*, *Nat. Commun.* **4**, 2837 (2013).
33. E. O. Nsoesie *et al.*, *PLOS ONE* **8**, e67164 (2013).
34. A. Hannak *et al.*, in *Proceedings of the 22nd International World Wide Web Conference*, Rio de Janeiro, 13 to 17 May 2013 (Association for Computing Machinery, New York, 2013), pp. 527–538.
35. A. J. Berinsky *et al.*, *Polit. Anal.* **20**, 351–368 (2012).

Acknowledgments: This research was funded, in part, by NSF grant no. 1125095 Army Research Office (ARO) grant no. W911NF-12-1-0556, and, in part, by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract D12PC00285. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF ARO/IARPA, DoI/NBE, or the U.S. government. See SM for data and methods.

Supplementary Materials

www.sciencemag.org/content/343/6176/page/suppl/DC1

10.1126/science.1248506

POLITICAL SCIENCE

Fake news on Twitter during the 2016 U.S. presidential election

Nir Grinberg[1,2]*, Kenneth Joseph[3]*, Lisa Friedland[1]*, Briony Swire-Thompson[1,2], David Lazer[1,2]†

The spread of fake news on social media became a public concern in the United States after the 2016 presidential election. We examined exposure to and sharing of fake news by registered voters on Twitter and found that engagement with fake news sources was extremely concentrated. Only 1% of individuals accounted for 80% of fake news source exposures, and 0.1% accounted for nearly 80% of fake news sources shared. Individuals most likely to engage with fake news sources were conservative leaning, older, and highly engaged with political news. A cluster of fake news sources shared overlapping audiences on the extreme right, but for people across the political spectrum, most political news exposure still came from mainstream media outlets.

In 1925, *Harper's Magazine* published an article titled "Fake news and the public," decrying the ways in which emerging technologies had made it increasingly difficult to separate rumor from fact (*1*). Nearly a century later, fake news has again found its way into the limelight, particularly with regard to the veracity of information on social media and its impact on voters in the 2016 U.S. presidential election. At the heart of these concerns is the notion that a well-functioning democracy depends on its citizens being factually informed (*2*). To understand the scope and scale of misinformation today and most effectively curtail it going forward, we need to examine how ordinary citizens experience misinformation on social media platforms.

To this end, we leveraged a panel of Twitter accounts linked to public voter registration records to study how Americans on Twitter interacted with fake news during the 2016 election season. Of primary interest are three simple but largely unanswered questions: (i) How many stories from fake news sources did individuals see and share on social media? (ii) What were the characteristics of those who engaged with these sources? (iii) How did these individuals interact with the broader political news ecosystem? Initial reports were alarming, showing that the most popular fake news stories in the last 3 months of the presidential campaign generated more shares, reactions, and comments on Facebook than the top real news stories (*3*). However, we do not yet know the scope of the phenomenon, in part because of the difficulty of reliably measuring human behavior from social media data (*4*). Existing studies of fake news on social media have described its spread within platforms (*5*, *6*) and highlighted the disproportionate role played by automated accounts (*7*), but they have been unable to make inferences about the experiences of ordinary citizens.

Outside of social media, fake news has been examined among U.S. voters via surveys and web browsing data (*8*, *9*). These methods suggest that the average American adult saw and remembered one or perhaps several fake news stories about the 2016 election (*8*), that 27% of people visited a fake news source in the final weeks before the election, and that visits to these sources constituted only 2.6% of hard news site visits (*9*). They also show a persistent trend of conservatives consuming more fake news content, with 60% of fake news source visits coming from the most conservative 10% of Americans (*9*). However, because social media platforms have been implicated as a key vector for the transmission of fake news (*8*, *9*), it is critical to study what people saw and shared directly on social media.

Finally, social media data also provide a lens for understanding viewership patterns. Previous studies of the online media ecosystem have found evidence of insulated clusters of far-right content (*10*), rabbit holes of conspiratorial content (*11*), and tight clusters of geographically dispersed content (*12*). We wish to understand how fake news sources were positioned within this ecosystem. In particular, if people who saw content from fake news sources were isolated from mainstream content, they may have been at greater risk of adopting misinformed beliefs.

Data and definitions

Fake news sources

We follow Lazer *et al.* (*13*), who defined fake news outlets as those that have the trappings of legitimately produced news but "lack the news media's editorial norms and processes for ensuring the accuracy and credibility of information." The attribution of "fakeness" is thus not at the level of the story but at that of the publisher [similar to (*9*)].

We distinguished among three classes of fake news sources to allow comparisons of different operational definitions of fake news. The three classes correspond to differences in methods of generating lists of sources as well as perceived differences in the sites' likelihoods of publishing misinformation. We labeled as "black" a set of websites taken from preexisting lists of fake news sources constructed by fact-checkers, journalists, and academics (*8*, *9*) who identified sites that published almost exclusively fabricated stories [see supplementary materials (SM) section S.5 for details]. To measure fake news more comprehensively, we labeled additional websites as "red" or "orange" via a manual annotation process of sites identified by Snopes.com as sources of questionable claims. Sites with a red label (e.g., Infowars.com) spread falsehoods that clearly reflected a flawed editorial process, and sites with an orange label represented cases where annotators were less certain that the falsehoods stemmed from a systematically flawed process. There were 171 black, 64 red, and 65 orange fake news sources appearing at least once in our data.

Voters on Twitter

To focus on the experiences of real people on Twitter, we linked a sample of U.S. voter registration records to Twitter accounts to form a panel (see SM S.1). We collected tweets sent by the 16,442 accounts in our panel that were active during the 2016 election season (1 August to 6 December 2016) and obtained lists of their followers and followees (accounts they followed).

[1]Network Science Institute, Northeastern University, Boston, MA, USA. [2]Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA. [3]Department of Computer Science and Engineering, University at Buffalo, SUNY, Buffalo, NY, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: d.lazer@northeastern.edu

We compared the panel to a representative sample of U.S. voters on Twitter obtained by Pew Research Center (*14*) and found that the panel is largely reflective of this sample in terms of age, gender, race, and political affiliation (see SM S.2).

We estimated the composition of each panel member's news feed from a random sample of the tweets posted by their followees. We called these tweets, to which an individual was potentially exposed, their "exposures." We also analyzed the panel's aggregate exposures, in which, for example, a tweet from an account followed by five panel members was counted five times. We restricted our analysis to political tweets that contained a URL for a web page outside of Twitter (SM S.3 and S.4). Because we expected political ideology to play a role in engagement with fake news sources, we estimated the similarity of each person's feed to those of registered Democrats or Republicans. We discretized the resulting scores to assign people into one of five political affinity subgroups: extreme left (L*), left (L), center (C), right (R), and extreme right (R*). Individuals with less than 100 exposures to political URLs were assigned to a separate "apolitical" subgroup (SM S.10).

Results

Prevalence and concentration

When totaled across all panel members and the entire 2016 U.S. election season, 5.0% of aggregate exposures to political URLs were from fake news sources. The fraction of content from fake news sources varied by day (Fig. 1A), increasing (in all categories) during the final weeks of the campaign (SM S.7). Similar trends were observed in content sharing, with 6.7% of political URLs shared by the panel coming from fake news sources.

However, these aggregate volumes mask the fact that content from fake news sources was highly concentrated, both among a small number of websites and a small number of panel members. Within each category of fake news, 5% of sources accounted for more than 50% of exposures (Fig. 1B). There were far more exposures to red and orange sources than to black sources (2.4, 1.9, and 0.7% of aggregate exposures, respectively), and these differences were largely driven by a handful of popular red and orange sources. The top seven fake news sources—all red and orange—accounted for more than 50% of fake news exposures (SM S.5).

Figure 1, C and D, shows that content was also concentrated among a small fraction of panel members for all categories of fake news sources. A mere 0.1% of the panel accounted for 79.8% of shares from fake news sources, and 1% of panel members consumed 80.0% of the volume from fake news sources. These levels of concentration were not only high in absolute terms, they were also unusually high relative to multiple baselines both within and beyond politics on Twitter (SM S.15).

The "supersharers" and "superconsumers" of fake news sources—those accountable for 80% of fake news sharing or exposure—dwarfed typical users in their affinity for fake news sources and, furthermore, in most measures of activity. For example, on average per day, the median supersharer of fake news (SS-F) tweeted 71.0 times, whereas the median panel member tweeted only 0.1 times. The median SS-F also shared an average of 7.6 political URLs per day, of which 1.7 were from fake news sources. Similarly, the median superconsumer of fake news sources had almost 4700 daily exposures to political URLs, as compared with only 49 for the median panel member (additional statistics in SM S.9). The SS-F members even stood out among the overall supersharers and superconsumers, the most politically active accounts in the panel (Fig. 2). Given the high volume of posts shared or consumed by superspreaders of fake news, as well as indicators that some tweets were authored by apps, we find it likely that many of these accounts were cyborgs: partially automated accounts controlled by humans (*15*) (SM S.8 and S.9). Their tweets included some self-authored content, such as personal commentary or photos, but also a large volume of political retweets. For subsequent analyses, we set aside the supersharer and superconsumer outlier accounts and focused on the remaining 99% of the panel.

Who was exposed to fake news sources?

Excluding outliers, panel members averaged 204 potential exposures [95% confidence interval (CI): 185 to 224] to fake news sources during the last month of the campaign. If 5% of potential exposures were actually seen (*16*), this would translate to, on average, about 10 exposures (95% CI: 9 to 11) to fake news sources during that month. The average proportion of fake news sources (among political URLs) in an individual's feed was 1.18% (95% CI: 1.12 to 1.24%). However, there was a large and significant discrepancy between

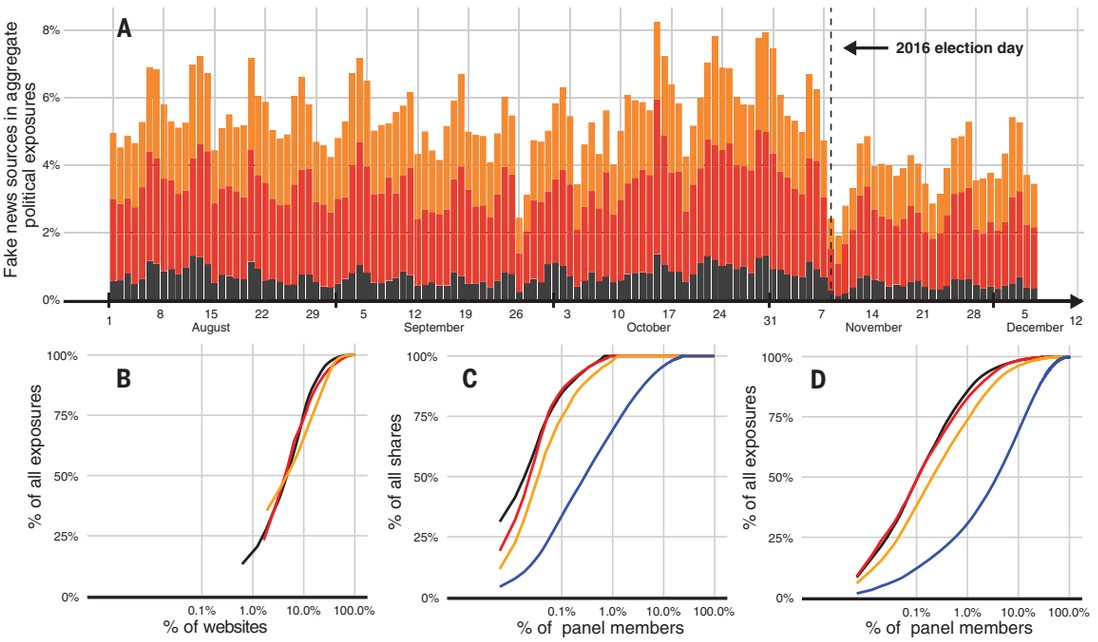


Fig. 1. Prevalence over time and concentration of fake news sources. (**A**) Daily percentage of exposures to black, red, and orange fake news sources, relative to all exposures to political URLs. Exposures were summed across all panel members. (**B** to **D**) Empirical cumulative distribution functions showing distribution of exposures among websites (B), distribution of shares by panel members (C), and distribution of exposures among panel members (D). The *x* axis represents percentage of websites or panel members responsible for a given percentage (*y* axis) of all exposures or shares. Black, red, and orange lines represent fake news sources; blue line denotes all other sources. This distribution was not comparable for (B) because of the much larger number of sources in its tail and the fundamentally different selection process involved.

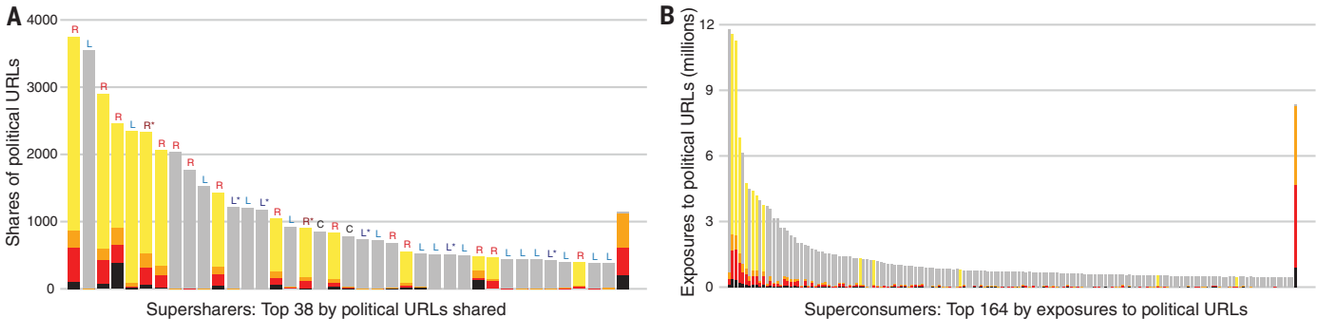


Fig. 2. Shares and exposures of political URLs by outlier accounts, many of which were also SS-F accounts. (**A**) Overall supersharers: top 1% among panelists sharing any political URLs, accounting for 49% of all shares and 82% of fake news shares. Letters above bars indicate political affinities. (**B**) Overall superconsumers: top 1% among panelists exposed to any political URLs, accounting for 12% of all exposures and 74% of fake news exposures. Black, red, and orange bars represent content from fake news sources; yellow or gray bars denote nonfake content (SS-F accounts are shown in yellow). The rightmost bar shows, for scale, the remainder of the panel's fake news shares (A) or exposures (B).

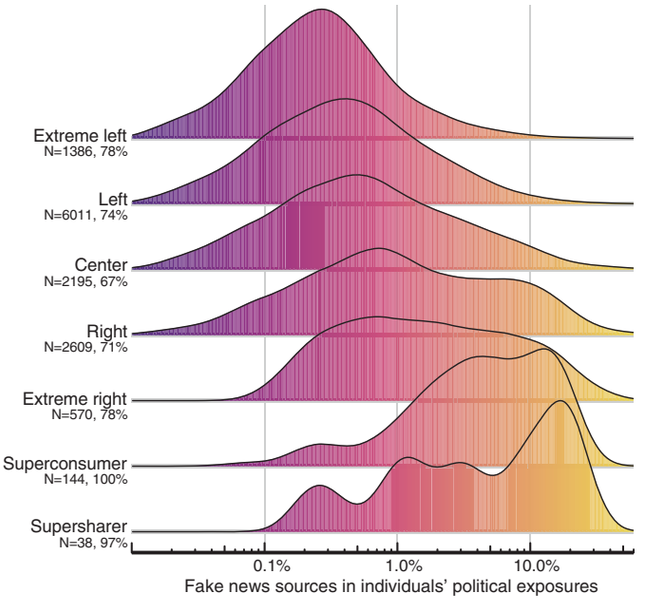


Fig. 3. Probability density estimates for the percentage of content from fake news sources in people's news feeds (for people with any fake news exposures). The number of individuals in each subgroup (N) and the percent with nonzero exposures to fake news sources are shown.

left and right ($P < 0.0001$). For example, people who had 5% or more of their political exposures from fake news sources constituted 2.5% of individuals on the left (L and L*) and 16.3% of the right (R and R*). See Fig. 3 for the distribution among all political affinity groups and SM S.10 for additional statistics.

According to binomial regressions fit separately to each political affinity group, the strongest predictors of the proportion of fake news sources in an individual's feed were the individual's age and number of political URLs in the individual's feed (Fig. 4, A and B, and SM S.11). A 10-fold increase in overall political exposures was associated with doubling the proportion of exposures to fake news sources (Fig. 4A)—that is, a 20-fold increase in the absolute number of exposures to fake news sources. This superlinear relationship holds for all five political affinity groups and suggests that a stronger selective exposure process exists for individuals with greater interest in politics. Figure 4B shows that age was positively and significantly associated with increased levels of exposure to fake news sources across all political groups.

Other factors were also associated with small increases in exposures to fake news sources: Men and whites had slightly higher rates, as did voters in swing states and voters who sent more tweets (excluding political URLs analyzed here). These findings are in line with previous work that showed concentration of polarizing content in swing states (*17*) and among older white men (*18*). However, effects for the above groups were small (less than one percentage point increase in proportion of exposures) and somewhat inconsistent across political groups.

Who shared fake news sources?

Political affinity was also associated with the sharing of content from fake news sources. Among those who shared any political content on Twitter during the election, fewer than 5% of people on the left or in the center ever shared any fake news content, yet 11 and 21% of people on the right and extreme right did, respectively ($P < 0.0001$) (see SM S.10). A logistic regression model showed that the sharing of content from fake news sources (as a binary variable) was positively associated with tweeting about politics, exposure to fake news sources, and political affinity, although the disparity across the political spectrum was smaller than suggested by univariate statistics (Fig. 4, C to E, and SM S.12). Other factors such as age and low ratio of followers to followees were also positively associated with sharing fake news sources, but effect sizes were small.

Next, we examined rates of sharing per exposure, modeling the likelihood that an individual would share a URL after being exposed to it (SM S.13). Conditioned on exposure to a politically congruent source, there were no significant differences in sharing rates between liberals and conservatives and across fake and nonfake sources (Fig. 4, F to I). Incongruent sources were shared at significantly lower rates than congruent sources ($P < 0.01$), with two exceptions. First, conservatives shared congruent and incongruent nonfake sources at similar rates. Second, we lacked the statistical power to assess sharing rates of conservatives exposed to liberal fake news, owing to the rarity of these events.

These findings highlight congruency as the dominant factor in sharing decisions for political news. This is consistent with an extensive body of work showing that individuals evaluate belief-incongruent information more critically than belief-congruent information (*19*). Our results suggest that fake news may not be more viral than real news (*5*).

Fake news and the media ecosystem

We extracted relationships between news websites according to a measure of shared audience and identified distinct group structure in a network of these relationships (Fig. 5). In a manner similar to other analyses of media coconsumption (*12*), we constructed this coexposure network

by using a technique that identifies statistically significant connections between sites (*20*) (SM S.14). We identified four groups of websites in this network: Groups 1 to 3 were large subsets of nodes consistently grouped together by three different clustering algorithms, whereas Group 4 comprised the remaining nodes.

Group 1, a collection of nationally relevant mainstream media outlets, contained only 18.4% of sites but was responsible for the vast majority of individuals' political URL exposures, ranging from an average of 72% (extreme right) to 86% (extreme left) across political affinity subgroups. These sites were mostly centric in political leaning, whereas Group 2 was significantly more conservative, and Group 3 was significantly more liberal. Many sites in Group 2 were fake news sources (68.8%), substantially more than in Group 1 (3.6%), Group 3 (2.0%), or Group 4 (11.4%). Exposure to Group 2, unlike to Group 1, varied considerably by political affinity. For individuals on the extreme right, Group 2 generated the majority (64.2%) of exposures outside of Group 1, as compared with 38.6, 22.2, 13.9, and 8.0% for those on the right, center, left, and extreme left, respectively.

Further, the high network density within Group 2 (Fig. 5) reflects that consumers of content from fake news sources were often exposed to multiple fake news sources. Of the 7484 nonoutlier panel members exposed to at least two fake news URLs, 95.6% of them saw URLs from at least two fake news sources, and 56.4% encountered URLs from at least five. In summary, fake news sources seem to have been a niche interest: Group 2 accounted for only a fraction of the content seen by even the most avid consumers of fake news but nonetheless formed a distinct cluster of sites, many of them fake, consumed by a heavily overlapping audience of individuals mostly on the right.

Discussion

This study estimated the extent to which people on Twitter shared and were exposed to content from fake news sources during the 2016 election season. Although 6% of people who shared URLs with political content shared content from fake news sources, the vast majority of fake news shares and exposures were attributable to tiny fractions of the population. Though previous work has shown concentration of volume both in political conversations on Twitter (*21*) and in fake news consumption (*9*), the extreme levels we observed are notable. For the average panel member, content from fake news sources constituted only 1.18% of political exposures, or about 10 URLs during the last month of the election campaign. These averages are of similar magnitude to estimates from previous work (*8*, *9*), which is noteworthy given the vastly different study populations and methodologies. As in these studies, we found that the vast majority of political exposures, across all political groups, still came from popular nonfake news sources. This is reassuring in contrast to claims of political echo chambers (*22*) and fake news garnering more engagement than real news during the election (*3*).

We identified several subpopulations that deserve particular attention when devising interventions. Within human populations, it will be important to understand the mechanisms that lead different groups to engage with fake news sources. For example, heightened engagement by older adults could result from cognitive decline (*23*), digital media literacy, stronger motivated reasoning, or cohort effects. Among Twitter accounts, despite stringent measures to exclude bot accounts, our remaining sample still included partially automated cyborgs. Unlike bots, which may struggle to attract human audiences, these cyborg accounts are embedded within human social networks. Given the increasing sophistication of automation tools available to the average user and the increasing volume and coordination of online campaigns, it will be important to study cyborgs as a distinct category of automated activity (*14*, *22*). Regarding limitations to this study, note that our findings derive from a particular sample on Twitter, so they might not generalize to other platforms (*24*). Additionally,

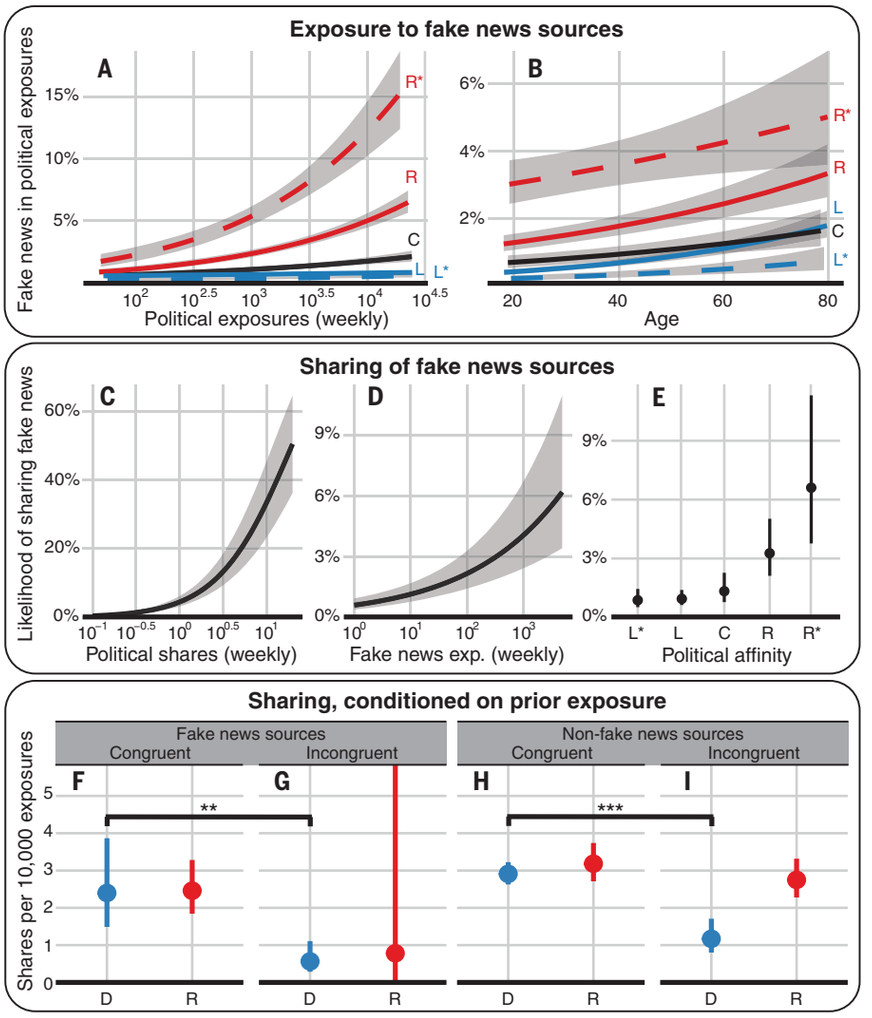


Fig. 4. Key individual characteristics associated with exposure to and sharing of fake news sources. The proportion of an individual's political exposures coming from fake news sources as a function of (**A**) number of political exposures, excluding fake news sources, and (**B**) age. Estimates are based on binomial regression models fitted separately to each political affinity subgroup. Blue, liberal; black, center; red, conservative. (**C** to **E**) An individual's likelihood of sharing one or more URLs from fake news sources as a function of (C) number of shares of political URLs, (D) number of exposures to fake news sources, and (E) political affinity. (**F** to **I**) Likelihood of a liberal (D) or conservative (R) individual sharing a political URL to which they have been exposed, depending on the political congruency and veracity of the source: (F) congruent and fake, (G) incongruent and fake, (H) congruent and nonfake, and (I) incongruent and nonfake. Brackets indicate significantly different pairs: $^{**}P < 0.01$, $^{***}P < 0.001$. All estimates and 95% CIs [gray shaded regions in (A) to (D); line segments in (E) to (I)] are based on regression models specified in SM S.11 to S.13, with the remaining model variables held constant to their median or most common level.

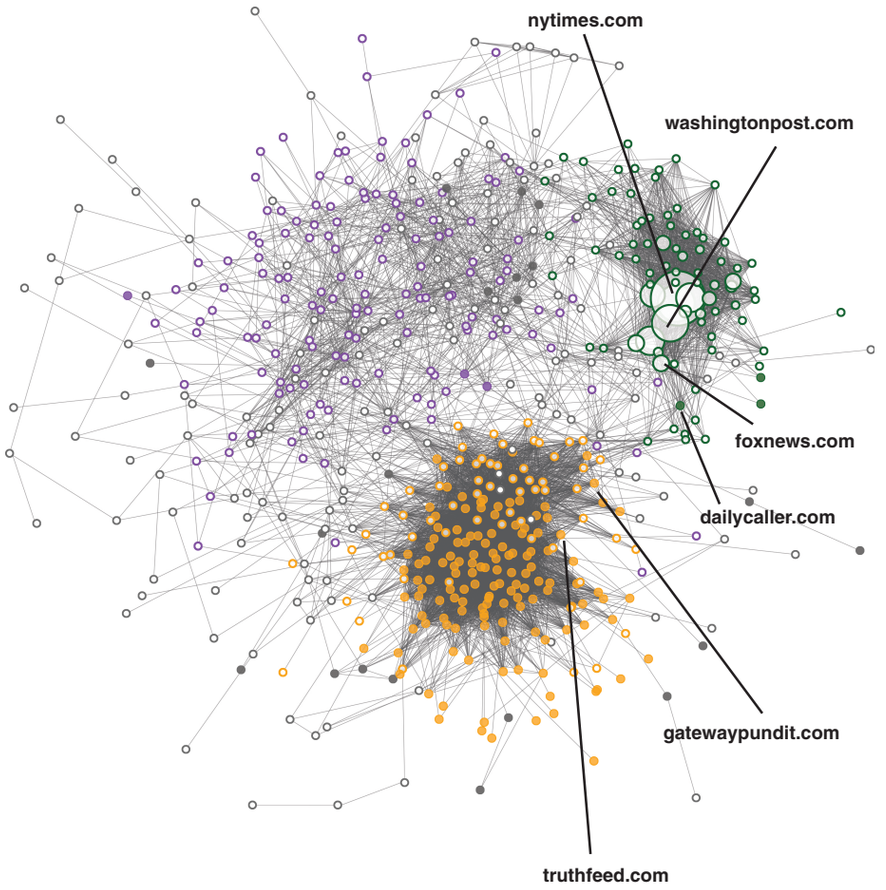


Fig. 5. Coexposure network. Each node is a political news, blog, or fact-checking website. Edges link pairs of sites where an unusually high number of (nonoutlier) panel members were exposed to content from both sites, controlling for the popularity of each site. Filled nodes represent fake news sources. Node colors indicate groups (1, green; 2, orange; 3, purple; 4, gray) identified via an ensemble of clustering algorithms. Sites with the highest exposures are sized slightly larger. See fig. S10 for node labels.

although our sample roughly reflects the demographics of registered voters on Twitter, it might systematically differ from that population in other ways.

Our findings suggest immediate points of leverage to reduce the spread of misinformation. Social media platforms could discourage users from following or sharing content from the handful of established fake news sources that are most pervasive. They could also adopt policies that disincentivize frequent posting, which would be effective against flooding techniques (*25*) while affecting only a small number of accounts. For example, platforms could algorithmically demote content from frequent posters or prioritize users who have not posted that day. For illustrative purposes, a simulation of capping political URLs at 20 per day resulted in a reduction of 32% of content from fake news sources while affecting only 1% of content posted by nonsupersharers. (SM S.15). Finally, because fake news sources have shared audiences, platforms could establish closer partnerships with fact-checking organizations to proactively watch top producers of misinformation and examine content from new sites that emerge in the vicinity of fake news sources in a coexposure network. Such interventions do raise the question of what roles platforms should play in constraining the information people consume. Nonetheless, the proposed interventions could contribute to delivering corrective information to affected populations, increase the effectiveness of corrections, foster equal balance of voice and attention on social media, and more broadly enhance the resiliency of information systems to misinformation campaigns during key moments of the democratic process.

REFERENCES AND NOTES

1. E. McKernon, *Harper's Magazine* **151**, 528–536 (1925).
2. J. H. Kuklinski, P. J. Quirk, J. Jerit, D. Schwieder, R. F. Rich, *J. Polit.* **62**, 790–816 (2000).
3. C. Silverman, "This analysis shows how viral fake election news stories outperformed real news on Facebook" (2016); www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.
4. D. Ruths, J. Pfeffer, *Science* **346**, 1063–1064 (2014).
5. S. Vosoughi, D. Roy, S. Aral, *Science* **359**, 1146–1151 (2018).
6. M. Del Vicario *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 554–559 (2016).
7. C. Shao *et al.*, *Nat. Commun.* **9**, 4787 (2018).
8. H. Allcott, M. Gentzkow, *J. Econ. Perspect.* **31**, 211–236 (2017).
9. A. Guess, B. Nyhan, J. Reifler, "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign" (2018); www.dartmouth.edu/~nyhan/fake-news-2016.pdf.
10. Y. Benkler, R. Faris, H. Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford Univ. Press, 2018).
11. K. Starbird, in *Proceedings of the 11th International AAAI Conference on Web and Social Media* (AAAI, 2017), pp. 230–239.
12. A. L. Schmidt *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 3035–3039 (2017).
13. D. M. J. Lazer *et al.*, *Science* **359**, 1094–1096 (2018).
14. S. Greenwood, A. Perrin, M. Duggan, "Social media update 2016" (Pew Research Center, 2016).
15. Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, *IEEE Trans. Depend. Secure Comput.* **9**, 811–824 (2012).
16. L. X. Wang, A. Ramachandran, A. Chaintreau, in *Workshops of the 10th International AAAI Conference on Web and Social Media* (AAAI, 2016), pp. 108–113.
17. P. N. Howard, B. Kollanyi, S. Bradshaw, L.-M. Neudert, "Social media, news and political information during the US election: Was polarizing content concentrated in swing states?" *Data Memo* 2017.8 (Oxford Project on Computational Propaganda, 2017).
18. S. Bhatt, S. Joglekar, S. Bano, N. Sastry, in *Proceedings of the 27th International World Wide Web Conference* (ACM, 2018), pp. 545–554.
19. Z. Kunda, *Psychol. Bull.* **108**, 480–498 (1990).
20. N. Dianati, *Phys. Rev. E* **93**, 012304 (2016).
21. E. Mustafaraj, S. Finn, C. Whitlock, P. T. Metaxas, in *Proceedings of the 3rd International Conference on Social Computing* (IEEE, 2011), pp. 103–110.
22. C. R. Sunstein, *#Republic: Divided Democracy in the Age of Social Media* (Princeton Univ. Press, 2018).
23. B. Swire, U. K. H. Ecker, S. Lewandowsky, *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 1948–1961 (2017).
24. Z. Tufekci, in *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (AAAI, 2014), pp. 505–514.
25. M. E. Roberts, *Censored: Distraction and Diversion Inside China's Great Firewall* (Princeton Univ. Press, 2018).
26. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Public Replication Package for Fake news on Twitter. Zenodo (2019); https://doi.org/10.5281/zenodo.2483311.
27. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Protected Replication data for Fake news on Twitter. Zenodo (2019); https://doi.org/10.5281/zenodo.2485428.

ACKNOWLEDGMENTS

We thank L. Adamic, Y. Benkler, S. McCabe, and the three anonymous reviewers for thoughtful feedback on the manuscript and TargetSmart for access to voter data. The research was approved by Northeastern University's Institutional Review Board. All opinions expressed in this article are those of the authors alone. **Funding:** D.L. acknowledges support by the ESRC ES N012283/1 and ARO W911NF-12-1-0556. **Author contributions:** D.L. conceived of the study. N.G., K.J., and L.F. collected and processed data, carried out statistical modeling, and produced visualizations. N.G., K.J., L.F., and B.S.-T. performed literature review and annotated data. All authors devised analyses and wrote and revised the paper. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Aggregate data and code from this study are freely available at Zenodo (*26*). Deidentified individual-level data are also available at Zenodo (*27*) upon signing a usage agreement stating that: (i) you shall not attempt to identify, reidentify, or otherwise deanonymize the dataset and (ii) you shall not further share, distribute, publish, or otherwise disseminate the dataset without Northeastern University's prior written approval.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/363/6425/374/suppl/DC1
Supplementary Text
Figs. S1 to S14
Tables S1 to S7
References (*28*–*68*)

23 May 2018; accepted 2 January 2019
10.1126/science.aau2706

Don't Let Me Be #Misunderstood: Predicting User Word Preferences According to Cognitive, Physical and Domain Constraints

ABSTRACT

Twitter hashtags can be viewed as neologisms, new phrases coined by the user community as the need arises. Given the 'need' for a new phrase, a number of optional hashtags can be coined, some of which gain popularity while others remain marginal. Assuming an inherent tension between the 'principle of least effort' and the speakers goal to be well understood, Twitter is an ideal domain for the investigation of these dynamics between constraints as it also imposes a length constraint of 140 characters, encouraging the user to try and strike a balance between clearness and efficiency.

By modeling the interplay between the various constraints we get a glimpse into the processes that shape the way we produce language and coin new words. In order to learn the interactions between these constraints we cast the problem as a ranking task. We use Gradient Boosted Trees adapted for learning ranking functions in order to predict the neologisms to be accepted. Our results outperform several baseline algorithms including SVM-rank, while maintaining high interpretability, thus our model's prediction power can be easily translated into linguistic insights, facilitating an improved user experience and maximizing diffusion processes of commercial and ideological online campaigns.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

ACM proceedings, LaTeX, text tagging

1. INTRODUCTION

Neologisms are newly coined words or phrases commonly used by a discourse community. Some neologisms that are already lexicalized and absorbed into the standard language are 'brunch', 'cyberspace', 'laser', 'mcjob', 'internet' and 'meme'[1]. While neologisms pose many problems to Natural Language Processing (NLP) systems [7], the formation and acceptance patterns of neologisms are of a major interest from pure (socio/psycho-) linguistics perspective [3, 11].

Traditionally, neologism are studied under a qualitative paradigm in which scholars examine the etymology of selected words [1, 2]. Quantitative approaches to language change are recently gaining popularity as resources like Google Books are becoming available [18, 21, 23]. All these approaches examine the language change in hundreds of years perspective.

The evolution of language, however, is continuous and takes place in much shorter time frames [11]. The massive data stream available in social networks is giving rise to the possibility of studying language change in real time.

Neologisms, similarly to the communication process in general, embody inherent tension between two main forces: the speaker's goal to be well understood and the 'principle of least effort', which argues that a word's frequency is inversely proportional to its length [33], coinciding with the theory of effective communication [28]. This approach to language was criticized by Chomsky [5, 6] among others. In effect, in the past decades the effective communication framework was not widely applied to natural language. Recently, however, this framework regained interest, demonstrating that content is a good predictor for word length [24] and that speakers choose shorter words in predictive contexts [20].

In this paper we examine to what extent the principle of least effort holds and under what circumstances it is violated. Furthermore, we wish to model how various constraints interact. We cast these cognitive linguistics questions as a prediction task in which we train an algorithm to predict the preferential ranking of competing words and strings describing the same concept. Specifically, we look at many small sets of competing 'interchangeable' Twitter hashtags, predicting which of the hashtags in each set will be used more frequently.

Twitter hashtags can be viewed as new phrases coined by the user community as the need arises. Given the need for such a new phrase, a number of optional hashtags can be

[1]The term *meme* was actually coined by Richard Dawkins back in 1976, referring to an idea that spreads itself via imitations with some variations, such as "tunes, ideas, catchphrases, clothes fashions, ways of making pots or of building arches" [10].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

coined (e.g. *#bindersFullOfWomen*[2], *#bindersfullofwomen*[3], *#romneysBinders*, etc.), some of which gain popularity while others remain marginal. By modeling the interplay between various linguistic constraints we get a glimpse into the processes that shape the way we produce language and coin new words.

An immediate practical application of these linguistic insights and the learned models is the contribution to artificial engineering of catchier hashtags for various purposes such as commercial and ideological campaigns.

In order to train the ranking model we adapt the GBrank framework proposed by [32] to the task at hand. In the GBrank framework, a Gradient Boosted Trees (GBT) algorithm, typically used to fit a regression function, is modified to enable ranking of preference pairs. The conversion of learning a regression function to learning a ranking function is achieved via a modified update mechanism in the learning process. Using the GBrank framework, we achieve superior results in the prediction task. The algorithm produces nonlinear models in various levels of complexity while maintaining interpretability, that can be easily translated into linguistic insights such as the contexts in which efficiency is preferred and what language conventions are more prone of violation.

2. RELATED WORK

Neologisms are understudied from a computational perspective. Cook and Stevenson show that knowledge of etymology and word formation can be encoded and exploited for automatic identification of neologisms in blog corpora and text messages [8, 7].

The acceptance of a neologism can be viewed as a propagation process. Information diffusion in online social networks has earned much interest in the past decade; however, the main foci of this body of works are the graph topology, the strength of connections between graph nodes and the activity level of individual nodes [16, 4, 31, 17, 26]. In this work we shift the focus from the graph topology to the linguistic traits that facilitate efficient diffusion of the information.

A few works address some linguistic aspects that affect the information spread. Analyzing Twitter hashtags of three topics, Chuna et al. [9] find correlation between a hashtag's length and its frequency. Tsur and Rappaport [30] predict hashtag popularity in a range of time frames given some linguistic features (e.g. length in characters and words, lexicality, semantic class) along with graph topology, and early temporal trends. Lin et al. [19] study "winners" vs. "also–rans" hashtags in the context of exogenous shocks.

While in the setting promoted by Tsur and Rappoport each hashtag "competes" for popularity with each and every other hashtag, we believe that small sets of interchangeable hashtags should be addressed separately in order to gain linguistic insights. We therefore propose a ranking task in which we learn the user's preference between similar candidates. Unlike Chuna et al. [9] and Lin et al. [19], we analyze thousands of such sets and consider a much broader set of linguistic features. Unlike the triad [9, 30, 19], we allow the creation of non linear models, hypothesizing that the relations between linguistic constraints are not linear. This hypothesis is well supported by our results.

We also note that our work is loosely inspired by the linguistic Optimaliy Theory paradigm [25] in which observed forms of language are the result of a complex dynamic between competing constraints. Our formulation and algorithmic approach takes a different path, though.

The few quantitative works related to neologism and word evolution are surveyed in the Section 1. To the best of our knowledge, this is the first computational work to address textual neologisms in the light of competition between linguistic constraints.

3. RANKING ALGORITHM

There are two requirements we want our algorithmic framework to meet: (1) The model should allow accurate prediction of speakers preferences. and (2) The model should be as interpretable as possible in order to gain some (psycho /cognitive)-linguistic insights about the processes governing the production of the text.

In order to meet these two criteria, we train a model based on linguistic and domain-constraint features. The model should allow us to learn the relations between various constraints. In the light of these requirements we formulate the problem as a ranking task.

Order and preferences: Given a set of hashtags HT and a corresponding set of target values Y,we define a preference order $ht_i \succ ht_j$ if $y_i > y_j$. We denote the preference pair as an ordered set $\langle ht_i, ht_j \rangle$. lAthough the values in Y can introduce a preference order over all pairs in $HT \times HT$, only a partial set of pairs is acknowledged, namely only pairs in which both elements belong to the same subset. Given C, a partition of HT to subsets of vectors (hashtags) such that $\cup_k c^k = C$, we define the set of relevant pairs as:

$$S = \{(ht_i^c, ht_j^c) | i \neq j, ht_i^c \succ ht_j^c, \quad c \in C\} \tag{1}$$

Notation: For ease of reading, where context allows, we use ht_i to represent both the $i'th$ hashtag in a set and its corresponding vector. We also simplify the notation when possible, omitting the c from notations like ht_i^c, given that the preference relation is only meaningful for hashtags from the same set c.

Ranking function:

We are interested in learning a function $h \in H$ such that h projects $h(ht_i^c) \geq h(ht_j^c)$ for any given preference pair $ht_i^c \succ ht_j^c$

Given a set of preferences S as defined in Equation 1, the objective loss function is then defined as

$$L(h) = \frac{1}{2} \sum_{\langle ht_i, ht_j \rangle \in S} (max\{0, h(ht_j) - h(ht_i)\})^2 \tag{2}$$

meaning there is a cost only for explicit violation of the true order[4].

[2]The phrase 'binders full of women' was used by Mitt Romney in the second presidential debate of 2012 in his response to a question about pay equity. The phrase quickly turned viral.

[3]We view different capitalizations of the same sequence of characters as different hashtags as they require different typing efforts (from the writer) and different cognitive efforts (from the reader).

[4]Theoretically, an artificially optimal solution could be achieved by fixing $h(ht) = b$ for any constant b, however, this solution is impossible given the proposed algorithm.

In the classic boosting framework, a set of base (weak) learners can be combined to create a single stronger learner [27]. Zheng et al. [32] outline a regression framework for learning of ranking functions using relative relevance judgments and click through data. They present GBrank – an algorithm based on gradient boosting that is used for reranking the web pages returned for a search engine query.

Our task can be viewed in a similar way: given a concept, impose a preference order on the possible hashtags that may be used in lieu of that concept. In this paper we adapt and modify the GBrank framework to suit the task at hand.

GBrank iterates between two phases: (1) fitting a regression function, and (2) rescaling the regression target values in order to minimize the ranking loss function. Each iteration adds a new base learner thus the combination of the base learners is expected to boost performance. Rescaling is performed in a gradient descent manner.

It is important to note that each phase of the algorithm optimizes a different objective function. Phase one fits a regression function by minimizing MSE (or any suitable loss function). This optimization can actually work against our goal of learning ordered pairs. The gradient rescaling in phase two is expected to calibrate the target values so that the new regression function keeps the given order.

Given a preference pair $\langle ht_i, ht_j \rangle$, the corresponding gradients of $L(h)$ (Equation 2) are: $max\{0, ht_j - ht_i\}$ and $max\{0, ht_i - ht_j\}$, hence the gradient is different than zero only if there is explicit violation of the order of the preference pair.

In order to use these gradients in the ranking scheme we add (or subtract) some fixed values τ to the regression target values. Each consecutive base learner is a regression tree that maximizes the fit with the new target values.

Different base learners can be assigned different weights in order to introduce regularization and prevent overfitting.

3.1 Modified GBrank algorithm

Our modified algorithm is presented in Table 1.

The reader should note the difference between sets. While T, T^+ and T^- that are used for learning a regression function (g_k) contain pairs, each pair consists of a vector and its (adjusted) currently predicted value, sets S, S^+ and S^- are sets of pairs of vectors and are used in order to adjust the currently predicted values to reflect the desired ranking.

In our experiments we use $\alpha = 1/2$ and $\beta = 1$, thus the training set used in the construction of each new tree is composed of the hashtags from half of the correctly ranked pairs and hashtags from half of the wrongly ranked pairs.

The major modification introduced to GBrank is that in stage 4 we use a weighted majority vote while [32] use the following linear combination of the $i = 1, ..., k$ regression trees $g_i(ht)$ as a 'pseudo' ranking function:

$$h_k(ht) = \frac{k \cdot h_{k-1}(ht) + \eta_k \cdot g_k(ht)}{k+1}$$

Experimenting with this pseudo ranking function, the learner indeed managed to minimize the regression loss function, however as model complexity grows (more trees/levels) the results get worse due to over correction by subsequent trees trying to compensate for the growing target values (stage 3). Convergence to sub optimal solution is finally achieved due to the shrinkage factor η_k, practically preventing further changes in the prediction. We therefore propose the weighted majority function $Rank(\cdot)$ in which each tree provides its own preference prediction as outlined above. In our experiments the modified GBrank achieved improvement of about 40% over the original GBrank proposed by Zheng et al. Another modification we introduced is the subsampling in stage 3. Subsampling is used as it is reported to improve results significantly in many settings [12]. The training set of each new base learner is combined from a sampled subset of the wrongly ordered pairs and a sampled subset of the the original training set (with the adjusted target values).

Input: A set of vectors and their corresponding target values $HT : Y$; a set of preferences S defined in Equation 1.

Output: A function $Rank : HT^c \times HT^c \longrightarrow \{1, -1\}$, where $Rank(ht, ht') = 1 \;\; iff \;\; ht \succ ht'$.

Initialize: "Guess" initial target values for g_0, e.g. $g_0(ht) = 0$ for each ht.

For $k = 1, 2, 3, ...$

1. $\hat{Y}^{k-1} = \langle \hat{y}_1^{k-1}, \hat{y}_2^{k-1}, \hat{y}_3^{k-1}, ... \rangle$
   where, $\hat{y}_l^{k-1} = g_{k-1}(ht_l)$
2. Divide S to S^+ and S^- as follows:
   $S^+ = \{(ht_i, ht_j) \in S | \hat{y}_i^{k-1} - \hat{y}_j^{k-1} > \beta\}$
   $S^- = \{(ht_i, ht_j) \in S | \hat{y}_j^{k-1} - \hat{y}_i^{k-1} > \beta\}$
   where β is some non-negative value.
3. Use Gradient Boosted Trees to fit a regression function $g_k(ht)$ on the following training set
   $T = sample(T^+, \alpha) \cup sample(T^-, \alpha)$
   where,
   $T^+ = \{(ht_i, \hat{y}_i^{k-1}), (ht_j, \hat{y}_j^{k-1}) | (ht_i, ht_j) \in S^+\}$
   $T^- = \{(ht_i, \hat{y}_j^{k-1} + \tau), (ht_j, \hat{y}_i^{k-1} - \tau) | (ht_i, ht_j) \in S^-\}$
4. Update the ensemble $Rank_k(ht, ht')$ with r_k:

$$Rank_k(ht, ht') = \begin{cases} 1 & \sum_{i=1}^{k} \eta_k r_i(ht, ht') > 0 \\ -1 & \sum_{i=1}^{k} \eta_k r_i(ht, ht') < 0 \end{cases}$$

where

$$r_k(ht, ht') = \begin{cases} 1 & g_k(ht) > g_k(ht') \\ -1 & g_k(ht) < g_k(ht') \\ 0 & otherwise \end{cases}$$

Table 1: Modified GB rank algorithm.

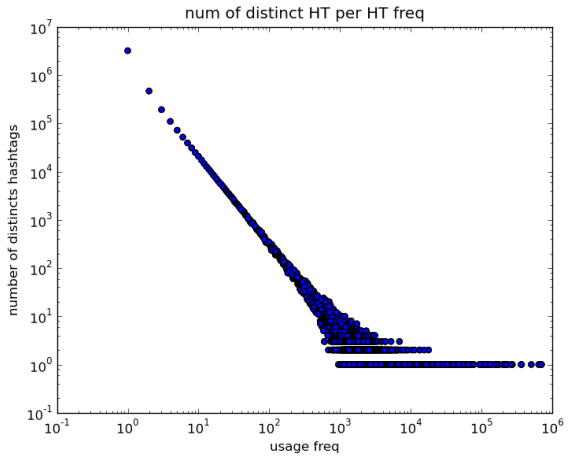
4. DATA

4.1 Twitter Corpus and statistics

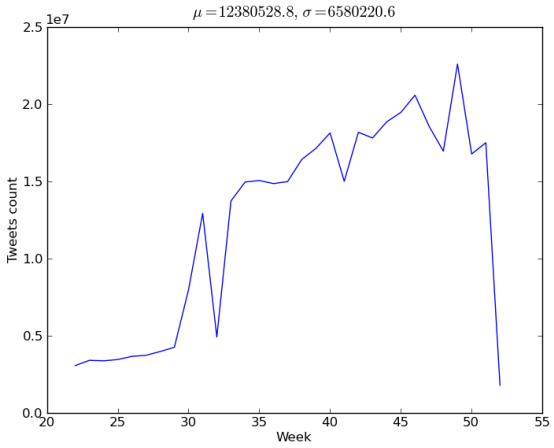
A Twitter posting is called a *tweet*. A tweet is restricted to 140 characters in length. This length constraint makes characters "expensive", hence tweets present an informal language and introduce many abbreviations. Twitter allows the use of two meta characters: '@' marking a user name (e.g. @BarackObama) , and '#' marking a hashtag: a sequence

of non whitespace characters preceded by the hash character (e.g. #healthCareReform).

Hashtags are used extensively and are adopted organically as part of the dynamic communication process. The use of hashtags is a popular way for a tweeter to provide his readers with some context[5], an important function due to the length constraint. For example the hashtag *#savethenhs*, reads as 'save the national health service', gives the context relevant to the tweet "Speaker refers to #Lanseys 'abysmal ignorance' as demonstrated on alcohol strategy; this SoS #notfitforpurpose #nhs #savethenhs #healthbill". It is important to note that from a functional perspective (indexing/search) capitalization does not play any role in the hashtag format, therefore *#NHS, #nhs* and *#nhS* are equivalent.



(a) Unique hashtags by number of occurrences.



(b) Number of sampled tweets per week.

Figure 1: Hashtag frequency and weekly sample size.

Our corpus consists of over 417 million tweets from Twitter, collected from June 2009 through December 2009. This corpus is a sample of approximately 15% of the Twitter stream in this six month' period. Over three million unique hashtags were observed in our data in over 49 million occurrences, an average of 0.11 hashtags per tweet. The hashtag frequency presents a long tail distribution where the 1000 most frequent hashtags (0.003% of the unique hashtags) cover 43% of hashtag occurrences (see Figure 1a). 67% of the unique hashtags appear only once.

4.2 Competing interchangeable hashtags

As described in the introduction, we are interested in predicting preference order within sets of interchangeable hashtags. Identifying small clusters of similarly used hashtags is a great challenge. For example, given the following hashtags: *#iranElection, #freeIran, #greenRevolution, #IranElection, #usElection, #iranelection, #followFriday*, it is clear that some of the hashtags should be classified under the general class 'Iran' or even the more specific class 'Iran election', however it is hard to determind whether *#freeIran* is interchangeable with *#iranElection* and/or with *#greenRevolution*[6] or not.

The fine granularity of clusters required in our task is beyond state of the art clustering algorithms for sparse data [29]. For that reason we back to two relatively conservative formulations of sets. In the first setting (CAPS) we cluster together hashtags that only differ in capitalization. In the second setting (STEM) we cluster together hashtags that share similar stems of the words composing them.

Table 2 presents a few examples of such clusters. The ranking of the hashtags within a group is based on the normalized frequency of the hashtag in the corpus. Normalization is required since the the released daily sample size was changed from time to time and since the overall volume of Twitter messages grew considerably during the six month sampling period (see Figure 1b).

In order to have meaningful ranking, we only looked at hashtags that appeared more than a hundred times assuming small counts are prone to bias and many time indicate unintentional typos. Filtering out the most infrequent hashtags, our experimental data contains a total of 15046 hashtags. In the conservative CAPS setting, sets size are of 2 – 11 tags, with the average of 2.27 hashtags per set and a standard deviation of 0.6.

5. MODEL FEATURES

In order to learn the model we represent each hashtag as a feature vector. Since our motivation is to model the dynamics between the various constraints we use features that may play a role in the language production process.

Our simplified hypothesis is that a speaker wishes to be well understood by her audience while minimizing the effort of producing the speech act [7]. We consider four major

[5]The exact functionality of a hashtag is defined by the practice of the community that uses it. The most frequent uses of hashtags are as a topic marker, as a **bold** typeface intended to draw attention to an important concept and a marker for the agenda of the tweeter.

[6]In fact, these hashtags are rather similar, being promoted by similar users in similar contexts. #freeIran and #iranElection are used in tweets related to the allegedly corrupted election in Iran and 'green' (as in #greenRevolution) is the color of the Iranian opposition activists and became the color of the demonstrations against the results of the official election.

[7]Obviously this hypothesis is simplified and there may be many utterances that do not directly support this hypothesis, e.g. sarcasm, politeness and other pragmatic uses of

Rank	**Iran election**	**Gaga VMAs**	**jobs**
1	#iranelection	#GagaVMAs	#jobs
2	#IranElection	#GaGaVMAs	#Jobs
3	#Iranelection	#GagaVMAS	#job*
4	#iranElection	#gagavmas	#Job*
5	#iranelections*	–	#JOBS
6	#IRANELECTION	–	#JOB*
7	#IranElections*	–	–
8	#iranelect*	–	–
9	#IRANelection	–	–
10	#IRanElection	–	–

Table 2: Three ordered sets (1 is the most popular). Iran Election refers to the allegedly corrupted elections of 2009, Gaga VMAs refers to the Lady Gaga's nomination for MTV's Video Music Awards, and jobs refers to the job market after the 2008 crisis. Asterisks indicate hashtags with shared stems that are added to the conservative clusters.

generic forces in our model: (1) production effort, (2) comprehension effort, (3) linguistic habits and conventions (possibly shaped by the production and comprehension efforts), and (4) Domain-incurred constraints, in our case the tweet's length restriction of 140 characters. These four forces are introduced to our model as families of features in a feature vector.

We expect these features to interact in a non-linear way, for example, consider the following four hashtags:(1) *#savethenationalhealthservice*, (2) *#savethenhs*, (3) *#saveTheNHS* and (4) *#saveTheNationalHealthService*. (1) and (2) require less typing effort than the (3) and (4) (respectively), while the later two are easier to comprehend since tokenization is made easier by the capitalization of word initials. Assuming that the community is already familiar with the abbreviation (NHS), it is clear that not only that typing *#saveTheNHS* is more efficient than typing *#saveTheNationalHealthService*, it is also easier to read and comprehend. Moreover, we expect (3) to be preferred by the users for its brevity in the light of the length constraint imposed by Twitter (although increased brevity, e.g. *#stnhs* will probably fall short for its high comprehension effort).

In the remainder of this section we list the features and provide a brief explanation for the intuition behind using these specific features.

Number of characters

The length of a hashtag plays an important role due to the 140 characters length constraint as well as to its effect on correct encoding of the concept the hashtag represents, ease of typing and ease or reading and comprehension.

Number of words

Hashtags can be combined from several words. The tokenization of the hashtag into distinct words is a demanding cognitive task on the reader's side.

Number of shift keystrokes

the number of strokes on the shift key represents the estimated production effort in the textual domain. The actual production effort can vary according to the device used (e.g. a full size keyboard vs. mobile device) and the application interface. Some noise is incurred by retweeting other messages(see below). While the use of special characters such as the underscore (_) or capital letters requires more typing effort, the use of capital and special characters can improve comprehension, compare *#savethenationalhealthservices* to *#saveTheNationalHealthService* or even to *#save_the_national_health_services*.

language. For some discussion see [13].

Retweet rate

A 'retweet' is a tweet that is shared ("forwarded", promoted) by another user. Retweeting is easy as the user does not have to retype the whole message but only click the 'retweet' key - minimizing language production effort even in cases where the original typing is hard[8]. While the retweet rate is not a linguistic feature per se, we expect it to contribute to the model, interacting with other features. As retweet rate is a global feature, we verified it is not correlated with the hashtag frequency and cannot be used as a single feature for accurate prediction of the ranking (correlation is 0.03).

Named Entities

In standard English named entities are capitalized. The user may or may not "import" this writing convention into Twitter even though it require some extra effort.

Proper name

In Standard English proper names are capitalized therefore users may "import" this habit to Twitter even though it requires extra key strokes and has no explicit contribution (e.g. *#obama* vs. *#Obama*).

Location

In standard English locations are typically capitalized (Washington). Many locations are combined from more than one word (New York, New York City) and they are often abbreviated (UK, NY, NY City).

Acronyms

Acronyms are typically written with capital letters for enhanced comprehension.

All caps

All caps can be used in order to mark an acronym or stress out part of the sentence/hashtag. From text production perspective, it is easier to capitalize a sequence of letters than capitalize only some of the letters (leaving a finger on the Shift key or using the Caps Lock).

Named entities (including Location and Proper Name) and acronym are binary features with the value 1 iff (part of) the hashtag belongs to the respective class. All-caps is 1 iff all hashtag letters are capitalized.

The following features are only relevant to the STEMS setting as they differentiate between hashtags with different characters.

[8]Depending on the application interface used, a retweet button might not be available, in that case the user copy and paste the original message, adding the token 'RT' (or 'RT @[user name]') at the beginning of the message, e.g. RT @support Adding a mobile number to your account can help you recover your password down the road is a retweet of the original message posted by Twitter's @support.

δ	1	10	50	100	1000
pairs in CAPS	10575	9487	6753	5266	1324
pairs in STEMS	22762	20201	14278	10816	2445

Table 3: Number of remaining preference pairs in each setting, given various values of δ as a preference threshold.

Levenshtein distance

The Levenshtein distance between the hashtag and its stemmed version. For example, the stemmed version of the hashtag *#NetworkingWitches* is *network,witch* and the Levenshtein distance is 5 (not counting capitalization differences). This feature is expected to represent the distance of the hashtags in the similarity set from a canonical form of the set's concept.

Suffixes

This group of features contains three different binary features, indicating whether the hashtag contains either *ed*, the *s* or the *ing* suffixes. This set of features is designed to capture the use of grammatical suffixes and the user's tendency to omit/use them with regards to other features.

Tokenization of hashtags to words, features extraction and annotation were all done automatically using standard algorithms and heuristics. Errors were corrected manually. Annotations are available at www.removed.due.anonymity.

6. EXPERIMENTAL FRAMEWORK

6.1 Preference pairs

Section 4 presents general corpus and domain statistics. In this section we provide some more details relevant to the specific experimental framework. Initially, 6636 clusters were extracted from our corpus, containing a total of 15046 hashtags that constitute 10757 preference pairs (see Section 4).

'Preference' is defined by the normalized frequency of the hashtags in the corpus, thus for a pair $\{ht_i, ht_j\}$ we say that $ht_i \succ ht_j$ (or $\langle ht_i, ht_j \rangle$) if $count(ht_i) > count(ht_j)$ and $ht_i \prec ht_j$ ($\langle ht_j, ht_i \rangle$) if $count(ht_i) < count(ht_j)$. However, since these preferences are decided by corpus statistics (usage counts), some of the preferences can be attributed to chance or to other non-linguistic factors such as retweeting. In order to prevent the randomness incurred by counts, we introduce δ: a preference-strictness threshold, thus S, the set of preference pairs is slightly altered:

$$S = \{\langle ht_i, ht_j \rangle | d(ht_i, ht_j) \geq \delta\} \quad (3)$$

where $d(ht_i, ht_j)$ is defined by:

$$d(ht_i, ht_j) = count(ht_i) - count(ht_j)$$

Table 3 gives the number of pairs remaining after applying various values of δ in formula 3.

6.2 Baseline algorithms

Least-effort baseline

As this work is inspired by the 'least-effort' paradigm that is even more prevalent in the informal domain of Twitter, it is straight forward to use a 'least effort' algorithm as the baseline. Given a pair (ht_i, ht_j) the baseline predicts the shorter hashtag to be preferred. In the conservative context of this work *shorter* means the one requiring less typing effort where each key stroke counts. For example, the hashtag *#freeIran* will be preferred over *#FreeIran* due to the extra press on the Shift key producing a capital *F*. Similarly, *#truth* is expected to be preferred over *#Truth* and *#Truth* over *#TRUTH*.

SVM-rank

SVM-rank is a modification of the SVM classifier to use ordinal regression, indicating a rank instead of a nominal class [15].

7. RESULTS

Table 4 presents the error rates of the GBrank, the SVMrank[9] and the least-effort baseline in both STEMS and CAPS settings and with varying strictness of preference (δ values).

While the error rate of the SVMrank is slightly lower than the least-effort baseline (except for STEMS with $\delta = 1$), GBrank achieves significantly lower error rates in all settings. As expected, increased strictness in the preference definition (as defined by Formula 3) improves prediction results, supporting the intuition that the "real" preferences defined by the frequency of mentions in social networks is prone to some noise that should be accounted for.

Comparing the STEMS and the CAPS settings, it seems that the STEMS setting poses a bigger challenge to all algorithms. Examining the data we verified that while the sets in conservative CAPS setting are noise-free (there is some noise between sets), the similarity sets in the STEMS setting bear some noise as hashtags with different suffixes, do not always conform to the exact same concept. In order to provide a cleaner discussion the rest of the paper refers mostly to the results of the CAPS setting.

It is well worth noting that although the CAPS setting seems restrictive, not only it provides a much cleaner look on the writing preferences, it may also reflect a fundamental linguistic phenomenon as it is somewhat resembles Optimality Theory approaches for modeling changes in the placement of the stressed syllable in a word and in the placement of expletive infixations [14].

The main motivation for our work is to learn how the interplay between different constraints affects neologisms coining and acceptance. GBrank produces nonlinear, though interpretable, models that allow us to "extract" the rules that govern this process. Figure 2 presents the first two trees produced by our GBrank for the setting described in Table 4 and for $\delta = 50$. Note that the role of tree T_k in the model is to compensate for errors and cases that were not handled by trees $T_{0,\ldots,k-1}$. In some cases, some of the trees will be idle, predicting 'no preference' since no decision node differentiates between the candidates in a preference pair. In case the ensemble is idle, preference is decided randomly. In the scenario of 7 trees, 3 levels and $\delta = 50$, the reported result was achieved with an average majority of 3.03 (of the seven trees) and standard deviation of 1.47.

[9]The SVMrank is used with a linear kernel function. All reported results are for 5-fold cross validation. Using polynomial or rbf kernel did not achieve convergence in a reasonable time.

Setting	Algorithm	$\delta = 1$	$\delta = 50$	$\delta = 100$
STEMS	**least effort**	0.383	0.34	0.314
	SVMrank	0.39	0.329	0.304
	GBrank	0.266	0.146	0.16
CAPS	**least effort**	0.32	0.25	0.214
	SVMrank	0.31	0.24	0.2
	GBrank	0.16	0.12	0.07

Table 4: Error rates of the least-effort baseline, the SVMrank and the GBrank in various levels of preference strictness (δ values). GBrank achieves the lowest error rate in all settings. GBrank setting: number of trees: 7, tree levels: 3, $\eta_k = 1$ and $\tau = 1$. Error rates are averaged for 5-fold cross validation.

#trees	1	2	3	4	5	6	7
STEMS	1	1.2	1.29	1.42	1.52	1.56	1.63
CAPS	1	1.3	1.78	2.07	2.37	2.72	3.03

Table 5: Average difference in the majority vote of trees (Stage 4 in the algorithm in Figure 1) for 5-fold cross validation $\delta = 50$ and $level = 3$.

Confidence and idleness

The average majority differences presented in Table 5 reflect the "confidence" of a trained ranker. A maximal confidence score would be the number of trees in the model, indicating that all trees agree on all pairs in all folds. Maximal score is never achieved in practice since the different trees may capture different linguistic phenomena and predict a preference accordingly. Some trees may also be idle for a given pair of hashtags as the features captured in these trees might not differentiate between the hashtags in the given pair.

Table 5 shows the average difference in the majority vote (stage 4 of the GBrank algorithm in Table 1). An average difference of 3.03 (7 trees, 3 levels, $\delta = 50$, CAPS setting) means that preference prediction was made by an average majority of 3.03 trees. The interpretation of this difference can indicate "decisiveness" on the range between two extreme cases: (a) four out of the seven trees are idle in the decision making (possibly different four for different pairs), and (b) prediction is based on five trees "preferring" a specific order and only two trees predicting the opposite, resulting in a difference of three in the majority vote. The average difference of 3.03 achieved by our algorithm indicates a relatively strong confidence even under interpretation (b).

It is notable that as the model is boosted by more trees the confidence increases. It is also worth noting that the confidence values in the CAPS setting are significantly higher than the respective values in the STEMS setting (for models with more than a single tree). This clear trend reflects the fact that the CAPS setting much cleaner than the noisy STEMS setting – similarly to the interpretation of the results presented in the previous subsection (and in Table 4).

Manipulating model complexity

A Gradient Boosted Trees method allows experimenting with models of various complexities. The complexity of a model is subject to three factors: the number of attributes spanning the vector space, the number of trees in the model and the maximal depth of the trees. More complex modes can capture subtler phenomena, under the risk of overfitting and/or loss of interpretability. Table 6 presents results for various complexities. While increased complexity seems to improve results, complex models (e.g. 15 trees of $depth = 5$) are hardly interpretable. Learning a single tree of depth 7 harms the results (0.22 error rate), clearly indicating overfitting. Growing additional trees with subsampling prevents overfitting while harming interpretability. We note that interpretability is possible even in the case of multiple (shallow) trees as most trees are idle regarding most hashtags.

Levels / Trees	1	3	5	7	10	15
1	0.26	0.22	0.2	0.19	0.18	0.16
2	0.25	0.19	0.175	0.14	0.1	0.1
3	0.24	0.15	0.12	0.12	0.1	0.09
4	0.22	0.13	0.12	0.1	0.08	0.07
5	0.21	0.13	0.1	0.8	0.08	0.07

Table 6: Error rates of models with different complexities according to the number of trees (columns) and the maximal depth of the trees (rows). Error rates are averaged for 5-fold cross validation for the CAPS setting with $\delta = 50$, $\eta_k = 1$ and $\tau = 1$.

An example of the trees constructed by the ranking algorithm is illustrated in Figures 2 *a* and *b* and Figures 3 *a* and *b*. The Sub-figures presents the first two trees with height restriction of three and four (respectively). As expected, the number of tree levels allowed in the model affects the trees structure and the order in which new features are introduced to the model.

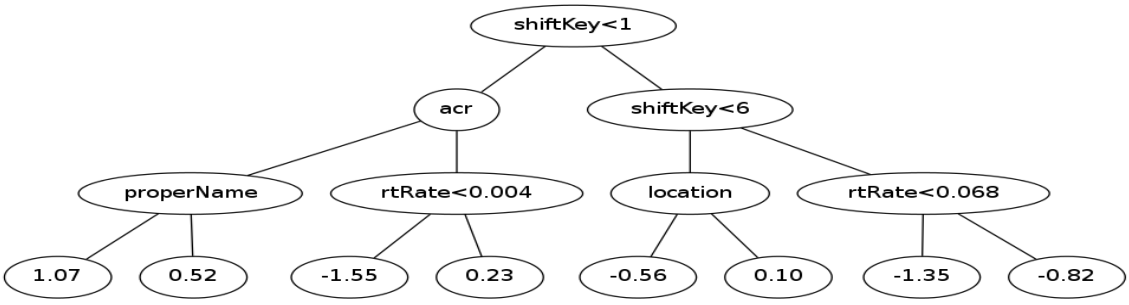
8. DISCUSSION

Brevity is the single most significant signal, considering that the 'least effort' baseline has an error rate of 25% ($\delta = 50$). User efficiency is indeed justifiable in many cases as 40% of the hashtags are single-word hashtags that are not named-entities. As expected, the models learned by our modified GBrank effectively model the interaction between nonlinear constraints (such as the number of words, characters, abbreviations etc.).

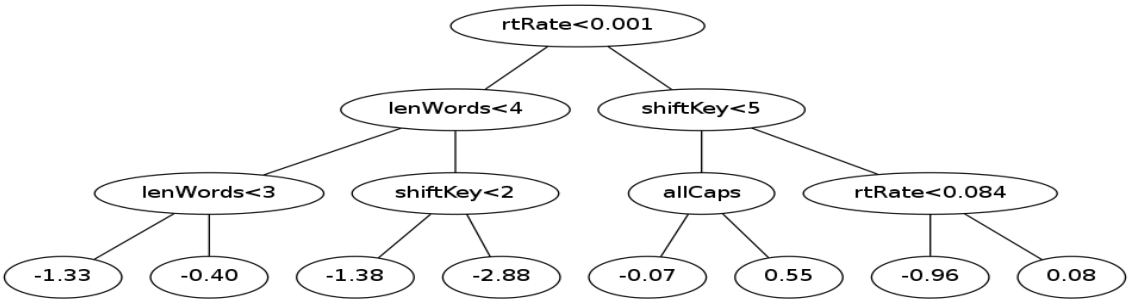
Two simple illustrative cases of correct prediction are $\#obama \succ \#Obama$ and $\#London \succ \#london$. Both preference pairs are single-word named-entities. The correct preference order predicted by our algorithm suggests that named entities should be broken down to types (e.g. personal names and locations) and that some types (of limited length), e.g. personal names, are more prone to "violate" the capitalization of proper names. This very simple yet interesting example is illustrated by the first tree in Figure 2. We note that this observation can be refined by other tree(s), modeling a nonlinear phenomenon of capitalization tendencies with regards to other features of the hashtags (length, presence of acronym, etc.).

Some other less trivial, yet correct, predictions made by our algorithm are: $\#tcot \succ \#TCOT \succ \#Tcot$ (acronym of 'top conservative', used by American republicans), $\#TV \succ \#Tv$, $\#BadRomanceiTunes \succ \#badromanceitunes$ and $\#GagaVMAs \succ \#GaGaVMAs \succ \#gagavmas'$ (compare to order at Table 2).

In the remainder of the section we discuss two of the possible causes for error, beyond the noise naturally occurring with real world data: ambiguity and canonization.



(a) Three levels, Tree 1



(b) Three levels, Tree 2

Figure 2: An example of the first two trees created by GBrank, $levels = 3$. The values at the leaves are the regression prediction values. The labels at the decision nodes indicate the feature decided upon. Nodes without explicit condition means an existence condition, i.e. in a $< properName >$ node the *existance* of a proper name in the hashtag is checked. Satisfying the condition in nodes of the form $< feature < n >$ leads to the left branch, satisfying the condition in nodes of the form $< label >$ leads to the right branch.

Ambiguity

Our data collection contains many ambiguous words, mai single-word named-entities. Two examples of such cases the hashtags *#jobs* (see Table 2) and *#dolphins*. Wh 'jobs' can refer to the job market, thus capitalization d not contribute to comprehension, it can also refer to App former CEO Steve Jobs, in which case capitalization may may not occur. Similarly, *#dolphins* can refer to the ec ogist battle cry "save the #dolphins" or refer to the Mia Dolphins NFL team. Disambiguation of hashtags is beyc the scope of this paper and will be addressed in future wo

Canonization

Some hashtags gain such popularity that they become p of the standard (Twitter) language. In these cases, the ha tag goes through an evolutionary process, eventually sta bilizing on its least-effort form. The easier recognition of canonical forms is in line with cognitive findings by [22] and [20]. An extreme example is the evolution of *#FF*[10] – the most frequent hashtag in our collection, used hundreds of thousands of times. While *#ff* and *#followfriday* are cryptic and incomprehensible to the untrained eye, starting as *#FollowFridays* late 2008, it quickly gained popularity, thus it no longer required much effort to comprehend and turned into #followfriday, which in turn evolved to #FF and #ff.

[10] *#ff*, abbreviation for 'follow Friday', is used by the Twitter community on Fridays as a mean for a tweeter to introduce new tweeters to her followers. A typical *#ff* tweet looks like "#ff @user1 @user2 @user3 @user4".

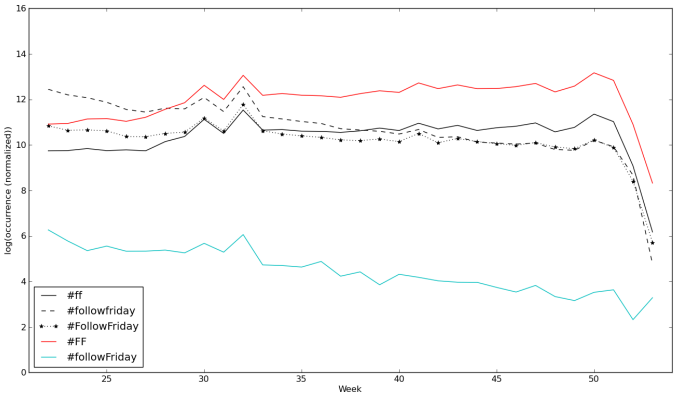
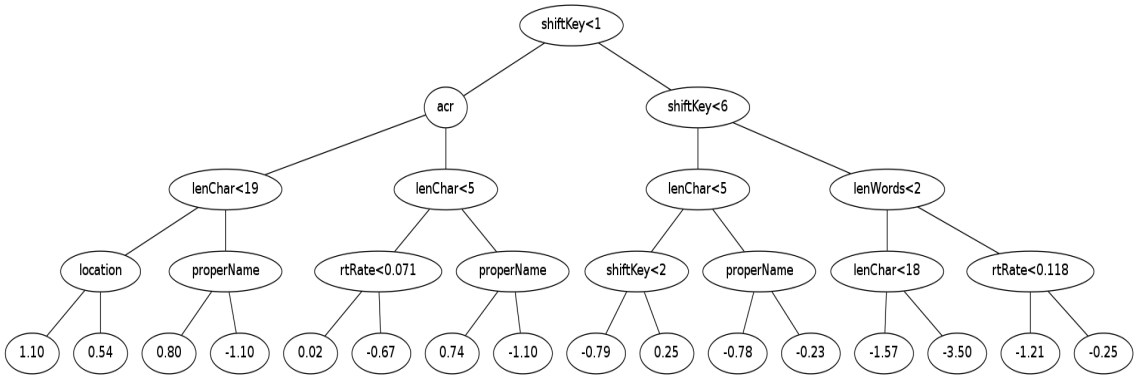


Figure 4: Trend lines of the popular *#followFriday* hashtag in its various incarnations.

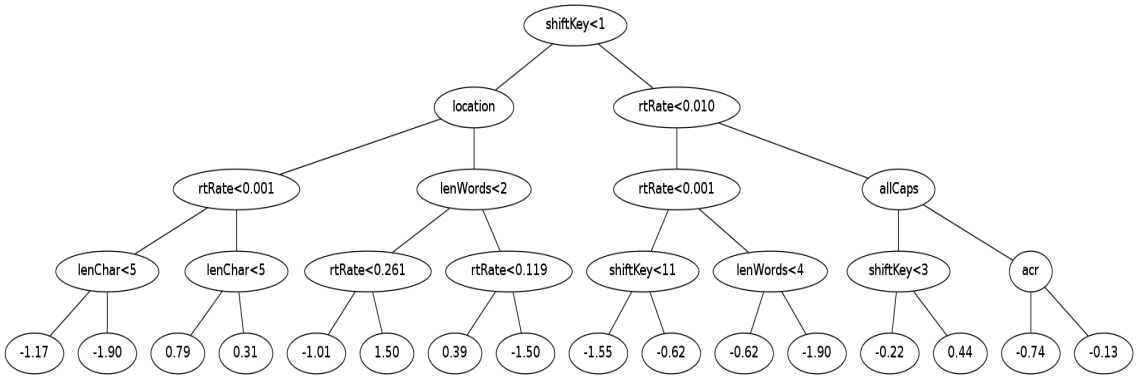
The evolutionary process of the *#ff* is illustrated in Figure 4[11]. The canonization process is evident by the trend line of *#ff* and *#FF* compared to the other options.

Identifying the evolutionary process and the canonization of some hashtags is not a trivial task and is left for future work. This phenomenon, to the extent it is present in our

[11] Our dataset does not contain tweets posted as early as 2008 but the trend can be interpolated easily.



(a) Tree 1



(b) Tree 2

Figure 3: An example of the first two trees created by GBrank, $levels = 4$. The values at the leaves are the regression prediction values. The labels at the decision nodes are interpreted similarly to the nodes in the trees in Figure 2.

data, may be another source of noise affecting our results.

9. CONCLUSION AND FUTURE WORK

Viewing Twitter hashtags as neologisms, we presented an algorithm for predicting user preferences in using hashtags. The algorithmic framework is a modification of the GBrank algorithm that is based on Gradient Boosted Trees adapted for ranking. Using this algorithm we outperform a naive, yet strong, baseline and an SVMrank algorithm. Another major advantage of our algorithmic framework is its interpretability which provides insights to the underlying processes governing language production. These insights, in turn, can be used for the artificial creation of more suitable hashtags, maximizing their diffusion potential.

There are two main directions for future work, beyond the issues mentioned in the section above: (1) achieve a suitable fine grained clustering of hashtags/neologisms that will allow expanding the system to handle less conservative sets of interchangeable competing candidates, and (2) introduce magnitude preserving metrics to the framework, thus gaining better understanding of the linguistic constraints reflected by the prediction model.

10. ACKNOWLEDGEMENTS

Removed in order to preserve anonymity.

11. REFERENCES

[1] J. Algeo. Blends, a structural and systemic view. *American Speech*, 52(1/2):47–64, 1977.

[2] J. Algeo. Where do all the new words come from? *American Speech*, 55(4):264–277, 1980.

[3] L. J. Brinton and E. C. Traugott. *Lexicalization and language change*. Cambridge University Press, 2005.

[4] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM, 2010.

[5] N. Chomsky. *Reflections on language*. Temple Smith London, 1976.

[6] N. Chomsky. *The minimalist program*, volume 28. Cambridge University Press, 1995.

[7] C. P. Cook. *Exploiting linguistic knowledge to infer properties of neologisms*. PhD thesis, University of Toronto, 2010.

[8] P. Cook and S. Stevenson. Automagically inferring the source words of lexical blends. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, pages 289–297, Melbourne, Australia, September 2007.

[9] E. Cunha, G. Magno, G. Comarela, V. Almeida,

M. A. Gonçalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 58–65, 2011.

[10] R. Dawkins. *The Selfish Gene*. Oxford paperbacks. Oxford University Press, 1976.

[11] G. Deutscher. *The unfolding of language*. Arrow, 2006.

[12] J. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[13] H. P. Grice. Logic and conversation. *1975*, pages 41–58, 1975.

[14] M. Hammond. Binarity in english nouns. *Nordlyd*, 24:97–110, 1996.

[15] T. Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM, 2006.

[16] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[17] J. Kleinberg. The flow of on-line information in global networks. In *Proceedings of the 2010 international conference on Management of data*, pages 1–2. ACM, 2010.

[18] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang, and M. A. Nowak. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–716, 2007.

[19] Y.-R. Lin, D. Margolin, B. Keegan, A. Baronchelli, and D. Lazer. #bigbirds never die: Understanding social dynamics of emergent hashtags. In *ICWSM*, 2013.

[20] K. Mahowald, E. Fedorenko, S. Piantadosi, and E. Gibson. Info/information theory: speakers choose shorter words in predictive contexts. *Cognition*, In press.

[21] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[22] R. M. Nosofsky and T. J. Palmeri. An exemplar-based random walk model of speeded classification. *Psychological Review; Psychological Review*, 104(2):266, 1997.

[23] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports*, 2, 2012.

[24] S. Piantadosi, H. Tily, and E. Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011.

[25] A. Prince and P. Smolensky. Optimality: From neural networks to universal grammar. *Science*, 275(5306):1604–1610, 1997.

[26] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

[27] R. E. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[28] C. Shannon, W. Weaver, R. Blahut, and B. Hajek. *The mathematical theory of communication*, volume 117. University of Illinois press Urbana, 1949.

[29] O. Tsur, A. Littman, and A. Rappoport. Efficient clustering of short messages into general domains. In *Proceedings of the seventh international AAAI conference on weblogs and social media*, 2013.

[30] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 643–652. ACM, 2012.

[31] J. Yang and J. Leskovec. Modeling Information Diffusion in Implicit Networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608. IEEE, 2010.

[32] Z. Zheng, K. Chen, G. Sun, and H. Zha. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 287–294. ACM, 2007.

[33] G. Zipf. Human behavior and the principle of least effort. 1949.

Measuring Price Discrimination and Steering on E-commerce Web Sites

Aniko Hannak
Northeastern University
Boston, MA
ancsaaa@ccs.neu.edu

Gary Soeller
Northeastern University
Boston, MA
soelgary@ccs.neu.edu

David Lazer
Northeastern University
Boston, MA
d.lazer@neu.edu

Alan Mislove
Northeastern University
Boston, MA
amislove@ccs.neu.edu

Christo Wilson
Northeastern University
Boston, MA
cbw@ccs.neu.edu

ABSTRACT

Today, many e-commerce websites personalize their content, including Netflix (movie recommendations), Amazon (product suggestions), and Yelp (business reviews). In many cases, personalization provides advantages for users: for example, when a user searches for an ambiguous query such as "router," Amazon may be able to suggest the woodworking tool instead of the networking device. However, personalization on e-commerce sites may also be used to the user's disadvantage by manipulating the products shown (price steering) or by customizing the prices of products (price discrimination). Unfortunately, today, we lack the tools and techniques necessary to be able to detect such behavior.

In this paper, we make three contributions towards addressing this problem. First, we develop a methodology for accurately measuring when price steering and discrimination occur and implement it for a variety of e-commerce web sites. While it may seem conceptually simple to detect differences between users' results, accurately attributing these differences to price discrimination and steering requires correctly addressing a number of sources of noise. Second, we use the accounts and cookies of over 300 real-world users to detect price steering and discrimination on 16 popular e-commerce sites. We find evidence for some form of personalization on nine of these e-commerce sites. Third, we investigate the effect of user behaviors on personalization. We create fake accounts to simulate different user features including web browser/OS choice, owning an account, and history of purchased or viewed products. Overall, we find numerous instances of price steering and discrimination on a variety of top e-commerce sites.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IMC'14, November 5–7, 2014, Vancouver, BC, Canada.
Copyright 2014 ACM 978-1-4503-3213-2/14/11 ...$15.00.
http://dx.doi.org/10.1145/2663716.2663744 .

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: Online Services—*commercial services, web-based services*; H.5.2 [**Information interfaces and presentation**]: User Interfaces—*evaluation/methodology*

Keywords

Search; Personalization; E-commerce; Price discrimination

1. INTRODUCTION

Personalization is a ubiquitous feature on today's top web destinations. Search engines such as Google, streaming-media services such as Netflix, and recommendation sites such as Yelp all use sophisticated algorithms to tailor content to each individual user. In many cases, personalization provides advantages for users: for example, when a user searches on Google with an ambiguous query such as "apple," there are multiple potential interpretations. By personalizing the search results (e.g., by taking the user's history of prior searches into account), Google is able to return results that are potentially more relevant (e.g., computer products, rather than orchards).

Recently, researchers and Internet users have uncovered evidence of personalization on e-commerce sites [1, 29, 30]. On such sites, the benefits of personalization for users are less clear; e-commerce sites have an economic incentive to use personalization to induce users into spending more money. For example, the travel website Orbitz was found to be personalizing the results of hotel searches [28]. Unbeknownst to users, Orbitz "steered" Mac OS X users towards more expensive hotels in select locations by placing them at higher ranks in search results. Orbitz discontinued the use of this personalization algorithm after one month [7].

At first blush, detecting personalization on e-commerce sites seems conceptually simple: have two users run the same search, and any differences in the results indicate personalization. Unfortunately, this approach is likely to have many false positives, as differences between users' results may exist for a number of reasons not related to personalization. For example, results may differ due to changes in product inventory, regional tax differences, or inconsistencies across data centers. As a result, accurately detecting personalization on e-commerce sites remains an open challenge.

In this paper, we first develop a methodology that can accurately measure personalization on e-commerce sites. We then use this methodology to address two questions: first, *how widespread is personalization on today's e-commerce web sites?* This includes *price discrimination* (customizing prices for some users) as well as *price steering* (changing the order of search results to highlight specific products). Second, *how are e-commerce retailers choosing to implement personalization?* Although there is anecdotal evidence of these effects [46, 48] and specific instances where retailers have been exposed doing so [1, 28], the frequency and mechanisms of e-commerce personalization remain poorly understood.

Our paper represents the first comprehensive study of e-commerce personalization that examines price discrimination and price steering for 300 real-world users[1], as well as synthetically generated fake accounts. We develop a measurement infrastructure that is able to distinguish genuine personalization of e-commerce sites from other sources of noise; this methodology is based on previous work on measuring personalization of web search services [17]. Using this methodology, we examine 16 top e-commerce sites covering general retailers as well as hotel and rental car booking sites. Our real-world data indicates that eight of these sites implement personalization, while our controlled tests based on fake accounts allow us to identify specific user features that trigger personalization on seven sites. Specifically, we observe the following personalization strategies:

- Cheaptickets and Orbitz implement price discrimination by offering reduced prices on hotels to "members".
- Expedia and Hotels.com engage in A/B testing that steers a subset of users towards more expensive hotels.
- Home Depot and Travelocity personalize search results for users on mobile devices.
- Priceline personalizes search results based on a user's history of clicks and purchases.

In addition to positively identifying price discrimination and steering on several well-known e-commerce sites, we also make the following four specific contributions. *First*, we introduce control accounts into all of our experiments, which allows us to differentiate between inherent noise and actual personalization. *Second*, we develop a novel methodology using information retrieval metrics to identify price steering. *Third*, we examine the impact of purchase history on personalization by reserving hotel rooms and rental cars, then comparing the search results received by these users to users with no history. *Fourth*, we identify a never-before-seen form of e-commerce personalization based on A/B testing, and show that it leads to price steering. Finally, we make all of our crawling scripts, parsers, and raw data available to the research community.

2. BACKGROUND AND MOTIVATION

We now provide background on e-commerce personalization and overview the terminology used in this paper.

2.1 A Brief History of Personalization

Online businesses have long used personalized recommendations as a way to boost sales. Retailers like Amazon and Target leverage the search and purchase histories of users to identify products that users may be interested in. In some cases, companies go to great lengths to obfuscate the fact that recommendations are personalized, because users sometimes find these practices to be creepy [10].

In several cases, e-commerce sites have been observed performing *price discrimination*: the practice of showing different prices to different people for the same item. Several years ago, Amazon briefly tested an algorithm that personalized prices for frequent shoppers [1]. Although many consumers erroneously believe that price discrimination on the Internet is illegal and are against the practice [8, 36], consumers routinely accept real-world price discrimination in the form of coupons, student discounts, or members-only prices [5].

Similarly, e-commerce sites have been observed performing *price steering*: the practice of re-ordering search results to place expensive items towards the top of the page. For example, the travel web site Orbitz was found to be promoting high-value hotels specifically to Apple users [28]. Although the prices of individual items do not change in this case, customers are likely to purchase items that are placed towards the top of search results [25], and thus users can be nudged towards more expensive items.

2.2 Scope of This Study

Throughout the paper, we survey a wide variety of e-commerce web sites, ranging from large-scale retailers like Walmart to travel sites like Expedia. To make the results comparable, we only consider products returned via *searches*—as opposed to "departments", home page offers, and other mechanisms by which e-commerce sites offer products to users—as searching is a functionality supported by most large retailers. We leave detecting price discrimination and steering via other mechanisms to future work. Additionally, we use *products* and their advertised price on the search result page (e.g., a specific item on Walmart or hotel on Expedia) as the basic unit of measurement; we leave the investigation of effects such as bundle discounts, coupons, sponsored listings, or hidden prices to future work as well.

2.3 Definitions

Personalization on web services comes in many forms (e.g., "localization", per-account customization, *etc.*), and it is not entirely straightforward to declare that an inconsistency between the product search results observed by two users is due to personalization. For example, the two users' search queries may have been directed to different data centers, and the differences are a result of data center inconsistency rather than intentional personalization.

For the purposes of this paper, we define personalization to be taking place *when an inconsistency in product search results is due to client-side state associated with the request.* For example, a client's request often includes tracking cookies, a User-Agent identifying the browser and Operating System (OS), and the client's source IP address. If any of these lead to an inconsistency in the results, we declare the inconsistency to be personalization. In the different-datacenter example from above, the inconsistency between the two results is not due to any client-side state, and we therefore declare it not to be personalization.

More so than other web services [17], e-commerce retailers have a number of different dimensions available to personal-

[1]Our study is conducted under Northeastern Institutional Review Board protocol #13-04-12.

ize on. In this paper, we focus on two of the primary vectors for e-commerce personalization:

Price steering occurs when two users receive different product results (or the same products in a different order) for the same query (e.g., Best Buy showing more-expensive products to user A than user B when they both query for "laptops"). Price steering can be similar to personalization in web search [17], i.e., the e-commerce provider may be trying to give the user more relevant products (or, they may be trying to extract more money from the user). Steering is possible because e-commerce sites often do not sort search results by an objective metric like price or user reviews by default; instead, results can be sorted using an ambiguous metric like "Best Match" or "Most Relevant".

Price discrimination occurs when two users are shown inconsistent prices for the same product (e.g., Travelocity showing a select user a higher price for a particular hotel). Contrary to popular belief, price discrimination in general is not illegal in the United States [13], as the Robinson–Patman Act of 1936 (*a.k.a.* the Anti-Price Discrimination Act) is written to control the behavior of product manufacturers and distributors, not consumer-facing enterprises. It is unclear whether price discrimination targeted against protected classes (e.g., race, religion, gender) is legal.

Although the term "price discrimination" evokes negative connotations, it is actually a fundamental concept in economic theory, and it is widely practiced (and accepted by consumers) in everyday life. In economic theory, *perfect price discrimination* refers to a pricing strategy where each consumer is charged the maximum amount that they are willing to pay for each item [37]. *Elastic* consumers can afford to pay higher prices, whereas *inelastic* (price-constrained) consumers are charged less. In practice, strategies like *direct* and *indirect segmentation* are employed by companies to charge different prices to different segments of consumers [32]. One classic (and extremely popular [5]) example of indirect segmentation is coupons: consumers that take the time to seek out and use coupons explicitly reveal to the retailer that they are price-inelastic. Another accepted form of indirect segmentation is Saturday-night stay discount for airfares. Accepted forms of direct segmentation include discounts for students and seniors.

3. METHODOLOGY

Both price steering and discrimination are difficult to reliably detect in practice, as differences in e-commerce search results across users may be due to factors other than personalization. In this section, we describe the methodology we develop to address these measurement challenges. First, we give the high-level intuition that guides the design of our experiments, and describe how we accurately distinguish personalization of product search results from other causes of inconsistency. Second, we describe the implementation of our experiments. Third, we detail the sites that we choose to study and the queries we select to test for personalization.

3.1 Measuring Personalization

The first challenge we face when trying to measure personalization on e-commerce sites is dealing with the variety of sites that exist. Few of these sites offer programmatic APIs, and each uses substantially different HTML markup to implement their site. As a result, we collect data by visiting the various sites' web pages, and we write custom HTML parsers to extract the products and prices from the search result page for each site that we study. We make these scripts and parsers available to the research community.

The second challenge we face is *noise*, or inconsistencies in search results that are *not* due to personalization. Noise can be caused by a variety of factors:

- **Updates to the e-commerce site:** E-commerce services are known to update their inventory often, as products sell out, become available, or prices are changed. This means that the results for a query may change even over short timescales.
- **Distributed infrastructure:** Large-scale e-commerce sites are often spread across geographically diverse datacenters. Our tests show that different datacenters may return different results for the same search, due to inconsistencies across data centers.
- **Unknown:** As an e-commerce site is effectively a black-box (i.e., we do not know their internal architecture or algorithms), there may be other, unknown sources of noise that we are unaware of.

To control for all sources of noise, we include a *control* in each experiment that is configured identically to one other treatment (i.e., we run one of the experimental treatments twice). Doing so allows us to measure the noise as the level of inconsistency between the control account and its twin; since these two treatments are configured identically, any inconsistencies between them *must* be due to noise, not personalization. Then, we can measure the level of inconsistency between the different experimental treatments; if this is higher than the baseline noise, the increased inconsistencies are due to personalization. As a result, we cannot declare any particular inconsistency to be due to personalization (or noise), but we can report the overall rate.

To see why this works, suppose we want to determine if Firefox users receive different prices than Safari users on a given site. The naïve experiment would be to send a pair of identical, simultaneous searches—one with a Firefox `User-Agent` and one with a Safari `User-Agent`—and then look for inconsistencies. However, the site may be performing A/B testing, and the differences may be due to requests given different A/B treatments. Instead, we run an additional control (e.g., a third request with a Firefox `User-Agent`), and can measure the frequency of differences due to noise.

Of course, running a single query is insufficient to accurately measure noise and personalization. Instead, we run a large set of searches on each site over multiple days and report the aggregate level of noise and personalization across all results.

3.2 Implementation

Except for when we use real users' accounts (our Amazon Mechanical Turk users in § 4), we collect data from the various e-commerce sites using custom scripts for PhantomJS [35]. We chose PhantomJS because it is a full implementation of the WebKit browser, meaning that it executes JavaScript, manages cookies, *etc.* Thus, using PhantomJS is significantly more realistic than using custom code that does not execute JavaScript, and it is more scalable than automating a full web browser (e.g., Selenium [41]). Phan-

Retailer	Site	Category
Best Buy	http://bestbuy.com	Electronics
CDW	http://cdw.com	Computers
HomeDepot	http://homedepot.com	Home-improvement
JCPenney	http://jcp.com	Clothes, housewares
Macy's	http://macys.com	Clothes, housewares
Newegg	http://newegg.com	Computers
Office Depot	http://officedepot.com	Office supplies
Sears	http://sears.com	Clothes, housewares
Staples	http://sears.com	Office supplies
Walmart	http://walmart.com	General retailer

Table 1: The general retailers we measured in this study.

tomJS has been proven to allow large-scale measurements in previous work [17].

IP Address Geolocation. In our experiments, all PhantomJS instances issued queries from a /24 block of IP addresses located in Boston. Controlling the source IP address is critical since e-commerce sites may personalize results based on a user's geolocation. We do not examine geolocation-based personalization in this paper; instead, we refer interested readers to work by Mikians et al. that thoroughly examines this topic [29, 30].

Browser Fingerprinting. All of our PhantomJS instances are configured identically, and present identical HTTP request headers and DOM properties (e.g., `screen.width`) to web servers. The only exceptions to this rule are in cases where we explicitly test for the effect of different `User-Agent` strings, or when an e-commerce site stores state on the client (e.g., with a cookie). Although we do not test for e-commerce personalization based on browser fingerprints [34], any website attempting to do so would observe identical fingerprints from all of our PhantomJS instances (barring the two previous exceptions). Thus, personalization based on currently known browser fingerprinting techniques is unlikely to impact our results.

3.3 E-commerce Sites

We focus on two classes of e-commerce web sites: general e-commerce retailers (e.g., Best Buy) and travel retailers (e.g., Expedia). We choose to include travel retailers because there is anecdotal evidence of price steering among such sites [28]. Of course, our methodology can be applied to other categories of e-commerce sites as well.

General Retailers. We select 10 of the largest e-commerce retailers, according to the Top500 e-commerce database [43], for our study, shown in Table 1. We exclude Amazon, as Amazon hosts a large number of different merchants, making it difficult to measure Amazon itself. We also exclude sites like `apple.com` that only sell their own brand.

Travel Retailers. We select six of the most popular web-based travel retailers [44] to study, shown in Table 2. For these retailers, we focus on searches for *hotels* and *rental cars*. We do not include airline tickets, as airline ticket pricing is done transparently through a set of Global Distribution Systems (GDSes) [6]. Furthermore, a recent study by Vissers et al. looked for, but was unable to find, evidence of price discrimination on the websites of 25 major airlines [45].

In all cases, the prices of products returned in search results are in US dollars, are pre-tax, and do not include shipping fees. Examining prices before taxes and shipping are applied helps to avoid differences in pricing that are due to the location of the business and/or the customer.

Retailer	Site	Hotels	Cars
Cheaptickets	http://cheaptickets.com	☑	☑
Expedia	http://expedia.com	☑	☑
Hotels.com	http://hotels.com	☑	☐
Orbitz	http://orbitz.com	☑	☑
Priceline	http://priceline.com	☑	☑
Travelocity	http://travelocity.com	☐	☑

Table 2: The travel retailers we measured in this study.

3.4 Searches

We select 20 searches to send to each target e-commerce site; it is the results of these searches that we use to look for personalization. We select the searches to cover a variety of product types, and tailor the searches to the type of products each retailer sells. For example, for JCPenney, our searches include "pillows", "sunglasses", and "chairs"; for Newegg, our searches include "flash drives", "LCD TVs", and "phones".

For travel web sites, we select 20 searches (location and date range) that we send to each site when searching for hotels or rental cars. We select 10 different cities across the globe (Miami, Honolulu, Las Vegas, London, Paris, Florence, Bangkok, Cairo, Cancun, and Montreal), and choose date ranges that are both short (4-day stays/rentals) and long (11-day stays/rentals).

4. REAL-WORLD PERSONALIZATION

We begin by addressing our first question: *how widespread are price discrimination and steering on today's e-commerce web sites?* To do so, we have a large set of real-world users run our experimental searches and examine the results that they receive.

4.1 Data Collection

To obtain a diverse set of users, we recruited from Amazon's Mechanical Turk (AMT) [3]. We posted three Human Intelligence Tasks (HITs) to AMT, with each HIT focusing on e-commerce, hotels, or rental cars. In the HIT, we explained our study and offered each user $1.00 to participate.[2] Users were required to live in the United States, and could only complete the HIT once.

Users who accepted the HIT were instructed to configure their web browser to use a Proxy Auto-Config (PAC) file provided by us. The PAC file routes all traffic to the sites under study to an HTTP proxy controlled by us. Then, users were directed to visit a web page containing JavaScript that performed our set of searches in an `iframe`. After each search, the Javascript grabs the HTML in the `iframe` and uploads it to our server, allowing us to view the results of the search. By having the user run the searches within their browser, any cookies that the user's browser had previously been assigned would automatically be forwarded in our searches. This allows us to examine the results that the user would have received. We waited 15 seconds between each search, and the overall experiment took ≈45 minutes to complete (between five and 10 sites, each with 20 searches).

[2]This study was conducted under Northeastern University IRB protocol #13-04-12; all personally identifiable information was removed from our collected data.

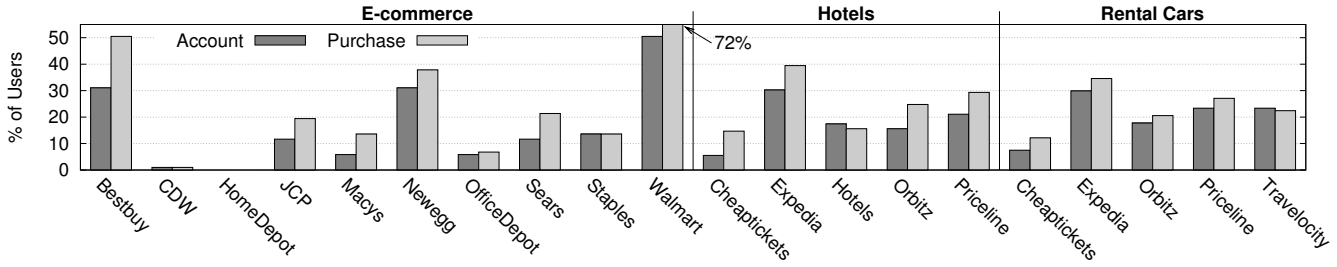


Figure 1: Previous usage (i.e., having an account and making a purchase) of different e-commerce sites by our AMT users.

The HTTP proxy serves two important functions. *First*, it allows us to quantify the baseline amount of noise in search results. Whenever the proxy observes a search request, it fires off *two* identical searches using PhantomJS (with no cookies) and saves the resulting pages. The results from PhantomJS serve as a *comparison* and a *control* result. As outlined in § 3.1, we compare the results served to the comparison and control to determine the underlying level of noise in the search results. We also compare the results served to the comparison and the real user; any differences between the real user and the comparison above the level observed between the comparison and the control can be attributed to personalization.

Second, the proxy reduces the amount of noise by sending the experimental, comparison, and control searches to the web site at the same time and from the same IP address. As stated in § 3.2, sending all queries from the same IP address controls for personalization due to geolocation, which we are specifically not studying in this paper. Furthermore, we hard-coded a DNS mapping for each of the sites on the proxy to avoid discrepancies that might come from round-robin DNS sending requests to different data centers.

In total, we recruited 100 AMT users in each of our retail, hotel, and car experiments. In each of the experiments, the participants first answered a brief survey about whether they had an account and/or had purchased something from each site. We present the results of this survey in Figure 1. We observe that many of our users have accounts and a purchase history on a large number of the sites we study.[3]

4.2 Price Steering

We begin by looking for *price steering*, or personalizing search results to place more- or less-expensive products at the top of the list. We do not examine rental car results for price steering because travel sites tend to present these results in a deterministically ordered grid of car types (e.g., economy, SUV) and car companies (with the least expensive car in the upper left). This arrangement prevents travel sites from personalizing the order of rental cars.

To measure price steering, we use three metrics:

Jaccard Index. To measure the overlap between two different sets of results, we use Jaccard index, which is the size of the intersection over the size of the union. A Jaccard Index of 0 represents no overlap between the results, while 1 indicates they contain the same results (although not necessarily in the same order).

[3]Note that the fraction of users having made purchases can be higher than the fraction with an account, as many sites allow purchases as a "guest".

Kendall's τ. To measure reordering between two lists of result, we use Kendall's τ rank correlation coefficient. This metric is commonly used in the Information Retrieval (IR) literature to compare ranked lists [23]. The metric ranges between -1 and 1, with 1 representing the same order, 0 signifying no correlation, and -1 being inverse ordering.

nDGG. To measure how strongly the ordering of search results is correlated with product prices, we use Normalized Discounted Cumulative Gain (nDCG). nDCG is a metric from the IR literature [20] for calculating how "close" a given list of search results is to an ideal ordering of results. Each possible search result r is assigned a "gain" score $g(r)$. The DCG of a page of results $R = [r_1, r_2, \ldots, r_k]$ is then calculated as $DCG(R) = g(r_1) + \sum_{i=2}^{k}(g(r_i)/\log_2(i))$. Normalized DCG is simply $DCG(R)/DCG(R')$, where R' is the list of search results with the highest gain scores, sorted from greatest to least. Thus, nDCG of 1 means the observed search results are the same as the ideal results, while 0 means no useful results were returned.

In our context, we use product prices as gain scores, and construct R' by aggregating all of the observed products for a given query. For example, to create R' for the query "ladders" on Home Depot, we construct the union of the results for the AMT user, the comparison, and the control, then sort the union from most- to least-expensive. Intuitively, R' is the most expensive possible ordering of products for the given query. We can then calculate the DCG for the AMT user's results and normalize it using R'. Effectively, if $nDCG(AMTuser) > nDCG(control)$, then the AMT user's search results include more-expensive products towards the top of the page than the control.

For each site, Figure 2 presents the average Jaccard index, Kendall's τ, and nDCG across all queries. The results are presented comparing the comparison to the control searches (Control), and the comparison to the AMT user searches (User). We observe several interesting trends. *First*, Sears, Walmart, and Priceline all have a lower Jaccard index for AMT users relative to the control. This indicates that the AMT users are receiving different products at a higher rate than the control searches (again, note that we are *not* comparing AMT users' results to each other; we only compare each user's result to the corresponding comparison result). Other sites like Orbitz show a Jaccard of 0.85 for Control and User, meaning that the set of results shows inconsistencies, but that AMT users are not seeing a higher level of inconsistency than the control and comparison searches.

Second, we observe that on Newegg, Sears, Walmart, and Priceline, Kendall's τ is at least 0.1 lower for AMT users, i.e., AMT users are consistently receiving results in a differ-

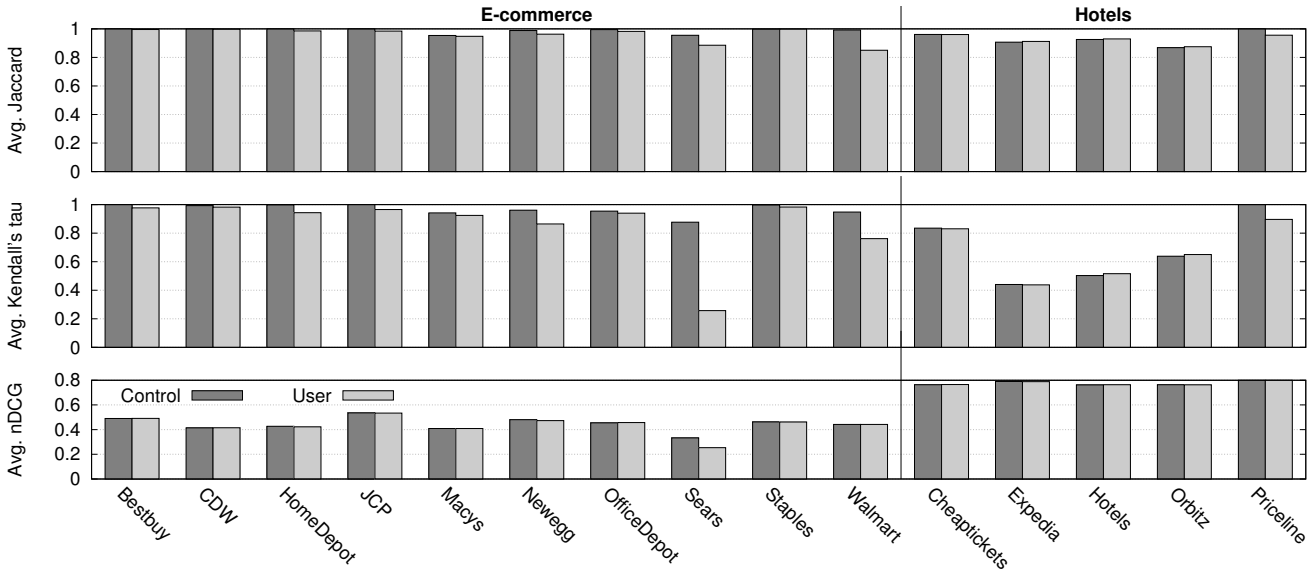


Figure 2: Average Jaccard index (top), Kendall's τ (middle), and nDCG (bottom) across all users and searches for each web site.

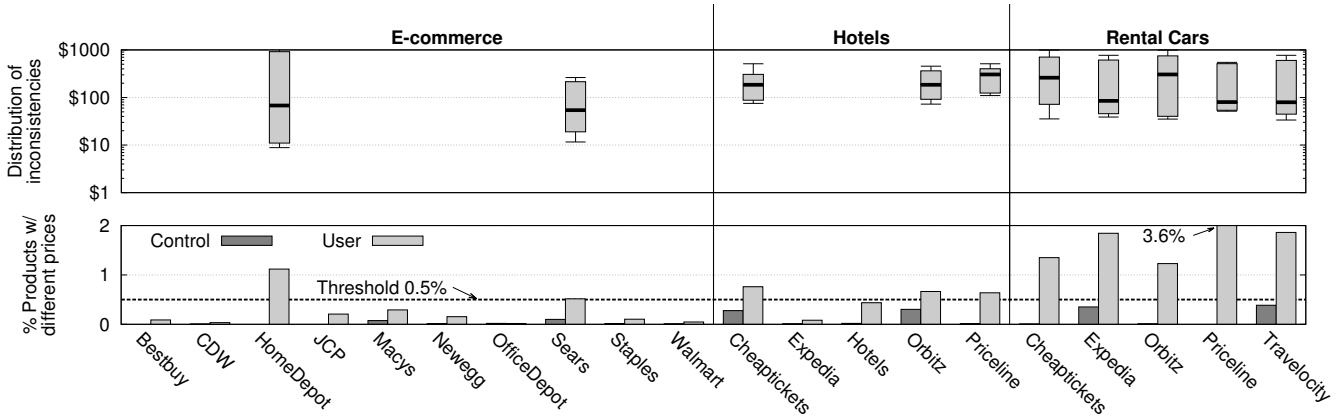


Figure 3: Percent of products with inconsistent prices (bottom), and the distribution of price differences for sites with $\geq$0.5% of products showing differences (top), across all users and searches for each web site. The top plot shows the mean (thick line), 25th and 75th percentile (box), and 5th and 95th percentile (whisker).

ent order than the controls. This observation is especially true for Sears, where the ordering of results for AMT users is markedly different. *Third*, we observe that Sears alone appears to be ordering products for AMT users in a price-biased manner. The nDCG results show that AMT users tend to have *cheaper* products near the top of their search results relative to the controls. Note that the results in Figure 2 are only useful for uncovering price steering; we examine whether the target sites are performing price discrimination in § 4.3.

Besides Priceline, the other four travel sites do not show significant differences between the AMT users and the controls. However, these four sites do exhibit significant noise: Kendall's τ is $\leq$0.83 in all four cases. On Cheaptickets and Orbitz, we manually confirm that this noise is due to randomness in the order of search results. In contrast, on Expedia and Hotels.com this noise is due to systematic A/B testing on users (see § 5.2 for details), which explains why we see equally low Kendall's τ values on both sites for all users. Unfortunately, it also means that we cannot draw any conclusions about personalization on Expedia and Hotels.com from the AMT experiment, since the search results for the comparison and the control rarely match.

4.3 Price Discrimination

So far, we have only looked at the set of products returned. We now turn to investigate whether sites are altering the prices of products for different users, i.e., price discrimination. In the bottom plot of Figure 3, we present the fraction of products that show price inconsistencies between the user's and comparison searches (User) and between the comparison and control searches (Control). Overall, we observe that most sites show few inconsistencies (typically <0.5% of products), but a small set of sites (Home Depot, Sears, and many of the travel sites) show both a significant fraction of price inconsistencies *and* a significantly higher fraction of inconsistencies for the AMT users.

To investigate this phenomenon further, in the top of Figure 3, we plot the distribution of price differentials for all sites where >0.5% of the products show inconsistency. We plot the mean price differential (thick line), 25th and

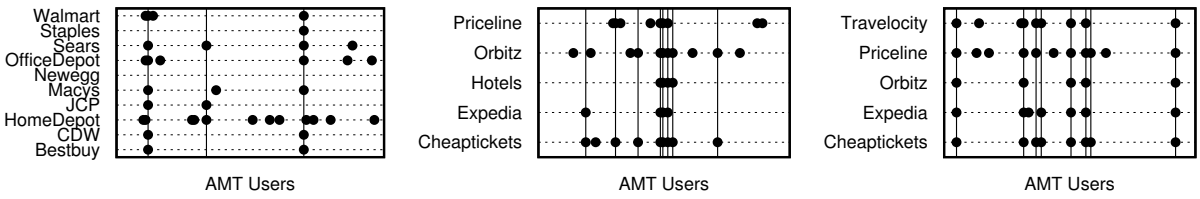


Figure 5: AMT users that receive highly personalized search results on general retail, hotels, and car rental sites.

75th percentile (box), and 5th and 95th percentile (whisker). Note that in our data, AMT users always receive higher prices than the controls (on average), thus all differentials are positive. We observe that the price differentials on many sites are quite large (up to hundreds of dollars). As an example, in Figure 4, we show a screenshot of a price inconsistency that we observed. Both the control and comparison searches returned a price of $565 for a hotel, while our AMT user was returned a price of $633.

4.4 Per-User Personalization

Next, we take a closer look at the subset of AMT users who experience high levels of personalization on one or more of the e-commerce sites. Our goal is to investigate whether these AMT users share any observable features that may illuminate why they are receiving personalized search results. We define highly personalized users as the set of users who see products with inconsistent pricing >0.5% of the time. After filtering we are left with between 2-12% of our AMT users depending on the site.

First, we map the AMT users' IP addresses to their geolocations and compare the locations of personalized and non-personalized users. We find no discernible correlation between location and personalization. However, as mentioned above, in this experiment all searches originate from a proxy in Boston. Thus, it is not surprising that we do not observe any effects due to location, since the sites did not observe users' true IP addresses.

Next, we examine the AMT users' browser and OS choices. We are able to infer their platform based on the HTTP headers sent by their browser through our proxy. Again, we find no correlation between browser/OS choice and high personalization. In § 5, we do uncover personalization linked to the use of mobile browsers, however none of the AMT users in our study did the HIT from a mobile device.

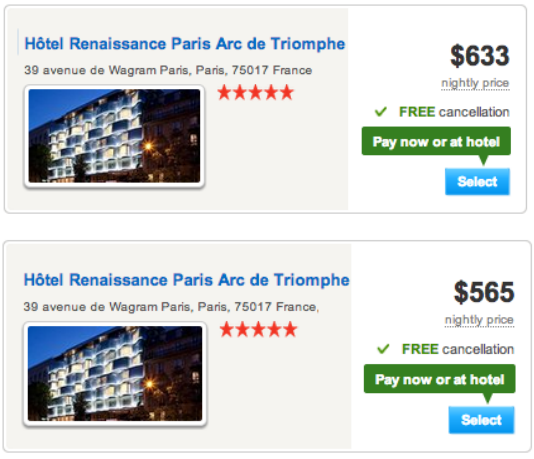


Figure 4: Example of price discrimination. The top result was served to the AMT user, while the bottom result was served to the comparison and control.

Finally, we ask the question: are there AMT users who receive personalized results on multiple e-commerce sites? Figure 5 lists the 100 users in our experiments along the x-axis of each plot; a dot highlights cases were a site personalized search results for a particular user. Although some dots are randomly dispersed, there are many AMT users that receive personalized results from several e-commerce sites. We highlight users who see personalized results on more than one site with vertical bars. More users fall into this category on travel sites than on general retailers.

The takeaway from Figure 5 is that we observe many AMT users who receive personalized results across multiple sites. This suggests that these users share feature(s) that all of these sites use for personalization. Unfortunately, we are unable infer the specific characteristics of these users that are triggering personalization.

Cookies. Although we logged the cookies sent by AMT users to the target e-commerce sites, it is not possible to use them to determine why some users receive personalized search results. First, cookies are typically random alphanumeric strings; they do not encode useful information about a user's history of interactions with a website (e.g., items clicked on, purchases, *etc.*). Second, cookies can be set by content embedded in third-party websites. This means that a user with a cookie from e-commerce site S may never have consciously visited S, let alone made purchases from S. These reasons motivate why we rely on survey results (see Figure 1) to determine AMT users' history of interactions with the target e-commerce sites.

4.5 Summary

To summarize our findings in this section: we find evidence for price steering and price discrimination on four general retailers and five travel sites. Overall, travel sites show price inconsistencies in a higher percentage of cases, relative to the controls, with prices increasing for AMT users by hundreds of dollars. Finally, we observe that many AMT users experience personalization across multiple sites.

5. PERSONALIZATION FEATURES

In § 4, we demonstrated that e-commerce sites personalize results for real users. However, we cannot determine *why* results are being personalized based on the data from real-world users, since there are too many confounding variables attached to each AMT user (e.g., their location, choice of browser, purchase history, *etc.*).

Category	Feature	Tested Values
Account	Cookies	No Account, Logged In, No Cookies
User-Agent	OS	Win. XP, Win. 7, OS X, Linux
	Browser	Chrome 33, Android Chrome 34, IE 8, Firefox 25, Safari 7, iOS Safari 6
Account History	Click	Low Prices, High Prices
	Purchase	Low Prices, High Prices

Table 3: User features evaluated for effects on personalization.

In this section, we conduct controlled experiments with fake accounts created by us to examine the impact of specific features on e-commerce personalization. Although we cannot test all possible features, we examine five likely candidates: browser, OS, account log-in, click history, and purchase history. We chose these features because e-commerce sites have been observed personalizing results based on these features in the past [1, 28].

We begin with an overview of the design of our synthetic user experiments. Next, we highlight examples of personalization on hotel sites and general retailers. None of our experiments triggered personalization on rental car sites, so we omit these results.

5.1 Experimental Overview

The goal of our synthetic experiments is to determine whether specific user features trigger personalization on e-commerce sites. To assess the impact of feature X that can take on values $x_1, x_2, \ldots, x_n$, we execute $n + 1$ PhantomJS instances, with each value of X assigned to one instance. The $n + 1$th instance serves as the control by duplicating the value of another instance. All PhantomJS instances execute 20 queries (see § 3.4) on each e-commerce site per day, with queries spaced one minute apart to avoid tripping security countermeasures. PhantomJS downloads the first page of results for each query. Unless otherwise specified, PhantomJS persists all cookies between experiments. All of our experiments are designed to complete in <24 hours.

To mitigate measurements errors due to noise (see § 3.1), we perform three steps (some borrowed from previous work [16, 17]): *first*, all searches for a given query are executed at the same time. This eliminates differences in results due to temporal effects. This also means that each of our treatments has exactly the same search history at the same time. *Second*, we use static DNS entries to direct all of our query traffic to specific IP addresses of the retailers. This eliminates errors arising from differences between datacenters. *Third*, although all PhantomJS instances execute on one machine, we use SSH tunnels to forward the traffic of each treatment to a unique IP address in a /24 subnet. This process ensures that any effects due to IP geolocation will affect all results equally.

Static Features. Table 3 lists the five features that we evaluate in our experiments. In the cookie experiment, the goal is to determine whether e-commerce sites personalize results for users who are logged-in to the site. Thus, two PhantomJS instances query the given e-commerce site without logging-in, one logs-in before querying, and the final account clears its cookies after every HTTP request.

In two sets of experiments, we vary the `User-Agent` sent by PhantomJS to simulate different OSes and browsers. The goal of these tests is to see if e-commerce sites personalize based on the user's choice of OS and browser. In the OS experiment, all instances report using Chrome 33, and Windows 7 serves as the control. In the browser experiment, Chrome 33 serves as the control, and all instances report using Windows 7, except for Safari 7 (which reports OS X Mavericks), Safari on iOS 6, and Chrome on Android 4.4.2.

Historical Features. In our historical experiments, the goal is to examine whether e-commerce sites personalize results based on users' history of viewed and purchased items. Unfortunately, we are unable to create purchase history on general retail sites because this would entail buying and then returning physical goods. However, it is possible for us to create purchase history on travel sites. On Expedia, Hotels.com, Priceline, and Travelocity, some hotel rooms feature "pay at the hotel" reservations where you pay at check-in. A valid credit card must still be associated with "pay at the hotel" reservations. Similarly, all five travel sites allow rental cars to be reserved without up-front payment. These no-payment reservations allow us to book reservations on travel sites and build up purchase history.

To conduct our historical experiments, we created six accounts on the four hotel sites and all five rental car sites. Two accounts on each site serve as controls: they do not click on search results or make reservations. Every night for one week, we manually logged-in to the remaining four accounts on each site and performed specific actions. Two accounts searched for a hotel/car and clicked on the highest and lowest priced results, respectively. The remaining two accounts searched for the same hotel/car and booked the highest and lowest priced results, respectively. Separate credit cards were used for high- and low-priced reservations, and neither card had ever been used to book travel before. Although it is possible to imagine other treatments for account history (e.g., a person who always travels to a specific country), price-constrained (inelastic) and unconstrained (elastic) users are a natural starting point for examining the impact of account history. Furthermore, although these treatments may not embody realistic user behavior, they do present unambiguous signals that could be observed and acted upon by personalization algorithms.

We pre-selected a destination and travel dates for each night, so the click and purchase accounts all used the same search parameters. Destinations varied across major US, European, and Asian cities, and dates ranged over the last six months of 2014. All trips were for one or two night stays/rentals. On average, the high- and low-price purchasers reserved rooms for $329 and $108 per night, respectively, while the high- and low-price clickers selected rooms for $404 and $99 per night. The four rental car accounts were able to click and reserve the exact same vehicles, with $184 and $43 being the average high- and low-prices per day.

Each night, after we finished manually creating account histories, we used PhantomJS to run our standard list of 20 queries from all six accounts on all nine travel sites. To maintain consistency, manual history creation and automated tests all used the same set of IP addresses and Firefox.

Ethics. We took several precautions to minimize any negative impact of our purchase history experiments on travel retailers, hotels, and rental car agencies. We reserved, at most, one room from any specific hotel. All reservations were made for hotel rooms and cars at least one month into the future, and all reservations were canceled at the conclusion of our experiments.

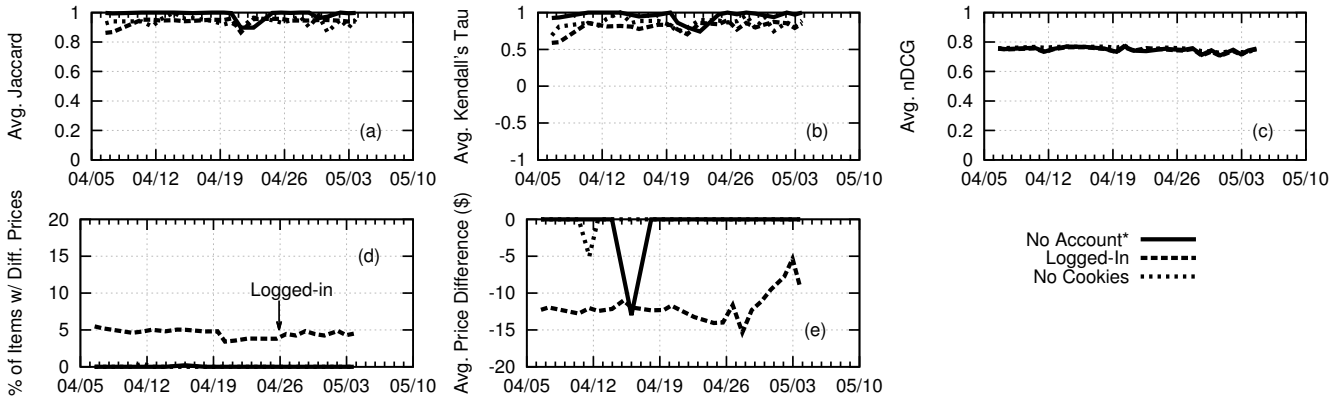


Figure 6: Examining the impact of user accounts and cookies on hotel searches on Cheaptickets.

Analyzing Results. To analyze the data from our feature experiments, we leverage the same five metrics used in § 4. Figure 6 exemplifies the analysis we conduct for each user feature on each e-commerce site. In this example, we examine whether Cheaptickets personalizes results for users that are logged-in. The x-axis of each subplot is time in days. The plots in the top row use Jaccard Index, Kendall's τ, and nDCG to analyze steering, while the plots in the bottom row use percent of items with inconsistent prices and average price difference to analyze discrimination.

All of our analysis is always conducted relative to a control. In all of the figures in this section, the starred (*) feature in the key is the control. For example, in Figure 6, all analysis is done relative to a PhantomJS instance that does not have a user account. Each point is an average of the given metric across all 20 queries on that day.

In total, our analysis produced >360 plots for the various features across all 16 e-commerce cites. Overall, most of the experiments do not reveal evidence of steering or discrimination. Thus, for the remainder of this section, we focus on the particular features and sites where we do observe personalization. None of our feature tests revealed personalization on rental car sites, so we omit them entirely.

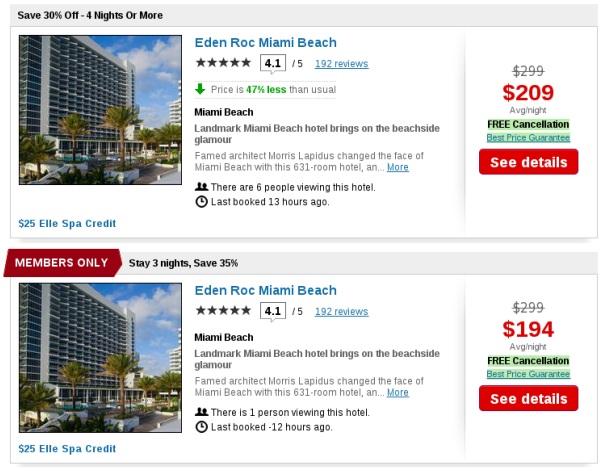


Figure 7: Price discrimination on Cheaptickets. The top result is shown to users that are not logged-in. The bottom result is a "Members Only" price shown to logged-in users.

5.2 Hotels

We begin by analyzing personalization on hotel sites. We observe hotel sites implementing a variety of personalization strategies, so we discuss each case separately.

Cheaptickets and Orbitz. The first sites that we examine are Cheaptickets and Orbitz. These sites are actually one company, and appear to be implemented using the same HTML structure and server-side logic. In our experiments, we observe both sites personalizing hotel results based on user accounts; for brevity we present the analysis of Cheaptickets and omit Orbitz.

Figures 6(a) and (b) reveal that Cheaptickets serves slightly different sets of results to users who are logged-in to an account, versus users who do not have an account or who do not store cookies. Specifically, out of 25 results per page, ≈2 are new and ≈1 is moved to a different location on average for logged-in users. In some cases (e.g., hotels in Bangkok and Montreal) the differences are much larger: up to 11 new and 11 moved results. However, the nDCG analysis in Figure 6(c) indicates that these alterations do not have an appreciable impact on the price of highly-ranked search results. Thus, we do not observe Cheaptickets or Orbitz steering results based on user accounts.

However, Figure 6(d) shows that logged-in users receive different prices on ≈5% of hotels. As shown in Figure 6(e), the hotels with inconsistent prices are $12 cheaper on average. This demonstrates that Cheaptickets and Orbitz implement price discrimination, in favor of users who have accounts on these sites. Manual examination reveals that these sites offer "Members Only" price reductions on certain hotels to logged-in users. Figure 7 shows an example of this on Cheaptickets.

Although it is not surprising that some e-commerce sites give deals to members, our results on Cheaptickets (and Orbitz) are important for several reasons. First, although members-only prices may be an accepted practice, it still qualifies as price discrimination based on direct segmentation (with members being the target segment). Second, this result confirms the efficacy of our methodology, i.e., we are able to accurately identify price discrimination based on automated probes of e-commerce sites. Finally, our results

		Account			Browser				OS			
		In	No*	Ctrl	FX	IE8	Chr*	Ctrl	OSX	Lin	XP	Win7*
OS	Ctrl	0.4	0.4	0.3	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
	Win7*	**1.0**	**1.0**	**1.0**	0.9	**1.0**	**1.0**	0.9	0.9	0.9	0.9	
	XP	0.9	0.9	0.9	**1.0**	0.9	0.9	**1.0**	**1.0**	**1.0**		
	Lin	0.9	0.9	0.9	**1.0**	0.9	0.9	**1.0**	**1.0**			
	OSX	0.9	0.9	0.9	**1.0**	0.9	0.9	**1.0**				
Browser	Ctrl	0.9	0.9	0.9	**1.0**	0.9	0.9					
	Chr*	**1.0**	**1.0**	**1.0**	0.9	**1.0**						
	IE8	**1.0**	**1.0**	**1.0**	0.9							
	FX	0.9	0.9	0.9								
Acct	Ctrl	**1.0**	**1.0**									
	No*	**1.0**										

Table 4: Jaccard overlap between pairs of user feature experiments on Expedia.

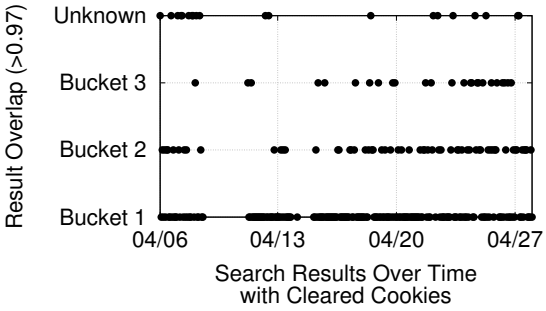


Figure 8: Clearing cookies causes a user to be placed in a random bucket on Expedia.

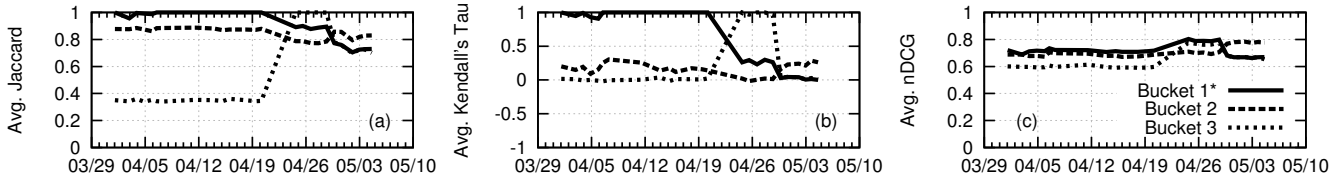


Figure 9: Users in certain buckets are steered towards higher priced hotels on Expedia.

reveal the actual differences in prices offered to members, which may not otherwise be public information.

Hotels.com and Expedia. Our analysis reveals that Hotels.com and Expedia implement the same personalization strategy: randomized A/B tests on users. Since these sites are similar, we focus on Expedia and omit the details of Hotels.com.

Initially, when we analyzed the results of our feature tests for Expedia, we noticed that the search results received by the control and its twin never matched. More oddly, we also noticed that 1) the control results did match the results received by other specific treatments, and 2) these matches were consistent over time.

These anomalous results led us to suspect that Expedia was randomly assigning each of our treatments to a "bucket". This is common practice on sites that use A/B testing: users are randomly assigned to buckets based on their cookie, and the characteristics of the site change depending on the bucket you are placed in. Crucially, the mapping from cookies to buckets is deterministic: a user with cookie C will be placed into bucket B whenever they visit the site unless their cookie changes.

To determine whether our treatments are being placed in buckets, we generate Table 4, which shows the Jaccard Index for 12 pairs of feature experiments on Expedia. Each table entry is averaged over 20 queries and 10 days. For a typical website, we would expect the control (*Ctrl*) in each category to have perfect overlap (1.0) with its twin (marked with a *). However, in this case the perfect overlaps occur between random pairs of tests. For example, the results for Chrome and Firefox perfectly overlap, but Chrome has low overlap with the control, which was also Chrome. This strongly suggests that the tests with perfect overlap have been randomly assigned to the same bucket. In this case, we observe three buckets: 1) {Windows 7, account control, no account, logged-in, IE 8, Chrome}, 2) {XP, Linux, OS X, browser control, Firefox}, and 3) {OS control}.

To confirm our suspicion about Expedia, we examine the behavior of the experimental treatment that clears its cookies after every query. We propose the following hypothesis: if Expedia is assigning users to buckets based on cookies, then the clear cookie treatment should randomly change buckets after each query. Our assumption is that this treatment will receive a new, random cookie each time it queries Expedia, and thus its corresponding bucket will change.

To test this hypothesis we plot Figure 8, which shows the Jaccard overlap between search results received by the clear cookie treatment, and results received by treatments in other buckets. The x-axis corresponds to the search results from the clear cookie treatment over time; for each page of results, we plot a point in the bucket (y-axis) that has >0.97 Jaccard overlap with the clear cookie treatment. If the clear cookie treatment's results do not overlap with results from any of the buckets, the point is placed on the "Unknown" row. In no cases did the search results from the clear cookie treatment have >0.97 Jaccard with more than a single bucket, confirming that the set of results returned to each bucket are highly disjoint (see Table 4).

Figure 8 confirms that the clear cookie treatment is randomly assigned to a new bucket on each request. 62% of results over time align with bucket 1, while 21% and 9% match with buckets 2 and 3, respectively. Only 7% do not match any known bucket. These results suggest that Expedia does not assign users to buckets with equal probability. There also appear to be time ranges where some buckets are not assigned, e.g., bucket 3 in between 04/12 and 04/15. We found that Hotels.com also assigns users to one of three buckets, that the assignments are weighted, and that the weights change over time.

Now that we understand how Expedia (and Hotels.com) assign users to buckets, we can analyze whether users in different buckets receive personalized results. Figure 9 presents

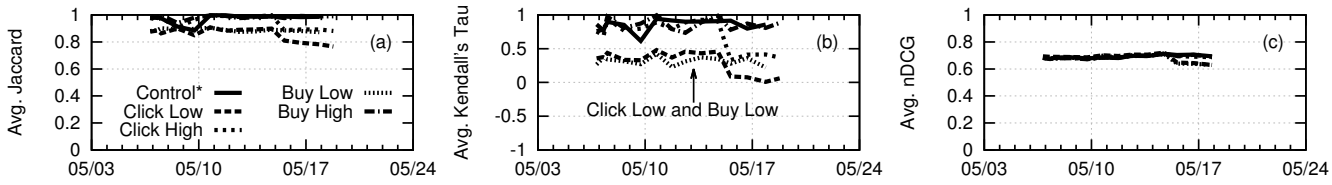


Figure 10: Priceline alters hotel search results based on a user's click and purchase history.

the results of this analysis. We choose an account from bucket 1 to use as a control, since bucket 1 is the most frequently assigned bucket.

Two conclusions can be drawn from Figure 9. First, we see that users are periodically shuffled into different buckets. Between 04/01 and 04/20, the control results are consistent, i.e., Jaccard and Kendall's τ for bucket 1 are $\approx$1. However, on 04/21 the three lines change positions, implying that the accounts have been shuffled to different buckets. It is not clear from our data how often or why this shuffling occurs.

The second conclusion that can be drawn from Figure 9 is that Expedia is steering users in some buckets towards more expensive hotels. Figures 9(a) and (b) show that users in different buckets receive different results in different orders. For example, users in bucket 3 see >60% different search results compared to users in other buckets. Figure 9(c) highlights the net effect of these changes: results served to users in buckets 1 and 2 have higher nDCG values, meaning that the hotels at the top of the page have higher prices. We do not observe price discrimination on Expedia or Hotels.com.

Priceline. As depicted in Figure 10, Priceline alters hotel search results based on the user's history of clicks and purchases. Figures 10(a) and (b) show that users who clicked on or reserved low-price hotel rooms receive slightly different results in a much different order, compared to users who click on nothing, or click/reserve expensive hotel rooms. We manually examined these search results but could not locate any clear reasons for this reordering. The nDCG results in Figure 10(c) confirm that the reordering is not correlated with prices. Thus, although it is clear that account history impacts search results on Priceline, we cannot classify the changes as steering. Furthermore, we observe no evidence of price discrimination based on account history on Priceline.

Travelocity. As shown in Figure 11, Travelocity alters hotel search results for users who browse from iOS devices. Figures 11(a) and (b) show that users browsing with Safari on iOS receive slightly different hotels, and in a much different order, than users browsing from Chrome on Android, Safari on OS X, or other desktop browsers. Note that we started our Android treatment at a later date than the other treatments, specifically to determine if the observed changes on Travelocity occurred on all mobile platforms or just iOS.

Although Figure 11(c) shows that this reordering does not result in price steering, Figures 11(d) and (e) indicate that Travelocity does modify prices for iOS users. Specifically, prices fall by $\approx$\$15 on $\approx$5% of hotels (or 5 out 50 per page) for iOS users. The noise in Figure 11(e) (e.g., prices increasing by \$50 for Chrome and IE 8 users) is not significant: this result is due to a single hotel that changed price.

The takeaway from Figure 11 is that we observe evidence consistent with price discrimination in favor of iOS users on Travelocity. Unlike Cheaptickets and Orbitz, which clearly mark sale-price "Members Only" deals, there is no visual cue on Travelocity's results that indicates prices have been changed for iOS users. Online travel retailers have publicly stated that mobile users are a high-growth customer segment, which may explain why Travelocity offers price-incentives to iOS users [26].

5.3 General Retailers

Home Depot. We now turn our attention to general retail sites. Among the 10 retailers we examined, only Home Depot revealed evidence of personalization. Similar to our findings on Travelocity, Home Depot personalizes results for users with mobile browsers. In fact, the Home Depot website serves HTML with different structure and CSS to desktop browsers, Safari on iOS, and Chrome on Android.

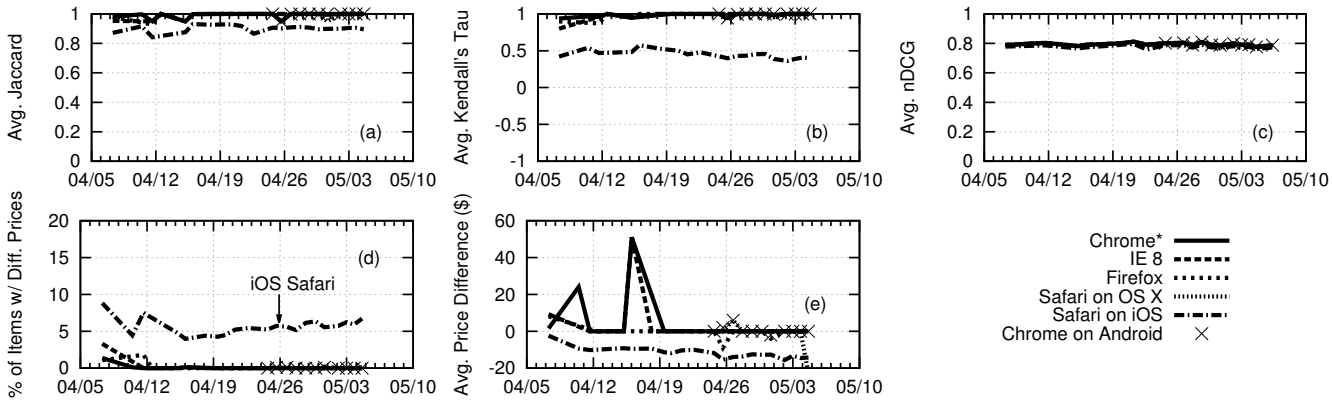


Figure 11: Travelocity alters hotel search results for users of Safari on iOS, but not Chrome on Android.

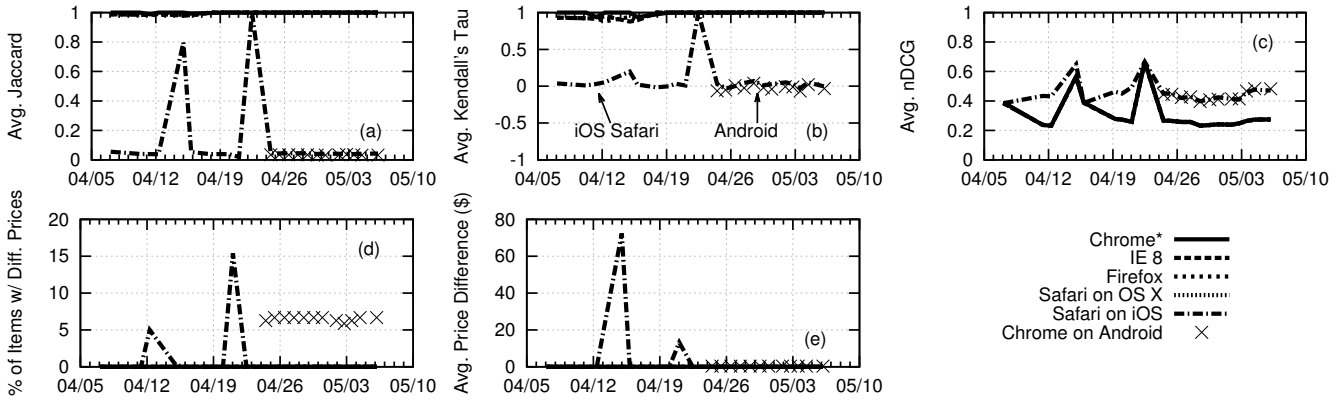


Figure 12: Home Depot alters product searches for users of mobile browsers.

Figure 12 depicts the results of our browser experiments on Home Depot. Strangely, Home Depot serves 24 search results per page to desktop browsers and Android, but serves 48 to iOS. As shown in Figure 12(a), on most days there is close to zero overlap between the results served to desktop and mobile browsers. Oddly, there are days when Home Depot briefly serves identical results to all browsers (e.g., the spike in Figure 12(a) on 4/22). The pool of results served to mobile browsers contains more expensive products overall, leading to higher nDCG scores for mobile browsers in Figure 12(c). Note that nDCG is calculated using the top k results on the page, which in this case is 24 to preserve fairness between iOS and the other browsers. Thus, Home Depot is steering users on mobile browsers towards more expensive products.

In addition to steering, Home Depot also discriminates against Android users. As shown in Figure 12(d), the Android treatment consistently sees differences on ≈6% of prices (one or two products out of 24). However, the practical impact of this discrimination is low: the average price differential in Figure 12(e) for Android is ≈$0.41. We manually examined the search results from Home Depot and could not determine why the Android treatment receives slightly increased prices. Prior work has linked price discrimination on Home Depot to changes in geolocation [30], but we control for this effect in our experiments.

It is possible that the differences we observe on Home Depot may be artifacts caused by different server-side implementations of the website for desktop and mobile users, rather than an explicit personalization algorithm. However, even if this is true, it still qualifies as personalization according to our definition (see § 2.3) since the differences are deterministic and triggered by client-side state.

6. RELATED WORK

In this section, we briefly overview the academic literature on implementing and measuring personalization.

Improving Personalization. Personalizing search results to improve Information Retrieval (IR) accuracy has been extensively studied in the literature [33, 42]. While these techniques typically use click histories for personalization, other features have also been used, including geolocation [2, 53, 54] and demographics (typically inferred from search and browsing histories) [18, 47]. To our knowledge, only one study has investigated privacy-preserving personalized search [52]. Dou et al. provide a comprehensive overview of techniques for personalizing search [12]. Several studies have looked at improving personalization on systems other than search, including targeted ads on the web [16,49], news aggregators [9, 24], and even discriminatory pricing on travel search engines [15].

Comparing Search Results. Training and comparing IR systems requires being able to compare ranked lists of search results. Thus, metrics for comparing ranked lists are an active area of research in the IR community. Classical metrics such as Spearman's footrule and ρ [11, 38] and Kendall's τ [22] both calculate pairwise disagreement between ordered lists. Several studies improve on Kendall's τ by adding per-rank weights [14, 40], and by taking item similarity into account [21, 39]. DCG and nDCG use a logarithmic scale to reduce the scores of bottom ranked items [19].

The Filter Bubble. Activist Eli Pariser brought widespread attention to the potential for web personalization to lead to harmful social outcomes; a problem he dubbed The Internet Filter Bubble [31]. This has motivated researchers to begin measuring the personalization present in deployed systems, such as web search engines [17, 27, 50] and recommender systems [4].

Exploiting Personalization. Recent work has shown that it is possible to exploit personalization algorithms for nefarious purposes. Xing et al. [51] demonstrate that repeatedly clicking on specific search results can cause search engines to rank those results higher. An attacker can exploit this to promote specific results to targeted users. Thus, fully understanding the presence and extent of personalization today can aid in understanding the potential impact of these attacks on e-commerce sites.

Personalization of E-commerce. Two recent studies by Mikians et al. that measure personalization on e-commerce sites serve as the inspiration for our own work. The first study examines price steering (referred to as *search discrimination*) and price discrimination across a large number of sites using fake user profiles [29]. The second paper extends the first work by leveraging crowdsourced workers to help detect price discrimination [30]. The authors identify several e-commerce sites that personalize content, mostly based on the user's geolocation. We improve upon these

studies by introducing duplicated control accounts into all of our measurements. These additional control accounts are necessary to conclusively differentiate inherent noise from actual personalization.

7. CONCLUDING DISCUSSION

Personalization has become a important feature of many web services in recent years. However, there is mounting evidence that e-commerce sites are using personalization algorithms to implement price steering and discrimination.

In this paper, we build a measurement infrastructure to study price discrimination and steering on 16 top online retailers and travel websites. Our method places great emphasis on controlling for various sources of noise in our experiments, since we have to ensure that the differences we see are actually a result of personalization algorithms and not just noise. *First*, we collect real-world data from 300 AMT users to determine the extent of personalization that they experience. This data revealed evidence of personalization on four general retailers and five travel sites, including cases where sites altered prices by hundreds of dollars.

Second, we ran controlled experiments to investigate what features e-commerce personalization algorithms take into account when shaping content. We found cases of sites altering results based on the user's OS/browser, account on the site, and history of clicked/purchased products. We also observe two travel sites conducting A/B tests that steer users towards more expensive hotel reservations.

Comments from Companies. We reached out to the six companies we identified in this study as implementing some form of personalization (Orbitz and Cheaptickets are run by a single company, as are Expedia and Hotels.com) asking for comments on a pre-publication draft of this manuscript. We received responses from Orbitz and Expedia. The Vice President for Corporate Affairs at Orbitz provided a response confirming that Cheaptickets and Orbitz offer members-only deals on hotels. However, their response took issue with our characterization of price discrimination as "anti-consumer"; we removed these assertions from the final draft of this manuscript. The Orbitz representative kindly agreed to allow us to publish their letter on the Web [7].

We also spoke on the phone with the Chief Product Officer and the Senior Director of Stats Optimization at Expedia. They confirmed our findings that Expedia and Hotels.com perform extensive A/B testing on users. However, they claimed that Expedia does not implement price discrimination on rental cars, and could not explain our results to the contrary (see Figure 3).

Scope. In this study we focus on US e-commerce sites. All queries are made from IP addresses in the US, all retailers and searches are in English, and real world data is collected from users in the US. We leave the examination of personalization on e-commerce sites in other countries and other languages to future work.

Incompleteness. As a result of our methodology, we are only able to identify positive instances of price discrimination and steering; we cannot claim the absence of personalization, as we may not have considered other dimensions along which e-commerce sites might personalize content. We observe personalization in some of the AMT results (e.g., on Newegg and Sears) that we cannot explain with the findings from our feature-based experiments. These effects might be explained by measuring the impact of other features, such as geolocation, HTTP Referer, browsing history, or purchase history. Given the generality of our methodology, it would be straightforward to apply it to these additional features, as well as to other e-commerce sites.

Open Source. All of our experiments were conducted in spring of 2014. Although our results are representative for this time period, they may not hold in the future, as the sites may change their personalization algorithms. We encourage other researchers to repeat our measurements by making all of our crawling and parsing code, as well as the raw data from § 4 and § 5, available to the research community at

`http://personalization.ccs.neu.edu/`

Acknowledgements

We thank the anonymous reviewers and our shepherd, Vijay Erramilli, for their helpful comments. This research was supported in part by NSF grants CNS-1054233 and CHS-1408345, ARO grant W911NF-12-1-0556, and an Amazon Web Services in Education grant.

8. REFERENCES

[1] Bezos calls Amazon experiment 'a mistake'. Puget Sound Business Journal, 2000. `http://www.bizjournals.com/seattle/stories/2000/09/25/daily21.html`.

[2] L. Andrade and M. J. Silva. Relevance Ranking for Geographic IR. *GIR*, 2006.

[3] Amazon mechanical turk. `http://mturk.com/`.

[4] F. Bakalov, M.-J. Meurs, B. König-Ries, B. Satel, R. G. Butler, and A. Tsang. An Approach to Controlling User Models and Personalization Effects in Recommender Systems. *IUI*, 2013.

[5] K. Bhasin. JCPenney Execs Admit They Didn't Realize How Much Customers Were Into Coupons. Business Insider, 2012. `http://www.businessinsider.com/jcpenney-didnt-realize-how-much-customers-\\were-into-coupons-2012-5`.

[6] P. Belobaba, A. Odoni, and C. Barnhart. *The Global Airline Industry*. Wiley, 2009.

[7] C. Chiames. Correspondence with the authors, in reference to a pre-publication version of this manuscript, 2014. `http://personalization.ccs.neu.edu/orbitz_letter.pdf`.

[8] R. Calo. Digital Market Manipulation. *The George Washington Law Review*, 82, 2014.

[9] A. Das, M. Datar, A. Garg, and S. Rajaram. Google News Personalization: Scalable Online Collaborative Filtering. *WWW*, 2007.

[10] C. Duhigg. How Companies Learn Your Secrets. The New York Times, 2012. `http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html`.

[11] P. Diaconis and R. L. Graham. Spearman's Footrule as a Measure of Disarray. *J. Roy. Stat. B*, 39(2), 1977.

[12] Z. Dou, R. Song, and J.-R. Wen. A Large-scale Evaluation and Analysis of Personalized Search Strategies. *WWW*, 2007.

[13] J. H. Dorfman. *Economics and Management of the Food Industry*. Routledge, 2013.

[14] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SODA*, 2003.

[15] A. Ghose, P. G. Ipeirotis, and B. Li. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*, 31(3), 2012.

[16] S. Guha, B. Cheng, and P. Francis. Challenges in Measuring Online Advertising Systems. *IMC*, 2010.

[17] A. Hannak, P. Sapiezyński, A. M. Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. Measuring Personalization of Web Search. *WWW*, 2013.

[18] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic Prediction Based on User's Browsing Behavior. *WWW*, 2007.

[19] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. *SIGIR*, 2000.

[20] K. Järvelin and J. Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. *ACM TOIS*, 20(4), 2002.

[21] R. Kumar and S. Vassilvitskii. Generalized Distances Between Rankings. *WWW*, 2010.

[22] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2), 1938.

[23] W. H. Kruskal. Ordinal Measures of Association. *Journal of the American Statistical Association*, 53(284), 1958.

[24] L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. *WWW*, 2010.

[25] B. D. Lollis. Orbitz: Mac users book fancier hotels than PC users. USA Today Travel Blog, 2012. `http://travel.usatoday.com/hotels/post/2012/05/orbitz-hotel-booking-mac-pc-/690633/1`.

[26] B. D. Lollis. Orbitz: Mobile searches may yield better hotel deals. USA Today Travel Blog, 2012. `http://travel.usatoday.com/hotels/post/2012/05/orbitz-mobile-hotel-deals/691470/1`.

[27] A. Majumder and N. Shrivastava. Know your personalization: learning topic level personalization in online services. *WWW*, 2013.

[28] D. Mattioli. On Orbitz, Mac Users Steered to Pricier Hotels. *The Wall Street Journal*, 2012. `http://on.wsj.com/LwTnPH`.

[29] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Detecting Price and Search Discrimination on the Internet. *HotNets*, 2012.

[30] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Crowd-assisted Search for Price Discrimination in E-Commerce: First results. *CoNEXT*, 2013.

[31] E. Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin Press, 2011.

[32] I. Png. *Managerial Economics*. Routledge, 2012.

[33] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *CACM*, 45(9), 2002.

[34] Panopticlick. `https://panopticlick.eff.org`.

[35] PhantomJS. 2013. `http://phantomjs.org`.

[36] A. Ramasastry. Web sites change prices based on customers' habits. CNN, 2005. `http://edition.cnn.com/2005/LAW/06/24/ramasastry.website.prices/`.

[37] C. Shapiro and H. R. Varian. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, 1999.

[38] C. Spearman. The Proof and Measurement of Association between Two Things. *Am J Psychol*, 15, 1904.

[39] D. Sculley. Rank Aggregation for Similar Items. *SDM*, 2007.

[40] G. S. Shieh, Z. Bai, and W.-Y. Tsai. Rank Tests for Independence—With a Weighted Contamination Alternative. *Stat. Sinica*, 10, 2000.

[41] Selenium. 2013. `http://selenium.org`.

[42] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. *KDD*, 2006.

[43] Top 500 e-retailers. `http://www.top500guide.com/top-500/`.

[44] Top booking sites. `http://skift.com/2013/11/11/top-25-online-booking-sites-in-travel/`.

[45] T. Vissers, N. Nikiforakis, N. Bielova, and W. Joosen. Crying Wolf? On the Price Discrimination of Online Airline Tickets. *HotPETs*, 2014.

[46] J. Valentino-Devries, J. Singer-Vine, and A. Soltani. Websites Vary Prices, Deals Based on Users' Information. *Wall Street Journal*, 2012. `http://online.wsj.com/news/articles/SB10001424127887323777204578189391813881534`.

[47] I. Weber and A. Jaimes. Who Uses Web Search for What? And How? *WSDM*, 2011.

[48] T. Wadhwa. How Advertisers Can Use Your Personal Information to Make You Pay Higher Prices. *Huffington Post*, 2014. `http://www.huffingtonpost.com/tarun-wadhwa/how-advertisers-can-use-y_b_4703013.html`.

[49] C. E. Wills and C. Tatar. Understanding What They Do with What They Know. *WPES*, 2012.

[50] X. Xing, W. Meng, D. Doozan, N. Feamster, W. Lee, and A. C. Snoeren. Exposing Inconsistent Web Search Results with Bobble. *PAM*, 2014.

[51] X. Xing, W. Meng, D. Doozan, A. C. Snoeren, N. Feamster, and W. Lee. Take This Personally: Pollution Attacks on Personalized Services. *USENIX Security*, 2013.

[52] Y. Xu, B. Zhang, Z. Chen, and K. Wang. Privacy-Enhancing Personalized Web Search. *WWW*, 2007.

[53] B. Yu and G. Cai. A query-aware document ranking method for geographic information retrieval. *GIR*, 2007.

[54] X. Yi, H. Raghavan, and C. Leggetter. Discovering Users' Specific Geo Intention in Web Search. *WWW*, 2009.

Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations

Aniko Hannak
Northeastern University
ancsaaa@ccs.neu.edu

Drew Margolin
Cornell University
dm658@cornell.edu

Brian Keegan
Northeastern University
b.keegan@neu.edu

Ingmar Weber
Qatar Computing Research Institute
iweber@qf.org.qa

Abstract

The prevalence of misinformation within social media and online communities can undermine public security and distract attention from important issues. Fact-checking interventions, in which users cite fact-checking websites such as Snopes.com and FactCheck.org, are a strategy users can employ to refute false claims made by their peers. While laboratory research suggests such interventions are not effective in persuading people to abandon false ideas, little work considers how such interventions are actually deployed in real-world conversations. Using approximately 1,600 interventions observed on Twitter between 2012 and 2013, we examine the contexts and consequences of fact-checking interventions. We focus in particular on the social relationship between the individual who issues the fact-check and the individual whose facts are challenged. Our results indicate that though fact-checking interventions are most commonly issued by strangers, they are more likely to draw user attention and responses when they come from friends. Finally, we discuss implications for designing more effective interventions against misinformation.

"Why you all in my ear?
Talkin' a whole bunch of sh*t that I ain't tryin' to hear
Get back mother**cker you don't know me like that"
— Ludacris, *Get Back* (2005)

Introduction

False rumor and other forms of misinformation are important issues of public concern. Many fear that new communication technologies may be accelerating the spread of misinformation, yet others point out that new technologies also offer the potential for a broader dissemination of facts to counter false claims (Garrett 2011). In previous eras, false information shared within casual conversations could only be confronted after an individual undertook to research the claim on their own, perhaps at great cost or delay. Today, individuals can draw on a wealth of well-sourced and widely recognized information sources including government websites, Wikipedia, and other traditional and non-traditional sources to engage in immediate *fact-checking interventions* against false claims issued in digital spaces.

Previous research suggests such interventions are likely to have minimal effects. Studies consistently show that individuals resist updating beliefs in the face of contradictory facts (Lewandowsky et al. 2012). However, the bulk of research has treated fact-checking as a largely asocial activity — something that professional media can do to for audiences — rather than a social phenomenon that occurs in real-world conversations. In particular, research has not considered how fact-checking might operate differently when when the check comes from a friend, a person with whom the individual has an ongoing relationship and debt of mutual accountability, or someone who has a particularly high or low status within their community. A fact-check is a form of criticism and a potential source of embarrassment. Research on social networks and influence backs up Ludacris's argument in the epigraph that individuals' receptiveness to such critiques are contingent upon the qualities of the social tie over which it takes place (Friedkin and Johnsen 1999; Cialdini and Goldstein 2004). Thus, it may not be surprising that corrections in the absence of such relationships only reinforce their beliefs in misinformation and/or dismiss the competing claim as unworthy of consideration.

In this paper we draw on the fine-grained records of discursive behavior available through Twitter to observe real-world fact-checking interventions in action. In particular, we observe cases where individuals deploy fact-checking websites (Snopes.com, PolitiFact.com, and FactCheck.org) in reply to statements made by others. These fact-checking replies, which we call *snopes*, provide a small but well defined sub-population of conversational fact-checks on the internet. By observing the dynamics around these explicit and recognizable interventions, we address three basic research questions. First, what are the typical social structural contexts of fact-checking interventions ("snopes") as they occur in the real-world? Second, what are the consequences of these snopes on subsequent user behavior? Third, what are the macro-level structural features of the networks formed by these differences in fact-checking behavior over individual relationships?

Our results reveal significant differences between the social status, behavioral consequences, and structural positions of users making ("snopers") and receiving ("snopees")

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

fact-checks. We find evidence that fact-checking behavior over weaker relationships involves significant differences in users' relative status. Consistent with prior findings, most fact-checks go un-acknowledged by the snopee. However, users who are snoped by "friends" sharing a reciprocated following relationship are over three times more likely to respond afterwards than over other relationship types. Finally, these fact-checking interventions coalesce into a network characterized by highly polarized partisan sniping with fact-checks by friends occurring almost exclusively on the periphery. Recognizing the social contexts and motivations for communicating misinformation can better inform the design of more effective fact-checking strategies.

Background and Theory

Misinformation can have dangerous consequences and yet false rumors are surprisingly stubborn in the face of challenge (Lewandowsky et al. 2012; Nyhan and Reifler 2010). Disturbingly, the presentation of "correct" information often has small effects on individual attitudes and beliefs, limited to particular contexts or specialized circumstances (Garrett 2011; Ecker, Lewandowsky, and Tang 2010). Prior work has repeatedly highlighted the social nature of rumor and misinformation, yet misinformation "correction" studies have, on the other hand, focused almost exclusively on asocial interventions in which individuals receive "corrections" from generic others or abstract institutions.

Sharing misinformation is a social activity, where-in the truth value of claims is often less important than the social and group related functions the claim helps perform (Allport and Postman 1947). Specious assertions can help maintain a group's identity and coherence as symbolic displays of a shared worldview (Foster 2004). Assertions can also help to communicate shared norms and values by implying the kinds of ideas that members of the group ought to hold (Baumeister, Zhang, and Vohs 2004). False information can have practical importance for a group, where speculative ideas can help individuals to collectively make sense of novel or highly uncertain situations (DiFonzo and Bordia 2007; Lewandowsky et al. 2012).

Despite the importance of social motivations in sharing misinformation, research on the effects of the impact of factual corrections has largely ignored this social component. The most commonly tested form of correction involves presenting an individual with facts from an asocial source, such as a generic authority or a large institution such as a news organization. The basic consensus from such research is that misinformation is "sticky." In particular, individuals are unlikely to shift attitudes or beliefs based on misinformation when this information is corrected, particularly if the original, false idea is consistent with a belief or worldview they already hold (Ecker, Lewandowsky, and Tang 2010; Nyhan and Reifler 2010). Even retractions, in which the facts are presented by the same source that provided the original misinformation, do not eliminate belief in the original (false) claim (Lewandowsky et al. 2012).

These findings suggest it is difficult to dislodge false information, and that providing accurate facts will not contribute much to such efforts. Given the highly social nature of misinformation communication, however, it is not surprising that these interventions have not been effective. We seek to investigate a different kind of correction which we term a fact-checking intervention or *snope*. We define a snope as an attempt by an individual to get another individual to pay attention to the facts on a given topic or within a conversation. Such snopes can include overt corrections of misinformation as well as broader attempts to inject factual information into online discussions.

We argue that "social snopes" should be more effective than "asocial snopes" for several reasons. Social networks are important carriers of influence regarding beliefs and attitudes (Friedkin and Johnsen 1999). Mechanisms such as the desire to reduce mental conflict and trust in the judgments of friends will shift individuals attitudes to be more similar to that of their friends, family and colleagues (Cialdini and Goldstein 2004). Snopes can be strong signals of such beliefs and attitudes, particularly on topics which friends may not frequently discuss (Garrett 2011; Garrett, Nisbet, and Lynch 2013). Friends are also likely to share common elements of worldviews which in turn makes snopes from these actors more believable and self-consistent (Weick 1995). Whereas snopes from strangers or generic authorities can be threatening, snopes from friends are more likely to provide the necessary context to integrate the facts within the shared worldview, making it palatable and worthy of consideration (Garrett, Nisbet, and Lynch 2013; Kahan 2010).

Social network structures can also carry normative restrictions on behavior. An individual's behavioral decisions are influenced not only by their beliefs but by their assessment of who and how they will be judged (Tetlock 2002). Individuals that share friends are structurally embedded within groups that can create pressures to conform (Granovetter 1985; Uzzi 1997). A snope from a friend embedded in the same group can signal that a general norm of factual reliability is in effect, or that speculations, whether position-consistent or not, will not be tolerated (Cialdini and Goldstein 2004).

Finally, status and leadership may play important roles in how individuals react to fact-checking interventions. Research on corrections suggests that individuals are more likely to update beliefs when they are corrected by leaders of their group (Lewandowsky et al. 2012). More broadly, popularity and social status within a social network may indicate the relative risks of scrutiny and reputation that individuals face. Leaders may feel the need to maintain credibility. Alternatively, low status members may feel the need to avoid mistakes and having status and reputation fall further.

Social network information has been used successfully in the automatic detection of misinformation on Twitter (Qazvinian et al. 2011). Consistent with our argument, credible tweets tend to come from individuals with a greater number of friends and followers (Castillo, Mendoza, and Poblete 2011), and once detected as a distributor of misinformation, social pressure can lead to Twitter accounts to be shut down (Ratkiewicz et al. 2011). At the same time, evidence suggests users are right to be wary of others' judgments of their credibility in the absence of a shared social

relationship. Users do a poor job of judging information accuracy based on content features alone, relying instead on user-level information such as user name and other contextual information (Morris et al. 2012; Liao et al. 2012). Thus, there is reason for users to expect that asocial snopes will themselves be misguided.

We thus suggest that understanding the social-structural contexts of snopes is critical to assessing and understanding their effectiveness. This motivates three research questions:

RQ 1 *Who snopes whom?* How does the popularity and status of snopers and snopees vary across relationship types? Are there differences in these users' accounts' ages or the latency of fact-checks?

RQ 2 *Do snopes matter?* What are the consequences of different kinds of social snoping? What social-structural features of snopes provoke the attention of snopees?

RQ 3 *Where do snopes happen?* How do snoping relationships vary across different communities? What does this structure reveal about the motivations and consequences of snoping as it occurs in real-world discourse?

Data and Methods

We expect that "snoping" may take many forms in online discourse. In particular, there are likely many cases where the parties involved are aware that one is snoping the other but this meaning is opaque to outsiders. Unfortunately, addressing our research questions requires cases where it is clear to outside observers that snoping is taking place. To obtain such cases we rely on conversational tweets containing links to fact-checking websites. Specifically, using a corpus of tweets collected between January 2012 and August 2013 from the Twitter "gardenhose" public stream, we identified a subset of 3,969 tweets meeting the following criteria.

Conversational The tweets had to be a reply from a "snoper" to a previous tweet made by a "snopee".

Link to a fact-check The tweets had to include a URL that resolved to one of three prominent fact-checking websites in the U.S.: Snopes.com, PolitiFact.com, or FactCheck.org.

We note that this and subsequent filtering steps create a very conservative sample as it excludes many instances of social snoping involving tweets containing no links, links to other websites, or are not recorded as being parts of conversations. This process generated an edgelist of tweets containing a fact-checking URL from snopers replying to or directed at the parent tweet (snopee) of the thread. For the analysis of changes in the snopee's behavior after the fact-checking intervention we collect the history for each of them using the public Twitter API. Concretely, for each snoper or snopee, we get (up to) 5,000 of their followers, (up to) 5,000 of their followees and (up to) their most recent 3,200 tweets. The follower/followee graph was extracted as of January 2014 and historic follower/followee information is not available through the API. Given that only the most recent 3,200 tweets for a given user could be obtained, we had to make sure we have enough history before the fact-checking

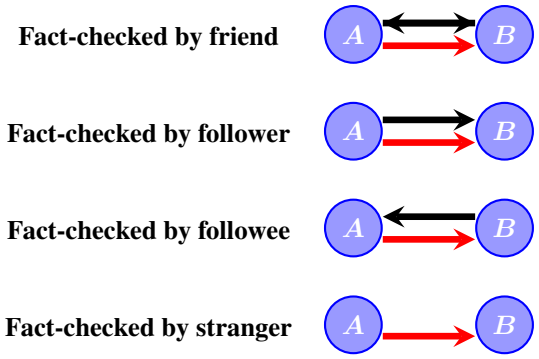


Figure 1: Four dyadic social contexts of fact-checking interventions. For all events where A fact-checks B (in red), B and A can follow each other (friend), A can follow B (follower), B can follow A (followee), or A and B do not follow each other.

intervention. Thus we picked out the users for whom we had historical data starting at least one day before the event. This left us with with 1,614 unique snoping events of which we selected 1,369 snopees who were snoped only once and 1,299 unique snopers who performed these snopes.

For each of the unique snopers and snopees identified above, we extracted the set of users they followed to create a directed following network. We then subsetted this following network to only include the users involved in fact-check interventions as well as the follower edges among these users.

We overlaid edge attributes on top of this subset network based on four distinct dyadic following contexts highlighted in Figure 1. Bob (B) was **snoped by a friend**[1] if Alice (A) fact-checked him and they mutually follow each other. The meaning of "friend" in this context is not meant to convey an offline relationship but simply the fact that they share some degree of mutual accountability by virtue of their reciprocated following relationship. Bob was **snoped by a follower** if Alice, who follows him, also fact-checks him. Unlike with friends, in this case Bob is much less likely to be aware of or concerned with who Alice is. Bob was **snoped by a followee** if he follows Alice and received a fact-check from her. Finally Bob was **snoped by a stranger** if Alice, who does not have a following relationship of any kind, fact-checks him. Intuitively the third scenario (fact-checking a followee) is very unlikely to happen and indeed we have very few data points in this category. We will use these four types of snoping relationships in subsequent analyses.

Direct observation of snoping tweets indicated that while many were used to challenge or correct a statement made by another user, some were intended in other ways, such as to answer a question regarding facts or to make a joke. To determine whether a snoping tweet was challenging the factual basis of the tweet it snoped we used coders from CrowdFlower.com. CrowdFlower is a crowdsourcing platform where paying clients can submit micro tasks, also re-

[1]We deviate from official terminology where a "friend" in Twitter's terminology is a "followee" in our terminology.

ferred to as Human Intelligence Tasks (HIT), and paid contributors/coders can volunteer to work on these tasks. Each task is typically a small set of multiple choice questions and the coders get paid a few dollar cents for completing it.

Coders were instructed to code as a "challenge" any snope in which the snoping tweet attempted to show that facts asserted in the snoped tweet were wrong or that the view expressed by the snopee was based on incorrect facts. For example, a user tweeted:

> ann romney is killing my world series buzz @oshanada http://t.co/LTjL1EKk

This link led to a website that suggested Ann Romney had questioned whether women deserve equal pay. This tweet was then "snoped" by another user who wrote:

> @snopee @otheruser Sorry folks, it's a hoax. The answer is here http://t.co/OmNS7Z19 Yes, we need equality, but facts need to be checked :)

This link led to a refutation of this story on Snopes.com. This pair of tweets was coded as a "challenge."

In contrast, not all links to the websites we identified were necessarily fact-checking interventions. For example, one user wrote:

> @snoper What is the fact checking website you showed us in class again?? It's so hard to keep up with whats the truth or not!

What we would have classified as a "snoper" replied with a link to base PolitiFact.com domain although the intent was clearly not to fact-check this "snopee". While challenges were relatively easy to identify, no clear set of alternative categories was readily apparent.

Relationships between snopers and snopees

Our first research question asks, "who snopes whom?" We begin by identifying the social relationships that typically exist between snoper and snopee. Using the complete list of snopes (rather than the unique snopee subset), we observe 1,290 total snopers, 1,310 snopees, and 36 users who both snoped and were snoped. For each of the four fact-checking relationship types identified in Figure 1, 773 snopes were made by strangers, 442 snopes by friends, 267 snopes by followers, and 51 snopes by followees. Because of their scarcity, we ignore snopes by followees in subsequent analyses. This first descriptive finding indicates that snopes between individuals who are likely to share some sense of mutual respect and accountability are in the minority (roughly 30% = friends), but are not unheard of. Stranger-stranger snopes are the most common form, however.

Structural position

We examined the relationship between snopers and snopes by considering the size and scope of their ego-networks. First, we analyzed the structural position of these users by identifying differences in absolute popularity. Second, we examined the average connectivity of their neighbors to differentiate "celebrities" with many poorly-connected followers from "elites" with fewer but better-connected followers.

In Figure 2, the number of followers for snopers (red) and snopees (green) are plotted across three types of fact-checking interventions. Fact-checks over friend relationships are marked by users' similarity in structural position. Within friend-to-friend snopes, both the snoper and the snopee have similarly-sized audiences. These audiences are also comparatively small. By contrast, fact-checks over follower and stranger relationships are marked by significant (one-way ANOVA, $F = 13.9, p < 0.01$) disparities in structural positions. Snopees in these asocial conditions where they are unlikely to "know their snoper" tend to have tens of thousands of followers on average while snopers are less popular with only hundreds of followers.

We next extend this analysis to the average connectivity of snopers' and snopees' followers as a measure of the user's status as a "celebrity" or an "elite." In Figure 3, the average number of followers' followers are plotted across the three types of fact-checking interventions. As before, fact-checks that occur over friend relationships pair users with similarly-sized audiences. By contrast, fact-checking interventions over following and stranger relationships involve significant asymmetries in users' audience characteristics. In both of these non-friendship cases, snopers' followers have significantly larger followings (one-way ANOVA, $F = 110.6, p < 0.01$) on average than the followers of snopees. In other words, snopers are comparatively "elite": they are followed by users who have more followers themselves than the users who follow snopees.

Finally, we examine the extent to which snopers' and snopees' following networks exhibit clustering. High levels of clustering indicate highly embedded networks in which users' followers also follow each other. In Figure 4, the clustering for snopers and snopees are plotted across the three types of fact-checking interventions. Again, fact-checks by friends and fact-checks by strangers both involve users with levels of clustering. In contrast, fact-checks by followers show a significant difference (one-way ANOVA, $F = 30.8, p < 0.01$): snopers have much more densely-connected local following networks than snopees.

Together these findings present a potential puzzle because snopers tend to have smaller direct audiences but these direct audiences are more tightly connected and have larger indirect audiences than snopees' indirect audiences. This suggests a specific social ecology in which activists embedded within local communities of well-connected actors potentially feel emboldened to call out celebrities.

Temporal relationships

We examine the temporal features surrounding fact-checking interventions to better understand the dynamics of attention to misinformation. Because users who joined Twitter earlier may be motivated to protect the platform from misinformation while users who joined later may lack the skills or shared norms against misinformation, we examine whether there are systematic differences in the ages of snopers' and snopees' accounts. However, we found no significant differences across relationship types between snopers' and snopees' ages.

Because information on Twitter is primarily shared via

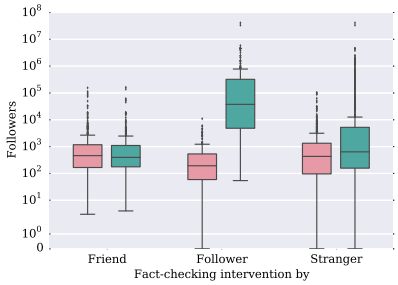


Figure 2: Number of followers.

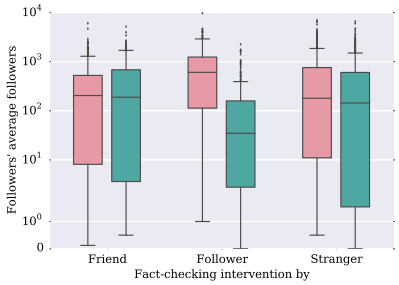


Figure 3: Average followers' followers.

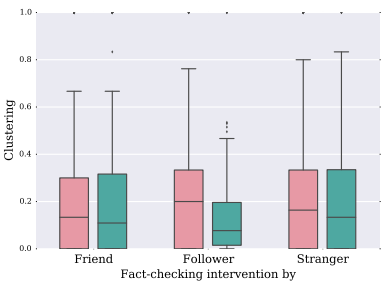


Figure 4: Clustering.

	Mentioned after			
	Yes		No	
Mentioned before	74	(58%)	54	(42%)
Not mentioned before	77	(7%)	1114	(94%)

Table 1: Cross-tabulation of a snopee mentioning the snoper 10 minutes before and after the identified snope.

following relationships, tweets from followees should be seen more quickly and, if misinformed, fact-checked more rapidly. However, there are no significant differences in the temporal lag between snopees' and snopers' messages across relationship types: the bulk of fact-checking interventions occur within an hour of the snopee's statement.

Behavioral consequences of fact-checking

Our second research question asks: How do individuals respond to being snoped? In this section we explore the responses that fact-checking interventions provoke from the snopees to whom they are directed. That is, the "snope" is considered as an experimental treatment provided by the snoper to the snopee. We then examine different qualities of these treatments, specifically, from whom they were issued and whether they "challenged" the respondent and affected their response. Based on the results reported above about the typical time lag between snoping tweets and the tweets to which they reply, we focus our report on snopees' responses over short time windows — 5 to 10 minutes. We have run analysis for up to 24-hour windows following the "snope" and found no substantial changes. All effects in longer time windows are either significant (but weaker) or no longer significant but in the same direction, indicating an expected decay in the effect of the treatment (snope) over time.

Do Snopes Get Recognized?

Our first analysis in this section considers whether snopes get the attention of the individuals to whom they are addressed. Individuals' Twitter streams, particularly amongst highly followed individuals, can be filled with a large number of replies. While the actual cognitive attention extended by snopees to individual tweets in their feed cannot be observed, we can observe how often a snoper is recognized by a snopee in subsequent tweets by considering whether and when the snopee mentions the snoper.

	Mentioned before		Mentioned after	
	Yes	No	Yes	No
Follower	4 (66.7%)	2 (33.3%)	6 (2.9%)	199 (97.1%)
Friend	33 (56.9%)	25 (43.1%)	45 (12.0%)	329 (88.0%)
Stranger	37 (57.8%)	27 (42.2%)	26 (4.2%)	586 (95.8%)

Table 2: Cross-tabulation of previous and subsequent mention of snopers by snopees across relationships types over 10-minute window.

Using this information, we divide our snoper-snopee conversations into two types. The first type are snopes "within ongoing conversation," where the snopee had mentioned the snoper in the minutes prior to the snoping event. The second type are "out-of-the-blue" snopes, where the snopee has not recently mentioned the snoper. As shown in Table 1, there is a substantial difference. The vast majority of snopes occur "out-of-the-blue." Moreover, the distinction between "ongoing" and "out-of-the-blue" snopes is important for whether the snope is subsequently recognized. If a snopee has already mentioned a snoper 10 minutes or less prior to getting snoped, they are likely to mention them after the snope. When the snoper has not been mentioned recently, however, the probability of being mentioned after a snope is significantly smaller (6.5%) ($\chi^2 = 300.60, df = 1, p < .001$). These patterns are robust to at least 24 hrs of conversation, where 2/3 of users mentioned within 1 day prior to the snope are mentioned within a day after the snope, but only 1 in 9 users not mentioned prior are mentioned subsequently.

We next consider how recognition relates to the dyadic relationship between the snoper and the snopee. Table 2 shows the difference between the likelihood of a response to friends as opposed to other relationship types when under the "out-of-the-blue" versus "ongoing conversation" snope conditions. For "ongoing conversations," there is no significant difference. No matter whom the snope comes from, snopees are roughly equally likely to respond. For "out-of-the-blue" snopes, however, there is a substantial difference.

	Unchallenged		Challenged	
Follower	6	(26.1%)	17	(73.9%)
Friend	38	(34.2%)	73	65.8%
Stranger	19	(15.0%)	108	(85.0%)

Table 3: Cross-tabulation of challenged and unchallenged snopes by relationship type.

	Mentioned after			
	Yes		No	
Follower	4	(29%)	10	(71%)
Friend	26	(45%)	32	(55%)
Stranger	20	(27%)	53	(73%)

Table 4: Cross-tabulation of challenged snopes by relationship type for 10-minute window and out-of-the-blue conversations.

The table shows that friends are about 3 times more likely to respond to "out-of-the-blue" snopes from friends than they are from strangers ($\chi^2 = 21.05, df = 1, p < .001$). This finding provides evidence that the ongoing relationship and mutual accountability that are part of a reciprocated following relationship on social media has impacts on how individuals react to snopes. That is, while users tend to ignore "out-of-the-blue" snopes in general, they are significantly more likely to response to "out-of-the-blue" snopes from friends — individuals whom they follow and who follow them — than from non-friend relationships.

One possible explanation for this phenomenon is that the friends are using fact-checking websites differently when replying to other friends. That is, friends may provide references to fact-checking websites as support for their friends point of view, but primarily to challenge or criticize strangers. To examine this possibility, we hand-coded snopee and snoper tweets for all cases where the snopee recognized the snoper within 24 hours after the snoping event. Four coders examined each pair of tweets, and the 261 pairs where at least 3 of the 4 coders agreed were retained for further analysis. Snoper–snopee tweet pairs were coded as either "factual challenges" (76%) or "other" (24%).

Table 3 shows a clear distinction. For friends and (non-mutual) followees, snopes are substantially less likely to be challenges. While the vast majority of snopes from strangers (85%) are challenges, only 66% of those from friends are challenges ($\chi^2 = 11.39, df = 1, p < .001$). This raises the question as to whether friends are recognizing one another's (out-of-the-blue) snopes more frequently because they are less challenging. We thus examine the rate of recognition for challenging snopes.

Table 4 shows the extent to which an individual recognizes challenges when they come from friends versus strangers within a 10 minute time window. Results show that the observed increase in responsiveness to friends' snopes is not explained by these snopes being less challenging, as the bias toward responding to friends over strangers remains even when only challenging snopes are considered ($\chi^2 = 4.31, df = 1, p < .05$).

A related concern is that individuals may simply attend more to friends' messages in general, thus leading to a greater rate of reply for any friend message. If this were true, it would not deny the importance of social relationships in fact-checking, but would suggest that the relevant mechanism is general attentiveness, not responses to the critical content of snopes. We thus consider non-challenging snopes as a base case as in these cases individuals received a reply with a snoping url but were not directly challenged. Though the samples are small ($n = 15$ for non-challenges by strangers; $n = 20$ for non-challenges by friends), we find that individuals reply to both challenges and non-challenges from strangers at virtually the same rate (27%). However, friends reply to 45% of challenges from friends, compared with only 25% of non-challenges from friends ($\chi^2 = 2.44, df = 1, p = .12$). This evidence suggests that the significant differences in rates of reply observed in our larger sample cannot be attributed to friends paying greater attention to replies from friends in general. In fact, though the sample is too small to be conclusive, there is evidence that friends are more likely to respond to challenges from friends than to general replies from friends.

In summary, even though strangers are more likely to snope with a challenge, they are significantly more likely to be ignored. These results provide substantial evidence of the critical role that established social relationships play in the dynamics of misinformation. Though recognition does not necessarily imply persuasion or acceptance, it suggests that the first step toward such persuasion, attention, is already skewed toward friends.

Table 5 shows examples of the tweet streams in which snopers challenged snopees and received a subsequent reply. Though we do not formally model it in this analysis, this discourse is representative of a general pattern in which stranger-stranger snopes contain more hostile or abrupt tones. In each of the first three cases, the snopee rejects the assertion of the snoper, that is, the snope appears to have no effect. Yet the friend-friend snope contains the acknowledgment "I took a look at the article" and then provides evidence to refute the original snope (a link to a website that suggests Snopes itself is biased). By contrast, the replies to strangers simply assert counter-claims without any real justification. The fourth example, a friend–friend snope, shows the idyllic deliberative form which scholars and commentators tend to praise in which the snopee revises his beliefs and publicly retracts his misinformed statement.

Do Snopes Change Discourse?

The preceding section explored whether there were differences in the extent to which snopees acknowledged snopes, and challenges in particular, within different relationships. In this section we examine how snopes change the subsequent discourse of the snopees. Our analyses again are divided into the categories used above: the source of the snope, whether it is recognized, and whether it is a challenge. We then use these as independent predictors of the rate at which the snopee sent tweets after being snoped.

Tweet rate is an important variable to consider because of the possibility that snopes have a "chilling effect" on tweet-

Order	User	Tweet
		Ongoing conversation, Snoped by stranger, Political
Original	Snopee	@Snoper what facts are those? You've been wrong ever since you opened your dumb mouth.
Snope	Snoper	@Snopee Red State Socialism' graphic says GOP-leaning states get lion's share of federal dollars http://t.co/dMGeGZlzNZ TOLD YOU DICKFACE!
Reply	Snopee	@Snoper they block bad bills. Like Obama care, cash for clunkers, and stimulus this is just more bureaucratic BS.
		Ongoing conversation, Snoped by friend, Political
Original	Snopee	@Snoper people have been talking about that for years before OBAMA came into office.
Snope	Snoper	@Snopee but its actually been passed now and Obama signed it http://t.co/Hkp2TztS March 23, 2013 Americans will be REQUIRED to have it!
Reply	Snopee	@Snoper I took a look at the article and "http://t.co/aeLzcNen Rumor has it" is not a credible source. http://t.co/JcyofiZO
		Out-of-the-blue snope, Snoped by stranger, Political
Original	Snopee	Does @SenTedCruz know that #FastandFurious began under George W. Bush?
Snope	Snoper	@snopee @SenTedCruz Check facts B4 running mouth.http://t.co/t7jcBppbLF
Reply	Snopee	@User2 @SenTedCruz Same exact gun-walking program, with a different name. Didn't hear right-wing outrage then.
		Out-of-the-blue snope, Snoped by friend, Non-political
Original	Snopee	Here's some advice Bill Gates recently dished out at a high school speech about 11 things they did not really learn http://t.co/kPgnpdXn
Snope	Snoper	@snopee no, no it isn't... http://t.co/x8m8TxBE
Reply	Snopee	@Snoper; @self no, no it isn't. http://t.co/d6Oochta well spotted Steve, ok it isn't. Some advice from Charles J. Sykes..

Table 5: Four example Twitter conversations coded as challenges containing the original tweet from the snopee, the snoping tweet from the snoper, and the reply tweet from the snopee.

ing, that is, when individuals are challenged or rebuked, they may shy away from further tweeting (at least temporarily). Measuring changes in tweet rate appropriately is difficult, however. By definition, snopees have been active users prior to the snoping event, as it is through the production of discourse that they become candidates for receiving replies. Thus, we know that all tweeters sent at least 1 tweet before being snoped. In a sense, a metronome that sent 1 tweet every hour would appear to reduce its tweet rate following any snopes that responded to it within a short time window.

A conservative way to control for this effect is to consider whether any reduction in tweet rate is greater than 1 tweet per minute of the observed time window. We first test whether tweet rates drop across all cases. We find that rates are significantly lower for time windows of 5 minutes, 10 minutes, and 60 minutes. However, the extent to which they are reduced is not greater than that which may be due to the presence of the snoped tweet. For example, the 95 percent confidence interval in the difference between tweet rates for the 60 minutes prior to the snope versus the 60 minutes after the snope is between .001 and .014 ($t = 2.30, p < .05$). But 1 tweet in 60 minutes is worth .017, so this difference cannot rule out the influence of this "guaranteed" effect.

We next consider whether "challenging" snopes have a significant impact that is distinct from other kinds of snopes. Using a linear model predicting tweet rate from a dummy code for whether the snope is a challenge, we find no significant relationship for any time window (5, 10, 60 or 120 minutes), and the regression explains less than 2% of the variance in tweet rate in each case. We then test for whether the friendship relationship makes a difference. Using a linear model, we estimate the relationship between whether a tweet is a challenge and whether it comes from a friend and find no significant results or improvement in explained variance. This may be due to the fact that, as shown in the analysis of replies, individuals are somewhat motivated to reply to being snoped with a challenge, thus offsetting the possible chilling effect. A more sophisticated model that conditions on an expected number of potentially defensive replies before estimating change in tweet rate may be required but is beyond the scope of the current analysis.

Community structure

Our third research question asks: How do snoping relationships vary across different communities? The preceding analyses were focused on the relationships and interactions between the snoper and the snopee. Here we examine the interactions within the population of users in our dataset aggregated into a large-scale, complex network of snopers and snopees interacting via following relationships as well as snoping communications. We extract and visualize the giant connected component of the follower graph in two steps.

First, we visualize the follower graph in isolation in Figure 5 to illustrate the underlaying social "substrate". Ev-

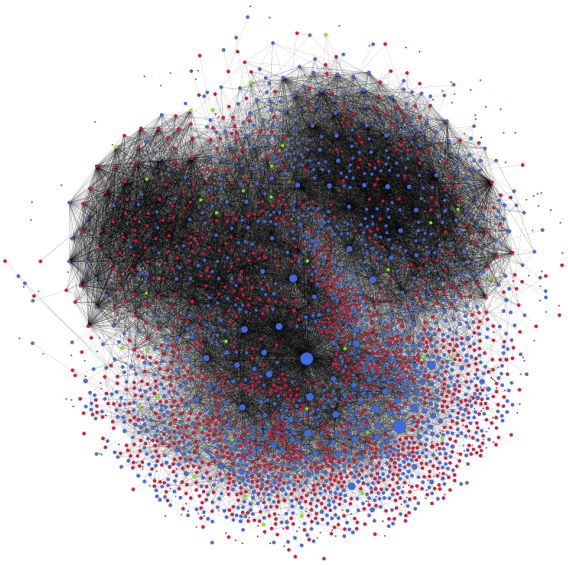


Figure 5: The largest connected component of the follower graph containing 2,636 nodes and 25,618 edges. Snopees are in blue ($n = 1290$), snopers ($n = 1310$) are in red, users who were both snopers and snopees in green ($n = 36$).

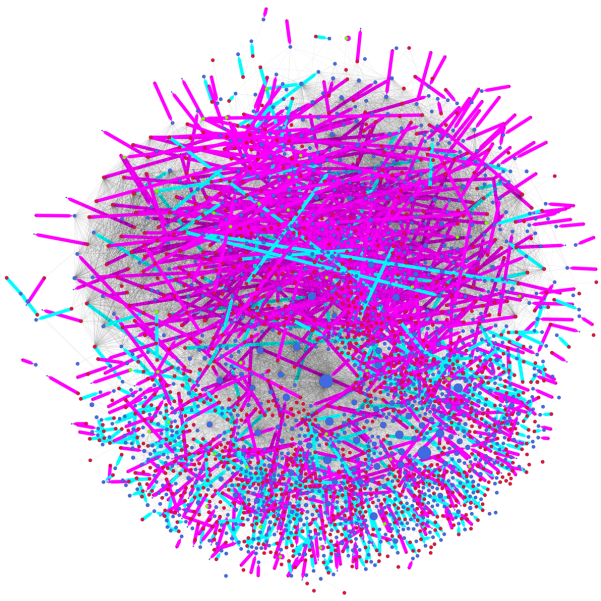


Figure 6: The same graph overlaid with snoping relationships. Pink edges were "snoped by strangers" ($n = 773$) and cyan edges were "snoped by friends" ($n = 442$). "Snoped by followers" ($n = 267$) and "snoped by followees" ($n = 51$) edges are not shown for clarity.

ery node in this network was either involved as a snoper or a snopee from the corpus above. The network is laid out using the ForceAtlas2 spring-embedding algorithm in Gephi by following relationships among users. This follower network exhibits a classic polarization (Adamic and Glance 2005; Conover et al. 2011) between two densely-knit communities. Manual inspection suggest they correspond to U.S. political parties, with conservative political officials and activists (top right) such as `@mittromney` and `@karlrove` and liberal officials and activists (top left) such as `@whitehouse` and `@joebiden`.

Second, we superimpose the snoping relationships from strangers (pink) and from friends (cyan) over this network in Figure 6. "Snoped by followers" and "snoped by followees" are omitted from the visualization for the sake of parsimony. The distribution of these different kinds of snoper–snopee relationships reveal that snoping takes place primarily between these clusters where, by definition, there are fewer friendship relationships. The graph is heavily populated by stranger dyads (pink) while snopes via friend relationships (cyan) are very rare.

This disparity suggests a perverse dynamic that sheds new light on commonly voiced concerns regarding partisan division and bickering. Whereas prior research has established that attention to information is disproportionately focused within political groups (Adamic and Glance 2005), this figure suggests that snoping is explicitly directed outwardly as criticism of individuals that snopers normally do not care to hear. Snopers spiritedly "snipe" opponents' messages despite having no interest in being a member for their messages generally, and perhaps unsurprisingly, these stranger-stranger snopes are subsequently ignored.

By contrast, within-community corrections are rare. This macro-social analysis suggests that fact-checking may play a role not generally foreseen in the discussion of democratic deliberative practice. Snoping tweets go to a unique audience — the snopee and all of the followers of the snoper. However, the polarized nature of the following graph as well as the specific affordances by which Twitter shares reply tweets to only those users who follow both the "replier" and "repliee" suggests few users in snopers' audiences see these tweets. Though not definitive, the results here suggest that snoping is used less as an invitation to others to behave more deliberatively and more as a performance or ritual to signal one's loyalty to a community and willingness to make public declarations against opponents.

There is also a more diffuse third community at the bottom of Figure 6 containing celebrities such as `@rihanna` and `@neiltyson`. Fact-checking interventions within this "popular culture" community have substantially more "snoped by friend" relationships (cyan) than "snoped by stranger" relationships. This may simply be an effect of celebrities reciprocating following relationships unselectively, but it also points to the toxic consequences of polarization and selective exposure. In conjunction with the prior findings about structural position, these users are equals to each other but peripheral to celebrities and elites within the population. This portends a more democratic ideal for deliberative speech and may be opportunities for users to revise their beliefs in or support for misinformation.

Discussion

This research proposed to examine the fact-checking as a social activity. Though exploratory in nature, the results support the conclusion that within real-world conversations, the social attributes of a snope are important to understanding its intent and impact.

Through our examination of the social component of fact-checking, our findings suggest that snopes are often performative acts intended more for the snoper's audience. Nonetheless, good-faith communicative acts actually intended for the snopee's consideration can be effective in prompting a response if it occurs over existing friendship relationships.

Clearest among these findings is the importance of pre-existing relationships. The majority of observed snopes, drawn at random from public discourse, come from strangers — individuals with no existing social relationship between them on Twitter. Fewer than 30% of snopes come from friends. Yet "friendly" snopes are much more likely to get an individual's attention. The difference appears to be explained not by greater attentiveness to friends in general, but by a tendency for individuals to respond to friends' challenges. Whether these responses are acknowledgments of error or attempts at self-defense to save face, it is clear that the existing social relationship between the parties encourages the individual to consider and address the challenge. The importance of ongoing social relations is also highlighted by differences in the way individuals react to snopes within ongoing conversations and snopes that occur "out-of-the-blue." Individuals do not appear to avoid snopers when they present their facts within an ongoing discussion. When snopes are lobbed in from outside such discussions, however, they are very likely to be ignored, especially when sent in by strangers.

Ironically, these stranger–stranger "out-of-the-blue" corrections are most akin to those created in many laboratory fact-checking experiments. Our results suggest that it is unsurprising that these laboratory corrections fail to produce results. However, the culprit is not individuals' resistance to facts, but their resistance to being challenged through asocial channels by faceless strangers. This result is promising for the study of strategies to combat misinformation, however, our data also show that, unfortunately, this least influential kind of snope is also the most typical in real world discourse. Challenges from strangers comprise the most common kind of snope, but they have the lowest chance of garnering a reply. Challenges from friends actually appear to increase discussion, suggesting they may make a contribution to cooperative deliberation.

Status and popularity also appear to play an interesting role in real-world snoping. Individuals with dense networks of highly-connected followers tend to snope popular individuals with sparse networks of poorly-connected followers. Once again, the typical form of snope is the least effective. Bivariate correlations suggest that snopes from high status users tend to be ignored more so than snopes from others, though the effect is weaker when considering friendship.

Are individuals intimidated out of responding to high status individuals, or do they suspect their intentions are not to foster deliberation in the first place? The community structure analysis suggests the latter. Elites appear to snope across communities, attacking the popular members of the groups of which they are not a part. This suggests these activities may be performative rather than deliberative, with elites garnering and maintaining status through snopes that are visible not only to the snopee but their entire set of followers.

Limitations and future work

First, this corpus is only a sample of fact-checking interventions and does not include tweets from before 2012, follower graph information at the time of the intervention itself, or fact-checks using other websites and data sources. Future work might rely on tie-strength observed prior to and after the snoping event (Bak, Kim, and Oh 2012). Second, the fact-checking interventions we identified were English language and predominantly involved content related to the context of U.S. politics. The variance in cultural norms about political deliberations and social media practices across linguistic and national contexts prevents us from making general claims about the effectiveness of socially-mediated fact-checking interventions. Third, the interventions were used as natural experimental treatments, but nevertheless lacked the independence and random assignment to conditions to definitively establish the effectiveness of socially-mediated fact-checking interventions in encouraging individuals to rethink their views and statements. These limitations might be addressed in the future by automated or semi-automated classifications of misinformation and random selection from different types of interventions to test their effectiveness. Finally, more advanced qualitative content analysis and natural language processing methods are required to better the conversational contexts, user dispositions, and topical features that shape decisions to both transmit misinformation and attempt corrections through fact-checks.

Conclusions

New communication technologies have increased the capacity to spread both misinformation and challenges to it. We argued that because the communication of misinformation is a fundamentally social process, the presentation and acknowledgment of facts must also be considered in its social context. Using a corpus of 1,614 conversational tweets referencing well-known fact-checking websites, we identified the social contexts and behavioral consequences of fact-checking interventions. Our results suggest these interventions occur over diverse social contexts and these contexts play a crucial factor in explaining changes in subsequent user behavior. The theoretical framework we have outlined and tested likewise have implications for designing strategies to respond to misinformation in both online and offline contexts.

Acknowledgments

We would like to thank Alan Mislove and Kiran Garimella for assistance with data preparation as well as three anonymous reviewers and the members of the Lazer Lab at Northeastern University for their feedback.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proc. 3rd Int'l Workshop on Link Discovery*, LinkKDD '05, 36–43. New York, NY, USA: ACM.

Allport, G. W., and Postman, L. 1947. *The psychology of rumor.* Russell & Russell.

Bak, J. Y.; Kim, S.; and Oh, A. 2012. Self-disclosure and relationship strength in Twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 6064.

Baumeister, R. F.; Zhang, L.; and Vohs, K. D. 2004. Gossip as cultural learning. *Review of General Psychology* 8(2):111–121.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on Twitter. In *Proceedings of the 20th international conference on World wide web*, 675684.

Cialdini, R. B., and Goldstein, N. J. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.* 55:591–621.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on Twitter. In *ICWSM*.

DiFonzo, N., and Bordia, P. 2007. *Rumor psychology: Social and organizational approaches.* American Psychological Association.

Ecker, U. K. H.; Lewandowsky, S.; and Tang, D. T. W. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition* 38(8):1087–1100.

Foster, E. K. 2004. Research on gossip: Taxonomy, methods, and future directions. *Review of General Psychology* 8(2):78–99.

Friedkin, N. E., and Johnsen, E. C. 1999. Social influence networks and opinion change. *Advances in Group Processes* 16(1):1–29.

Garrett, R. K.; Nisbet, E. C.; and Lynch, E. K. 2013. Undermining the corrective effects of media-based political fact checking? the role of contextual cues and nave theory: Undermining corrective effects. *Journal of Communication* 63(4):617–637.

Garrett, R. K. 2011. Troubling consequences of online political rumoring. *Human Communication Research* 37(2):255–274.

Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American journal of sociology* 481–510.

Kahan, D. 2010. Fixing the communications failure. *Nature* 463(7279):296–297.

Lewandowsky, S.; Ecker, U. K. H.; Seifert, C. M.; Schwarz, N.; and Cook, J. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest* 13(3):106–131.

Liao, Q. V.; Wagner, C.; Pirolli, P.; and Fu, W.-T. 2012. Understanding experts' and novices' expertise judgment of Twitter users. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 24612464.

Morris, M. R.; Counts, S.; Roseway, A.; Hoff, A.; and Schwarz, J. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 441450.

Nyhan, B., and Reifler, J. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32(2):303–330.

Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 15891599.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gon\ccalves, B.; Patil, S.; Flammini, A.; and Menczer, F. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, 249252.

Tetlock, P. E. 2002. Social functionalist frameworks for judgment and choice: intuitive politicians, theologians, and prosecutors. *Psychological review* 109(3):451.

Uzzi, B. 1997. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative science quarterly* 35–67.

Weick, K. E. 1995. *Sensemaking in organizations*, volume 3. Sage.

The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior

Yabing Liu[†] **Chloe Kliman-Silver**[‡] **Alan Mislove**[†]

[†]Northeastern University [‡]Brown University

Abstract

The microblogging site Twitter is now one of the most popular Web destinations. Due to the relative ease of data access, there has been significant research based on Twitter data, ranging from measuring the spread of ideas through society to predicting the behavior of real-world phenomena such as the stock market. Unfortunately, relatively little work has studied the changes in the Twitter ecosystem itself; most research that uses Twitter data is typically based on a small time-window of data, generally ranging from a few weeks to a few months. Twitter is known to have evolved significantly since its founding, and it remains unclear whether prior results still hold, and whether the (often implicit) assumptions of proposed systems are still valid.

In this paper, we take a first step towards answering these question by focusing on the evolution of Twitter's users and their behavior. Using a set of over 37 billion tweets spanning over seven years, we quantify how the users, their behavior, and the site as a whole have evolved. We observe and quantify a number of trends including the spread of Twitter across the globe, the rise of spam and malicious behavior, the rapid adoption of tweeting conventions, and the shift from desktop to mobile usage. Our results can be used to interpret and calibrate previous Twitter work, as well as to make future projections of the site as a whole.

Introduction

Online social networks(OSNs) are now a popular way for users to connect, communicate, and share content; many serve as the de-facto Internet portal for millions of users (Post 2014). Because of the massive popularity of these sites, data about the users and their communication offers unprecedented opportunities to examine how human society functions at scale. As a result, significant recent research has focused on these sites, with a particular emphasis on Twitter due to the relative ease of accessing a large amount of data. For example, recent papers have ranged from studying Twitter-specific behavior (e.g., the patterns of retweeting (Macskassy and Michelson 2011), deletion of tweets (Almuhimedi et al. 2013) and usage of different entities in tweet contents (Macskassy 2012; Yang et al. 2012)), to examining how privacy leaks and violations can occur (Mao, Shuai, and Kapadia 2011; Meeder et al. 2010), and even using aggregated data from Twitter to predict the behavior of real-world phenomena such as the stock market (Gilbert and Karahalios 2010).

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

While the set of research using Twitter data has expanded rapidly, there has been relatively little work that has studied the evolution of the Twitter ecosystem *itself*. For example, Twitter has grown from thousands of users in 2007 to millions in 2009 to hundreds of millions in 2013. In parallel with this growth, we have seen a significant maturation of the Twitter platform: Twitter today is used by many organizations and individuals as a primary way of communicating with others. Most research that uses Twitter data is typically based on a small time-window of data—generally ranging from a few weeks to a few months—making it difficult to quantify long-term trends. Twitter is known to have evolved significantly since its founding, and it remains unclear how much the user base and behavior has evolved, whether prior results still hold, and whether the (often implicit) assumptions of proposed systems are still valid.

In this paper, we take a first step towards answering these questions by directly examining the evolution of Twitter itself, focusing on the Twitter users and their behavior. Using a set of over 37 billion tweets spanning between 2006 and 2013, we quantify how the users, their behavior, and the site as a whole have evolved. We observe and quantify a number of interesting patterns, including

- The spread of Twitter across the globe, both in terms of users in different regions and tweets containing different languages (e.g., the percentage of U.S./Canada users drops from over 80% to 32%, and the percentage of users tweeting in English falls from 83% to 52%).
- The percentage of tweets that are no longer available due to a user's or Twitter's action increases to over 20% for some time ranges.
- The percentage of Twitter user accounts today that are inactive shows rapid growth; over 32.5% of accounts have not tweeted for over a year.
- The increase of malicious behavior on Twitter beginning in 2009, including fake followers, fake accounts, and hashtag promotion; over 6% of all accounts are now suspended.

Dataset	Date range	Users	Tweets	Date collected	Coverage of all Tweets	Coverage of all Users
Crawl	21/03/2006 – 14/08/2009	25,437,870	1,412,317,185	14/08/2009	∼100%	∼100%
Gardenhose	15/08/2009 – 31/12/2013	376,876,673	36,495,528,785	Time of tweet	∼10–15%	∼30.61%
UserSample	21/03/2006 – 31/12/2013	1,210,077	—	12/31/2013	∼0.1%	∼0.1%
Total	21/03/2006 – 31/12/2013	388,796,600	37,907,845,970	—	—	—

Table 1: Source and basic statistics for the Twitter data used in this study.

- The switch from a primarily-mobile system (via SMS) to a primarily-desktop system (via the web) and back to a primarily-mobile system (via mobile applications). Today, over half of all tweets are created on mobile devices.

Our results can be used to interpret and calibrate previous Twitter studies, as well as to make future projections of the site as a whole. We make all of our analysis available to the research community (to the extent allowed by Twitter's Terms of Service) to aid other researchers and to stimulate further research in this area; researchers can access it at

`http://twitter-research.ccs.neu.edu/`

Background and Data Source

Twitter is a "micro-blogging" service that allows users to multicast short messages (called *tweets*). Each user has a set of other users (called *followers*) who receive their messages; those who a user follows are called *friends*. The follow relationship in Twitter is directed, and requires authorization from the followee only when the followee has elected to make their account *protected*. Each tweet can only be up to 140 characters in length, and the default setting in Twitter is to allow all tweets to be publicly visible.

Twitter data We obtain our Twitter data from two sources; basic statistics of these datasets is provided in Table 1.[1] First, we use an almost-complete[2] collection of all tweets issued between March 21, 2006 and August 14, 2009 collected by previous work (Cha et al. 2010); we refer to this as the *Crawl* dataset. This dataset was collected in August 2009 by iteratively downloading all of the tweets of all public users alive at the time.

Second, we collect data from the Twitter "gardenhose" public stream[3] between August 15, 2009 and December 31, 2013; we refer to this as the *Gardenhose* dataset. Our measurement infrastructure was down for the 10 weeks between October 18, 2010 and December 31, 2010, so we do not have data for that time period. Twitter states that the gardenhose is a random sample of all public tweets. Each *Gardenhose* tweet includes information about the user who created the tweet (e.g., the user's location and count of total tweets) that is current as-of the time of the tweet.

[1]This study was conducted under Northeastern University Institutional Review Board protocol #10-03-26.

[2]The dataset does not include any tweets deleted before August 14, 2009, and only includes the 3,200 most recent tweets (as of August 14, 2009) for each user.

[3]https://stream.twitter.com/1.1/statuses/sample.json, with elevated access.

Gardenhose sampling rate We briefly estimate the sampling rate of the *Gardenhose* dataset; Twitter states that the gardenhose is a random sample, but does not state the rate. We estimate the sampling rate by relying on the `statuses_count` field of the user in each tweet; the `statuses_count` field represents the total number of tweets (statuses) that the user has issued at the time of the tweet. Each month, we determine the first observed value of `statuses_count` (sc_{first}), the last observed value of `statuses_count` (sc_{last}), and the number of tweets we observed (obs). We can then estimate the sampling rate for that user with

$$rate = \frac{obs}{sc_{last} - sc_{first}}$$

We plot the average value of $rate$ across all users with $sc_{last} - sc_{first} > 1000$ in Figure 1, observing a sampling rate of ∼15% until 07/2010, and ∼10% since then (the "dips" observed in the graph are due to short periods of time when our measurement infrastructure was down).

Limitations Because the *Crawl* dataset was collected in August 2009 (as opposed to the *Gardenhose* dataset, which was collected over a period of years, as tweets were issued), the user information for the *Crawl* dataset is as-of August 2009. This limitation will occasionally present itself during our analysis, and we discuss these limitations in-line.

We also face two limitations with the *Gardenhose* dataset. First, the *Gardenhose* dataset ends up containing biased sample of users, with a bias towards more active users. The reason for this is that users who tweet very often are extremely likely to show up in our dataset; a user who tweets only once has a ∼10% chance of appearing. Second, Twitter

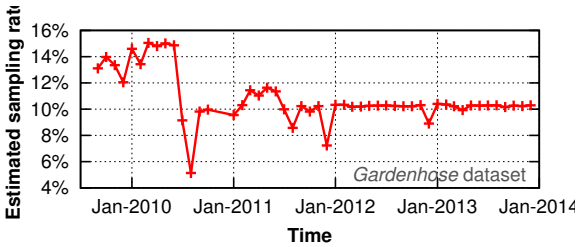


Figure 1: The estimated sampling rate (the average value of *rate* for users with more than 1,000 statues in a month) of the *Gardenhose* dataset over time. The occasional drops in sampling rate are due to times that our collection infrastructure was down for short periods of time.

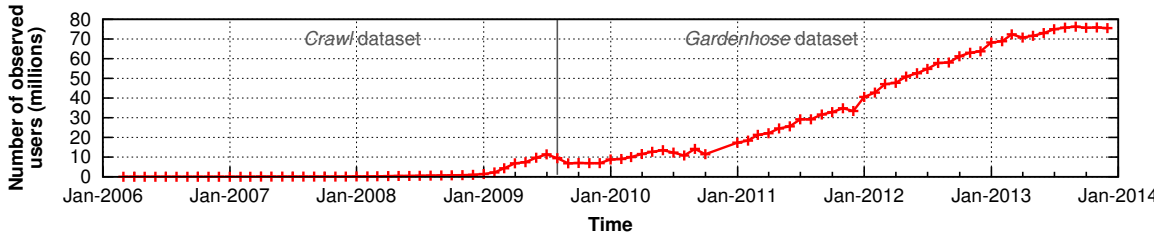


Figure 2: The number of users we observed tweeting in each month in the *Crawl* and *Gardenhose* datasets. The "dip" in August 2009 is due to the switch from a complete sample to a 15% sample. Our numbers are much lower that Twitter's announced "active users" numbers due to our data sample and Twitter's definition of an "active" user.

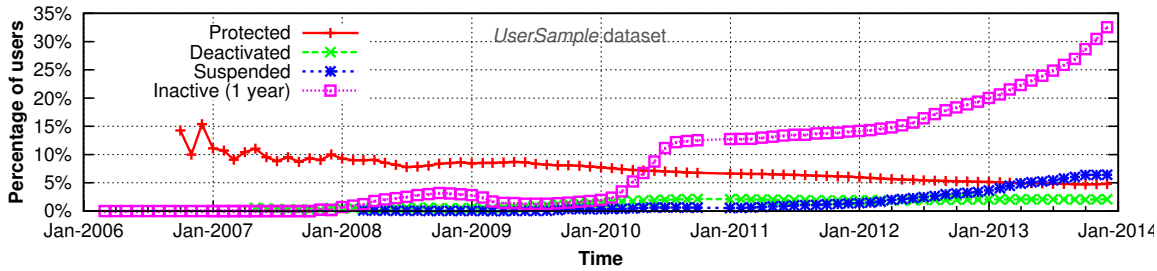


Figure 3: The percentage of the entire Twitter user base over time whose accounts are protected, deactivated, suspended, or inactive (for at least a year), based on the *UserSample* dataset. We observe a dramatic increase in both inactive and suspended accounts starting in 2010.

does not inform us when users leave the network, so we are unable to determine when users mark their accounts as protected (thereby hiding their tweets), are *suspended* by Twitter for violating the Terms of Service, or *deactivate* their accounts (i.e., manually delete their account).

To address these limitations of the *Gardenhose* dataset, we collect a third and final dataset *UserSample* that represents a random sample of *users* instead of *tweets*. Specifically, we generate 2 million random `user_id`s between 1 and 1,918,524,009 (the largest `user_id` that we ever observed), representing a ∼0.1% sample of all Twitter users. We then query Twitter (both via the API and the web site) in January 2014 for the most recent information on each of these users, allowing us to determine if the `user_id` was ever assigned, has been suspended, or is protected. We infer if the `user_id` has been deactivated by the user if Twitter says the user does not exist but we observed a tweet from the user in our *Crawl* or *Gardenhose* datasets. Overall, we find that at least 1,210,077 (60.51%) of these 2 million `user_id`s were ever assigned to a user.

Throughout our analysis, we use the most appropriate dataset(s) for each question at hand. Additionally, we label each graph with the dataset(s) that it uses.

User characteristics

We begin our analysis by studying how the user population of Twitter has changed since its inception.

User growth and activity We first examine the characteristics of the Twitter user population. Figure 2 shows the total number of users that we observed over time; we observe massive growth that is in-line with Twitter's reported number of monthly active users (Weil 2010). For example, we observed over 73 million users tweet in June 2013; while Twitter reports 218 million active users for that time period (SEC 2013); our number is lower due to the fact that we have a random 10% sample and Twitter's definition of an active user is based on login activity, not tweeting activity. We also observe rapid growth from 2009 through 2012, with a leveling-off of the number of active users in 2013; this suggests that Twitter's user population growth may be slowing down.

Next, we briefly examine the aggregate level of user ac-

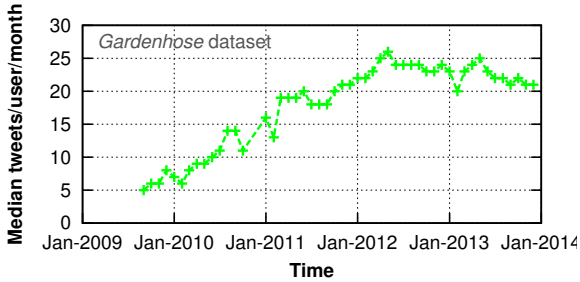


Figure 4: The median number of tweets per user per month over time, based on the first and last `statuses_count` observed for each user. Note that this result is based on *Gardenhose* dataset, which is biased towards more active users.

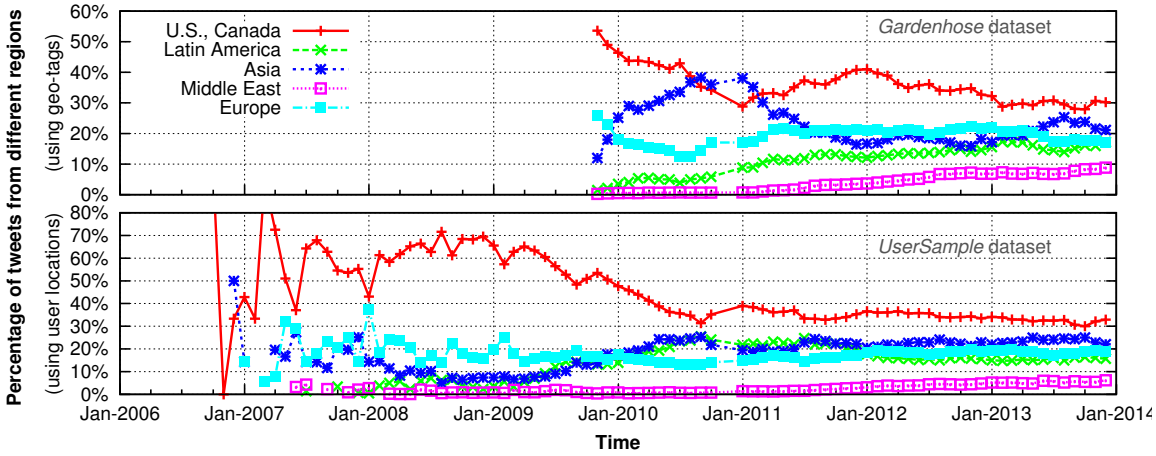


Figure 5: The percentage of tweets created in different geographical regions over time. Shown are locations inferred from self-reported user locations (*UserSample* dataset, bottom) and geo-tags (*Gardenhose* dataset, top). Geo-tags were introduced in November 2009, so we only show data since then. Africa and Oceania are both almost always less that 1%, and are not shown for clarity. A number of interesting trends can be observed, quantifying the spread of Twitter across the globe.

tivity as Twitter has evolved. To do so, we use observed tweets for each user in the *Gardenhose* dataset and use the first and last `statues_count` field in the user profile (we note that using the *Gardenhose* dataset introduces a bias towards more active users). The median value of the number of tweets per user per month is presented in Figure 4. We can see a general rise in activity from later 2009 to 2012, with the rate stabilizing and then decreasing.

Users leaving Twitter While Twitter has seen massive growth, a non-trivial percentage of users leave Twitter, either by deactivating their account or by being suspended by Twitter. Additionally, many users mark their account as *protected*, meaning only their approved followers can view their tweets. Finally, many users simply leave Twitter and become *inactive*, meaning they have not tweeted for over a year. In Figure 3, we use the *UserSample* dataset to plot the percentage of the entire Twitter user base whose accounts are deactivated, suspended, protected, or inactive at different points in time (e.g., in January 2013, 2% of the user population at that time were deactivated, 4% had been suspended, 5% had marked their accounts as protected, and 20% had not tweeted for over a year).[4]

We first observe the massive percentage of inactive accounts, representing up to 32.5% of all accounts by the end of 2013. The increasing nature of this trend suggests that the leveling off of active users per month (Figure 2) may soon cause the majority of accounts to be inactive. We also observe that up to 15% of users who joined in Twitter before December 2007 protected their accounts, while the percentage goes down to 4.8% by 2013; this implies that most users who have joined Twitter recently have kept their accounts public. We further observe a dramatic increase in the percentage of suspended users, with over 6% of the entire Twitter user population suspended by late 2013; this is inline with studies on the rise of malicious activity (Thomas et al. 2013; Yeung 2013). Finally, a relatively stable 2% of users who have deactivated their accounts, indicating that relatively few users are leaving Twitter by explicitly deleting their accounts.[5]

User location Now, we examine the geographical distribution of the users over time. To do so, we rely on two pieces of information: (a) the self-reported, unformatte t the location field, we query Bing Maps with each unique location string, and only consider results that Bing returns with "high confidence". To interpret the geo-tags, we use country GIS shape files to translate latitude/longitudes into countries and administrative districts. Since geo-tags were introduced in November 2009, we only report geo-tags for months afterwards. We find 42.4% of users provide a location string interpretable by Bing, and 1.23% of tweets have included geo-tags.

We present the results of this analysis in Figure 5, showing the percentage of users located to different regions of the world using both self-reported locations (bottom) and geo-tags (top). We observe a number of interesting trends that quantify the spread of Twitter across the globe: First, we observe a steep decline of the percentage of the tweets from the U.S. and Canada from a high of above 80% to 32%; most of this decline comes in 2009 and 2010. At the same time, we observe a substantial increase in the percentage of tweets from the Middle East (starting in early 2011, corresponding

[4]Since Twitter does not provide the date a user's account becomes unavailable, we define the date of being suspended, deactivated, protected as the last date on which we observed a tweet from the user.

[5]We may underestimate the percentage of users who deactivated their accounts, as we can only infer that deactivated `user_ids` ever existed if we observed a tweet from them.

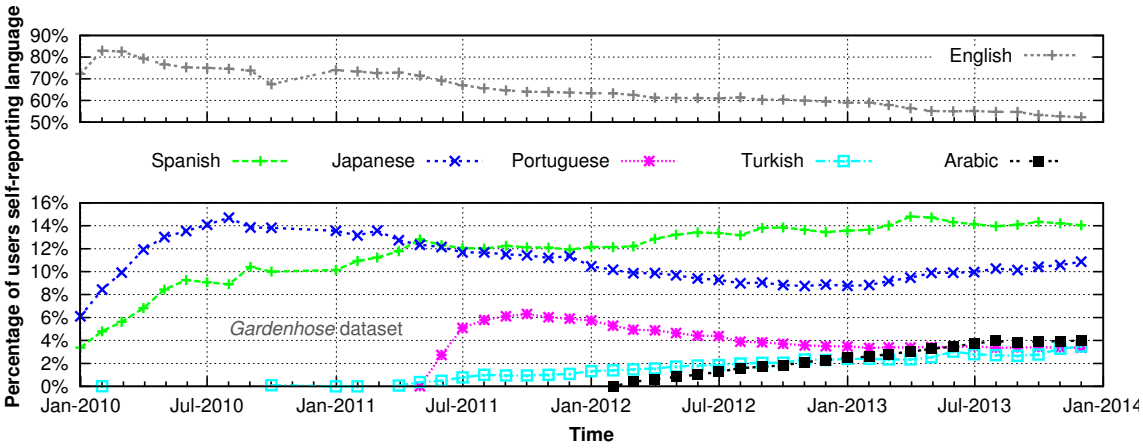


Figure 6: The percentage of users self-reporting the six most popular languages over time. We plot English separately in the top graph in order to increase the readability; note that the scale is different between the two graphs. English shows a mostly linear decrease from 83% in January 2010 to 52% in December 2013.

to the Arab Spring) and Latin America. However, Europe is relatively stable over the course of Twitter's evolution, generally representing around 20% of the tweets. Finally, comparing the two graphs, we can observe a difference in the spread of geo-tags, with Asia users being over-represented relative to the entire user population in 2009 and 2010. This is likely due to the popularity of smartphones equipped with GPS in Asia, which have since become popular globally.

Languages We now examine the languages Twitter users report over time. To do so, we rely on the self-reported `lang` field that Twitter allows users to specify in their profile; this field first appeared on January 12, 2010, so we report data since then in Figure 6 (English in the top graph, other languages in the bottom graph). We observe a steady (and continuing) decrease of users reporting English, from a high of 83% to 52% in December 2013. Spanish and Japanese show a steady representation of approximately 10%, while the decrease of English is correlated with an increase in a large number of other languages, including Turkish, Portuguese, and Arabic (correlating well with the results in the previous section showing the distribution of user locations). Overall, these results underscore the fact that Twitter's user population is continuing to become more diverse and global.

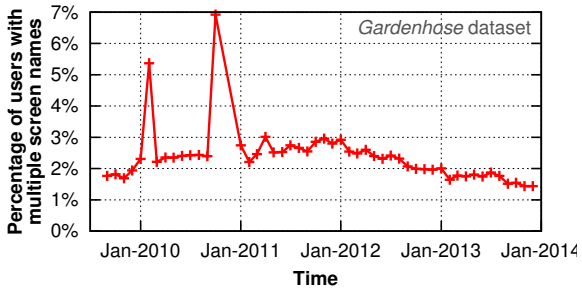


Figure 7: The percentage of users who have used more than one screen names in each month. The "spikes" correspond to times where Twitter released old, inactive screen names to be reclaimed.

Screen name changes A little-known feature of Twitter is that users can easily change their screen name (e.g., changing @Barack to @BarackObama), meaning tweets from the same user may show up under different screen names.[6] Using our *Gardenhose* dataset, we can observe these changes happening by looking for tweets from the same `user_id` with different screen names. Figure 7 plots the percentage of the user population that we observe to have used multiple screen names each month; the "spikes" in February and October 2010 correspond to time periods where Twitter opened up old, inactive screen names to be reclaimed by active users (Bryant 2010). Previous results have suggested that these users are more likely to be spammers (Chowdhury 2010), and our results correlate with prior studies showing an increase in the level of Twitter spam in 2010 and 2011 (Thomas et al. 2011; Acohido 2010). Overall, our results show that a non-trivial percentage (up to 3%) of users change their screen names each month; this suggests that researchers should internally refer to users with Twitter's (unchangeable) `user_id`s to ensure that users are accurately tracked over time.

Social characteristics We now turn to examine the social characteristics of Twitter users. Recall that Twitter users can *follow* each other, and following a user means the followed

[6] If a user changes their screen name, their `user_id` remains the same; this allows us to track screen name changes. There is no limit to the number of times a user can change their screen name.

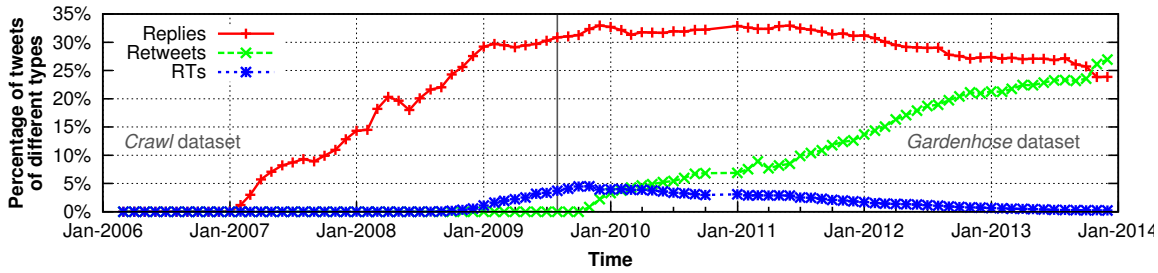


Figure 8: The percentage of tweets of different types over time. Both *RTs* and *Retweets* represent retweets, with the former capturing manually created retweets by users. Native retweets were supported by Twitter starting in November 2009.

user's tweets will show up when you log in. Following a user only requires permission if the followed user's profile is protected.

Using the *Gardenhose* dataset, we calculate the median number of followers (i.e., those following a user) and friends (i.e., those who a user follows) for all observed users over time, and present the results in Figure 9 (top). Similar to Figure 4, these results are biased towards more active users due to the use of the *Gardenhose* dataset. We observe a dramatic increase in the median followers/friends count of almost 400% from 2009 to 2013. This trend underscores Twitter's importance as a information dissemination platform; today, many celebrities, companies, and organizations use Twitter as one of the primary mechanisms to communicate with others (Christoforos 2011).

We also examine the average ratio of friends-to-followers in Figure 9 (bottom), and find an interesting trend: the ratio increases from 1.50 to a high of 1.77 in January 2012 before returning to its previous value. We make two observations. First, the fact that the ratio is much higher than 1 indicates that the distribution of followers is much more biased than the distribution of friends (i.e., most users have many more friends that followers) indicating that Twitter is disassortative; similar observations have been made about Twitter (Cha et al. 2010) and other social networks (Mislove et al. 2007). Second, the increase corresponds well with the rise of Twitter follower spam in 2010 and 2011 (Stringhini et al. 2012); we posit that the subsequent decrease is likely due to Twitter's more active role in suspending and deleting malicious accounts (Thomas et al. 2013; 2011).

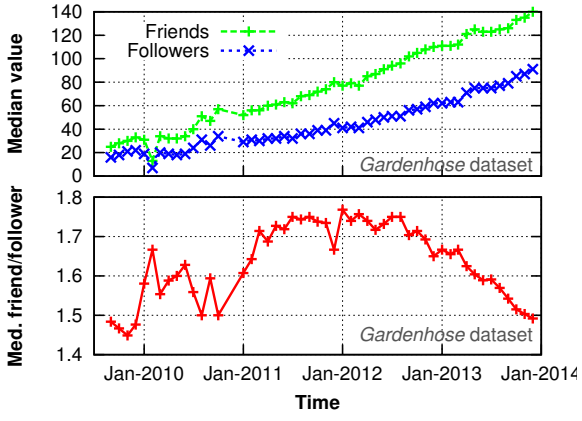


Figure 9: The median number of friends and followers across all users (top), and the median ratio of friends to followers (bottom) as derived from the *Gardenhose* dataset. We observe a dramatic densification of the Twitter social graph. Similar to Figure 4, using the *Gardenhose* dataset causes a bias towards more active users.

Tweeting behavior

In the above section, we examined the changing patterns in Twitter user population, now we turn to take a look at the changes in users' tweeting behavior over time.

Tweet causes We begin by examining internal-to-Twitter actions that cause tweets, focusing on two mechanisms: retweets (i.e., a user re-sharing one of his friends' tweets with his own followers) and replies (i.e., a user replying to a tweet authored by one of his friends). While the Twitter API allowed users to create replies natively starting in early 2007, Twitter did not natively support creating retweets until November 2009 (Meeder et al. 2010; Parr 2009). Instead, users who wished to retweet a tweet manually copied the tweet and added a "RT @username" at the beginning to indicate a retweet. As a result, when calculating retweets, we need to look for both native retweets (*Retweets*) as well as manual retweets (*RTs*).

Using the *Crawl* and *Gardenhose* datasets, we calculate the percentage of tweets that are replies, retweets, and RTs over time in Figure 8. We observe a number of interesting trends. First, we can see a rapid adoption of the reply mechanism, peaking at almost 35% of all tweets in 2010 and declining slightly afterwards. Second, we observe that retweets are initially a small percentage of all tweets, presumably due to the manual effort required to create a retweet before November 2009. However, the percentage of native retweets increases rapidly afterwards, likely due to the native retweet support that many Twitter clients provide. In fact, in late 2013, the percentage of retweets is larger than the percentage of replies. Overall, the decline in replies indicates that there is declining person-to-person communication on Twitter, suggesting significant changes in users' tweeting behavior.

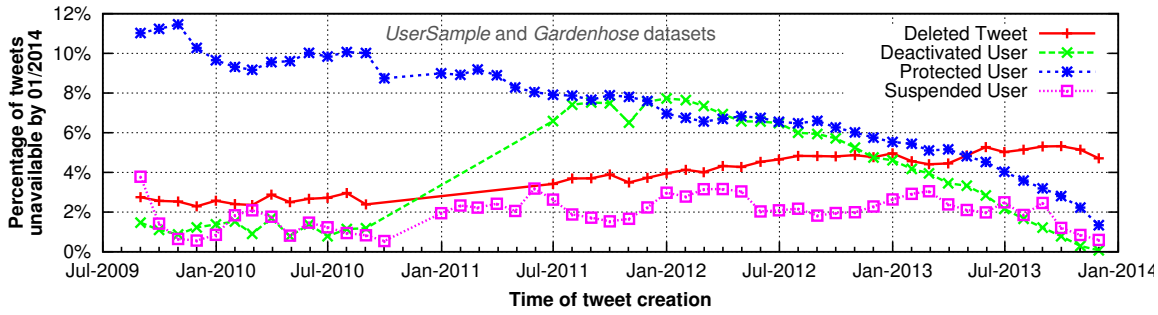


Figure 10: The percentage of tweets in the *Gardenhose* dataset that are unavailable as-of January 2014, considering only tweets issued by users in our *UserSample* dataset. Up to 10% of tweets are issued by users who later change their account to be protected, and up to 5% of tweets are explicitly deleted by users. The rise of tweets unavailable due to suspended and deleted users corresponds strongly with the increase of spam on Twitter (Thomas et al. 2011).

Unavailability of tweets Twitter's Terms of Service (Twitter 2012) requires that any data shared about tweets is shared only in the form of a `tweet_id`; the recipient then must query Twitter to obtain the actual tweet data. Presumably, this policy is in-place so that Twitter can respect users' privacy by preventing further access to tweets that a user deletes or marks as protected. This policy significantly impacts researchers, however, as researchers wishing to reproduce prior findings may not be able to obtain the entire data sets used by others.

In order to understand the impact that this policy has, we study the percentage of tweets that become unavailable over time. In general, there are four mechanisms that could lead a public tweet to later be unavailable: (1) the tweet could be explicitly deleted by the user, (2) the user could switch their account to be "protected", thereby making their tweets only available to their approved followers, (3) the user's account could be suspended by Twitter, and (4) the user could deactivate their entire account. To study how tweets become unavailable, we need to re-query Twitter for the current status of users; since we cannot query Twitter for all 388 million users, we instead perform this analysis with the *UserSample* dataset.

In Figure 10, we present the percentage of tweets created over time that are no longer available as-of January 2014 (i.e., we look for tweets in the *Gardenhose* dataset issued by users in the *UserSample* dataset). We observe that up to 20% of these users' publicly issued tweets, in aggregate, can no longer be accessed; this significantly impacts the ability for researchers to reproduce prior results. We also observe a number of other trends. First, we observe a natural decline in the percentage of tweets unavailable due to protected and deactivated users; this is expected, as more recent tweets have had less "time" for the issuing user to become protected or deactivated. Second, we observe that the percentage of tweets that are unavailable due to suspended users is fairly constant around 2%; this suggests that Twitter quickly suspends malicious users, so time passed is not significant factor in the likelihood of a tweet being unavailable due to suspension. Third, as has been show in previous work (Almuhimedi et al. 2013), most deleted tweets are deleted quickly after being issued. The increase in the percentage of deleted tweets therefore suggests a change in user behavior over time, with users becoming more likely to delete tweets.

Tweet contents Next, we take a closer look at the contents of tweets. Examining the contents of over 37 billion tweets is quite challenging, so we focus on three types of internal tweet *entities* that Twitter natively supports: tweets with URLs (i.e., a user sharing a link), tweets with hashtags such as `#fail` (i.e., a user stating the topic of the tweet), and the tweets with mentions[7] such as @BarackObama (i.e., a user mentioning another user in the tweet). Similar to the RT syntax for retweets, the syntax for specifying both hashtags and mentions was created by the Twitter users themselves, and only later natively supported by Twitter (Cooper 2013; Stone 2009).

Figure 11 (bottom) presents the percentage of tweets that have at least one of the various entity types over time. We can observe that both mentions and URLs were popular by 2009, but that hashtags only began appearing in more than 10% of tweets in 2010. Surprisingly, since 2009, the percentage of tweets with mentions has increased substantially, while the percentage of tweets with URLs has actually decreased to stabilize at ∼12%. Overall, these results suggest that Twitter has become more "conversational", with users mentioning other users in over 50% of tweets today.

We note that users can choose to include more than one of a given entity type in a tweet (e.g., a single tweet can include multiple hashtags). To understand this behavior, Figure 11 (top) plots the average number of entities in tweets that have at least one such entity (i.e., for the hashtags line, we only consider tweets with at least one hashtag). We observe that URLs and mentions have largely stabilized around 1.0 and 1.3, respectively, but that the average number of hashtags shows a continuing increase beyond 1.6. This trend is likely explained by an increasing level of hashtag spam that has

[7]Note that replies are by definition a type of mention, as the replying user includes the username they are replying to in their reply.

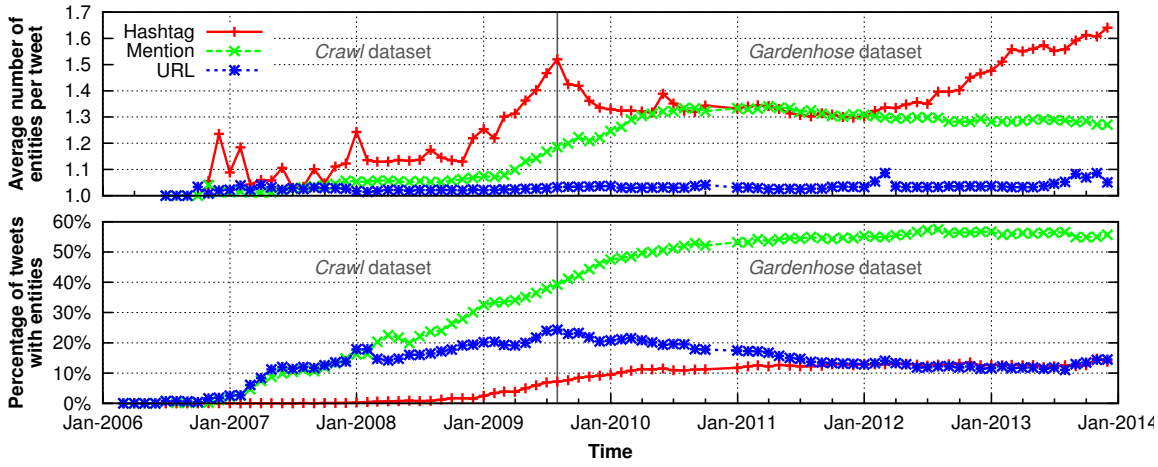


Figure 11: The percentage of tweets with different types of entities (bottom), and average number of entities for such tweets (top) over time. We observe increasing adoption of mentions, and an increasing likelihood of many hashtags per tweet.

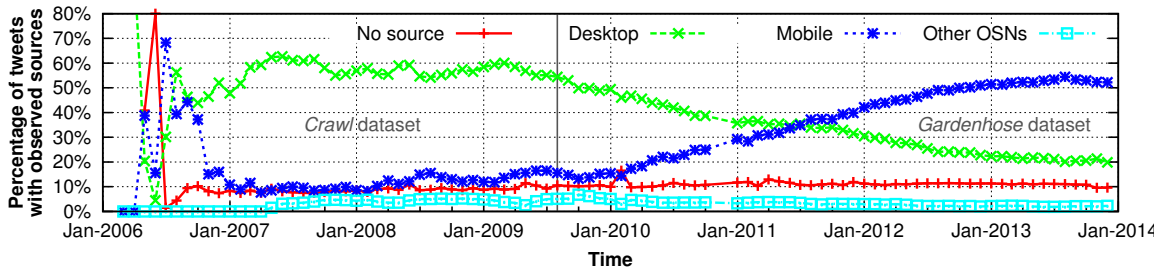


Figure 12: The percentage of tweets created with different `sources` (i.e., different clients) over time.

been observed (Vaas 2013; Ostrow 2009), where malicious users issue tweets with many hashtags in an attempt to make the hashtags appear popular.

Twitter clients Twitter was originally designed to be used on mobile devices by sending SMSes (hence the 140 character limit). From 2006 to the present, we have witnessed an explosion of popularity of smartphones and other mobile devices like iPads; using Twitter applications is now a popular activity on these devices. Additionally, over this time period, Twitter has become closely intertwined with other social networks such as Facebook, with many users automatically cross-posting their updates between multiple sites. As a final experiment, we explore how tweets are created by taking advantage of the `source` field that Twitter attaches to each tweet.

The `source` field is different for each different Twitter client, so we begin by manually classifying all 54 unique `sources` that represented at least 1% of tweets in any month. We classify the `sources` in categories: Desktop, Mobile, Other OSNs. Certain `sources` exist in multiple categories (e.g., Echofon has both mobile applications and desktop applications), so we do not include these.

We present the breakdown of tweets from different sources in Figure 12, and make a number of observations. Overall, there is a consistently decreasing trend for desktop clients (including web, API, WebClient and other applications for the desktop computers), and a corresponding increasing trend for mobile clients (including iPhone, Android, BlackBerry, iPad, and other applications for the mobile devices). We also observe that mobile devices are briefly a popular source in 2006 (using SMS), but quickly drop before rising again. These trends quantify the shift towards mobile devices, with mobile devices representing the majority of tweets starting in 2013. Surprisingly, we observe that tweets automatically created by other OSNs (including TwitterFeed, Facebook, and Tumblr) consistently represent approximately 3% of the overall tweet volume.

Related Work

In this section, we briefly detail previous studies of Twitter users and their behavior.

Sample Coverage Most studies to understand user activities on Twitter utilize the public Twitter API to collect the information of users and tweets. To measure the limitations and representation of sample data, recent work (Morstatter et al. 2013) examined how Twitter selected tweets to re-

turn, and found that many API queries are not representative. However, we use the Twitter gardenhose in our study, which Twitter explicitly states is a random selection of public tweets.

By comparing US Twitter users to the US population using census data, previous work (Mislove et al. 2011) found that US users are a highly non-uniform sample of the population in terms of race/ethnicity, gender, and geographical distribution. Our work here is complementary, as we are examining the evolution of the Twitter userbase.

Tweeting Behavior A number of researchers have examined different aspects of the user tweeting behavior. To understand why users retweet, some papers analyze the patterns and causes of retweeting, suggesting that the primary motivations for retweeting are new information for the user (Macskassy and Michelson 2011) and statements of support (Recuero, Araujo, and Zago 2011). Our results do help clarify some prior work: for example, Macskassy et al. (Macskassy and Michelson 2011) report that 32% of tweets are retweets, contradicting our measurement of 10% at the same time. The mismatch is likely caused by the authors' snowball sampling method.

Several studies have focused on the deletion of tweets. Tweets are deleted for numerous reasons; for example, they can be spam, or need to be expressed in a different way, or can cause the user regret later on (Petrovic, Osborne, and Lavrenko 2013). The most comprehensive study found (Almuhimedi et al. 2013) that the majority of tweets get deleted within an hour, and the fastest tweets to be deleted generally have typos or need rephrasing. Both studies find that around 2–3% of tweets were deleted in their 2012 dataset, which is consistent with our results (2.35%) for the same time period.

Recently, researchers have worked to gain a better understanding of how users socialize with others in the network: the conversation generated by mentioning another user (Macskassy 2012), and the top trending topics generated using hashtag (Huang, Thornton, and Efthimiadis 2010; Yang et al. 2012).

Today, people can tweet by texting, use a mobile application, use an agent application to tweet (either non mobile or mobile), among other ways. By tracking the `source` of tweets, prior work (Perreault and Ruths 2011) has shown that mobile Twitter users are more likely to be active than non-mobile users, and the tweets made on a mobile device tend to be more conversational and personal. In our paper, we find an interesting similar trending that texting (one of the few ways non-smartphone users can access Twitter) has decreased in popularity steadily since 2006; whereas smartphone devices and applications have become very popular.

User Demographics Geo-locating users has become a prominent area in the study of Twitter data. With regard to the `location` field in the user profile, prior work (Hecht et al. 2011) found that 34% of Twitter users had entered fake locations in their profile that provided no geographical information or location and 11.5% entered geotags; the rest entered valid geographical data. Our results are consistent with this study, but more comprehensive (as we shown the evolution of these trends, instead of a single snapshot).

In terms of users' self-reported `lang`, our findings supports the previous findings by Krishnamurthy (Krishnamurthy, Gill, and Arlitt 2008) about the top 10 languages on Twitter in 2008. However, we also show that this situation has changed significantly in the intervening time, with English today covering barely half of the user population.

Prediction of User Characteristics While it is largely orthogonal to our work, there has been significant work in prediction of user profile fields, including user's location, gender, race/ethnicity, and age (Mislove et al. 2011; Pennacchiotti and Popescu 2011a; 2011b; Bergsma et al. 2013; Nguyen et al. 2013). For example, recent work (Bergsma et al. 2013) found that the user-provided last name, first name, and location can provide information on the potential country of origin, the language, gender, ethnicity, and race with between 83% and 90% accuracy. Others (Nguyen et al. 2013) have shown that the specific language, sentiment, linguistic style of the tweets, as well as the tweeting behavior (retweets, hashtags) can give some idea about the age or interests of the user.

Researchers have examined how to infer the user's location from the tweet contents. For example, some studies (Hecht et al. 2011; Chandra, Khan, and Muhaya 2011; Cheng, Caverlee, and Lee 2010) showed how to predict the user's location based on the tweet contents, by looking for location-specific words in their tweets. As a result, as the amount of tweets increases, the accuracy of correctly predicting the user's location is likely to increase as well.

Concluding discussion

Twitter has seen significant interest from the research community over the past few years. However, most studies have used Twitter data as a means to an end, such as to predict aspects of the real world or to better understand information flow through society. There has been relatively little work that has studied the evolution of Twitter itself. Given that Twitter has changed significantly, it becomes unclear how to interpret prior results and whether the assumptions made in the past are still valid.

Using a dataset of over 37 billion tweets from seven years, we presented a close examination of the evolution of the Twitter user population and their behavior. We are able to strongly quantify a number of trends, including the spread of Twitter across the globe, the shift from a primarily-desktop to a primarily-mobile system, the rise of malicious behavior, and the changes in users' tweeting behavior. We hope that our findings will aid researchers in understanding the Twitter platform and interpreting prior results. As Twitter continues to develop, we plan to repeat our analysis to study the future evolution of the Twitter user population and behavior.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported in part by NSF grants

CNS-1054233 and CNS-1319019, ARO grant W911NF-12-1-0556, and an Amazon Web Services in Education Grant.

References

Acohido, B. 2010. Facebook, Twitter see more spam attacks. http://usat.ly/1gKhYiY.

Almuhimedi, H.; Wilson, S.; Liu, B.; Sadeh, N.; and Acquisti, A. 2013. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *CSCW*.

Bergsma, S.; Dredze, M.; Durme, B. V.; Wilson, T.; and Yarowsky, D. 2013. Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter. In *NAACL*.

Bryant, M. 2010. Twitter goldrush on the way. Old usernames to be released. http://tnw.co/1ozpteR.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*.

Chandra, S.; Khan, L.; and Muhaya, F. B. 2011. Estimating Twitter User Location Using Social Interactions-a Content Based Approach. In *SocialCom*.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In *CIKM*.

Chowdhury, A. 2010. State of Twitter Spam. https://blog.twitter.com/2010/state-twitter-spam.

Christoforos, A. 2011. All these celebrities on Twitter. http://exm.nr/1epC2X5.

Cooper, B. B. 2013. The surprising history of Twitter's hashtag and 4 ways to get the most out of them. http://bit.ly/1euTB8z.

Gilbert, E., and Karahalios, K. 2010. Widespread Worry and the Stock Market. In *ICWSM*.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *CHI*.

Huang, J.; Thornton, K. M.; and Efthimiadis, E. N. 2010. Conversational Tagging in Twitter. In *HT*.

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about Twitter. In *WOSN*.

Macskassy, S. A., and Michelson, M. 2011. Why do People Retweet? Anti-Homophily Wins the Day! In *ICWSM*.

Macskassy, S. A. 2012. On the Study of Social Interactions in Twitter. In *ICWSM*.

Mao, H.; Shuai, X.; and Kapadia, A. 2011. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *WPES*.

Meeder, B.; Tam, J.; Kelley, P. G.; and Cranor, L. F. 2010. RT@ IWantPrivacy: Widespread violation of privacy settings in the Twitter social network. In *W2SP*.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and Analysis of Online Social Networks. In *IMC*.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the demographics of Twitter users (Poster Paper).

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. How Old Do You Think I Am? A Study of Language and Age in Twitter. In *ICWSM*.

Ostrow, A. 2009. Twitter Spam Invades Trending Topics. http://on.mash.to/1dfFANu.

Parr, B. 2009. PROJECT RETWEET: Twitter's Plan to Officially Support Retweeting. http://on.mash.to/1fJK9tQ.

Pennacchiotti, M., and Popescu, A.-M. 2011a. A Machine Learning Approach to Twitter User Classification. In *ICWSM*.

Pennacchiotti, M., and Popescu, A.-M. 2011b. Democrats, republicans and starbucks afficionados: user classification in twitter. In *KDD*.

Perreault, M., and Ruths, D. 2011. The Effect of Mobile Platforms on Twitter Content Generation. In *ICWSM*.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2013. I Wish I Didn't Say That! Analyzing and Predicting Deleted Messages in Twitter. *CoRR*.

Post, H. 2014. Twitter Statistics. http://bit.ly/18KIwd2.

Recuero, R.; Araujo, R.; and Zago, G. 2011. How Does Social Capital Affect Retweets? In *ICWSM*.

SEC. 2013. Twitter, Inc. Common Stock. http://1.usa.gov/1gVlt2o.

Stone, B. 2009. Replies Are Now Mentions. https://blog.twitter.com/2009/replies-are-now-mentions.

Stringhini, G.; Egele, M.; Kruegel, C.; and Vigna, G. 2012. Poultry Markets: On the Underground Economy of Twitter Followers. In *WOSN*.

Thomas, K.; Grier, C.; Song, D.; and Paxson, V. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *IMC*.

Thomas, K.; McCoy, D.; Grier, C.; Kolcz, A.; and Paxson, V. 2013. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. In *USENIX Security*.

2012. Twitter Terms of Service. https://twitter.com/tos/.

Vaas, L. 2013. Spammers take over Twitter Trends with sexy hashtags. http://bit.ly/OArYAX.

Weil, K. 2010. Measuring Tweets. https://blog.twitter.com/2010/measuring-tweets.

Yang, L.; Sun, T.; Zhang, M.; and Mei, Q. 2012. We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption? In *WWW*.

Yeung, K. 2013. Spammers are back on Twitter, and this time they're coming after you through lists. http://tnw.co/1epCTqJ.

The Tweets They are a-Changin': Evolution of Twitter Users and Behavior

Yabing Liu†, Chloe Kliman-Silver§, Alan Mislove†

†Northeastern University
§Brown University

ICWSM 2014

Twitter

Twitter: Popular microblogging platform

- Started in 2006 as SMS service
- Over 200 million monthly active users today
- Used by many organizations and individuals

Result: Significant amounts of Twitter research

- Twitter makes data easy to access
- Significant public data available

Examine how human society functions at scale

What have people studied?

Tweeting behavior

over 768,000 tweets in 1 month -- retweets [Macskassy and Michelson, ICWSM'11]

over 650,000 tweets over 1 month -- tweet contents [Macskassy, ICWSM'12]

over 476 million tweets over 7 months -- hashtags [Yang et al., WWW'12]

1.6 million deleted tweets over 1 week -- deletion of tweets [Almuhimedi, et al., CSCW'13]

Twitter user demographics

about 100,000 users from 3 datasets -- user lang [Krishnamurthy, et al., WOSN'08]

about 32 million English tweets over 1 month -- user location [Hecht et al., CHI'11]

The talk

Goal: How Twitter changes over time?

Collect over 37 billion tweets spanning over 7 years

Examine the evolution of the (public) Twitter ecosystem

- Whether prior results still hold
- Whether the (often implicit) assumptions of proposed systems are still valid

Outline

~~1 Motivation~~

~~2 Goals~~

3 Twitter Datasets

4 User characteristics

5 Tweeting behavior

First Twitter dataset (2006-2009)

Dataset	Date range	Users	Tweets	Date collected	Tweets	Users
Crawl	21/03/2006 – 14/08/2009	25,437,870	1,412,317,185	14/08/2009	~100%	~100%

Crawl:

Collected by previous work [Cha et al. 2010]

Iteratively download the 3,200 most recent tweets of all public users alive at the time

Notes:

Does not include any tweets deleted before August 14, 2009

The user information is as-of August 2009.

Second Twitter dataset

Dataset	Date range	Users	Tweets	Date collected	Tweets	Users
Gardenhose	15/08/2009 – 31/12/2013	376,876,673	36,495,528,785	Time of tweet	~10–15%	~30.61%

Gardenhose:

Twitter 'Gardenhose' public stream

https://stream.twitter.com/1.1/statuses/sample.json, with elevated access.

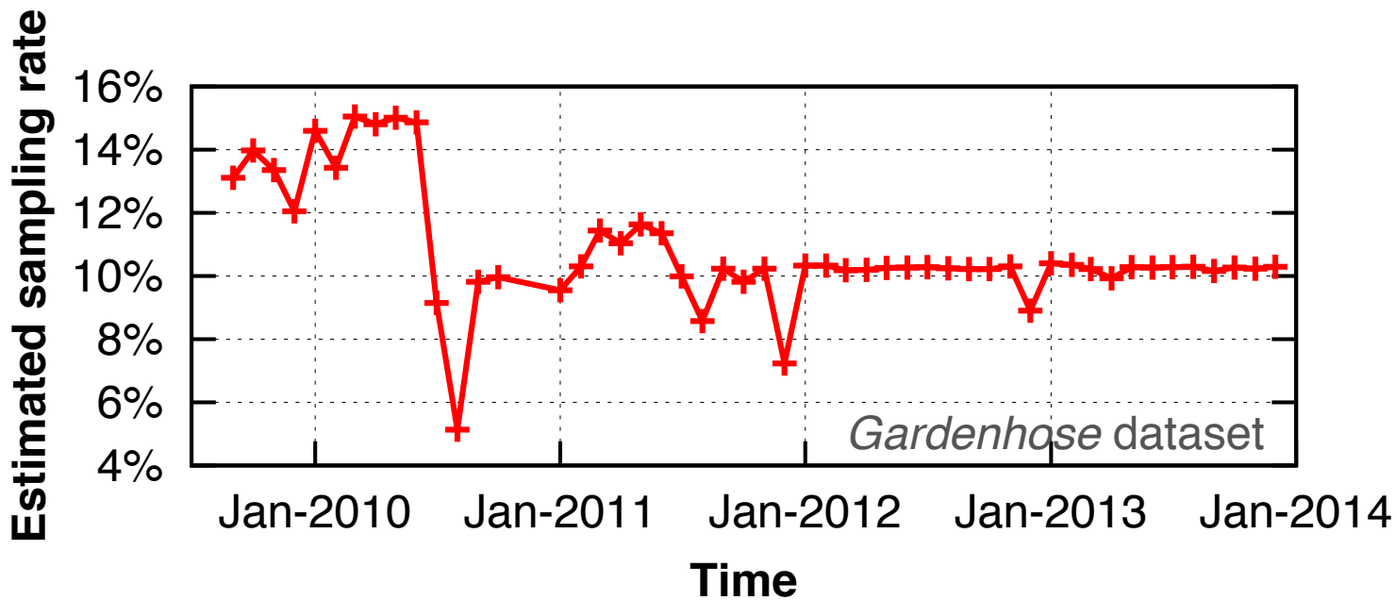
A random sample of all public tweets(tweet + user)

Notes:

With a bias towards more active users

Twitter does not inform us when user leave the network.

The sampling rate of



Notes:

Reason: Twitter does not state the rate.

A sampling rate of ~15% until July 2010, and ~10% since then

Our measurement infrastructure was down between Oct. 18, 2010 and Dec. 31, 2010.

Third Twitter dataset

Dataset	Date range	Users	Tweets	Date collected	Tweets	Users
UserSample	21/03/2006 – 31/12/2013	1,210,077	–	12/31/2013	~0.1%	~0.1%

UserSample:

- A random sample of users
- Generate 2 million random user_ids between 1 and 1,918,524,009
- Query Twitter in Jan 2014 for the most recent info on each user
- Both via the Twitter API and the web site
- 1,210,077 (60.51%) user_ids were ever assigned to a user.

Together:

- We have over 388 million unique users and over 37 billion tweets.
- For each analysis, we use the most appropriate dataset.

Outline

~~1 Motivation~~

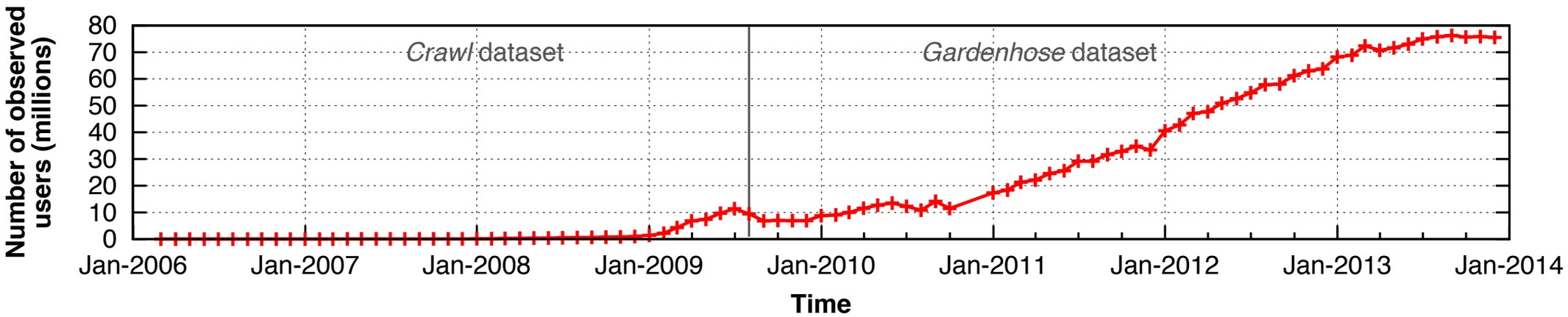
~~2 Goals~~

~~3 Twitter Datasets~~

4 User characteristics

5 Tweeting behavior

How is Twitter growing?



Observations:

Rapid growth from 2009 through 2012 and a leveling-off of the number in 2013

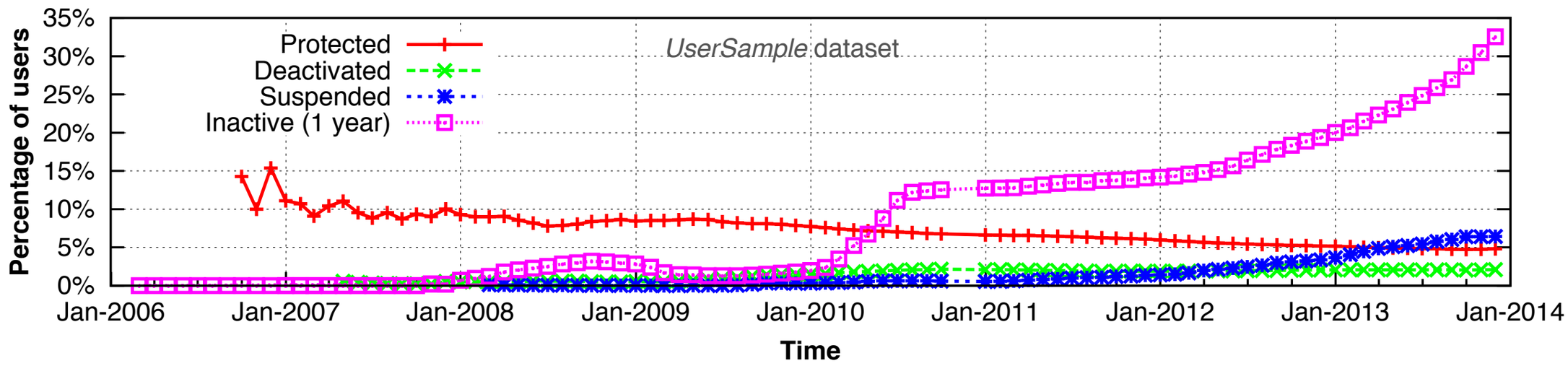
June 2013: Over 73 million users tweet VS. 218 million reported active users

Reasons:

Users from a random 10% sample of tweets

Twitter's definition of an active user: login activity, not tweeting activity

How many users are leaving



Observations:

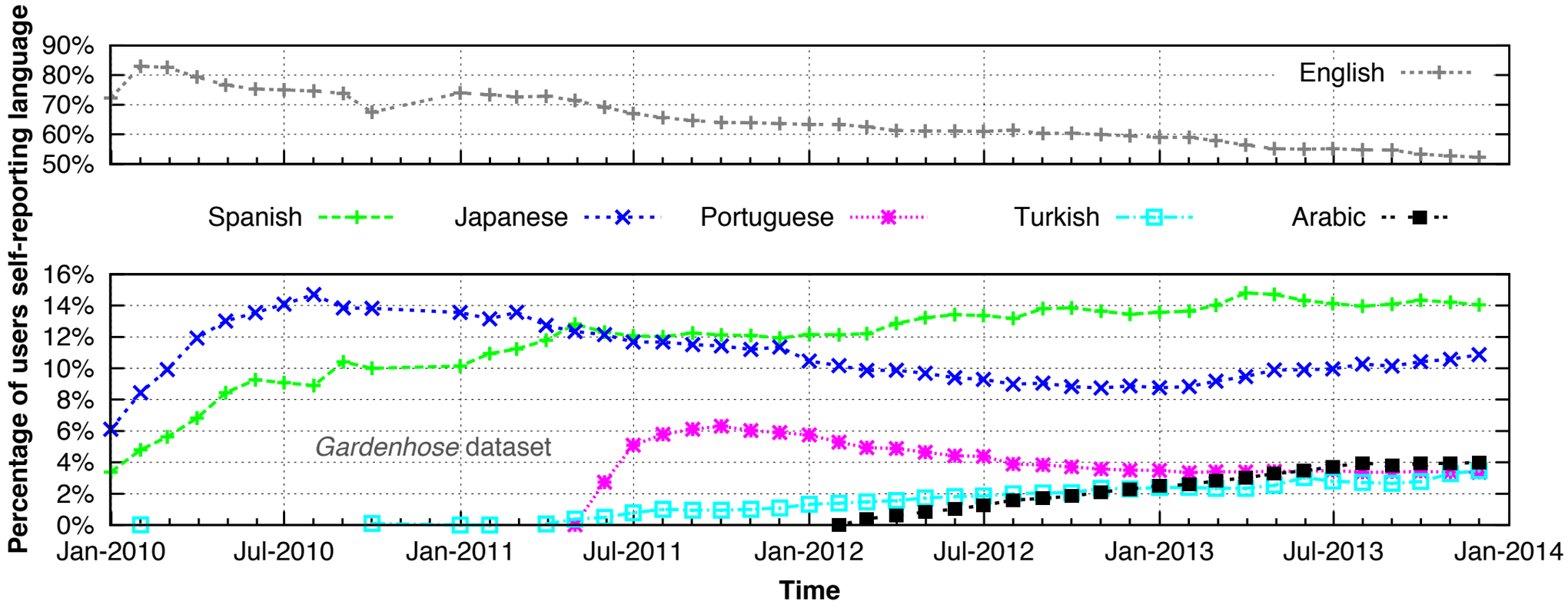
Protected accounts: goes down to 4.8% by 2013 -- most new accounts are public

Deactivated accounts: a relatively stable 2% of users

Suspended accounts: over 6% of entire Twitter users by 2013

Inactive accounts: up to 32.5% of all accounts by the end of 2013

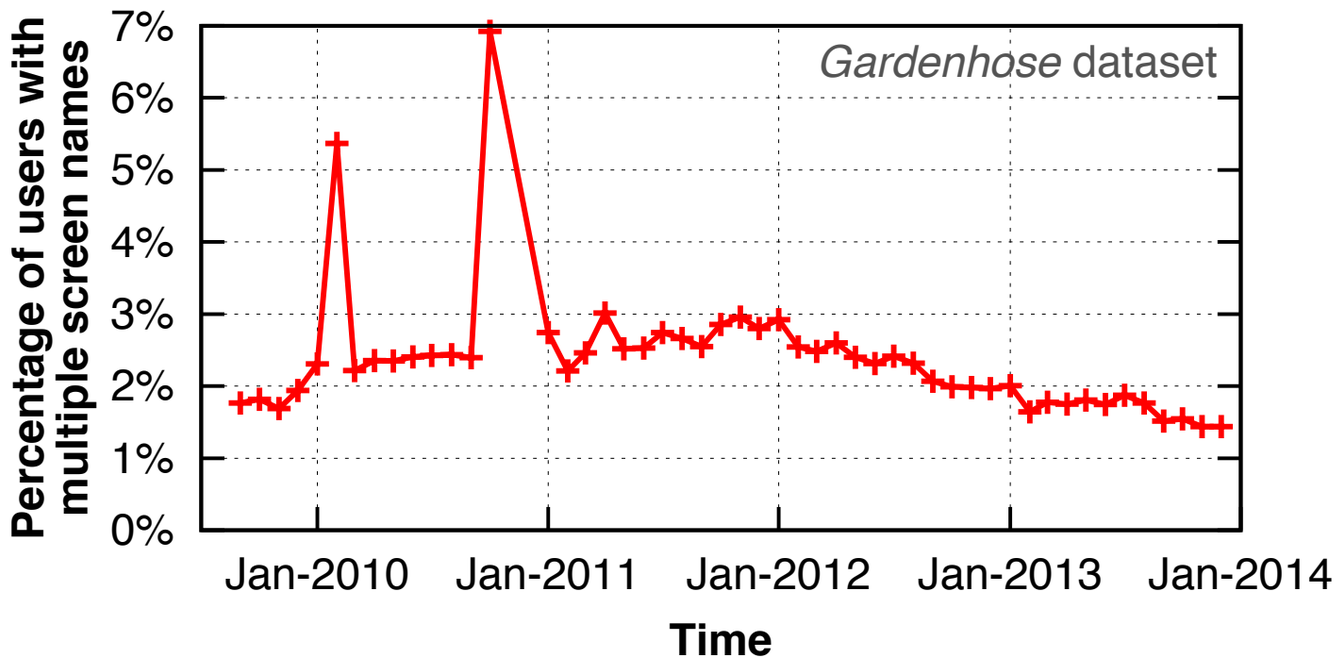
What languages do users speak?



Observations:

- The self-reported lang field since Jan.12, 2010
- English: a steady and continuing decrease of users from 83% to 52%
- Spanish and Japanese: approximately 10%
- More diverse and global

When do users change screen name?



Observations:

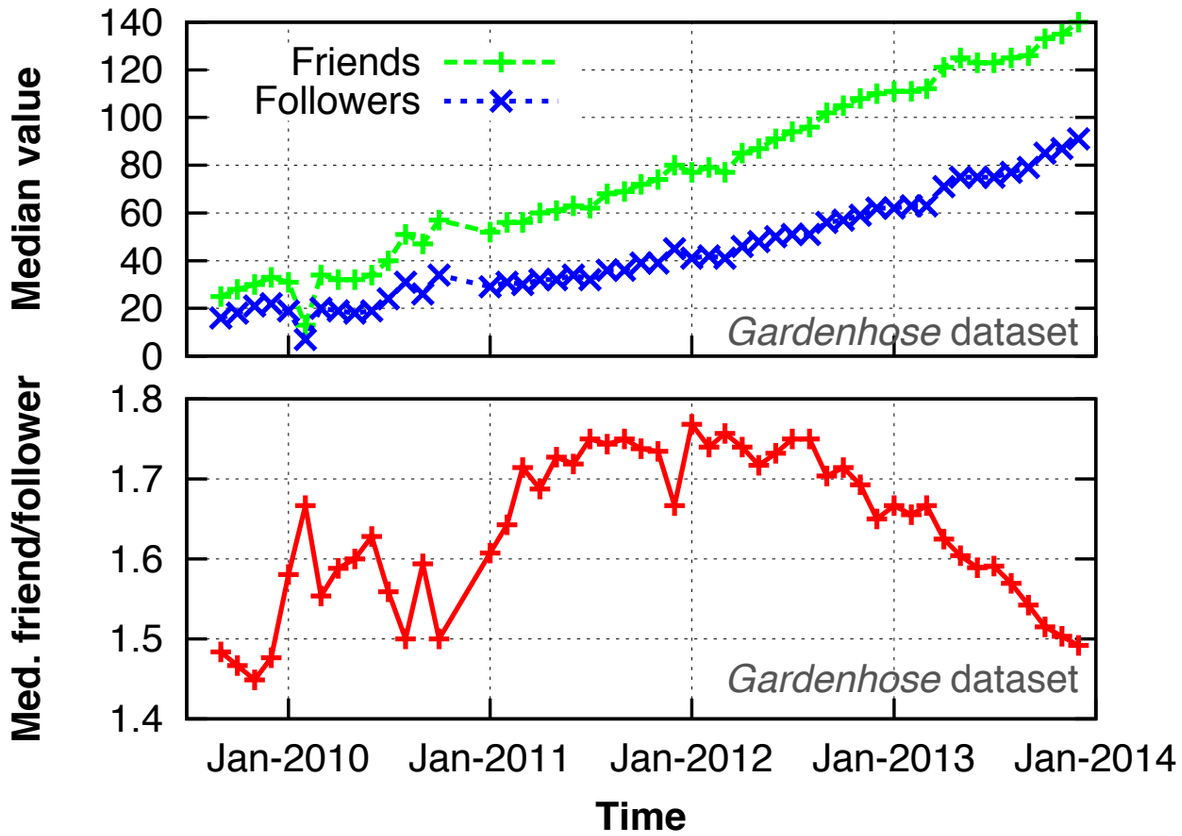
Up to 3% of users change their screen names every month.

Example: @Barack to @BarackObama

The "spikes" in Feb and Oct 2010: Twitter opened up old, inactive screen names to be reclaimed by active users.

To track users: user_id

How social are Twitter users?



Observations:

A dramatic increase in the median followers/friends count of almost 400% from 2009 to 2013

The distribution of followers is much more biased than the distribution of friends. => Twitter is disassortative.

The rise of Twitter follower spam in 2010 and 2011

Outline

~~1 Motivation~~

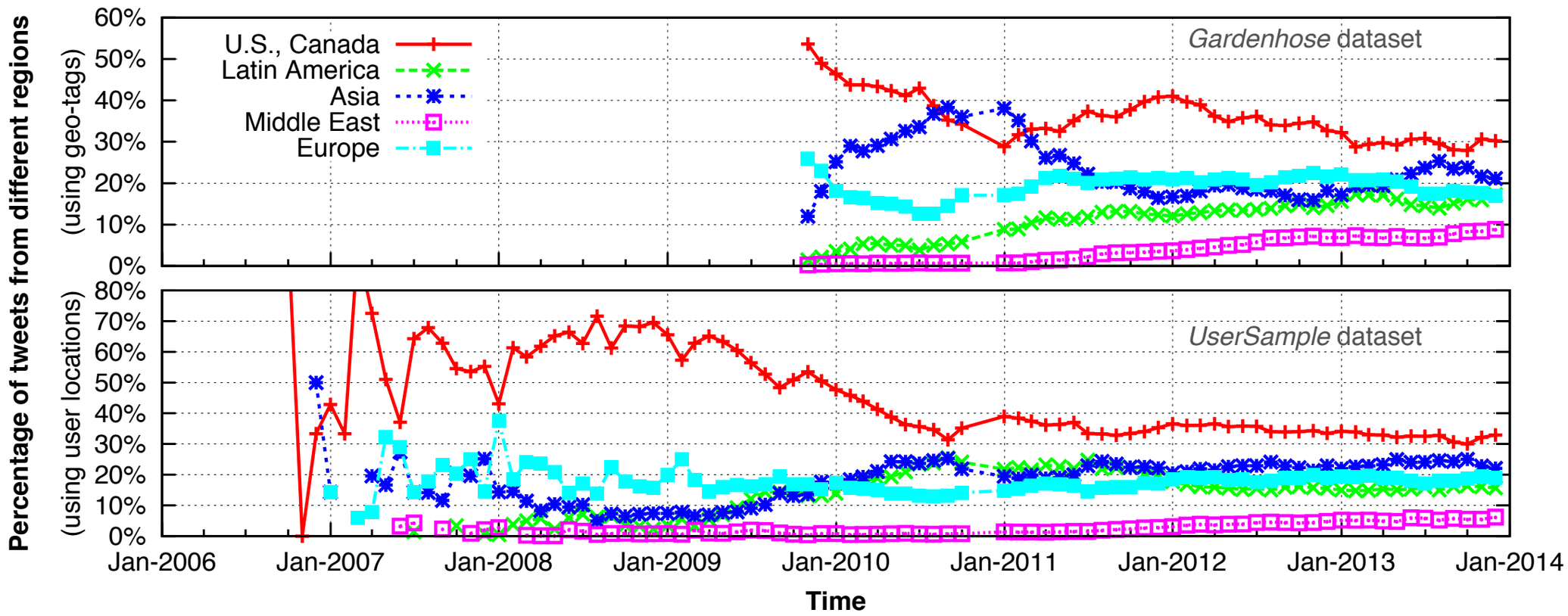
~~2 Goals~~

~~3 Twitter Datasets~~

~~4 User characteristics~~

5 Tweeting behavior

Where are the tweets coming



Information:

The self-reported, unformatted location field attached to user profile [Bing Maps]

The geo field(lat/lon) attached to some tweets since Nov. 2009 [GIS shape files]

42.4% of users provide a location string interpretable by Bing.

1.23% of tweets have included geo-tags.

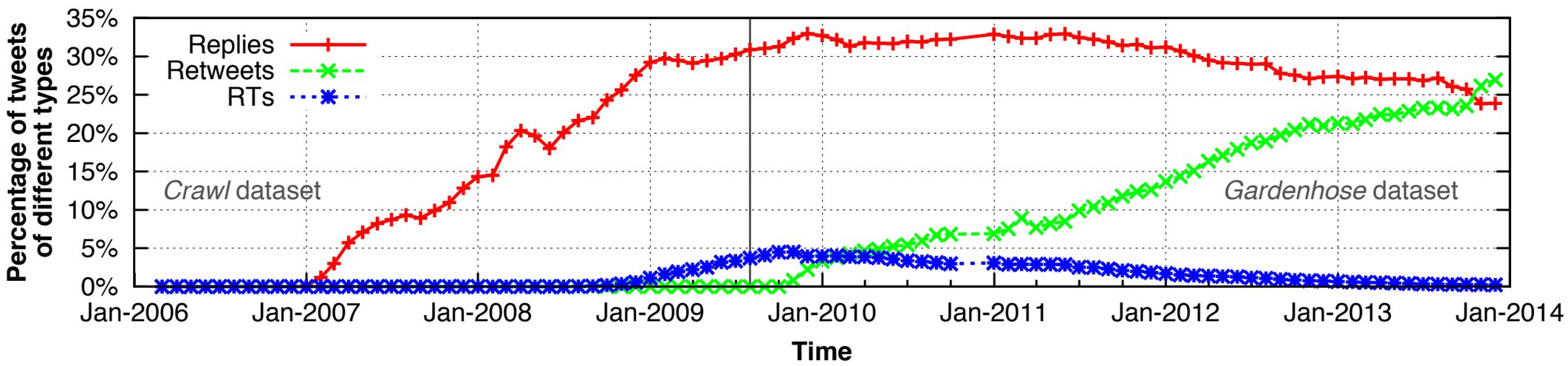
Observations:

U.S. and Canada: decline from 80% to 32%

Middle East and Latin America: a substantial increase of tweets

Europe: stable at 20%

What induces users to tweet?



Information:

Retweets: natively supported by Twitter since Nov 2009

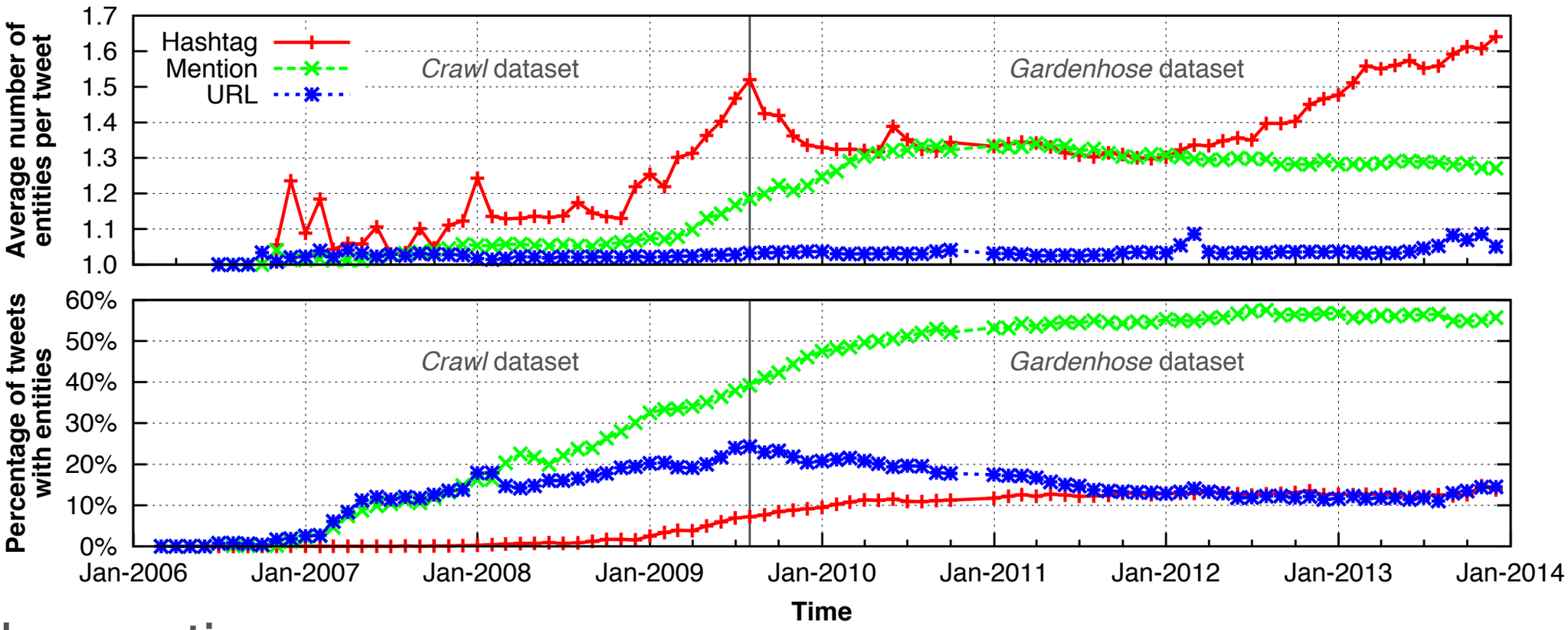
RTs: manually copied the tweet and added a "RT @username" at the beginning

Observations:

Retweets: the percentage increases rapidly afterwards.

Reply: a rapid adoption of the mechanism, peaking at ~35% of all tweets in 2010 and declining slightly afterwards

What do tweets contain?



Observations:

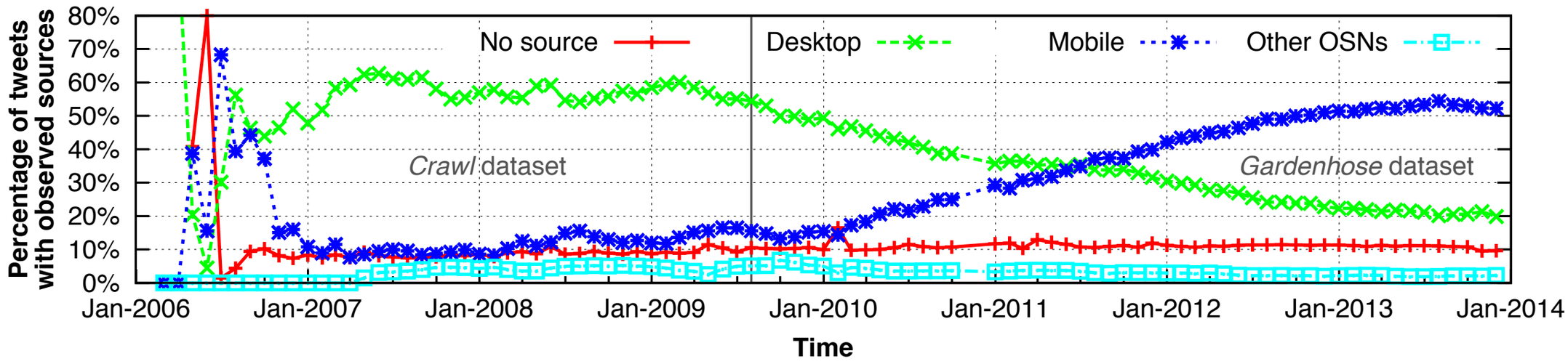
The percentage of tweets with mentions has increased substantially since 2009.

The percentage of tweets with URLs has decreased to stabilize at 12%.

URLs and mentions have stabilized around 1.0 and 1.3, respectively.

The average number of hashtags shows a continuing increase beyond 1.6.

What device are users tweeting from?



Information:

The source field attached to each tweet

Manually classify all 54 unique sources that represented at least 1% of tweets in any month

Observations:

A consistently decreasing trend for desktop clients and a corresponding increasing trend for mobile clients

Tweets created by Other OSNs: consistently ~3% of the overall tweets

Conclusions

Collect dataset of over 37 billion tweets from 7 years

- Examine the evolution of Twitter itself
- Focus on the Twitter users and their behavior

Quantify a number of trends

- the spread of Twitter across the globe
- the shift from a primarily-desktop to a primarily-mobile system
- the rise of spam and malicious behavior
- the changes in users' tweeting behavior

Aid researchers in understanding the Twitter platform and interpreting prior results

Questions?

We make all of our analysis available to the research community (to the extent allowed by Twitter's Terms of Service) at

http://twitter-research.ccs.neu.edu/

Email: ybliu@ccs.neu.edu

Backup slides

Determine user_id status in *UserSample* dataset

Query Twitter in Jan 2014 for the most recent info on each user

Both via the Twitter Rest API and the web site

https://twitter.com/intent/user?user_id="+userid

User_Id Statuses	Via
Public	Twitter API
Protected	Twitter API
Suspended	Web Site
Deactivated	Web Site + Tweet
Unknown	Web Site + NoTweet

Comparison of findings

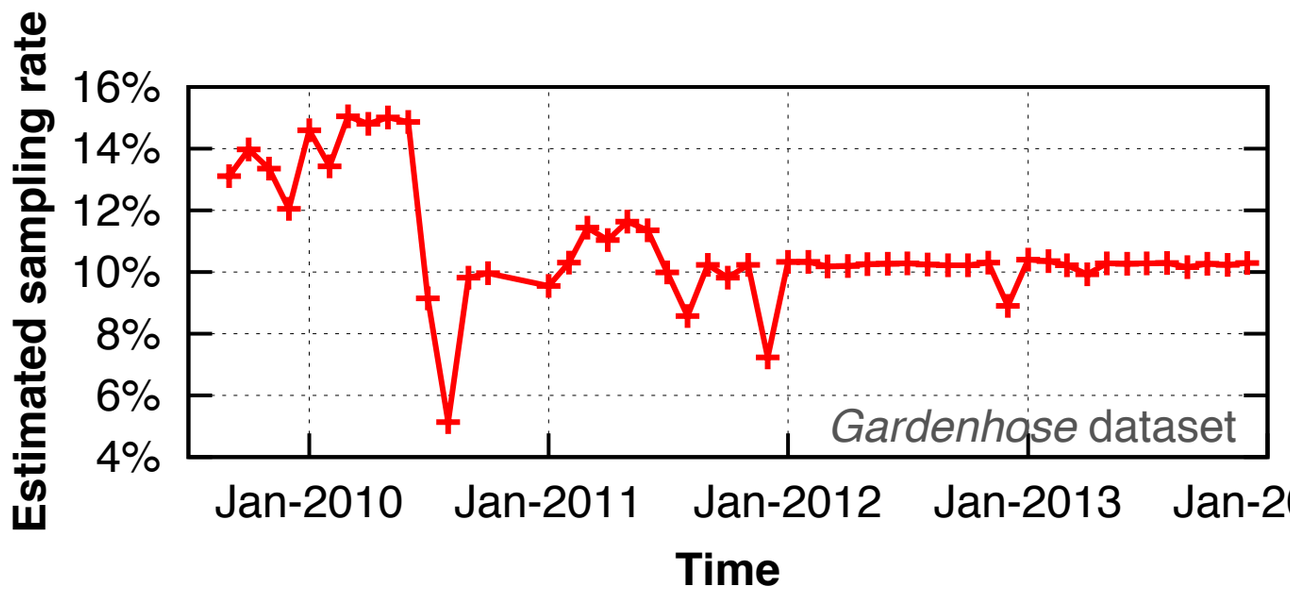
Examples:

[Macskassy and Michelson 2011] report that 32% of tweets are retweets, contradicting our measurement of 10% at the same time. The mismatch is likely caused by the authors' snowball sampling method.

[Petrovic, Osborne, and Lavrenko 2013] and [Almuhimedi et al. 2013] find that around 2-3% of tweets were deleted in their 2012 dataset, which is consistent with our results (2.35%) for the same time period.

In terms of lang, our findings supports the previous findings by [Krishnamurthy, Gill, and Arlitt 2008] about the top 10 languages on Twitter in 2008. However, we also show that this situation has changed significantly, with English today covering barely half of the user population.

The sampling rate of



$$rate = \frac{obs}{sc_{last} - sc_{first}}$$

The average value of rate across all users with $SC_{last} - SC_{first} > 1000$

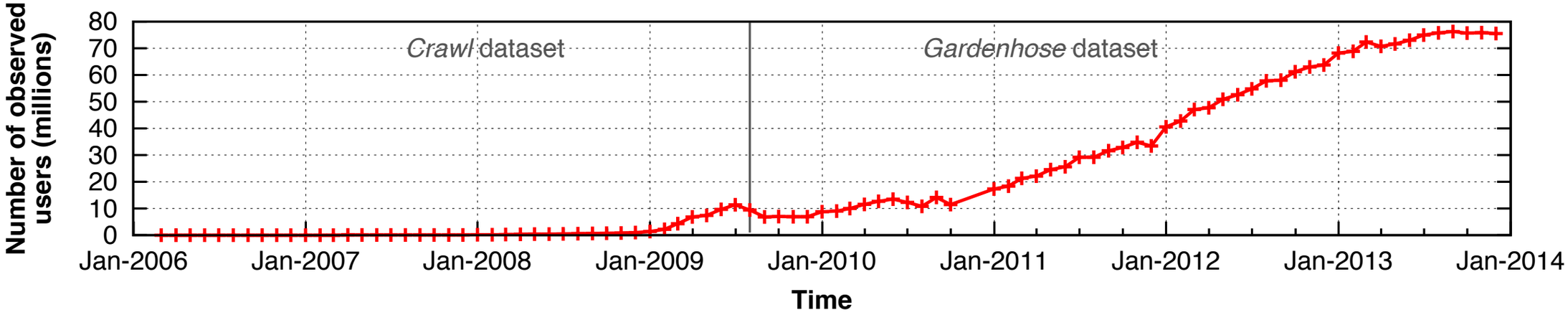
The first observed value of statuses_count SC_{first}

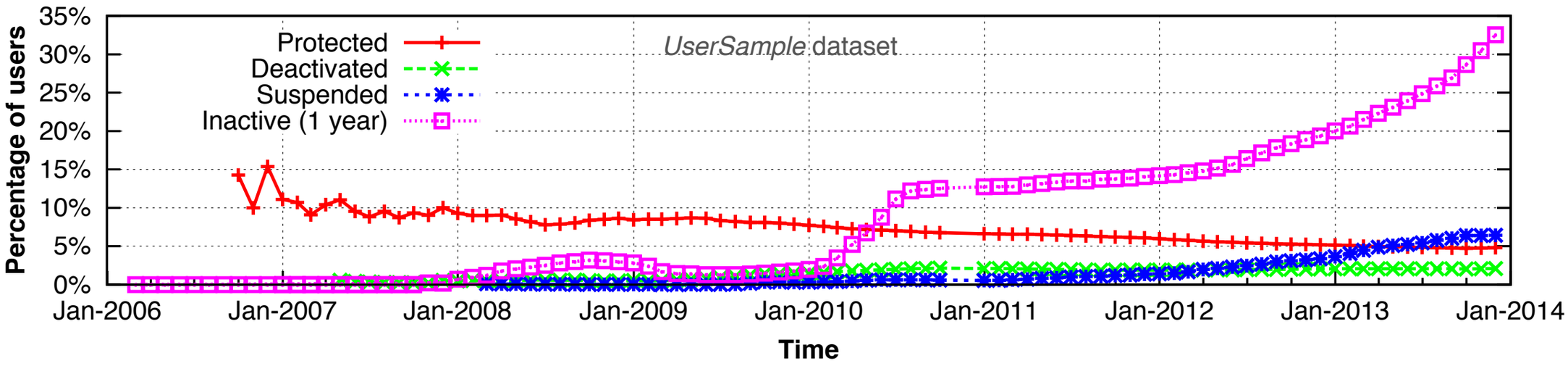
The last observed value of statuses_count SC_{last}

The number of tweets we observed obs

JSON Example: {"created_at":"Fri Nov 01 00:00:40 +0000 2013","id":
396064209307303936,"text":"RT @HentaiUchi: 17 Like it? RTVRetweet it!
http:\/\/t.co\/VKiS2ceBuvf",user":{"id":
1639501730,"id_str":"1639501730","name":"Momo Velia
Deviluke","screen_name":"MomoVeliia","followers_count":

Users joining and leaving





The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior

Yabing Liu[†] **Chloe Kliman-Silver**[‡] **Alan Mislove**[†]

[†]Northeastern University [‡]Brown University

Abstract

The microblogging site Twitter is now one of the most popular Web destinations. Due to the relative ease of data access, there has been significant research based on Twitter data, ranging from measuring the spread of ideas through society to predicting the behavior of real-world phenomena such as the stock market. Unfortunately, relatively little work has studied the changes in the Twitter ecosystem itself; most research that uses Twitter data is typically based on a small time-window of data, generally ranging from a few weeks to a few months. Twitter is known to have evolved significantly since its founding, and it remains unclear whether prior results still hold, and whether the (often implicit) assumptions of proposed systems are still valid.

In this paper, we take a first step towards answering these question by focusing on the evolution of Twitter's users and their behavior. Using a set of over 37 billion tweets spanning over seven years, we quantify how the users, their behavior, and the site as a whole have evolved. We observe and quantify a number of trends including the spread of Twitter across the globe, the rise of spam and malicious behavior, the rapid adoption of tweeting conventions, and the shift from desktop to mobile usage. Our results can be used to interpret and calibrate previous Twitter work, as well as to make future projections of the site as a whole.

Introduction

Online social networks(OSNs) are now a popular way for users to connect, communicate, and share content; many serve as the de-facto Internet portal for millions of users (Post 2014). Because of the massive popularity of these sites, data about the users and their communication offers unprecedented opportunities to examine how human society functions at scale. As a result, significant recent research has focused on these sites, with a particular emphasis on Twitter due to the relative ease of accessing a large amount of data. For example, recent papers have ranged from studying Twitter-specific behavior (e.g., the patterns of retweeting (Macskassy and Michelson 2011), deletion of tweets (Almuhimedi et al. 2013) and usage of different entities in tweet contents (Macskassy 2012; Yang et al. 2012)), to examining how privacy leaks and violations can occur (Mao, Shuai, and Kapadia 2011; Meeder et al. 2010), and even using aggregated data from Twitter to predict the behavior of real-world phenomena such as the stock market (Gilbert and Karahalios 2010).

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

While the set of research using Twitter data has expanded rapidly, there has been relatively little work that has studied the evolution of the Twitter ecosystem *itself*. For example, Twitter has grown from thousands of users in 2007 to millions in 2009 to hundreds of millions in 2013. In parallel with this growth, we have seen a significant maturation of the Twitter platform: Twitter today is used by many organizations and individuals as a primary way of communicating with others. Most research that uses Twitter data is typically based on a small time-window of data—generally ranging from a few weeks to a few months—making it difficult to quantify long-term trends. Twitter is known to have evolved significantly since its founding, and it remains unclear how much the user base and behavior has evolved, whether prior results still hold, and whether the (often implicit) assumptions of proposed systems are still valid.

In this paper, we take a first step towards answering these questions by directly examining the evolution of Twitter itself, focusing on the Twitter users and their behavior. Using a set of over 37 billion tweets spanning between 2006 and 2013, we quantify how the users, their behavior, and the site as a whole have evolved. We observe and quantify a number of interesting patterns, including

- The spread of Twitter across the globe, both in terms of users in different regions and tweets containing different languages (e.g., the percentage of U.S./Canada users drops from over 80% to 32%, and the percentage of users tweeting in English falls from 83% to 52%).
- The percentage of tweets that are no longer available due to a user's or Twitter's action increases to over 20% for some time ranges.
- The percentage of Twitter user accounts today that are inactive shows rapid growth; over 32.5% of accounts have not tweeted for over a year.
- The increase of malicious behavior on Twitter beginning in 2009, including fake followers, fake accounts, and hashtag promotion; over 6% of all accounts are now suspended.

Dataset	Date range	Users	Tweets	Date collected	Coverage of all Tweets	Coverage of all Users
Crawl	21/03/2006 – 14/08/2009	25,437,870	1,412,317,185	14/08/2009	∼100%	∼100%
Gardenhose	15/08/2009 – 31/12/2013	376,876,673	36,495,528,785	Time of tweet	∼10–15%	∼30.61%
UserSample	21/03/2006 – 31/12/2013	1,210,077	—	12/31/2013	∼0.1%	∼0.1%
Total	21/03/2006 – 31/12/2013	388,796,600	37,907,845,970	—	—	—

Table 1: Source and basic statistics for the Twitter data used in this study.

- The switch from a primarily-mobile system (via SMS) to a primarily-desktop system (via the web) and back to a primarily-mobile system (via mobile applications). Today, over half of all tweets are created on mobile devices.

Our results can be used to interpret and calibrate previous Twitter studies, as well as to make future projections of the site as a whole. We make all of our analysis available to the research community (to the extent allowed by Twitter's Terms of Service) to aid other researchers and to stimulate further research in this area; researchers can access it at

`http://twitter-research.ccs.neu.edu/`

Background and Data Source

Twitter is a "micro-blogging" service that allows users to multicast short messages (called *tweets*). Each user has a set of other users (called *followers*) who receive their messages; those who a user follows are called *friends*. The follow relationship in Twitter is directed, and requires authorization from the followee only when the followee has elected to make their account *protected*. Each tweet can only be up to 140 characters in length, and the default setting in Twitter is to allow all tweets to be publicly visible.

Twitter data We obtain our Twitter data from two sources; basic statistics of these datasets is provided in Table 1.[1] First, we use an almost-complete[2] collection of all tweets issued between March 21, 2006 and August 14, 2009 collected by previous work (Cha et al. 2010); we refer to this as the *Crawl* dataset. This dataset was collected in August 2009 by iteratively downloading all of the tweets of all public users alive at the time.

Second, we collect data from the Twitter "gardenhose" public stream[3] between August 15, 2009 and December 31, 2013; we refer to this as the *Gardenhose* dataset. Our measurement infrastructure was down for the 10 weeks between October 18, 2010 and December 31, 2010, so we do not have data for that time period. Twitter states that the gardenhose is a random sample of all public tweets. Each *Gardenhose* tweet includes information about the user who created the tweet (e.g., the user's location and count of total tweets) that is current as-of the time of the tweet.

[1]This study was conducted under Northeastern University Institutional Review Board protocol #10-03-26.

[2]The dataset does not include any tweets deleted before August 14, 2009, and only includes the 3,200 most recent tweets (as of August 14, 2009) for each user.

[3]https://stream.twitter.com/1.1/statuses/sample.json, with elevated access.

Gardenhose sampling rate We briefly estimate the sampling rate of the *Gardenhose* dataset; Twitter states that the gardenhose is a random sample, but does not state the rate. We estimate the sampling rate by relying on the `statuses_count` field of the user in each tweet; the `statuses_count` field represents the total number of tweets (statuses) that the user has issued at the time of the tweet. Each month, we determine the first observed value of `statuses_count` (sc_{first}), the last observed value of `statuses_count` (sc_{last}), and the number of tweets we observed (obs). We can then estimate the sampling rate for that user with

$$rate = \frac{obs}{sc_{last} - sc_{first}}$$

We plot the average value of $rate$ across all users with $sc_{last} - sc_{first} > 1000$ in Figure 1, observing a sampling rate of ∼15% until 07/2010, and ∼10% since then (the "dips" observed in the graph are due to short periods of time when our measurement infrastructure was down).

Limitations Because the *Crawl* dataset was collected in August 2009 (as opposed to the *Gardenhose* dataset, which was collected over a period of years, as tweets were issued), the user information for the *Crawl* dataset is as-of August 2009. This limitation will occasionally present itself during our analysis, and we discuss these limitations in-line.

We also face two limitations with the *Gardenhose* dataset. First, the *Gardenhose* dataset ends up containing biased sample of users, with a bias towards more active users. The reason for this is that users who tweet very often are extremely likely to show up in our dataset; a user who tweets only once has a ∼10% chance of appearing. Second, Twitter

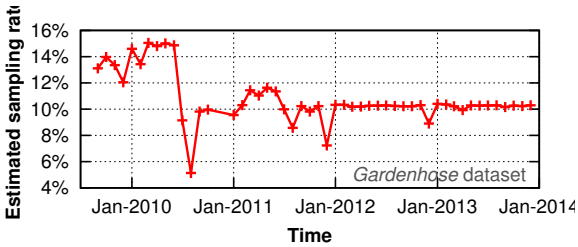


Figure 1: The estimated sampling rate (the average value of *rate* for users with more than 1,000 statues in a month) of the *Gardenhose* dataset over time. The occasional drops in sampling rate are due to times that our collection infrastructure was down for short periods of time.

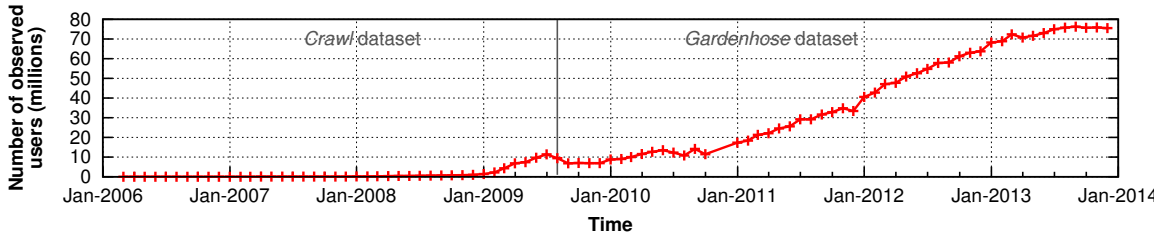


Figure 2: The number of users we observed tweeting in each month in the *Crawl* and *Gardenhose* datasets. The "dip" in August 2009 is due to the switch from a complete sample to a 15% sample. Our numbers are much lower that Twitter's announced "active users" numbers due to our data sample and Twitter's definition of an "active" user.

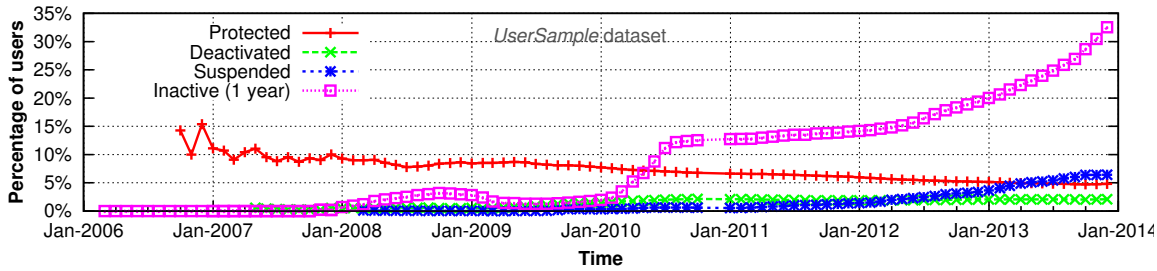


Figure 3: The percentage of the entire Twitter user base over time whose accounts are protected, deactivated, suspended, or inactive (for at least a year), based on the *UserSample* dataset. We observe a dramatic increase in both inactive and suspended accounts starting in 2010.

does not inform us when users leave the network, so we are unable to determine when users mark their accounts as protected (thereby hiding their tweets), are *suspended* by Twitter for violating the Terms of Service, or *deactivate* their accounts (i.e., manually delete their account).

To address these limitations of the *Gardenhose* dataset, we collect a third and final dataset *UserSample* that represents a random sample of *users* instead of *tweets*. Specifically, we generate 2 million random `user_id`s between 1 and 1,918,524,009 (the largest `user_id` that we ever observed), representing a ∼0.1% sample of all Twitter users. We then query Twitter (both via the API and the web site) in January 2014 for the most recent information on each of these users, allowing us to determine if the `user_id` was ever assigned, has been suspended, or is protected. We infer if the `user_id` has been deactivated by the user if Twitter says the user does not exist but we observed a tweet from the user in our *Crawl* or *Gardenhose* datasets. Overall, we find that at least 1,210,077 (60.51%) of these 2 million `user_id`s were ever assigned to a user.

Throughout our analysis, we use the most appropriate dataset(s) for each question at hand. Additionally, we label each graph with the dataset(s) that it uses.

User characteristics

We begin our analysis by studying how the user population of Twitter has changed since its inception.

User growth and activity We first examine the characteristics of the Twitter user population. Figure 2 shows the total number of users that we observed over time; we observe massive growth that is in-line with Twitter's reported number of monthly active users (Weil 2010). For example, we observed over 73 million users tweet in June 2013; while Twitter reports 218 million active users for that time period (SEC 2013); our number is lower due to the fact that we have a random 10% sample and Twitter's definition of an active user is based on login activity, not tweeting activity. We also observe rapid growth from 2009 through 2012, with a leveling-off of the number of active users in 2013; this suggests that Twitter's user population growth may be slowing down.

Next, we briefly examine the aggregate level of user ac-

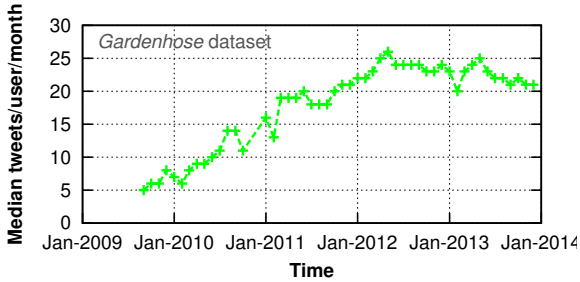


Figure 4: The median number of tweets per user per month over time, based on the first and last `statuses_count` observed for each user. Note that this result is based on *Gardenhose* dataset, which is biased towards more active users.

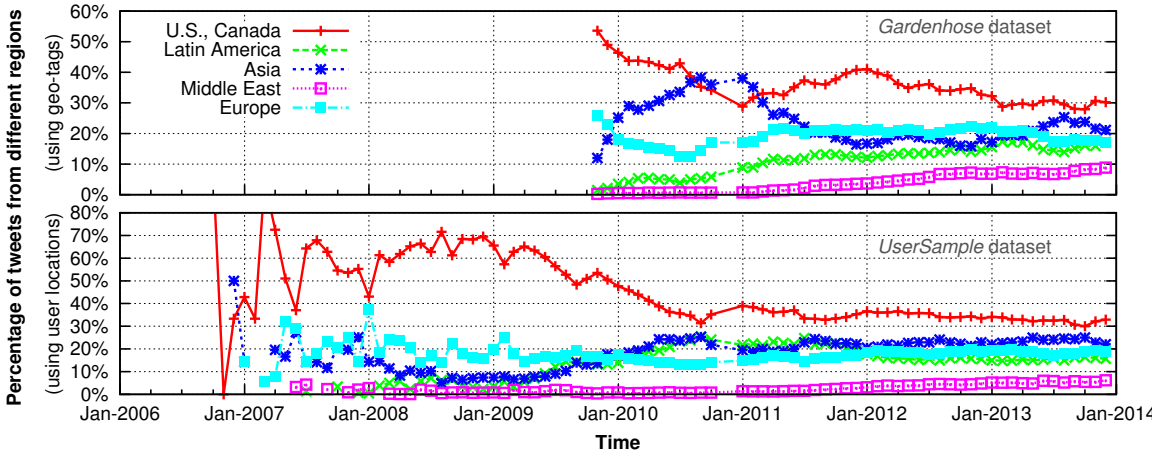


Figure 5: The percentage of tweets created in different geographical regions over time. Shown are locations inferred from self-reported user locations (*UserSample* dataset, bottom) and geo-tags (*Gardenhose* dataset, top). Geo-tags were introduced in November 2009, so we only show data since then. Africa and Oceania are both almost always less that 1%, and are not shown for clarity. A number of interesting trends can be observed, quantifying the spread of Twitter across the globe.

tivity as Twitter has evolved. To do so, we use observed tweets for each user in the *Gardenhose* dataset and use the first and last `statues_count` field in the user profile (we note that using the *Gardenhose* dataset introduces a bias towards more active users). The median value of the number of tweets per user per month is presented in Figure 4. We can see a general rise in activity from later 2009 to 2012, with the rate stabilizing and then decreasing.

Users leaving Twitter While Twitter has seen massive growth, a non-trivial percentage of users leave Twitter, either by deactivating their account or by being suspended by Twitter. Additionally, many users mark their account as *protected*, meaning only their approved followers can view their tweets. Finally, many users simply leave Twitter and become *inactive*, meaning they have not tweeted for over a year. In Figure 3, we use the *UserSample* dataset to plot the percentage of the entire Twitter user base whose accounts are deactivated, suspended, protected, or inactive at different points in time (e.g., in January 2013, 2% of the user population at that time were deactivated, 4% had been suspended, 5% had marked their accounts as protected, and 20% had not tweeted for over a year).[4]

We first observe the massive percentage of inactive accounts, representing up to 32.5% of all accounts by the end of 2013. The increasing nature of this trend suggests that the leveling off of active users per month (Figure 2) may soon cause the majority of accounts to be inactive. We also observe that up to 15% of users who joined in Twitter before December 2007 protected their accounts, while the percentage goes down to 4.8% by 2013; this implies that most users who have joined Twitter recently have kept their accounts public. We further observe a dramatic increase in the percentage of suspended users, with over 6% of the entire Twitter user population suspended by late 2013; this is inline with studies on the rise of malicious activity (Thomas et al. 2013; Yeung 2013). Finally, a relatively stable 2% of users who have deactivated their accounts, indicating that relatively few users are leaving Twitter by explicitly deleting their accounts.[5]

User location Now, we examine the geographical distribution of the users over time. To do so, we rely on two pieces of information: (a) the self-reported, unformatte t the location field, we query Bing Maps with each unique location string, and only consider results that Bing returns with "high confidence". To interpret the geo-tags, we use country GIS shape files to translate latitude/longitudes into countries and administrative districts. Since geo-tags were introduced in November 2009, we only report geo-tags for months afterwards. We find 42.4% of users provide a location string interpretable by Bing, and 1.23% of tweets have included geo-tags.

We present the results of this analysis in Figure 5, showing the percentage of users located to different regions of the world using both self-reported locations (bottom) and geo-tags (top). We observe a number of interesting trends that quantify the spread of Twitter across the globe: First, we observe a steep decline of the percentage of the tweets from the U.S. and Canada from a high of above 80% to 32%; most of this decline comes in 2009 and 2010. At the same time, we observe a substantial increase in the percentage of tweets from the Middle East (starting in early 2011, corresponding

[4]Since Twitter does not provide the date a user's account becomes unavailable, we define the date of being suspended, deactivated, protected as the last date on which we observed a tweet from the user.

[5]We may underestimate the percentage of users who deactivated their accounts, as we can only infer that deactivated `user_ids` ever existed if we observed a tweet from them.

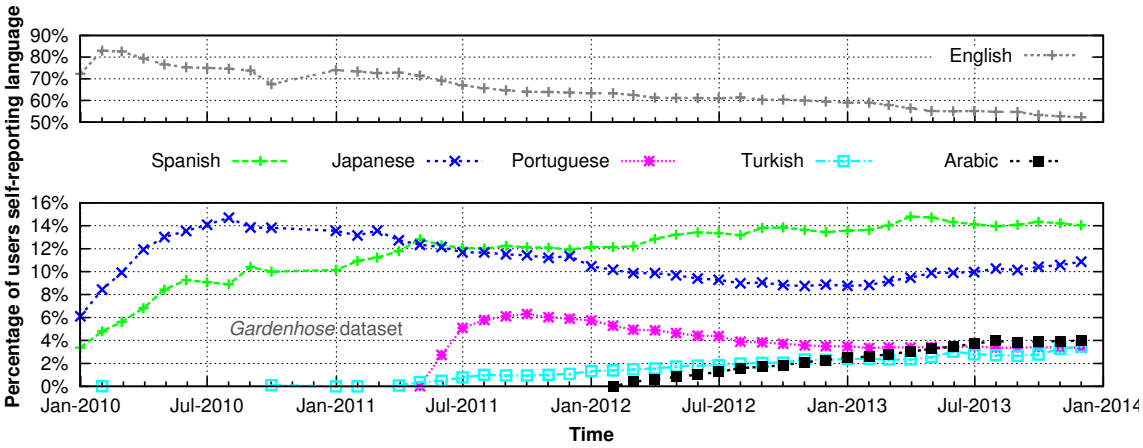


Figure 6: The percentage of users self-reporting the six most popular languages over time. We plot English separately in the top graph in order to increase the readability; note that the scale is different between the two graphs. English shows a mostly linear decrease from 83% in January 2010 to 52% in December 2013.

to the Arab Spring) and Latin America. However, Europe is relatively stable over the course of Twitter's evolution, generally representing around 20% of the tweets. Finally, comparing the two graphs, we can observe a difference in the spread of geo-tags, with Asia users being over-represented relative to the entire user population in 2009 and 2010. This is likely due to the popularity of smartphones equipped with GPS in Asia, which have since become popular globally.

Languages We now examine the languages Twitter users report over time. To do so, we rely on the self-reported `lang` field that Twitter allows users to specify in their profile; this field first appeared on January 12, 2010, so we report data since then in Figure 6 (English in the top graph, other languages in the bottom graph). We observe a steady (and continuing) decrease of users reporting English, from a high of 83% to 52% in December 2013. Spanish and Japanese show a steady representation of approximately 10%, while the decrease of English is correlated with an increase in a large number of other languages, including Turkish, Portuguese, and Arabic (correlating well with the results in the previous section showing the distribution of user locations). Overall, these results underscore the fact that Twitter's user population is continuing to become more diverse and global.

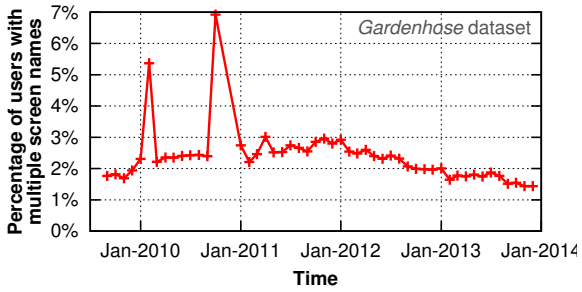


Figure 7: The percentage of users who have used more than one screen names in each month. The "spikes" correspond to times where Twitter released old, inactive screen names to be reclaimed.

Screen name changes A little-known feature of Twitter is that users can easily change their screen name (e.g., changing @Barack to @BarackObama), meaning tweets from the same user may show up under different screen names.[6] Using our *Gardenhose* dataset, we can observe these changes happening by looking for tweets from the same `user_id` with different screen names. Figure 7 plots the percentage of the user population that we observe to have used multiple screen names each month; the "spikes" in February and October 2010 correspond to time periods where Twitter opened up old, inactive screen names to be reclaimed by active users (Bryant 2010). Previous results have suggested that these users are more likely to be spammers (Chowdhury 2010), and our results correlate with prior studies showing an increase in the level of Twitter spam in 2010 and 2011 (Thomas et al. 2011; Acohido 2010). Overall, our results show that a non-trivial percentage (up to 3%) of users change their screen names each month; this suggests that researchers should internally refer to users with Twitter's (unchangeable) `user_id`s to ensure that users are accurately tracked over time.

Social characteristics We now turn to examine the social characteristics of Twitter users. Recall that Twitter users can *follow* each other, and following a user means the followed

[6] If a user changes their screen name, their `user_id` remains the same; this allows us to track screen name changes. There is no limit to the number of times a user can change their screen name.

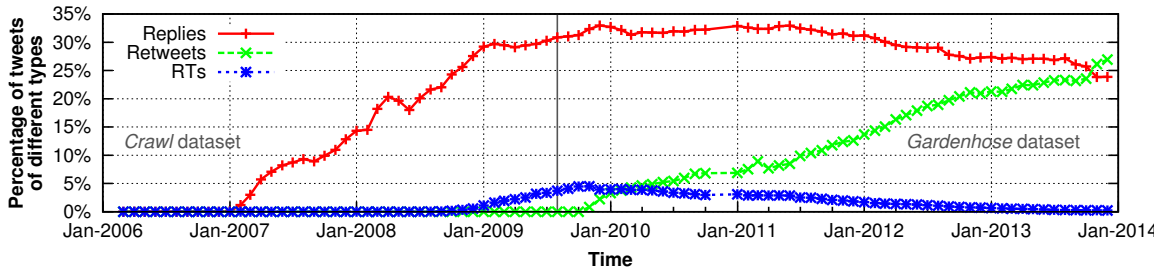


Figure 8: The percentage of tweets of different types over time. Both *RTs* and *Retweets* represent retweets, with the former capturing manually created retweets by users. Native retweets were supported by Twitter starting in November 2009.

user's tweets will show up when you log in. Following a user only requires permission if the followed user's profile is protected.

Using the *Gardenhose* dataset, we calculate the median number of followers (i.e., those following a user) and friends (i.e., those who a user follows) for all observed users over time, and present the results in Figure 9 (top). Similar to Figure 4, these results are biased towards more active users due to the use of the *Gardenhose* dataset. We observe a dramatic increase in the median followers/friends count of almost 400% from 2009 to 2013. This trend underscores Twitter's importance as a information dissemination platform; today, many celebrities, companies, and organizations use Twitter as one of the primary mechanisms to communicate with others (Christoforos 2011).

We also examine the average ratio of friends-to-followers in Figure 9 (bottom), and find an interesting trend: the ratio increases from 1.50 to a high of 1.77 in January 2012 before returning to its previous value. We make two observations. First, the fact that the ratio is much higher than 1 indicates that the distribution of followers is much more biased than the distribution of friends (i.e., most users have many more friends that followers) indicating that Twitter is disassortative; similar observations have been made about Twitter (Cha et al. 2010) and other social networks (Mislove et al. 2007). Second, the increase corresponds well with the rise of Twitter follower spam in 2010 and 2011 (Stringhini et al. 2012); we posit that the subsequent decrease is likely due to Twitter's more active role in suspending and deleting malicious accounts (Thomas et al. 2013; 2011).

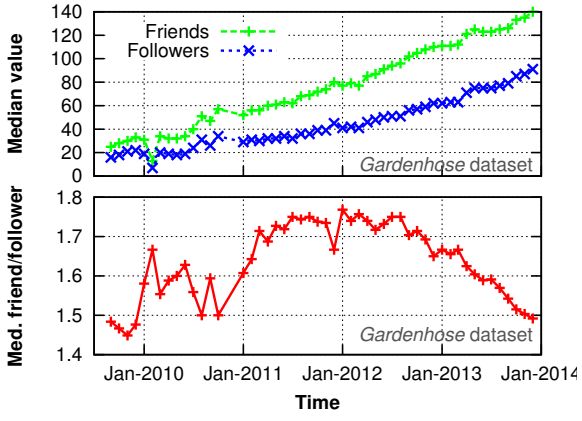


Figure 9: The median number of friends and followers across all users (top), and the median ratio of friends to followers (bottom) as derived from the *Gardenhose* dataset. We observe a dramatic densification of the Twitter social graph. Similar to Figure 4, using the *Gardenhose* dataset causes a bias towards more active users.

Tweeting behavior

In the above section, we examined the changing patterns in Twitter user population, now we turn to take a look at the changes in users' tweeting behavior over time.

Tweet causes We begin by examining internal-to-Twitter actions that cause tweets, focusing on two mechanisms: retweets (i.e., a user re-sharing one of his friends' tweets with his own followers) and replies (i.e., a user replying to a tweet authored by one of his friends). While the Twitter API allowed users to create replies natively starting in early 2007, Twitter did not natively support creating retweets until November 2009 (Meeder et al. 2010; Parr 2009). Instead, users who wished to retweet a tweet manually copied the tweet and added a "RT @username" at the beginning to indicate a retweet. As a result, when calculating retweets, we need to look for both native retweets (*Retweets*) as well as manual retweets (*RTs*).

Using the *Crawl* and *Gardenhose* datasets, we calculate the percentage of tweets that are replies, retweets, and RTs over time in Figure 8. We observe a number of interesting trends. First, we can see a rapid adoption of the reply mechanism, peaking at almost 35% of all tweets in 2010 and declining slightly afterwards. Second, we observe that retweets are initially a small percentage of all tweets, presumably due to the manual effort required to create a retweet before November 2009. However, the percentage of native retweets increases rapidly afterwards, likely due to the native retweet support that many Twitter clients provide. In fact, in late 2013, the percentage of retweets is larger than the percentage of replies. Overall, the decline in replies indicates that there is declining person-to-person communication on Twitter, suggesting significant changes in users' tweeting behavior.

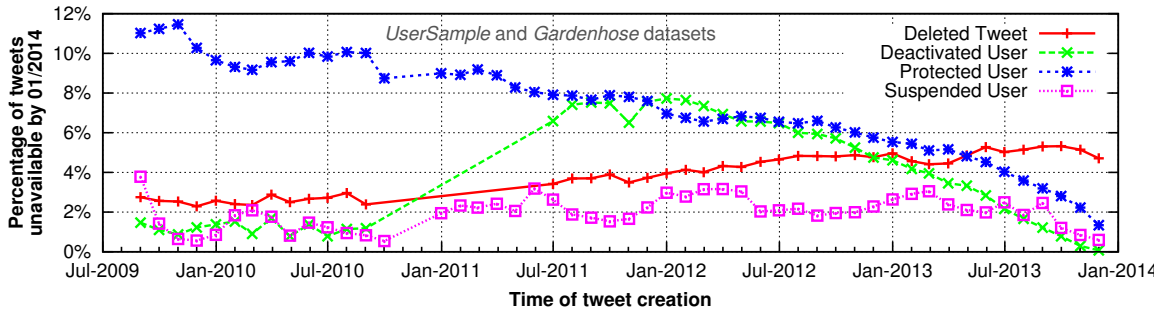


Figure 10: The percentage of tweets in the *Gardenhose* dataset that are unavailable as-of January 2014, considering only tweets issued by users in our *UserSample* dataset. Up to 10% of tweets are issued by users who later change their account to be protected, and up to 5% of tweets are explicitly deleted by users. The rise of tweets unavailable due to suspended and deleted users corresponds strongly with the increase of spam on Twitter (Thomas et al. 2011).

Unavailability of tweets Twitter's Terms of Service (Twitter 2012) requires that any data shared about tweets is shared only in the form of a `tweet_id`; the recipient then must query Twitter to obtain the actual tweet data. Presumably, this policy is in-place so that Twitter can respect users' privacy by preventing further access to tweets that a user deletes or marks as protected. This policy significantly impacts researchers, however, as researchers wishing to reproduce prior findings may not be able to obtain the entire data sets used by others.

In order to understand the impact that this policy has, we study the percentage of tweets that become unavailable over time. In general, there are four mechanisms that could lead a public tweet to later be unavailable: (1) the tweet could be explicitly deleted by the user, (2) the user could switch their account to be "protected", thereby making their tweets only available to their approved followers, (3) the user's account could be suspended by Twitter, and (4) the user could deactivate their entire account. To study how tweets become unavailable, we need to re-query Twitter for the current status of users; since we cannot query Twitter for all 388 million users, we instead perform this analysis with the *UserSample* dataset.

In Figure 10, we present the percentage of tweets created over time that are no longer available as-of January 2014 (i.e., we look for tweets in the *Gardenhose* dataset issued by users in the *UserSample* dataset). We observe that up to 20% of these users' publicly issued tweets, in aggregate, can no longer be accessed; this significantly impacts the ability for researchers to reproduce prior results. We also observe a number of other trends. First, we observe a natural decline in the percentage of tweets unavailable due to protected and deactivated users; this is expected, as more recent tweets have had less "time" for the issuing user to become protected or deactivated. Second, we observe that the percentage of tweets that are unavailable due to suspended users is fairly constant around 2%; this suggests that Twitter quickly suspends malicious users, so time passed is not significant factor in the likelihood of a tweet being unavailable due to suspension. Third, as has been show in previous work (Almuhimedi et al. 2013), most deleted tweets are deleted quickly after being issued. The increase in the percentage of deleted tweets therefore suggests a change in user behavior over time, with users becoming more likely to delete tweets.

Tweet contents Next, we take a closer look at the contents of tweets. Examining the contents of over 37 billion tweets is quite challenging, so we focus on three types of internal tweet *entities* that Twitter natively supports: tweets with URLs (i.e., a user sharing a link), tweets with hashtags such as `#fail` (i.e., a user stating the topic of the tweet), and the tweets with mentions[7] such as @BarackObama (i.e., a user mentioning another user in the tweet). Similar to the RT syntax for retweets, the syntax for specifying both hashtags and mentions was created by the Twitter users themselves, and only later natively supported by Twitter (Cooper 2013; Stone 2009).

Figure 11 (bottom) presents the percentage of tweets that have at least one of the various entity types over time. We can observe that both mentions and URLs were popular by 2009, but that hashtags only began appearing in more than 10% of tweets in 2010. Surprisingly, since 2009, the percentage of tweets with mentions has increased substantially, while the percentage of tweets with URLs has actually decreased to stabilize at ∼12%. Overall, these results suggest that Twitter has become more "conversational", with users mentioning other users in over 50% of tweets today.

We note that users can choose to include more than one of a given entity type in a tweet (e.g., a single tweet can include multiple hashtags). To understand this behavior, Figure 11 (top) plots the average number of entities in tweets that have at least one such entity (i.e., for the hashtags line, we only consider tweets with at least one hashtag). We observe that URLs and mentions have largely stabilized around 1.0 and 1.3, respectively, but that the average number of hashtags shows a continuing increase beyond 1.6. This trend is likely explained by an increasing level of hashtag spam that has

[7]Note that replies are by definition a type of mention, as the replying user includes the username they are replying to in their reply.

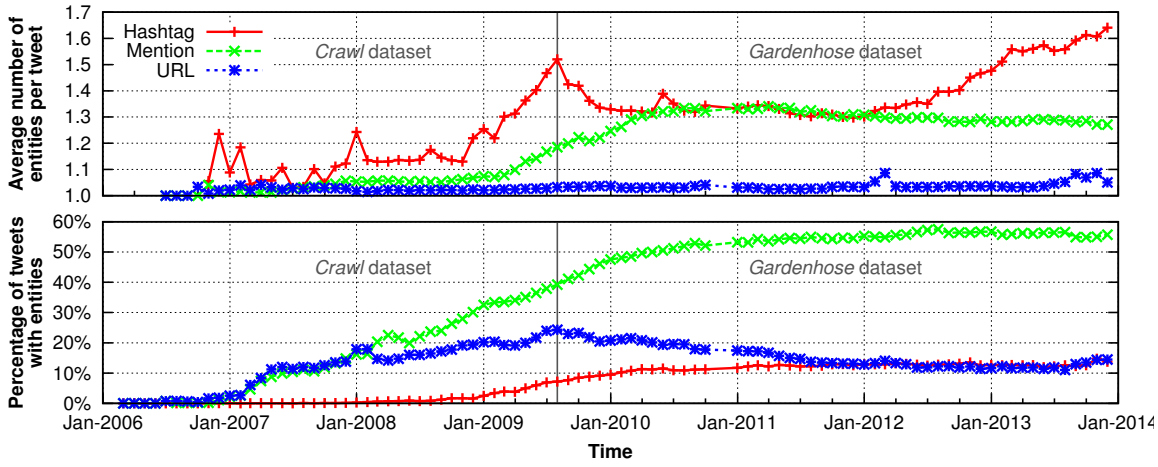


Figure 11: The percentage of tweets with different types of entities (bottom), and average number of entities for such tweets (top) over time. We observe increasing adoption of mentions, and an increasing likelihood of many hashtags per tweet.

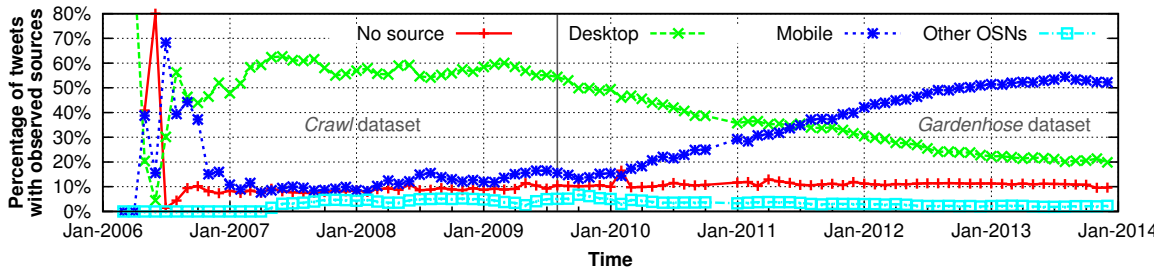


Figure 12: The percentage of tweets created with different `sources` (i.e., different clients) over time.

been observed (Vaas 2013; Ostrow 2009), where malicious users issue tweets with many hashtags in an attempt to make the hashtags appear popular.

Twitter clients Twitter was originally designed to be used on mobile devices by sending SMSes (hence the 140 character limit). From 2006 to the present, we have witnessed an explosion of popularity of smartphones and other mobile devices like iPads; using Twitter applications is now a popular activity on these devices. Additionally, over this time period, Twitter has become closely intertwined with other social networks such as Facebook, with many users automatically cross-posting their updates between multiple sites. As a final experiment, we explore how tweets are created by taking advantage of the `source` field that Twitter attaches to each tweet.

The `source` field is different for each different Twitter client, so we begin by manually classifying all 54 unique `sources` that represented at least 1% of tweets in any month. We classify the `sources` in categories: Desktop, Mobile, Other OSNs. Certain `sources` exist in multiple categories (e.g., Echofon has both mobile applications and desktop applications), so we do not include these.

We present the breakdown of tweets from different sources in Figure 12, and make a number of observations. Overall, there is a consistently decreasing trend for desktop clients (including web, API, WebClient and other applications for the desktop computers), and a corresponding increasing trend for mobile clients (including iPhone, Android, BlackBerry, iPad, and other applications for the mobile devices). We also observe that mobile devices are briefly a popular source in 2006 (using SMS), but quickly drop before rising again. These trends quantify the shift towards mobile devices, with mobile devices representing the majority of tweets starting in 2013. Surprisingly, we observe that tweets automatically created by other OSNs (including TwitterFeed, Facebook, and Tumblr) consistently represent approximately 3% of the overall tweet volume.

Related Work

In this section, we briefly detail previous studies of Twitter users and their behavior.

Sample Coverage Most studies to understand user activities on Twitter utilize the public Twitter API to collect the information of users and tweets. To measure the limitations and representation of sample data, recent work (Morstatter et al. 2013) examined how Twitter selected tweets to re-

turn, and found that many API queries are not representative. However, we use the Twitter gardenhose in our study, which Twitter explicitly states is a random selection of public tweets.

By comparing US Twitter users to the US population using census data, previous work (Mislove et al. 2011) found that US users are a highly non-uniform sample of the population in terms of race/ethnicity, gender, and geographical distribution. Our work here is complementary, as we are examining the evolution of the Twitter userbase.

Tweeting Behavior A number of researchers have examined different aspects of the user tweeting behavior. To understand why users retweet, some papers analyze the patterns and causes of retweeting, suggesting that the primary motivations for retweeting are new information for the user (Macskassy and Michelson 2011) and statements of support (Recuero, Araujo, and Zago 2011). Our results do help clarify some prior work: for example, Macskassy et al. (Macskassy and Michelson 2011) report that 32% of tweets are retweets, contradicting our measurement of 10% at the same time. The mismatch is likely caused by the authors' snowball sampling method.

Several studies have focused on the deletion of tweets. Tweets are deleted for numerous reasons; for example, they can be spam, or need to be expressed in a different way, or can cause the user regret later on (Petrovic, Osborne, and Lavrenko 2013). The most comprehensive study found (Almuhimedi et al. 2013) that the majority of tweets get deleted within an hour, and the fastest tweets to be deleted generally have typos or need rephrasing. Both studies find that around 2–3% of tweets were deleted in their 2012 dataset, which is consistent with our results (2.35%) for the same time period.

Recently, researchers have worked to gain a better understanding of how users socialize with others in the network: the conversation generated by mentioning another user (Macskassy 2012), and the top trending topics generated using hashtag (Huang, Thornton, and Efthimiadis 2010; Yang et al. 2012).

Today, people can tweet by texting, use a mobile application, use an agent application to tweet (either non mobile or mobile), among other ways. By tracking the `source` of tweets, prior work (Perreault and Ruths 2011) has shown that mobile Twitter users are more likely to be active than non-mobile users, and the tweets made on a mobile device tend to be more conversational and personal. In our paper, we find an interesting similar trending that texting (one of the few ways non-smartphone users can access Twitter) has decreased in popularity steadily since 2006; whereas smartphone devices and applications have become very popular.

User Demographics Geo-locating users has become a prominent area in the study of Twitter data. With regard to the `location` field in the user profile, prior work (Hecht et al. 2011) found that 34% of Twitter users had entered fake locations in their profile that provided no geographical information or location and 11.5% entered geotags; the rest entered valid geographical data. Our results are consistent with this study, but more comprehensive (as we shown the evolution of these trends, instead of a single snapshot).

In terms of users' self-reported `lang`, our findings supports the previous findings by Krishnamurthy (Krishnamurthy, Gill, and Arlitt 2008) about the top 10 languages on Twitter in 2008. However, we also show that this situation has changed significantly in the intervening time, with English today covering barely half of the user population.

Prediction of User Characteristics While it is largely orthogonal to our work, there has been significant work in prediction of user profile fields, including user's location, gender, race/ethnicity, and age (Mislove et al. 2011; Pennacchiotti and Popescu 2011a; 2011b; Bergsma et al. 2013; Nguyen et al. 2013). For example, recent work (Bergsma et al. 2013) found that the user-provided last name, first name, and location can provide information on the potential country of origin, the language, gender, ethnicity, and race with between 83% and 90% accuracy. Others (Nguyen et al. 2013) have shown that the specific language, sentiment, linguistic style of the tweets, as well as the tweeting behavior (retweets, hashtags) can give some idea about the age or interests of the user.

Researchers have examined how to infer the user's location from the tweet contents. For example, some studies (Hecht et al. 2011; Chandra, Khan, and Muhaya 2011; Cheng, Caverlee, and Lee 2010) showed how to predict the user's location based on the tweet contents, by looking for location-specific words in their tweets. As a result, as the amount of tweets increases, the accuracy of correctly predicting the user's location is likely to increase as well.

Concluding discussion

Twitter has seen significant interest from the research community over the past few years. However, most studies have used Twitter data as a means to an end, such as to predict aspects of the real world or to better understand information flow through society. There has been relatively little work that has studied the evolution of Twitter itself. Given that Twitter has changed significantly, it becomes unclear how to interpret prior results and whether the assumptions made in the past are still valid.

Using a dataset of over 37 billion tweets from seven years, we presented a close examination of the evolution of the Twitter user population and their behavior. We are able to strongly quantify a number of trends, including the spread of Twitter across the globe, the shift from a primarily-desktop to a primarily-mobile system, the rise of malicious behavior, and the changes in users' tweeting behavior. We hope that our findings will aid researchers in understanding the Twitter platform and interpreting prior results. As Twitter continues to develop, we plan to repeat our analysis to study the future evolution of the Twitter user population and behavior.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported in part by NSF grants

CNS-1054233 and CNS-1319019, ARO grant W911NF-12-1-0556, and an Amazon Web Services in Education Grant.

References

Acohido, B. 2010. Facebook, Twitter see more spam attacks. http://usat.ly/1gKhYiY.

Almuhimedi, H.; Wilson, S.; Liu, B.; Sadeh, N.; and Acquisti, A. 2013. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *CSCW*.

Bergsma, S.; Dredze, M.; Durme, B. V.; Wilson, T.; and Yarowsky, D. 2013. Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter. In *NAACL*.

Bryant, M. 2010. Twitter goldrush on the way. Old usernames to be released. http://tnw.co/1ozpteR.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*.

Chandra, S.; Khan, L.; and Muhaya, F. B. 2011. Estimating Twitter User Location Using Social Interactions-a Content Based Approach. In *SocialCom*.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In *CIKM*.

Chowdhury, A. 2010. State of Twitter Spam. https://blog.twitter.com/2010/state-twitter-spam.

Christoforos, A. 2011. All these celebrities on Twitter. http://exm.nr/1epC2X5.

Cooper, B. B. 2013. The surprising history of Twitter's hashtag and 4 ways to get the most out of them. http://bit.ly/1euTB8z.

Gilbert, E., and Karahalios, K. 2010. Widespread Worry and the Stock Market. In *ICWSM*.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *CHI*.

Huang, J.; Thornton, K. M.; and Efthimiadis, E. N. 2010. Conversational Tagging in Twitter. In *HT*.

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about Twitter. In *WOSN*.

Macskassy, S. A., and Michelson, M. 2011. Why do People Retweet? Anti-Homophily Wins the Day! In *ICWSM*.

Macskassy, S. A. 2012. On the Study of Social Interactions in Twitter. In *ICWSM*.

Mao, H.; Shuai, X.; and Kapadia, A. 2011. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *WPES*.

Meeder, B.; Tam, J.; Kelley, P. G.; and Cranor, L. F. 2010. RT@ IWantPrivacy: Widespread violation of privacy settings in the Twitter social network. In *W2SP*.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and Analysis of Online Social Networks. In *IMC*.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the demographics of Twitter users (Poster Paper).

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. How Old Do You Think I Am? A Study of Language and Age in Twitter. In *ICWSM*.

Ostrow, A. 2009. Twitter Spam Invades Trending Topics. http://on.mash.to/1dfFANu.

Parr, B. 2009. PROJECT RETWEET: Twitter's Plan to Officially Support Retweeting. http://on.mash.to/1fJK9tQ.

Pennacchiotti, M., and Popescu, A.-M. 2011a. A Machine Learning Approach to Twitter User Classification. In *ICWSM*.

Pennacchiotti, M., and Popescu, A.-M. 2011b. Democrats, republicans and starbucks afficionados: user classification in twitter. In *KDD*.

Perreault, M., and Ruths, D. 2011. The Effect of Mobile Platforms on Twitter Content Generation. In *ICWSM*.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2013. I Wish I Didn't Say That! Analyzing and Predicting Deleted Messages in Twitter. *CoRR*.

Post, H. 2014. Twitter Statistics. http://bit.ly/18KIwd2.

Recuero, R.; Araujo, R.; and Zago, G. 2011. How Does Social Capital Affect Retweets? In *ICWSM*.

SEC. 2013. Twitter, Inc. Common Stock. http://1.usa.gov/1gVlt2o.

Stone, B. 2009. Replies Are Now Mentions. https://blog.twitter.com/2009/replies-are-now-mentions.

Stringhini, G.; Egele, M.; Kruegel, C.; and Vigna, G. 2012. Poultry Markets: On the Underground Economy of Twitter Followers. In *WOSN*.

Thomas, K.; Grier, C.; Song, D.; and Paxson, V. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *IMC*.

Thomas, K.; McCoy, D.; Grier, C.; Kolcz, A.; and Paxson, V. 2013. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. In *USENIX Security*.

2012. Twitter Terms of Service. https://twitter.com/tos/.

Vaas, L. 2013. Spammers take over Twitter Trends with sexy hashtags. http://bit.ly/OArYAX.

Weil, K. 2010. Measuring Tweets. https://blog.twitter.com/2010/measuring-tweets.

Yang, L.; Sun, T.; Zhang, M.; and Mei, Q. 2012. We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption? In *WWW*.

Yeung, K. 2013. Spammers are back on Twitter, and this time they're coming after you through lists. http://tnw.co/1epCTqJ.

A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns

Oren Tsur
Harvard University
& Northeastern University
orentsur@seas.harvard.edu

Dan Calacci
Northeastern University
dcalacci@ccs.neu.edu

David Lazer
Northeastern University
& Harvard University
d.lazer@neu.edu

Abstract

Framing is a sophisticated form of discourse in which the speaker tries to induce a cognitive bias through consistent linkage between a topic and a specific context (frame). We build on political science and communication theory and use probabilistic topic models combined with time series regression analysis (autoregressive distributed-lag models) to gain insights about the language dynamics in the political processes. Processing four years of public statements issued by members of the U.S. Congress, our results provide a glimpse into the complex dynamic processes of framing, attention shifts and agenda setting, commonly known as 'spin'. We further provide new evidence for the divergence in party discipline in U.S. politics.

1 Introduction

Language is one of the main tools used by politicians to promote their agenda, gain popularity, win elections and drive societal change (Luntz, 2007). The growing availability of online archives of political data such as public statements, bill proposals, floor speeches, interviews or social media streams allows computational analysis of many aspects of the political process. The analysis performed can increase transparency, facilitate a better educated constituency and improve understanding of the political process.

In this paper we propose a framework for automatic analysis of a large collection of political texts. Specifically, we demonstrate how the use of Bayesian methods and time series analysis captures the different ways in which political parties control the political discourse. We show that topic ownership and framing strategies can be inferred using topic models. Moreover, we demonstrate how the models learned are used to construct time series of expressed agendas. These time series are fitted using autoregressive distributive-lag models in order to learn the partisan temporal relations between topics and expressed agendas.

This framework could also be applied in other domains such as ideology divergence in online forums of radical groups or for measuring the changes in public sentiment toward commercial brands.

Contribution (i) To the best of our knowledge this is the first work to analyze framing strategies on large scale in an unsupervised manner[1]. (ii) we combine topic models with regression analysis in recovering longitudinal trends. (iii) We further provide evidence for the dynamics of framing campaigns, commonly known as 'political spin[2]'. Finally, (iv) we show how this framework can shed new light on the broad scholarship on the divergence of party discipline.

2 Related Work

2.1 Political Communication Theory

Some of the theoretical constructs employed by Political Science scholars to describe features of the political communication mechanism include: *topic ownership*, *framing*, and *agenda setting*. Understanding these theoretical concepts is necessary in laying the ground for our computational approach. This subsection provides the key definitions and a brief survey of the relevant literature.

Topic/Issue Ownership We say that a candidate, a representative or a party *owns* a topic if this topic, set of ideas or the competence in handling specific issues are strongly associated with her/the

[1] We do use some meta data such as the speaker's party and its timestamp for the time series analysis.

[2] 'Political spin' may also refer to fact twisting and factual distractions promoted using various media outlets. We do not refer to these types of spin in this work.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1629–1638, Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

party (Petrocik, 1991; Petrocik, 1996; Damore, 2004). For example, environmental issues are traditionally associated with specific parties and not others (e.g. in U.S. politics, environmental issues are mostly associated with the Democratic party (Dunlap et al., 2001)).

Framing Framing is the psychological schema we use in order to organize and process our experiences. Politicians can use different contextual *frames* when referring to a specific topic, giving the public very different views on the topic at hand (Goffman, 1974; Gamson, 1989; Entman, 1993; Chong and Druckman, 2007). A notable example is the divisive partisan rhetoric used by U.S. politicians when referring to the legality of abortion. Democratic and Republican positions, framed as 'pro choice' and 'pro life', respectively, spin the abortion discourse as an issue of values of individual freedom (pro-choice) or validating the sanctity of life (pro-life). Similarly, Republicans refer to the inheritance tax by the overwhelmingly negative coinage 'death tax', while Democrats use 'estate tax'.

Framing strategies, however, go beyond the use of fixed phrases such as 'death tax' and 'pro-choice'. The Affordable Care Act (ACA) and the debate over raising the minimum wage can be framed as an issue of social justice or in the context of the economic burden it incurs on tax payers and by potential job loss.

Agenda Setting and shifting Agenda setting is achieved by framing and by increased or decreased *attention* (attention shifts) in order to set or change the political, media or public agenda (McCombs and Shaw, 1972; Scheufele and Tewksbury, 2007). Some examples of agenda setting campaigns are the repeated comments about the importance of child vaccination, highlighting the need for equal pay in the 2015 State of the Union Presidential Address, or, more broadly, repeatedly addressing the need for affordable healthcare.

2.2 Computational Analysis of Political Data

The availability of archives and streams of political data is driving a growing number of computational works that address a wide array of Political Science questions. Methods vary from simple word matching to more sophisticated Bayesian models and deep learning techniques.

Slant in news articles has been modeled by (Gentzkow and Shapiro, 2010) and (Lee, 2013), comparing word tokens and n-grams to predefined lists extracted from labeled data. Hidden Markov Models are used by (Sim et al., 2013) in order to measure ideological proportions in political speech, and (Iyyer et al., 2014) use recursive neural networks for a similar task.

Topic models have been used to detect connections between contributions and political agendas as expressed in microblogging platforms (Yano et al., 2013) and for reconstructing voting patterns based on the language in congressional bills (Gerrish and Blei, 2012). The flow of policy ideas has been modeled via measuring text reuse in different versions of bill proposals (Wilkerson et al., 2013).

Nguyen et al. (2013) use supervised hierarchical topic regression to improve prediction of political affiliation and sentiment.

Expressed agendas in press releases issued by U.S. Senators have been modeled by Grimmer using author topic models (Grimmer, 2010). It is important to point to some key differences between our work and Grimmer's work. While the model used by Grimmer allows attribution of a single topic per document, we are interested in a mixed membership model as we hypothesize possible correspondence between topics and frames. Moreover, while we are interested in partisan dynamics, Grimmer is interested in the expressed agendas of individuals thus focusing on an authorship model. Finally, unlike Grimmer, we also introduce autoregressive distributed-lag models in order to capture temporal dynamics between topics and parties as reflected in the data.

Another line of work can be found in the more traditional Political Science scholarship. The success of framing strategies is studied by the analysis of real time reactions to political debates (Boydstun et al., 2014). Autoregressive models are used for analyzing adjustment of issue positions with respect to news items during the Dutch national election campaign of 2006 (Kleinnijenhuis and de Nooy, 2013). This approach is based on manual annotation of data.

Logistic regression on manually coded campaign advertisements is used in order to learn the dynamics of issue ownership by individual candidates (Damore, 2004).

While some of the works above address related research questions (agenda setting, topic ownership) or use similar computational approaches (topic models, regression models), our work is the

Why Can't We Crack Down On Fraud Without A Big Government Takeover Of Health Care?
Congressman John Boehner (R-West Chester) released the following statement today on President Obama's health care speech in St. Louis, Missouri, in which he will announce a new plan to use private contractors to investigate Medicare and Medicaid fraud. While the President is visiting America's heartland today, I hope he will take the opportunity to finally listen to the American people, who are shouting, "stop" at the top of their lungs. They just don't want out-of-touch Washington Democrats' job-killing government takeover of health care. They don't want more than $500 billion in tax hikes. They don't want nearly $500 billion in Medicare cuts. They don't want these outrageous kickbacks, payoffs, and sweetheart backroom deals. We need to scrap this bill, and start with a clean piece of paper on real, step-by-step, commonsense reforms to lower health care costs. I support rooting out waste, fraud, and abuse in government programs like Medicare and Medicaid. This new proposal from the President may be worthy of bipartisan support, but why can't we crack down on fraud without a big-government takeover of health care?

Figure 1: Examples of a public statement released on March 10, 2010 by Republican minority leader – Congressman John Boehner (now speaker of the U.S. House of Representatives). The highlighted sequences illustrate the different topics/frames used - health care (green), economy/budget (yellow) and corruption (orange).

first to offer a complete framework for automatic detection of topic ownership and attention shifting on a large scale. Additionally, our partisan analysis provides a model for longitudinal partisan communication strategies without the need for encoding of external events and specific campaigns.

3 Data

A brief overview of the U.S. Congress The American political system is a bicameral legislature composed of the Senate (100 senators, two from each state) and the House of Representatives (435 voting members plus 6 non-voting representatives, number depends on the population of each state). Election is held every two years, in which one third of the Senators and all members of the House face reelection. Members are typically affiliated with either the Democratic Party or the Republican Party. Congressional election and Presidential election coincide every four years.

The Corpus We use a corpus of public statements released by members of Congress in both the Senate and The House of Representatives, collected by Project Vote Smart[3]. An example of a public statement is presented in Figure 1.

In this work we use all individual statements and press releases in a span of four years (2010-2013), a total of 134000 statements made by 641 representatives. This time span encompasses two Congressional elections (November 2010 and 2012). Table 1 gives the number of Democratic and Republican representatives in the three Congress terms (111-113) covered in our data.

[3]http://votesmart.org/

Chamber	Party	Congress Term		
		111th	112th	113th
Senate	DEM	57	51	53
	REP	41	47	45
House	DEM	257	193	199
	REP	178	242	234

Table 1: Majority shifts in the House in the 111-113 Congress terms. Independent representatives are omitted.

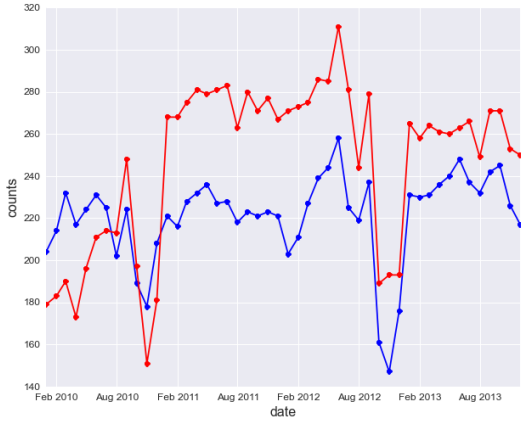


Figure 2: Monthly average number of statements by party.

While the administration was Democratic during all four years of our data, notice the Democratic loss of majority in the 112th Congress. We focus on the years 2010-2013 since Project Vote Smart has better coverage of the political discourse after 2009.

It is interesting to note that while the total number of statements per month reflects the change of majority in the November 2010 and 2012 elections (Table 1 and Figure 2), accounting for the number of seats per party it appears that the average Democrat is consistently more 'productive' with $\mu = 6.24$, $\sigma^2 = 2.1$ (Dem) and $\mu = 5.5$,

$\sigma^2 = 1.5$ (Rep) statements per month. We hence report all results after normalization by the number of seats each party posses at each timestamp.

4 Computational Framework

In order to automatically discover the correlated dynamics of attention shifts, we take a layered approach, consisting of the stages described below.

4.1 Topic Inference

In the first stage, we use topic models in order to learn topic distribution over words and identify the set of topics addressed in the corpus. Topic Modeling describes a general algorithmic framework for unsupervised discovery of a set of topics expressed in a collection of documents. The framework is based on the assumption that documents are generated by mixtures of k topics. It is therefore assumed that documents are generated by the following process: for each word in a document, we choose a topic from a given topic distribution, then choose a word from the distribution over words that the chosen topic specifies. Latent Dirichlet Allocation (LDA), the framework employed here, assumes that the distribution over topics has a Dirichlet prior. In practice, we assume a Dirichlet prior on topics and use variational Bayes (VB) optimization to infer topic distributions over words (Blei et al., 2003). In order to considerably improve efficiency, we use an online variational Bayes inference algorithm, shown to perform similarly to batch LDA (Hoffman et al., 2010). It is important to note that our goals and assumptions about the data do not lend themselvse to the use of dynamic or correlated topic models (Blei and Lafferty, 2006a; Blei and Lafferty, 2006b)[4].

4.2 Topic Assignment and Unification

The distribution of ranked topics over documents presents a "long tailed" distribution in which a few topics achieve a significant coverage of a document. This is a result of the mixed membership "generative" approach and the bag-of-words assumption. In a more realistic setting the number of topics per document is restricted. We wish to restrict the number of topics per document while still conforming to the mixture model assumption. We therefore reassign topics to each document (statement) d in the following manner:

1. Assign a topic to each word based on distribution of topics over words infferred in the previous stage.
2. Find a set T' of k' topics ($k' < k$) that cover $q\%$ of the document in a greedy way. The topic assingment for document d will then be $d \rightarrow T'$.

4.3 Data Slicing

We slice the data according to four parameters: topic (or topical cluster), time, party and document (statement). These slicing parameters allow us the flexibility required to thoroughly analyze the data. In the time parameter, we have four settings: no slicing (all data is treated as if it were produced simultaneously), monthly slicing, weekly slicing and daily slicing, each gives different granularity of ownership patterns.

4.4 Autoregressive-Distributed-Lag Models

A linear function $b + w^T X = b + \sum_j w_j^T X^j$ is a simple yet robust method for testing dependency between X and Y. Ordinary least square regression finds the coefficients that minimize the mean square error of $Y = b + \sum_j w_j^T X^j$ given (X, Y). In our case (X, Y) are time series. We argue that a lagged dependency between two time series suggests a framing or attention shifting campaign.

Regression analysis of time series assumes independence between error terms. This key statistical property is often violated in real world data as y_t often depends on y_{t-1} thus the time series residuals tend to correlate. The consequences of violating the independence of errors are threefold: i) Statistical tests of significance are uninformative and cannot be used to prove dependency between the model parameters, ii) The coefficients learned lose accuracy, and iii) error terms are correlated, and hence contain information that is lost in analysis instead of used to leverage the prediction power of the model. The importance of controlling for autoregressive properties and for seasonality effects was recently demonstrated in the error analysis of the Google Flu Trends algorithm (Lazer et al., 2014).

In order to control for error dependency we add the auto regressing component $\gamma^T Y^n$ to the ordinary regression, as shown in Equation 1:

$$y_t = \alpha + \beta^T X^m + \gamma^T Y^n + \epsilon_t \quad (1)$$

[4] We are interested in the change of the proportions of topics over time and not in the change of the word distribution within topics and we don't assume inherent correlation of topics.

Cluster	Topic ID	Top Words
Health	30	health care law will obamacare insurance repeal affordable americans costs new reform people president healthcare act coverage mandate american obama
	51	medicare seniors program social medicaid benefits fraud payments security programs cost services costs billion payment beneficiaries waste year savings million
Energy	38	project pipeline president obama keystone jobs climate energy xl construction state change permit administration approval oil will canada environmental create
	69	oil alaska gulf coast spill drilling offshore bp murkowski begich markey resources noaa said industry moratorium mexico gas administration sen
Security	34	day nation country today americans us american war world people america lives will honor families years men many th attacks
	89	nuclear united iran international israel foreign president states security weapons people world syria nations sanctions regime must government peace
Economy	68	budget spending debt president cuts fiscal government deficit will plan trillion obama house congress year federal cut economy washington billion
	88	jobs small businesses business job economy economic create will new growth work american america help creation act manufacturing can sector

Table 2: Top twenty words in selected topics in four topical clusters.

where, $\beta^T X^m$ indicates the distributed-lag terms:

$$\beta^T X^m = \sum_{i=0}^{m} \beta_i x_{t-i} \quad (2)$$

and $\gamma^T Y^n$ indicates the autoregressive component described by:

$$\gamma^T Y^n = \sum_{j=1}^{n} \gamma_j y_{t-j} \quad (3)$$

for some $n \leqslant t$ (notice that i ranges from 0 while j ranges from 1).

In order to control for seasonality (such as holidays and recess') we add a set of categorical variables indicating the weekday and the week-in-year of a statement, so the autoregressive model is:

$$y_t = \alpha + \beta^T X^m + \gamma^T Y^n + \sum_l W_l^T I^l(t) + \epsilon_t \quad (4)$$

Where $l \in \{day, week\}$ thus $I^l(t)$ is the identity matrix with the dimension of the seasonal granularity, in our case $I^{day} = I_{7\times7}$ for each day of the week and $I^{week} = I_{52\times52}$ for the week of the year. $I^l_{i,i} = 1$ iff t timestamp falls in the i-th day-of-week/week-in-year.

Finally, in practice it is usually sufficient restrict the autoregressive term to one parameter with $j = 1$ (accounting to the y value at the previous time stamp), this is consistent with the 24 hours news cycle reported by (Leskovec et al., 2009) among others. Since our goal is to find correlated attention shifts we can substitute the summation distributed-lag term by a single lagged term. Thus, we aim to minimize the MSE in the following model:

$$y_t = \alpha + \beta x_{t'} + \gamma y^n_{t-1} + \sum_l W_l^T I^l(t) + \epsilon_t \quad (5)$$

Where $t' = t-i$ and $i \in \{0, 1, 2, ..., 28\}$ indicating no lag, one day lag, 2 days lag, a week's lag, etc.

5 Results

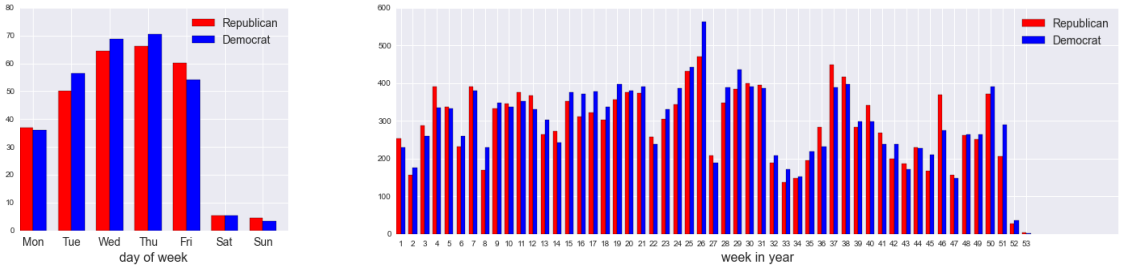
5.1 Topical Ownership and Framing

5.1.1 Inferred Topics

As an input for the topic modeling module (stage 1 of the system) we use a lexicon of the 10000 most frequent words in the corpus. We use $k = 100$ as the number of topics. Experiments with $k \in \{30, 50, 500, 1000\}$ produced topics that were either too general or too incoherent. Once the topic-word distributions were inferred, topics were validated by two annotators examining the top 50 words for each topic. Annotators used hierarchical labels – an energy related topic t_i could be annotated *energy / clean-tech*, while another topic t_j could be annotated *energy / economy / keystone-xl*. Annotations were consolidated to unify the coding[5]. After consolidation annotators agreed on all topic annotations. Some examples of topics labels are 'health', 'energy', 'economy', 'boiler-plate', 'political process', 'local' and a few 'random' topics.

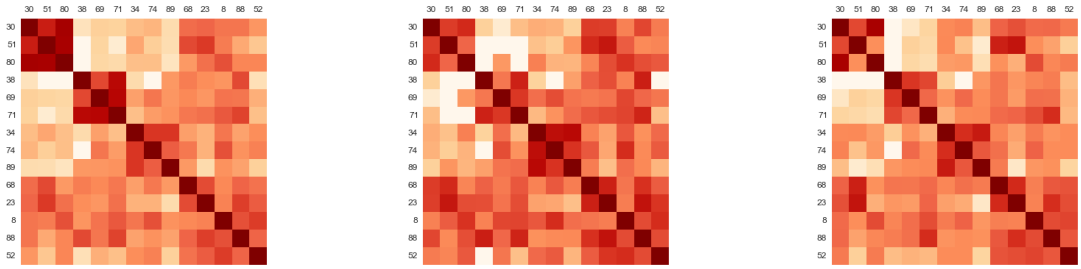
After topic assignment as described in Section 4.2 each document is associated with only 2–6 topics. In this work we focus on the 14 most salient (concise, general and frequent) topics in the corpus. These 14 topics fall under four topical clusters - Health, Energy, Army/Security and Economy/Budget. Table 2 contains examples of top words and labels for some of the topics from four topical clusters.

[5]For example, if topic t_i was labeled *energy, cleantech* by one annotator and *energy, green* by the other, the annotators would agree to use either *cleantech* or *green* consistently.



(a) (b)

Figure 3: Seasonality effect: average number of statements issued per day of week (a) and per week in year (b).



(a) All statements (b) Republican statements (c) Democrat statements

Figure 4: Normalized Pointwise Mutual Information (PMI) of topic cooccurrence of 14 topics of four topical clusters Health (30, 51, 80), Energy (38, 69,71), securtity (34, 74, 89) and Budget & Economy (68, 23, 8, 88, 52)

Cluster	Topic	DEM	REP	DEM	REP
Health	30	1679	4622	3169	5386
	51	746	233		
	80	898	437		
Energy	38	128	255	4042	3415
	69	1102	948		
	71	2859	2119		
Security	34	6239	5121	12393	11140
	74	3875	3071		
	89	3807	4138		
Economy	68	12260	19916	31604	31706
	23	5221	3742		
	8	6981	2456		
	88	12845	11139		
	52	3479	1154		

Table 3: Total number of statements by party in four topical clusters. DEM indicates the Democrat party, REP indicates the Republican party.

Cluster	Topic	DEM	REP	DEM*	REP*
Health	30	46	154	2	79
	51	151	22	34	1
	80	157	27	37	2
Energy	38	47	43	2	6
	69	114	56	18	5
	71	144	52	43	5
Security	34	141	63	52	9
	74	144	46	34	8
	89	80	113	10	22
Economy	68	32	174	7	127
	23	151	49	60	9
	8	205	2	165	0
	88	137	68	63	17
	52	190	12	123	0

Table 4: Number of *weeks* each party "owned" a topic by issuing more statements (DEM, REP) and number of weeks the party owned the topic with statistical significance $p < 0.05$ (DEM*, REP*).

5.1.2 Partisan Topic Ownership

Table 3 shows the partisan ownership by providing the number of statements issued by each party on each topic and for topical clusters. It also illustrates that different topical granularities portray different ownership patterns. For example, while it seems like the health cluster is owned by the Republican party (Table 3, cluster level), a closer look at specific topics in the cluster reveals a more complex picture – the Republicans actually own only topic 30, which turns to be the most dominant topic in the cluster. Similarly, while the statement counts in the Economy cluster are quite balanced (31604 vs. 31706), the counts of the individual topics in the cluster are polarized. Remember that these topical classes were all inferred by the LDA in an unsupervised way. These partisan ownership patterns were also confirmed by domain experts.

Longevity is a crucial factor in topic ownership. A weekly ownership of a topic is achieved by a party Q if it issued more statements on the topic than party R in that particular week. We compute the significance of the ownership assuming a null hypothesis that statements are issued by the parties

by two Bernoulli processes with the same parameters. Table 4 provides the number of weeks each party owned each topic and the number of weeks it had a significant ownership ($p < 0.05$)[6].

Topic 30 illustrates the different perspectives. The total statement count (see Table 3) reveals a clear ownership by the Republican party, issuing 73% of the statements. While turning to weekly ownership (Table 4) we get similar number (Republicans control 77% of the weeks); assuming only significance ownership, Republicans significantly own the discourse for 79 weeks while Democrats have significant ownership in only 2 weeks which means the Republicans own 97% of the significantly owned weeks.

5.1.3 Topic Cooccurrence

Topic cooccurrence could approximate the way topics are framed. A heatmap of *within* statement topic cooccurrence based on Pointwise Mutual Information (PMI) (Church and Hanks, 1990) is presented in Figure 4. The topical clusters are characterized by blocks along the diagonal. The blocks structure is to be expected due to the inherent topical similarity within clusters. It is interesting to see the inter-cluster differences in PMI between the two parties. At the cluster level, Republicans tend to use the Budget & Economy topics with topics in all other topical clusters, evident by the stronger colors in the five bottom (left) rows (columns) in 4b comparing to 4c.

A notable example is the way Republicans frame the controversial Keystone XL project (Energy, topic 38) with the impact on the job market and small businesses (Budget & Economy, topic 88), a topic traditionally owned by Democrats (see top topical words in Table 2 and topic ownership at Table 4).

5.2 Partisan Discipline

Party discipline is of great interest for political scientists (Crook and Hibbing, 1985; Bowler et al., 1999; Krehbiel, 2000; McCarty, 2001) . Typically, party discipline is examined by analysis of roll call votes on bills. Discipline, however can be also measured by adherence to party lines in talking points and agenda setting campaigns. Discipline, therefore, can be captured by conformity of language in public statements. While it is "common knowledge" among political scientists that Republicans are more adherent to "talking points" – to the best of our knowledge there are no large scale studies that support (or refute) that.

[6]The numbers do not necessarily add up to 208 (the number of weeks in four years) due to weeks with no significant ownership , e.g. the parties issued a similar number of statements (usually zero) on that topic.

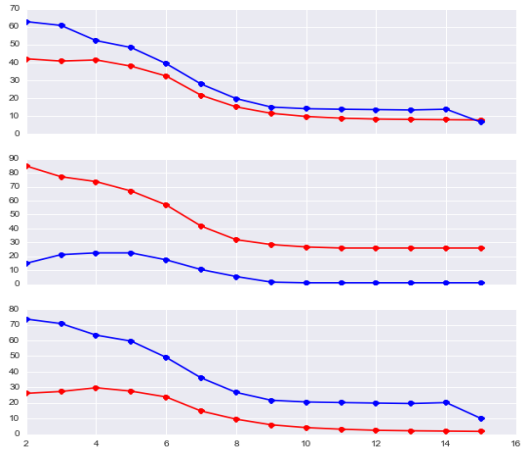


Figure 5: Average number of n-grams owned by each party on all topics (top), in Republican owned topics (middle) and in Democrat owned topics (bottom).

In the absence of official lists of "talking points", repeated use of similar phrases (n-grams) can provide an indication for the level of party discipline. In each topic, we looked at all n-grams ($n \in \{2, ..., 14\}$) that were used by more than five members of the Congress. For example, the tri-gram "the American people" (topic 38) appears in 81 statements made by 54 members of congress, only two of them were Democrats. Similarly, the tri-gram "social security benefits" (topic 51) appears in 123 statements, issued by 89 members, 71 of which were Democrats. Examining "ownership" of n-grams (per n-gram, per topic) reveals that that Republicans do tend to stick to talking points more than Democrats do.

Figure 5 provides the average number of n-grams owned by each party over all topics (top), over Republican owned topics (middle) and over Democratic owned topics (bottom). While on average Democrats own more n-grams than Republicans (Figure 5, top), the difference is marginal and is attributed to the fact that Democrats own more topics than the Republicans (10 vs. 4, see Table 3). Comparison between n-gram ownership in Democratic owned topics and Republican owned topics (Figure 5, middle and bottom) shows that while each party owns more n-grams in the *topics it owns*, Republicans present stronger ownership over the n-grams in their owned topics than Democrats in their respective owned topics. Moreover, Republicans present relative discipline even

in Democratic owned topics.

Manually looking at some sampled n-grams it appears that mid-length n-grams are shared "talking points" and longer n-grams are full citations from bill proposals and committee reports. These findings are in line with textual sharing semantics (Lin et al., 2015).

5.3 Time Series Analysis

To this end we create two time series for each topic $c \in T$: $S_c^{C_D}$ – daily normalized counts for Democrats and $S_c^{C_R}$ – daily normalized counts for Republicans. Normalization of counts is needed in order to account for the bias introduced by the difference in the number of seats each party holds and the changes in that number in the different terms as apparent from Table 1.

Our data exhibit two strong seasonality effects: a weekly cycle with the lowest point on the weekend and peaking on Thursday (Figure 3a), and a yearly cycle with low points at the weeks of 4th of July, Thanksgiving, August recess and Christmas (Figure 3b). These seasonality effects are captured by the added terms in Equation 4.

After time series are constructed we apply first-difference detrending (Enders, 2008) in order to transform the time series to stationary series and avoid trend-incurred correlations.

We fit autoregressive-distributed-lag models for all pairs in $\{X = S_{c,l}, Y = S_{c'}\}$, where $c, c' \in T$ (topics), $l \in \{0, 1, 2, 3, ..., 7, 14, ..., 28\}$.

In this setting we fit 5153 pairs of time series of which 718 pairs had a significant coefficient for X ($p < 0.05$). Artificial significance due to abundance of fitted models was accounted to by applying the strict Bonferroni correction (Dunn, 1961) on the significance level. The correction resulted in 103 significant correlations, most of them with lag of up to 3 days. Table 5 gives the number of intra/inter significant correlation for lags $l \in 0, 1, 2$.

One example for such correlation is the Republicans "responding" to Democratic topic 88 with with topic 8 (intra-cluster) in one and two days lag. We interpret this as a different spin on the budget issue. Another example is the Democratic party corresponds to Republican topic 30 with topic 88 (inter-cluster) on the same day (no lag). We interpret this as a way to place the Acordable Care Act in a specific economic frame. We note that while the lagged correlated time series do not imply a responsive pattern, a significance of lagged correlation may suggest such a pattern. We provide some evidence in the qualitative analysis in the next section.

Cluster	Dependent	Significant Correlations		
		$l = 0$	$l = 1$	$l = 2$
Intra-cluster	DEM	28	1	1
	REP	26	4	5
Inter-cluster	DEM	15	2	0
	REP	17	0	0

Table 5: Number of statistically significant ($p < .05$, Bonferroni corrected) daily lagged correlations between cross-party time series.

5.4 Discussion and Qualitative Analysis

Inter and intra-cluster correlations can be interpreted as manifestations of different types of framing strategies and campaigns for attention shifts. A detailed analysis of the interplay between the different frames is beyond the scope of this paper and is left for political scientists.

The majority of the significant correlations were found with no lag. It is important to note that these correlations are found significant even after accounting to autoregressive patterns. Zero-lag correlations could be interpreted in a number of ways. Two probable interpretations are (i) daily time series are too crude to model lag patterns, and (ii) the parties respond to some external event at the same time. While we cannot address (i) due to sparseness and noise[7], we can sample statements and examine them manually. Manual examination reveals a strong responsive pattern in peaking trends. One typical example is the Republican spike in topic 30 on March 10. The statement at Figure 1 is very illustrative as it *explicitly* refers to a statement by President Obama. Explicit references to statements made by the other side are found more frequently in Republican statements and reveal a clear responsive pattern that also suggest a strong party discipline, in line with the results in Section 5.2. This small scale qualitative analysis complements the quantitative results reported in Section 5.3 and provide evidence for a responsive pattern even in zero lag series.

6 Conclusion

We presented a statistical framework for the analysis of framing strategies and agenda setting campaigns in the political sphere. Combining topic models and time series analysis, we modeled topic

[7]The exact timestamp is sometimes missing, set to midnight or affected by external factors.

ownership and party discipline and analyzed responsive patterns in an unsupervised way and with no prior knowledge of the political system. Our work draws from political science theory, validating some theoretical constructs and shedding new light on others. The proposed framework and the results could be further used and interpreted by political scientists and communication scholars.

7 Acknowledgments

We thank Ryan Kennedy, Navid Dianati, Katherine Ognyanova, Stefan Vojcik and Shahar Even-Dar Mandel for fruitful discussions. We thank the anonymous reviewers for their helpful comments. This work was supported by the following grants: MURI #504026 and ARO #50433.

References

David Blei and John Lafferty. 2006a. Correlated topic models. *Advances in neural information processing systems*, 18:147.

David M Blei and John D Lafferty. 2006b. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Shaun Bowler, David M Farrell, and Richard S Katz. 1999. Party cohesion, party discipline, and parliaments. *Party discipline and parliamentary government*, pages 3–22.

Amber E Boydstun, Rebecca A Glazier, Matthew T Pietryka, and Philip Resnik. 2014. Real-time reactions to a 2012 presidential debate a method for understanding which messages matter. *Public Opinion Quarterly*, 78(S1):330–343.

Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10:103–126.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Sara Brandes Crook and John R Hibbing. 1985. Congressional reform and party discipline: The effects of changes in the seniority system on party loyalty in the us house of representatives. *British Journal of Political Science*, 15(02):207–226.

David F Damore. 2004. The dynamics of issue ownership in presidential campaigns. *Political Research Quarterly*, 57(3):391–397.

Riley E Dunlap, Chenyang Xiao, and Aaron M McCright. 2001. Politics and environment in america: Partisan and ideological cleavages in public support for environmentalism. *Environmental politics*, 10(4):23–48.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Walter Enders. 2008. *Applied econometric time series*. John Wiley & Sons.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

William A Gamson. 1989. News as framing. *American Behavioral Scientist*, 33(2):157–161.

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78:35–71.

Sean Gerrish and David M. Blei. 2012. How they vote: Issue-adjusted models of legislative behavior. In *Neural Information Processing Systems (NIPS)*, pages 2762–2770.

Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience*. Harvard University Press.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd meeting of the Association for Computational Linguistics*.

Jan Kleinnijenhuis and Wouter de Nooy. 2013. Adjustment of issue positions based on network strategies in an election campaign: A two-mode network autoregression model with cross-nested random effects. *Social Networks*, 35(2):168–177.

Keith Krehbiel. 2000. Party discipline and measures of partisanship. *American Journal of Political Science*, pages 212–227.

David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of google flu: Traps in big data analysis. *Science Magazine (AAAS)*.

Han Soo Lee. 2013. Do national economic and political conditions affect ideological media slant? *Political Communication*, 30:395–418.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM.

Y. Lin, D. Margolin, and D Lazer. 2015. Uncovering social semantics from textual traces: A theory-driven approach and evidence from public statements of u.s. members of congress. *Journal of the Association for Information Science and Technology*, page (forthcoming).

Frank I. Luntz. 2007. *Words That Work: It's Not What You Say, It's What People Hear*. Hyperion.

Nolan McCarty. 2001. The hunt for party discipline in congress. In *American Political Science Association*, volume 95, pages 673–687. Cambridge Univ Press.

Maxwell E McCombs and Donald L Shaw. 1972. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Neural Information Processing Systems*.

John R Petrocik. 1991. Divided government: Is it all in the campaigns? *The politics of divided government*, pages 13–38.

John R Petrocik. 1996. Issue ownership in presidential elections, with a 1980 case study. *American journal of political science*, pages 825–850.

Dietram A Scheufele and David Tewksbury. 2007. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, 57(1):9–20.

Yanchuan Sim, Brice Acree, Justin H Gross, and Noah A Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of EMNLP*.

John Wilkerson, David A. Smith, and Nick Stramp. 2013. Tracing the flow of policy ideas in legislatures: A text reuse approach. In *New Directions in Analyzing Text as Data*, September.

Tae Yano, Dani Yogotama, and Noah A. Smith. 2013. A penny for your tweets: Campaign contributions and capitol hill microblogs. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, July.

Beyond CPM and CPC: Determining the Value of Users on OSNs

Diego Saez-Trumper
Yahoo Labs
Barcelona, Spain
dsaez-trumper@acm.org

Yabing Liu
Northeastern University
Boston, USA
ybliu@ccs.neu.edu

Ricardo Baeza-Yates
Yahoo Labs
Barcelona, Spain
rbaeza@acm.org

Balachander Krishnamurthy
AT&T Labs–Research
Florham Park, USA
bala@research.att.com

Alan Mislove
Northeastern University
Boston, USA
amislove@ccs.neu.edu

ABSTRACT

Not all of the over one billion users of online social networks (OSNs) are equally valuable to the OSNs. The current business model of monetizing advertisements targeted to users does not appear to be based on any visible grouping of the users. The primary metrics remain CPM (cost per mille—i.e., thousand impressions) and CPC (cost per click) of ads that are shown to users. However, there is significant diversity in the actions of users—some users upload interesting content triggering additional views and comments leading to further cascades of action. Beyond direct impressions, a user's action can generate indirect impressions by actions induced on friends and other users. Identifying the valuable user segments requires examination of profile data, friendships, and most importantly, their activity. Here we explore an alternate approach for measuring the value of users in OSNs by proposing a framework from the viewpoint of a popular OSN. Using a real dataset on the social network and activities of users, we show that a small subset of actions are likely to be key indicators of a user's value. Additionally, by examining the current targeting demographics available in Facebook, we are able to explore the relative (monetary) value that different users represent to the OSN.

Categories and Subject Descriptors

I.6 [**Simulation and Modeling**]: Applications; J.4 [**Social and Behavioral Sciences**]: Sociology; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services.*

General Terms

Theory, Measurement, Economics.

Keywords

Online advertising; Online social networks; User value

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
COSN'14, October 1–2, 2014, Dublin, Ireland.
Copyright 2014 ACM 978-1-4503-3198-2/14/10 ...$15.00.
http://dx.doi.org/10.1145/2660460.2660477.

1. INTRODUCTION

Online advertising underlies much of the economic basis of the Web today; many sites provide free services supported by advertising. Google—far and away the largest advertising network on the Web—reported over $50 billion in 2013 [15] in advertising revenue alone. Typically, advertisers place ads on Google by specifying keywords of interest; different keywords have different values with their market price determined via a dynamic auction. The cost to advertisers when using networks like Google is generally expressed in terms of CPM (Cost per Mille, or the cost of 1,000 ad impressions) or CPC (Cost per Click, or the cost to receive a single ad click independent of the number of impressions).

Recently, online social networks (OSNs) such as Facebook have also seen an advertising market develop. For example, Facebook's most recently quarterly report indicated advertising revenue of over $7.8 billion in 2013 [12]. Advertising on OSNs works in a manner similar to advertising on Google: advertisers specify targeting parameters (i.e., attributes that the advertisers desire users to have) and a CPM/CPC bid price, and the OSN ranks the ads to select the ones to be shown. The ranking is typically based on the bids and the click-through-rate (CTR) of the ad.

The strong similarity between advertising on the Web and OSNs is surprising given that OSNs are significantly different from a typical website or a search engine answer page. On Web-search-based ads such as Google, the ad network, by default, knows relatively little about the user. Instead, the network must track users using cookies and other techniques, extracting more information about users through data mining. For example, Google provides a significant number of services (e.g., email, calendar, etc.) to users, presumably to be able to gather additional information. However, in the case of OSN-based ads, users must have an account and be logged in, in order to even see ads. As part of participating in the OSN, users provide information about themselves in profiles (interests, identities of friends, demographics, educational history, etc.) and through interactions with the site (posting updates, "checking in", installing applications, etc.). Moreover, because the OSN is run by a single centralized entity, the OSN observes all user actions on the site.

We posit that users on OSNs have sharply different values—in terms of the revenue they generate through ad impressions—to both the OSN itself and advertisers. For

example, influential users often have many friends, post significant amounts of content, and have their posts forwarded to many others; these users are likely to be more valuable than the average user. Such users are likely to bring more value to the OSN itself (as they afford more advertising opportunities) and to advertisers as well (as their activities offer more opportunities for the advertiser's message to be spread). However, little work has gone into studying how the value of users in OSNs varies, and determining the extent to which users' value contributions can be extracted, quantified, and presented to advertisers. Moreover, no one has come up with a framework for assigning different values to individual users; the focus is typically on aggregate user values.

We explore an alternate approach for measuring the value of users in OSNs such as Facebook via a framework for estimating their relative value. We do so from the perspective of the OSN operator, who both runs the advertising network and the OSN site. However, our analysis is constrained by the (very) limited visibility we have into the OSN's revenue (i.e., we are typically unable to know how many ad impressions each user receives, or how often they click on ads). Thus, we rely on the data available to us to parametrize our model, with the knowledge that the OSN operator likely has significantly better data available; this more accurate data could be used to further refine our model and the resulting predictions. It is worth noting that our goal is not to validate a specific model, but rather to propose a potential model and explore its feasibility given the limited data available to external observers.

We examine the wealth of information that the OSN operator receives about user activity on their site and present a methodology for reasoning about how different user actions correspond to revenue. We argue that a user's value can be divided into direct impressions (advertising opportunities that a user provides by browsing OSN site pages) and indirect impressions (advertising opportunities that a user provides by enticing others to browse OSN site pages). Indirect impressions can *cascade* via the network, where a user's actions ultimately cause other users to visit the OSN. For example, after one user uploads a photo, this action may cause her friends to log in to the OSN and view it (potentially causing more users to log in, etc.).

Having a proper understanding of the economic value of different users benefits both the OSN and advertisers. OSNs can enable targeting of such "more valuable" users, increasing revenues and making their advertising platform more attractive to advertisers. Additionally, uncovering the value of users will also benefit the users themselves, as they become aware of their relative value. In other types of networks, understanding which nodes are most central has proven quite valuable. For example, on the AS-level Internet topology, the peerings and traffic details of Internet Exchange Points (IXPs) were not widely known until recently [3]. It was revealed that each of largest European IXPs handles on a daily basis as much traffic as some of the global Tier-1 ISPs and supports a peering fabric that consists of more peering links than were previously believed to exist Internet-wide [32]. Locating high value users may likewise uncover a different microcosm in the OSN.

We explore our framework by leveraging a detailed data set from Facebook covering 90,269 users in the New Orleans metropolitan area [29]. We estimate the number of impressions attributable to users via activity, and show that users from our data set are likely to have sharply different values. We also show that our model can be extended to represent the user value in monetary terms beyond just the number of advertising impressions. We collect data from the Facebook Ads platform for each of the New Orleans users who have filled in basic demographic information, and use this as a basis for estimating revenue per user.

2. BACKGROUND

We now provide background on OSN advertising and cover work related to this study.

Facebook's Advertising Model. To study current advertising models in OSNs, we focus on Facebook, as it is the largest and most mature OSN. Facebook offers targeting parameters—such as location, gender, interests—and advertisers pay either per click or per impression for users who match the advertiser's specified targeting parameters. Although the parameters are quite detailed in terms of different demographics and interests, they currently do not relate directly to the target user's popularity or level of activity within the OSN (e.g., there is generally no mechanism for directly targeting "users who are influential").

The CPM and CPC ad prices are set through an auction, where each advertiser bids the maximum that she is willing to pay for impressions or clicks. Facebook selects the "best" ads to present to the user; while the ad selection algorithms are secret, the OSNs presumably use the highest bids in the case of CPM, or the highest expected revenues in the case of CPC (similar to the popular approach used in sponsored search advertisement auctions [17]). Thus, targeting parameters that are popular with advertisers are expected to have higher winning auction prices.

Information Diffusion. Much work has studied "influential" users in social networks; but deciding where to "seed" ads to reach the biggest possible audience remains a challenge. Social contagion has been studied from different angles: finding the set of users that maximize the probability of spreading [11], discovering topical authorities [31] or identifying trendsetters [26]. Other roles in the diffusion process include promoters [7], early adopters and imitators [5].

Beyond social contagion, cascading behavior is common in OSNs [9], and although the user's popularity does not necessarily create cascades [10], being popular is essential for direct influence [4] (popular users broadcast to a broader audience). Moreover, if we consider that cascades tend to be wider than deeper [25], the size of a user's audience (i.e., node in-degree) is key to estimating their value.

Auction mechanisms. Online advertising has been extensively studied due to its popularity and data availability [16, 30]. Researchers have studied the properties of advertisers [19, 23], and techniques for online advertisers to maximize their revenue [24]. Barford et al. [6] studied the features, mechanisms and dynamics of display advertising on the Web, demonstrating that a user's profile (i.e., browser and cookies) can have a significant impact on which ads are shown. Moreover, they demonstrated that the specific types of ads delivered generally correspond with the details of user profiles. Our work is complementary to these, as we focus on differentiating between more- and less-valuable users, rather than ad auctions or targeting. Beyond CPC

and CPM [21, 22], there are other proposed pricing models [13, 14] that combine both (called impression-plus-click pricing). Other work has explored predicting the number of clicks—effectively, the CTR—for new ads [2, 18].

3. USER VALUE FRAMEWORK

Each page on the OSN that the user visits gives an opportunity to the OSN provider to show advertisements. As discussed above, the value of a user in an OSN is directly proportional to the number of advertising impressions and clicks that the user generates by their actions. The actions may be visible (such as uploading content or commenting on a friend's content) or invisible (such as visiting a friend's profile or browsing a friend's photos without commenting).

We call the advertisements shown directly to the user *direct impressions*. Thus, users who browse the OSN frequently are likely to generate many direct impressions. When a user perform actions that have effects visible to others in the OSN, the user has the potential to also generate *indirect impressions*. For example, a user commenting on a friend's photo may trigger other users to return to the OSN (thereby generating additional, indirect impressions). Thus, when more people browse the OSN as a result of one user's action, more impressions can be attributed to the action.

We argue that different user actions are likely to generate different numbers of impressions. The "place" where the action has been done (e.g., in the user's profile, friends' profiles, or on group/community pages) can result in generating different numbers of impressions.

Below, we propose a potential framework that considers all these factors and uses them to compute a user's value (in terms of the advertising revenue of the OSN that can be attributed to the user). First, we analyze why different actions produce different numbers of indirect impressions and how external observers or the OSN provider can measure this. Next, we show how users' characteristics and the places where they perform their actions affect their value. Finally, we propose a comprehensive methodology for computing users' values that can be applied to many OSNs.

3.1 The Value of Actions

Consider an OSN where the most common action is to browse photos of friends, but articles posted by friends are rarely read. In such an OSN, when a user uploads a photo, the user is generating more indirect impressions than by posting an article. The value of an action is thus related to the new actions triggered.

The primary challenge in measuring the value of actions is that many of the impressions cannot be directly observed; only the OSN knows when they occur and few OSNs provide visibility into how often other users browse content. An external observer, however, can estimate invisible actions, for example, by considering visible actions as a proxy for invisible actions [33]. Another option is to extrapolate this information from previous studies that have access to (private) invisible actions and show that most user activities on OSNs consist of visiting friend profiles and photos. We take this approach and use two studies [8,28] that examined user's actions on popular OSNs. Each of these studies relied on clickstream data (e.g., records of requests to the OSN) in order to study how users spend their time in OSNs. Table 1 shows activity distribution of a large number of users in three different OSNs in these two studies. Although the actual distribution of user activities may vary with the OSN, we show in Section 4 that small variations in action value assignment are not likely to dramatically impact the final user value.

Facebook		**Orkut**		**Hi5**	
Category	Share	Category	Share	Category	Share
Home	35 %	Profile,Friends	41%	Photos	45%
Profile	16 %	Photos	31%	Profile	20%
Photos	16 %	Scrapbook	20%	Home	13%
Friends	4.7 %	Other	3%	Friends	13%
Groups	3 %	Communities	1%	Groups	1%

Table 1: Comparison of popular user activities across three OSN sites [8, 28].

3.2 Users Characteristics and Interactions

We next consider the information that external observers can collect to help estimate the value of users. Many OSN services make basic personal information provided by each user (gender, age, location, interests) public by default. It is also generally possible to obtain some information about the social graph, such as the number of friends and their identities. While the basic information is useful for targeting (e.g., an advertiser is targeting 30-year-old men in Barcelona), the latter is useful to estimate the indirect impressions. For example, when a user with thousands of friends posts an update, the user is broadcasting information to a wider audience than a user with only a few friends. As a user's friends tend to be similar in demographics and tastes to the user [1], it is conceivable that most of the indirect impressions would be shown to a similar target group.

3.3 Measuring User Value

We now present a framework for measuring the relative values of users, and begin by defining the value of user characteristics, activities, and friends' activities:

User characteristics (u_c): This term measures individual user characteristics and is composed of two elements: the targeting parameters t and the number of friends d (i.e., the user's degree). We can tailor t to a given target group; if the advertiser is seeking older demographics, t could be defined as being proportional to the age. If the target is related to a geographical location, t would be inversely proportional to (e.g., the logarithm of) the distance. All such targeting parameters can be combined depending on the advertiser requirements. Precision and granularity of targeting will depend on user's demographic information available on each OSN (for example gender, countries GDP, etc.). The second parameter d reflects the amplification of an action as a result of the direct audience reached by each user. To be conservative, we define d as the logarithm of the user's degree (as studies have shown that the fraction of a user's friends who the user interacts with is not a linear function of their degree [33]). Hence

$$u_c \propto t \cdot d \propto t \cdot \log\left(\#friends + 1\right) \quad (1)$$

Notice that this only captures the first hop on an activity cascade. The next hop will be captured by the activities of the users influenced by this user.

User activity in her own profile ($u_{a_{self}}$): This is a weighted sum of u actions (such as number of photos up-

loaded, number of articles posted, etc.) done by u in her own profile or home page.

$$u_{a_{self}} \propto \sum_i w_i \ \#action_i \quad (2)$$

where w_i is a weight proportional to the action value. Most likely, activities in a user's profile correspond to direct impressions.

Friends activity in a user's profile ($u_{a_{friends}}$): For all the users v that are friends with u, we measure their activity in u's profile. Given that each time v performs an action in u's profile this information is sent both to u and v friends by default (this can be changed through privacy settings but is rarely done), we weight all these actions by v's individual characteristics v_c:

$$u_{a_{friends}} \propto \sum_{v \in |u|} v_c \sum_i w_i \ \#action_i \quad (3)$$

As discussed earlier, most actions are reactive and thus most of them will correspond to indirect impressions.

User activity in their friends' profiles ($u_{a_{visitor}}$): When u carried out an action in her friends' profiles:

$$u_{a_{visitor}} \propto \sum_{v \in |u|} v_c \sum_i w_i \ \#action_i \quad (4)$$

As in the previous case, most of these activities will likely correspond to indirect impressions.

If we are targeting all users in the same way (that is, $t = 1$), then the final formula for the u_{value} is a function of her activity and her friends:

$$u_{value} \propto (u_{a_{self}} + u_{a_{friends}} + u_{a_{visitor}}) \ u_c \quad (5)$$

In Section 5, we analyze the case of the value of different users as per their different interests. Overall, the first weight captures mainly direct impressions while the other two capture mainly indirect impressions.

This simple definition could be extended to include different weights for friends based on tie strength (closer friends are more likely to see our actions and generate more impressions), privacy settings, and "circles" (groups of users see only partial information about our actions depending on our privacy configuration). Groups or community activity could also be included. However, an easy way to add group activity, is to consider the group as a friend, where all the group members would be friends of friends. Then, group activities can be included in the terms $u_{a_{friends}}$ and $u_{a_{visitor}}$.

4. APPLYING USER VALUE

Having defined a framework for reasoning about the value of users to the OSN, we now apply our model on a real OSN dataset. The goal is to explore how the relative value is distributed by our model across users, and how different strategies to measure invisible actions affect these results.

4.1 Dataset Description

We use a 2009 dataset collected from Facebook covering users in New Orleans [29]. We only consider the 50,564 users with public profiles out of the 90,269 users. As we are interested in classifying users by their interests and demographics, we use only those who share their age and gender, have at least one interest, and have at least one "post" on their wall. Most of the users divulge gender and age but only 23,950 users have at least one interest and an even smaller number (7,054) users have any posts.

4.2 Choosing Weights

Facebook's users can share different types of posts (status updates, posts, URLs, etc.), upload multimedia content (photos, videos), and perform actions within communities (join a group, event, fan page, etc.). To simplify our analysis, we group all these actions in three categories: posts, multimedia, and communities. We need a way to define the weights w_i described in Section 3 for each of these groups. This is difficult because the invisible actions (e.g. watching a video without leaving any comments) are unknown to external observers. Of course, the OSN operator is privy to much more detailed information and could likely derive many of the weights from the traces of user activity. Instead, we rely on externally-visible information in order to estimate the rough values of the weights.

Previous studies (see Table 1) have shown that visiting friend's profiles (corresponding to our "posts" category) is the most frequent activity, followed by browsing photos (multimedia), while group or community actions are infrequent. Although this ordering is consistent among different OSNs, the percentage of time spent in each category varies (e.g people spent 45% browsing photos in Hi5 vs only 16% on Facebook). Thus, we need to study the sensitivity of u_{value} to different w_i values. To understand how these weights affect the results, we can check to see if correlation among these three groups of actions is high, indicating that the weights are less important. If user's activities are equally distributed, weights are less important, because the numbers of actions ($\#action_i$) will be proportional.

We found a strong correlation between posts and multimedia actions but a lower correlation between multimedia and communities. Users uploading significant multimedia content are thus also creating many "wall posts" but do not necessarily engage in group activities. Thus, we normalize by the three categories of actions and group "home", "profile" and "friends" activities in the posts category; we then assign weights of 0.75 for posts, 0.21 for multimedia, and 0.04 for communities.

4.3 Value Distributions

Next, we want to study how value is distributed among users after applying our model and how they are related to user attributes (age, gender, and interests). Our hypothesis is that the number of impressions (i.e. value) generated by each user varies over a wide range. We expect a small fraction of users to create a lot of impressions and many users generating only a handful.

To test our hypothesis, we first compute the value for each user in our dataset. Next, we normalize these values, with 1 being the most valuable user. Not too surprisingly, we found users' value distribution is Zipfian, confirming our hypothesis that a small subset generates most of the impressions.

Next, we identify the high value users. In our experiment, we are considering all the users in our New Orleans dataset as the target group. Thus, t in equation 1 does not depend on age, gender or interests and so u_c only depends on the number of friends. We want to compute a generic value that allows us to compare the impressions generated by different demographic groups. For example, do women generate more

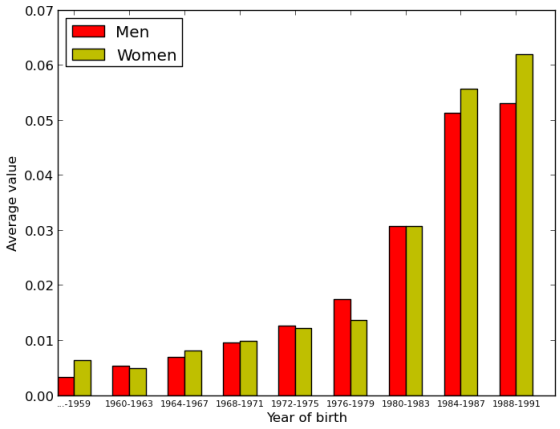


Figure 1: Users' relative value predicted by our model, broken down by age and gender. Younger users appear to be significantly more valuable, largely due to their higher level of activity.

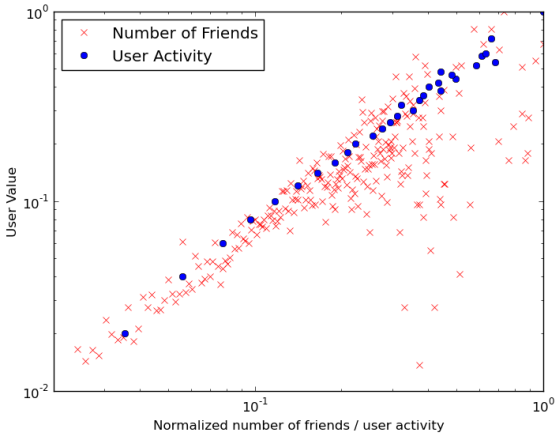


Figure 2: User's value vs. number of friends and activity.

impressions than men? We find that women are more valuable in our model than men, and young people generate more impressions than more mature users (see Figure 1). Given that the difference between the least valuable group (males born after 1959) and the most valuable group (women born between 1989-1991) is less than 10%—with large standard deviation in each group—we can assume that high and low value users are spread across different demographic groups.

A previous study [27] suggests that there is a strong correlation between the number of friends and user activity; users that post/upload more information have more friends. Our experiments show that popular users are more active and valuable (see Figure 2), which is partly a consequence of using friends and activity to compute the user's value. However the correlation is much higher for the activity than for the number of friends; activity produces impressions while the number of friends is only a potential amplifier of activity.

5. FROM VALUE TO REVENUE

The model that we have proposed and evaluated so far has considered ad impressions as the metric for valuing users, as OSNs are typically ad-supported. However, we would also like to be able to reason about the actual (monetary) value of each user so that the OSN provider and advertisers can determine which users are profitable. The mechanism we propose here is only one way of doing so; there may be others that are more attractive if additional data is available.

5.1 Using Ad Auction Data

To map user value to revenue we need to translate ad impressions to monetary revenue. An approach that assumes that all ad impressions provide equal revenue is unrealistic as certain user attributes are likely to be much more valuable (i.e., we know that there are certain demographics that are much more valuable to advertisers, and users in different countries are likely to provide significantly different revenue). Fortunately, Facebook provides advertisers with a "suggested bid" [20] for a given set of targeting parameters; this suggested bid is expressed as a range (min, median, and max) over the current bids for the target demographic. Thus, for a given user, we can determine the current revenue that Facebook receives for each ad impression shown to that user by querying for the suggested bid.

5.2 Collecting Data

To demonstrate this approach in practice, we use the same Facebook New Orleans data set from 2009 from Section 3. For each of the users, we extract as many profile attributes as possible, including basic attributes like gender and age, and free-form attributes like user interests. Recall that we only consider users who provided their age, gender, and at least one interest; leaving us with 7,054 users. We map each of the free-form interests to Facebook's "precise interests"; we found a match to a Facebook-supported interest category 60.4% of the time (4,265 users).

We query Facebook's Ads platform for the suggested bid of each of the 4,265 users, using *all* available targeting parameters [20]. For users with multiple interests that we could match, we queried multiple times (once for each interest).

Examining the bid data, we make two interesting observations. First, the distribution is remarkably even, with the CPC of 99% of users' interests ranging between \$0.62 and \$1.53 and the CPM for 99% of users' interests ranging between \$0.07 and \$0.31. This suggests that there are not specific interests that are significantly more highly valued by advertisers than others. Second, we observe that the prices are quite stable over time (see an example in Figure 3), indicating that our methodology is likely to hold over at least short periods of time.

5.3 Putting It All Together

We demonstrated in the previous section that a user's value (in terms of ad impressions) can be modeled based on the user's actions. Here we showed that these ad impressions can be translated into actual revenue received by the OSN. Now, we use the New Orleans data set to estimate the potential ad revenues brought in by the different users (i.e., we estimate the value of the users to Facebook). To do so, we first use the model of user value to assign credit for ad impressions to different users. Then, we use the advertising revenue for each user, and translate credit for ad impressions to credit for revenue (basically, using a different t per user).

We present the results in Figure 4. Recall that for users with multiple interests, we obtained multiple CPM/CPC

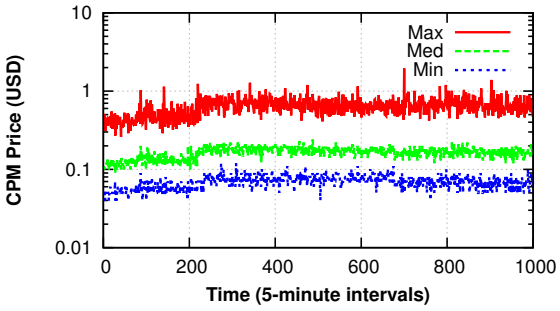


Figure 3: Facebook's suggested CPM bids from one particular set of parameters over three days, showing price stability over time (note y-axis in log-scale).

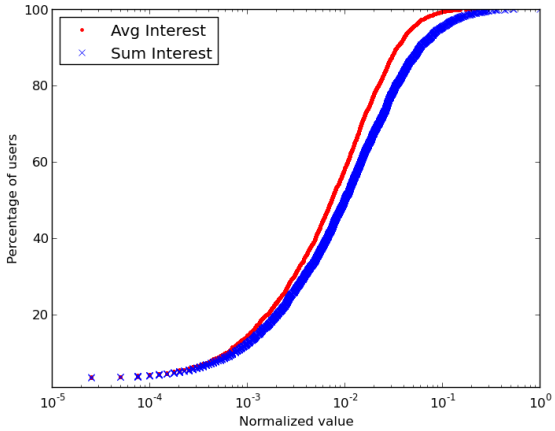


Figure 4: Average and potential (sum) value of users as per their interests and the advertising bids.

values (one for each interest). We present two separate means of aggregating these values together for such users: (a) we *average* of all of their interest prices together and (b) we *sum* all of their interest prices together (e.g. the overall potential value of a user). Since most users have only one matched interest, the two cases do not differ significantly.

We observe that there is a surprisingly wide variety in the user value to the OSN. For example, there are users who are over 10,000 times more valuable than other users. In fact, over 95% of the users have a value of less than one-tenth of the most valuable users. This result suggests that the value of different users to the OSN is quite different, and that it would be beneficial for both the OSN and advertisers to focus more on these highly valuable users.

6. CONCLUDING DISCUSSION

Popular OSNs provide "free" access to users in return for revenue generated by advertising impressions shown to them. We explored how different classes of actions result in different advertising impression counts and corresponding revenue. Implicitly, our goal was to demonstrate through our model that users on OSNs have an intrinsic value that varies with the extent of their participation on OSNs. Our two-stage approach first identifies the actions that are key to generating direct and indirect impressions. We then map these actions to actual revenue, showing the feasibility using real-world data sets. Our model is extensible, applicable to other OSNs and adaptable to alternate revenue mapping.

The results of our initial study are intriguing: a small subset of actions on OSNs are responsible for most of the advertising impressions, and a small fraction of users are key to the overall advertising revenue. Identifying these users can benefit OSNs (who may provide more services for such users), advertisers (who can target the more valuable users directly), and users (who will have understand how their actions are actually valued on OSNs). We can imagine an economic *modus vivendi* where there is an explicit trade between user's actions and profile information and the resulting service from the OSN.

Limitations. Our work has several limitations, starting with the heuristic nature of our proposed model and the assignment of parameters. Both of these are due to limited data availability on the actual views and clicks in the OSN. With better data, we could include additional activities and better parametrize the model. Additionally, validating the proposed model also requires data (such as more detailed breakdowns of the advertising revenue) that is currently in the possession of the OSNs. We leave addressing both of these to future work.

Leveraging user value. While advertisers are clearly interested in the value of users, we believe that it is actually the OSN provider who is best positioned to make direct use of this knowledge. This is due to three reasons: First, the OSN provider is clearly best able to make accurate estimates of user value; the provider observes all user activity, and typically only makes a portion (if any) available to advertisers or other third-parties. Second, the OSN provider is able to encourage "high value" behavior by its users by directly rewarding them or by more prominently featuring the more highly valued friends. By doing this, the OSN increases the value of all users and by proxy the OSN platform itself, and also encourages emulation by other users. Third, the OSN provider is able to charge higher prices for advertising to such users, providing benefits for both advertisers and the provider. Note that users can also benefit by becoming aware of their relative value.

7. REFERENCES

[1] P. Adams. *Grouped: How Small Groups of Friends Are the Key to Influence on the Social Web*. Voices That Matter. Pearson Education, 2011.

[2] V. Aksakallı. Optimizing direct response in internet display advertising. *Electron. Commer. Rec. Appl.*, 11(3):229–240, May 2012.

[3] B. Augustin, B. Krishnamurthy and W. Willinger. Ixps: Mapped? In *IMC*, 2009.

[4] E. Bakshy, J. M. Hofman, W. A. Mason and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, 2011.

[5] E. Bakshy, B. Karrer and L. Adamic. Social influence and the diffusion of user-created content. In *EC*, 2009.

[6] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma and S. Muthukrishnan. Adscape: Harvesting and analyzing online display ads. In *WWW*, 2014.

[7] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida and M. Gonçalves. Detecting spammers and content

promoters in online video social networks. In *SIGIR*, July 2009.

[8] F. Benevenuto, T. Rodrigues, M. Cha and V. Almeida. Characterizing user behavior in online social networks. In *IMC*, 2009.

[9] M. Cha, F. Benevenuto, Y.-Y. Ahn and K. P. Gummadi. Delayed information cascades in Flickr: Measurement, analysis, and modeling. *Computer Networks*, 56(3):1066–1076, 2012.

[10] M. Cha, H. Haddadi, F. Benevenuto and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*, 2010.

[11] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, 2001.

[12] J. Edwards. Facebook shares surge on first ever $1 billion mobile ad revenue quarter, 2013. `http://www.businessinsider.com/facebook-q4-2013-earnings-2014-1`.

[13] A. Goel and K. Munagala. Hybrid keyword search auctions. In *WWW*, 2009.

[14] S. Goel, S. Lahaie and S. Vassilvitskii. Contract auctions for sponsored search. In *WINE*, 2009.

[15] Google. 2013 financial tables, 2013. `http://investor.google.com/financial/2013/tables.html`.

[16] S. Guha, B. Cheng and P. Francis. Challenges in measuring online advertising systems. In *IMC*, 2010.

[17] B. J. Jansen, T. Flaherty, R. A. Baeza-Yates, L. Hunter, B. Kitts and J. Murphy. The components and impact of sponsored search. *IEEE Computer*, 42(5), 2009.

[18] A. C. König, M. Gamon and Q. Wu. Click-through prediction for news queries. In *SIGIR '09*, 2009.

[19] K. J. Lang, B. Moseley and S. Vassilvitskii. Handling forecast errors while bidding for display advertising. In *WWW*, 2012.

[20] Y. Liu, C. Kliman-Silver, B. Krishnamurthy, R. Bell and A. Mislove. Measurement and analysis of osn ad auctions. In *COSN*, 2014.

[21] Y. Moon and C. Kwon. Online advertisement service pricing and an option contract. *Electron. Commer. Rec. Appl.*, 10(1):38–48, Jan. 2011.

[22] H. Nazerzadeh, A. Saberi and R. Vohra. Dynamic cost-per-action mechanisms and applications to online advertising. In *WWW*, 2008.

[23] C. Perlich, B. Dalessandro, R. Hook, O. Stitelman, T. Raeder and F. Provost. Bid optimizing and inventory scoring in targeted online advertising. In *KDD*, 2012.

[24] A. Radovanovic and W. D. Heavlin. Risk-aware revenue maximization in display advertising. In *WWW*, 2012.

[25] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi and V. Almeida. On word-of-mouth based discovery of the Web. In *IMC*, 2011.

[26] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates and F. Benevenuto. Finding trendsetters in information networks. In *KDD*, 2012.

[27] D. Saez-Trumper, D. Nettleton and R. Baeza-Yates. High correlation between incoming and outgoing activity: a distinctive property of OSNs? In *ICWSM*, 2011.

[28] F. Schneider, A. Feldmann, B. Krishnamurthy and W. Willinger. Understanding online social network usage from a network perspective. In *IMC*, 2009.

[29] B. Viswanath, A. Mislove, M. Cha and K. P. Gummadi. On the Evolution of User Interaction in Facebook. In *WOSN*, 2009.

[30] Y. Wang, D. Burgener, A. Kuzmanovic and G. Maciá-Fernández. Understanding the network and user-targeting properties of web advertising networks. In *ICDCS*, 2011.

[31] J. Weng, E.-P. Lim, J. Jiang and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, 2010.

[32] W. Willinger, B. Ager, N. Chatzis, A. Feldmann, N. Sarrar and S. Uhlig. Anatomy of a large european ixp. In *SIGCOMM*, 2012.

[33] C. Wilson, A. Sala, K. P. Puttaswamy and B. Y. Zhao. Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on the Web*, 6(4), 2012.

Uncovering Social Semantics From Textual Traces: A Theory-Driven Approach and Evidence From Public Statements of U.S. Members of Congress

Yu-Ru Lin
School of Information Sciences, University of Pittsburgh, 710 IS Building, University of Pittsburgh 135 North Bellefield Avenue, Pittsburgh, PA 15260. E-mail: yurulin@pitt.edu

Drew Margolin
Department of Communication, Cornell University, Ithaca, NY 14580. E-mail: dm658@cornell.edu

David Lazer
Department of Political Science and College of Computer and Information Science, Northeastern University, Boston, MA 02115; John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138. E-mail: d.lazer@neu.edu

The increasing abundance of digital textual archives provides an opportunity for understanding human social systems. Yet the literature has not adequately considered the disparate social processes by which texts are produced. Drawing on communication theory, we identify three common processes by which documents might be detectably similar in their textual features—authors *sharing subject matter*, *sharing goals*, and *sharing sources*. We hypothesize that these processes produce distinct, detectable relationships between authors in different kinds of textual overlap. We develop a novel *n*-gram extraction technique to capture such signatures based on *n*-grams of different lengths. We test the hypothesis on a corpus where the author attributes are observable: the public statements of the members of the U.S. Congress. This article presents the first empirical finding that shows different social relationships are detectable through the structure of overlapping textual features. Our study has important implications for designing text modeling techniques to make sense of social phenomena from aggregate digital traces.

Introduction

The quantity of human behavior that is stored and observable as raw, unstructured text is rapidly increasing. One of the most promising methods for analyzing text is the use of discursive *isomorphism*—the identification of texts with similar features or structures—to infer relationships between the authors of the texts (Bramsen, Escobar-Molano, Patel, & Alonso, 2011; Bryden et al., 2013; DiMaggio & Powell, 1983). In this paper we propose a technique for using *n*-gram similarity to infer the kind of relationship that is likely to exist between authors; specifically, whether the authors are addressing the same subject matter, share rhetorical goals, or consult similar sources of information. We argue that while these relationships are not mutually exclusive each is produced by a distinct theoretical process that can vary and be influenced independently. We then argue that the presence and intensity of these distinct relationships is most detectable in similarities between authors at different lengths of phrases or *n*-grams. We then test this hypothesis by using the public statements of the members of the U.S. Congress from January 2011 through August, 2012.

Textual data can indicate "invisible networks"—underlying interpersonal, ideological, structural, or other relationships between people that would otherwise only be observable through close observation of and interaction with them (Lin, Margolin, & Lazer, 2014). A number of techniques have been developed to infer these social properties such as the extent to which individuals share an organizational culture (Barnett et al., 1981; Danowski, 1982), members of teams share mental models (Carley, 1997), ideological and attitudinal similarity between individuals (Adamic, Lento, Adar, & Ng, 2014; Danowski, 2011),

Received October 30, 2013; revised December 30, 2014; accepted February 20, 2015

© 2015 The Authors. Journal of the Association for Information Science and Technology published by Wiley Periodicals, Inc. on behalf of ASIS&T • Published online 2 June 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23540

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

common topics and ways of speaking (Griffiths, Steyvers, Blei, & Tenenbaum, 2005; McCallum, Wang, & Corrada-Emmanuel, 2007; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004), or group identities (Tamburrini, Cinnirella, Jansen, & Bryden, 2015).

Yet despite the proliferation of techniques there has been little work asking *how* or *why* observed isomorphism is achieved. The result is that much research uses similar techniques to infer different relationships and properties. For example, in some research, individuals who use the same words are inferred to share topics of discussion (Rosen-Zvi et al., 2004), whereas in others they are inferred to identify with and conform to similar group norms (Tamburrini et al., 2015). Authors may also produce similar texts because there are incentives for them to imitate one another or a common third party directly (Adamic et al., 2014), suggesting they share access to common sources of information and incentives to save cost or demonstrate a unified front (DiMaggio & Powell, 1983).

The distinction between these relationships is often important. Understanding which process produced the observable relationship between authors can help researchers to identify the relevant theories and boundary conditions for understanding the authors' social behavior. In political discussions, for example, textual isomorphism due to conformity is likely to indicate group-based polarization (Adamic & Glance, 2005), whereas textual isomorphism due to topic sharing can indicate that individuals are engaging with the needs and demands of the public (Boydstun, Glazier, & Pietryka, 2013; Gentzkow & Shapiro, 2010). Isomorphism due to imitation or *mimetic* processes can indicate a lack of research resources (Gandy, 1982; Grimmer, 2009) or deference to authority (Glazier & Boydstun, 2012) or norms that encourage copying (Cordell, in press). Similarly, within an organization isomorphism can occur because of a faction of like-minded individuals (Carley, 1997) or a group that participates on topics with highly regulated, restricted, or ritualized discourse (Hart, Childers, & Lind, 2013; Meyer & Rowan, 1991)?

The meaning and expected endurance of the observed relation depend on the social process that is assumed to produce the textual similarity. Clusters based on ideological polarization are unlikely to evolve in the same way as clusters produced by imitation driven by resource shortages, nor can they be influenced or corrected with the same remedies. A collective convergence on a common subject of attention may be difficult to sustain, whereas a collective effort to hold an ideological position may be difficult to overcome.

Current techniques do not adequately distinguish these relationships either theoretically nor empirically. We address this gap by articulating distinct social processes: *authors sharing subject matter, sharing rhetorical goals*, and *copying from shared sources*; each of which should lead to discursive isomorphism across authors and each of which has distinct implications for the way in which this isomorphism is manifest. Specifically, we argue that each process will produce similarity between authors most distinctly at a different length of *n*-gram. We then analyze texts produced by members of the U.S. Congress by comparing their observable discursive isomorphism to other, nondiscursive relationships between them. Our work thus takes a first step in computationally distinguishing different kinds of textual similarity. The main contributions of this work are:

1. We provide theoretical justification for distinguishing three processes by which similar texts are authored.
2. We develop a randomized *n*-gram extraction algorithm to capture text similarity based on *n*-grams of different lengths that effectively tackles the computational challenge of counting vast numbers of arbitrarily long *n*-grams based on a document-pair sampling strategy.
3. We conduct a study on a text corpus produced by members of the U.S. Congress, demonstrating that the distribution of *n*-gram similarity between the members' statements is tri-modal. We then show that each mode is most closely associated with different relationships known to exist between members of Congress: sharing topics, sharing rhetorical goals, and sharing sources.

The rest of the paper is organized as follows. After a review of related work we identify and define processes by which similar texts can be produced and describe the empirical implications of each process' operation. We then describe the randomized *n*-gram extraction procedure that allows the comparison of document texts on *n*-grams of arbitrary length. We present our empirical results for the public statements of U.S. Congress members. Finally, we provide a discussion and conclusion.

Related Work

Previous research has emphasized only a single textual production process at a time for a couple of reasons. First, analysis techniques have often been developed to be applied to specific kinds of textual data where one process is already known or assumed to be dominant. Second, most techniques tend to simplify the relationships between authors' thoughts and beliefs and the thoughts and beliefs of their audiences, thus underemphasizing both strategic processes (Atteveldt, Kleinnijenhuis, & Ruigrok, 2008; Eisenberg, 1984), where words are chosen for their effect, and mimetic processes, where they are chosen for their convenience (DiMaggio & Powell, 1983).

For several decades communication researchers have analyzed structural features of documents using semantic networks (Danowski, 1982; Doerfel & Barnett, 1999). A semantic network is derived by considering words as nodes and co-occurrences or meaning relations between the words as links (Carley & Kaufer, 1993). Early semantic network analysis was conducted on texts that were authored in response to surveys where subjects were asked by researchers to describe meaningful relationships between concepts (e.g., Barnett et al., 1981; Carley, 1997; Rice & Danowski, 1993). In this context, where subjects are asked to describe

DOI: 10.1002/asi

their cognitions to a researcher in a confidential survey, researchers reasonably assumed that if two individuals produced similar documents the similarity was due to these individuals sharing cognitions, beliefs, or worldviews, as summarized by Danowski (2011, p. 226): "individuals with more similar semantic networks based on their encoding of messages are likely to perceive the world more similarly and behave more similarly." Similarly, Doerfel and Barnett (1999, p. 589) state that semantic network analysis "focuses on the structure of a system based on shared meaning." Consistent with this assumption, this technique is generally used to measure how groups perceive the world, such as the "collective consciousness" (Barnett et al., 1981) within an organization or "mental models" used by team members (Carley, 1997).

More recently semantic network analysis has been applied to the observation of natural discourse (e.g., Doerfel & Connaughton, 2009; Rice, 2005; Yuan, Feng, & Danowski, 2013). In these natural contexts the assumption remains, however, that when authors create texts they are trying to clearly indicate meaningful relationships between words, that is, that their communication behavior is a "conduit" of their thoughts (Rogers & Kincaid, 1981). While this assumption is reasonable in some contexts, it is problematic in contexts where there are incentives for strategic communication. In these strategic contexts individuals have an incentive to construct their statements informed by the meanings and cognitions of their audience precisely because it is different from their own. For example, Danowski (2011) describes the creation of "optimal messages" that "steer" the audience "away from certain words or towards them" (p. 229). Authors may also use the fact that some meanings are not shared widely in the audience to create strategically ambiguous messages that mean one thing to one subgroup in the audience and another thing to another (Eisenberg, 1984). In these cases two authors may use the same combinations of words not because these combinations naturally reflect their own (shared) cognitions but because they each predict they will have a similar effect on their audience.

Work in information science and linguistics has also tended to treat texts as neutral "conduits" of conceptual meaning. Researchers have developed a variety of computational techniques for extracting the relationship between "concepts" and "word meanings" such as latent semantic analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997) and probabilistic models such as probabilistic latent semantic analysis (pLSA) (Hofmann, 2001) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). By design, these approaches seek stable, context-invariant relationships between concepts. Strategic uses of language, where these typical relations are exploited to lead audiences to atypical conclusions (Carley & Kaufer, 1993), are largely not considered (one exception is O'Connor [2012]).

The LSA approach has more recently been extended to the topic model technique (e.g., Blei, 2012; Blei et al., 2003; Hofmann, 2001; Wallach, 2006). The LDA (Blei et al., 2003) topic model introduces the idea that documents are mixtures of topics (like "concepts" in LSA or "aspects" in pLSA). Because the aim of topic modeling is the reliable classification of documents, the technique does not make assumptions about why words, or larger *n*-grams (Wang, McCallum, & Wei, 2007), co-occur (e.g., because they have meaningful relationship) but focuses more exclusively on the statistical fact of this co-occurrence.

Although topic models do not make assumptions about the relationship between word use and an author's cognition, they are vulnerable to mischaracterizing similarities due to mimetic processes. When a set of words (or longer *n*-grams) frequently co-occur it is possible topic models assume this is because of a systematic statistical tendency for these individual words to be drawn conditional on the presence of the "topic." But it is also possible that their co-occurrence is explained by the wholesale repetition of a chunk of text in which they are co-present. This mimetic isomorphism in text has been observed in numerous cases, in the blogosphere (Leskovec, Backstrom, & Kleinberg, 2009; Simmons, Adamic, & Adar, 2011), social media (Adamic et al., 2014), political news reporting (Grimmer, 2009), discussions on Wikipedia and the Supreme Court (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012), as well as within the U.S. Congress (Hassan et al., 2008).

An additional area of research focuses strictly on the detection of mimetic processes in the sense of "plagiarism" (Barrón-Cedeño & Rosso, 2009; Grozea, Gehl, & Popescu, 2009; Lyon, Barrett, & Malcolm, 2004; Lyon, Malcolm, & Dickerson, 2001). These techniques usually rely on the 2- to 7-gram feature analysis to identify candidate documents as the source of the copy based on the low statistical likelihood that independently written texts will have a large number of matches in long *n*-grams (Lyon et al., 2004) (or longer *n*-grams). Quinn, Monroe, Colaresi, Crespin, and Radev (2006) applied this technique to the U.S. Congressional Speech corpus and studied the importance of speakers based on how their speeches were similar to other speeches. These techniques presume that the repetition of features are the result of copying. There is evidence, however, that in some social contexts highly restrictive social norms can lead to texts that are highly repetitive even without the presence of copying (e.g., corporate annual reports [Hart et al., 2013]).

Distinguishing Processes

Conceptual Distinction

As described above, previous analysis techniques have tended to assume that similarities are the result of one particular process even though others can plausibly explain their results. One reason for this may be that a set of candidate *semantic organizing processes*, the processes by which a set of authors can produce similar texts, have not been conceptually distinguished. Here we distinguish these processes by focusing on the variables that most consistently

DOI: 10.1002/asi

trigger or influence the intensity of each process independent of the others. We expect that in real-world settings these processes often operate at the same time; however, these distinct triggers can lead them to be dominant or subordinate in different contexts. To illustrate these theoretical concepts we provide examples from the corpus we use in our analysis.

The first organizing process can be described as constrained *searching* (March & Simon, 1993; Margolin & Monge, 2013). In constrained searching, authors look for and consider only words, concepts, phrases, or longer *n*-grams from a restricted population. Examples of restricted populations include vocabularies that are appropriate to specific subject matter. A vocabulary is a preexisting set of terms and phrases with clear, stable definitions (Loewenstein, Ocasio, & Jones, 2012). Linguistic or cultural norms may also restrict authors to draw from particular sets of terms and phrases. In the context of a discussion of certain ideas, particular concepts may be taboo (Hart et al., 2013; Meyer & Rowan, 1991). Thus, in some settings the fact that a set of terms has been used in one statement may constrain those responding to that statement to use similar terms (Danescu-Niculescu-Mizil et al., 2012). For example, on June 28, 2012, the Supreme Court released its ruling that the Affordable Care Act (colloquially known as "Obamacare") does not violate the Constitution. On this day, 369 members of Congress (161 Democrats and 208 Republicans) released public statements that addressed this topic. In discussing the Supreme Court's decision regarding the ACA, members who both supported and opposed the law used healthcare-related terms such as "insurance," "doctor," and "affordable."

Searching processes should be distinguished from *selection* processes (Kauffman, 1993). In searching processes, authors seek to create distinct statements, but are constrained by external forces to construct these statements from common materials, leading them to be similar. In selection processes, authors are attempting to construct statements that are similar in terms of their meaning or effect. They construct statements and evaluate them in terms of their fit with their goals or intentions, editing the statements to produce those that most closely meet their criteria. Thus, similarity in selection reflects similarity in underlying attributes of the individuals, such as mental models or beliefs, whereas similarity in searching reflects similarity in a situation, context, or predicament that authors face.

In the case of the Supreme Court decision, we observe a clear partisan split in how the bill is referred to. Of the 161 Democrats who spoke about the Supreme Court's decision and thus referred to the law in some way, 102 (63%) explicitly named the law "the Affordable Care Act" (ACA) whereas only 18 (9%) of the Republicans used the official name. This difference is not due to a constraint on searching. Republicans did not eschew the vocabulary terms that comprise the name of the law: 52% say "affordable"; 100% say "care"; and 72% say "act" at least once. They simply avoid combining them to name the law. Instead, many Republicans (47, 23%) referred to the law by a novel construction, "the president's health care law." No Democrats used this term, despite the fact that they regularly constructed their statements from its component parts: president (25%); health (100%); care (100%); law (74%); again suggesting an intentional avoidance that belies their underlying position toward the bill.

A third way that authors can produce similar texts is through *mimetic* processes. In mimetic processes search and selection are short-circuited as authors use preassembled strategies they observe others using (DiMaggio & Powell, 1983). Authors may rely on mimetics because they lack the resources to independently research and construct their own statements (DiMaggio & Powell, 1983; Gandy, 1982) or because there are benefits to appearing to be in exact coordination with other authors (Benford & Snow, 2000; Lammers & Barbour, 2006). Although authors who imitate others are likely to share some constraints and attributes with those whom they mimic, the exactness with which they replicate others' statements can be misleading. The rationale for imitation is the achievement of a satisfactory but suboptimal solution to a complex search or coordination problem at low cost (DiMaggio & Powell, 1983; March & Simon, 1993). Thus, authors may often imitate others with whom they only partially agree or share constraints with, producing discourse that is only loosely similar to what they would have said had they spent the time and resources to produce their own independent statement. The strength of similarity produced by mimetic processes is thus more of an indicator of resource constraints (Grimmer, 2009), an authority relationship between the authors (Danescu-Niculescu-Mizil et al., 2012), or a highly demanding and complex context.

For example, in response to the Supreme Court decision the phrase "[Health centers in home state have received $X million to] create new health center sites in medically underserved areas, enable health centers to increase the number of patients served, expand preventive and primary health care services" was used by five different Democratic members of Congress. The last 26 words of this phrase can also be found in a document produced by the White House promoting the benefits of the ACA.[1] The high semantic similarity across these five individuals could be due to independent searches for and selection of high-quality arguments that independently arrived at identical 26-word constructions, but it appears more likely that this similarity indicates the absence of such processes, as these members were able to produce adequately credible statements that signaled loyalty to the White House with limited effort.

These processes may often be co-present in the construction of texts; however, their distinct antecedents allows them to vary independently. In a presidential debate two candidates may search within the same vocabulary to select phrases with opposite implications (Boydstun et al., 2013), while in local election campaigns candidates addressing different districts but from the same party may select similar

[1]http://www.whitehouse.gov/sites/default/files/docs/the_aca_helps_rural_america.pdf

DOI: 10.1002/asi

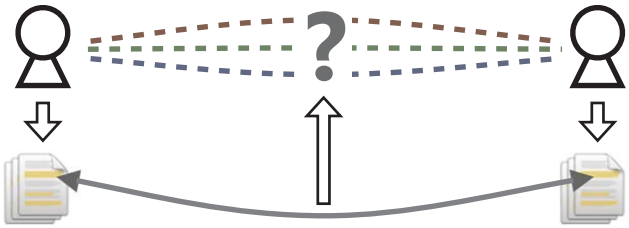


FIG. 1. Illustration of the problem: How can we uncover different social processes from observed similarities between documents produced by social actors? [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

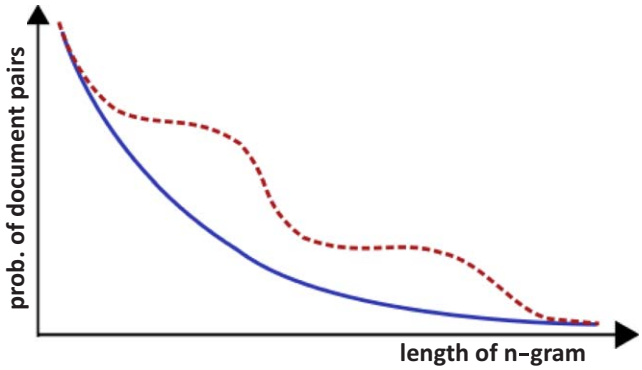


FIG. 2. Hypothetical similarity distribution over *n*-gram length. "Bumps" in the distribution of similarities (red dashed curve) represent the operation of a process of semantic organization, some mechanism that is pulling the documents to be similar to one another at that particular length. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

constructions while addressing different topics of local interest. Within an organization, departments with excess resources may produce documents with which others only loosely agree with but nonetheless heavily imitate and reuse (Lammers & Barbour, 2006; Monge & Poole, 2008), while departments working with highly technical and disciplined material (e.g., engineering, accounting) may spend substantial resources to produce independently similar statements (Hart et al., 2013). We thus suggest it is both possible and necessary to model a tendency for authors to produce similar texts as a function of the strength of these different processes. Figure 1 presents the challenge we address: distinguishing the appropriate relationship between authors based on an observation that they produce similar texts.

Empirical Distinction via N-*Gram Length*

In this section we explore the possibility of distinguishing these processes by calculating semantic similarity between authors at different lengths of *n*-grams—contiguous sequences of *n* words from a given sequence of text. The similarity between two documents can be described by the distribution of similarity at all *n*-gram lengths from 1 to *Nd* where *Nd* is the number of words in either document. Pairs of documents can show different distributions in their similarity over *n*-grams of different lengths. For example, one pair of documents might use many of the same words but few of the same longer constructions. At the other extreme, a pair of documents might be almost identical except for one or two edits or typos. Figure 2 shows a a space on which to plot the hypothetical similarity distribution of document pairs for the above scenario. In both extreme cases, the pair of documents share a large number of uni-grams. In the former case, the *n*-gram overlap declines precipitously as *n* increases, whereas in the latter case it remains close to constant as *n* increases over certain lengths.

The role of the semantic organizing processes described above is to pull the similarity between authors at a particular *n*-gram length "up," creating a "bump" in the similarity distribution at that length. By examining semantic similarity at multiple *n*-gram lengths it may be possible to observe the operation of different processes, that is, the multimodal similarity distribution displayed by the red dashed curve.

Below we identify three specific semantic organizing processes, each of which represents one of our families of processes: authors sharing subject matter (search process), sharing rhetorical goals (selection process), and copying from shared sources (mimetic process). We hypothesize that each of these different processes will have its strongest implications at a particular level of *n*-gram length.

Shared subject matter. Some sets of documents are similar because their authors search among a common pool of ideas to construct their statements. We define such sets of similar documents as sharing *subject matter*. The key demands of subject matter are comprehension and competence. Given a set of ideas to refer to authors must use terms that are recognizable and combinations that make sense (Clark & Wilkes-Gibbs, 1986; Krauss & Fussell, 1991). These terms are most commonly found in the vocabulary of a subject, a preexisting set of terms and phrases with clear, stable definitions that refer to aspects of the subject matter (Loewenstein et al., 2012).

Although public figures often try "set agendas" by focusing on subject matter advantageous to them (McCombs & Shaw, 1972), audiences frequently demand that they address topics of public concern (Boydstun et al., 2013; Jerit, 2008). There is evidence that audiences detect and reward those who use terms central to the discussion of these topics (Boydstun et al., 2013; Doerfel & Connaughton, 2009). In these cases, inappropriate use of subject-related concepts will cast doubt on the competence or intentions of the author. We argue that vocabularies will tend to be comprised of short *n*-grams and, therefore, the majority of the material found by authors via searches in a shared vocabulary will be of the short *n*-gram variety. Studies of the creation of vocabulary terms indicates that they emerge in relation to ideas and features that frequently co-occur in the external world (Loewenstein et al., 2012; Rosch, 1978). For example, a bird is an animal with "wings" and a "beak"

DOI: 10.1002/asi

(Rosch, 1978). Vocabulary terms have stable meanings with regard to these ideas and entities, and this need for frequency and stability is more easily achieved by short *n*-grams. Specifically, as a construction grows in *n*-gram length each new word added to it is likely to make it either more specific, reducing the likelihood that its referent occurs frequently in the external world, or more ambiguous, reduce the clarity of the consensus around its meaning (Aerts & Gabora, 2005; Carley & Kaufer, 1993; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Experts in scientific fields tend to have vocabulary for specific entities that they encounter frequently that lay individuals lack (Lynch, Coley, & Medin, 2000), while at a certain *n*-gram length, constructions move beyond noun phrases to include full clauses that have contextually contingent meaning (Corman et al., 2002).

Thus, we expect that when several authors intend to competently discuss a subject, they will tend to search within a common vocabulary and this vocabulary will be largely comprised of short *n*-grams. From this we derive:

> **H1 (Shared subject matter):** To the extent to which two authors share subject matter, they will be more similar in their use of short *n*-grams than in their use of longer *n*-grams.

Shared rhetorical goals. Some sets of documents are similar because the author(s) of the documents would like their audience(s) to come to a particular conclusion. We describe such documents as *sharing rhetoric*, that is, strategic or persuasive goals.

When authors share rhetorical goals they will select the phrases and constructions or "optimal messages" (Danowski, 2011) that make their point and eschew those that undermine it (Chong & Druckman, 2007; Entman, 1993). These selections may be conscious, such as in the case of a heavily edited speech from a political campaign (Chapel, 1976), or unconscious, as when individuals find certain expressions offensive and others pleasing as a result of being socialized into a particular institutional view (Lammers & Barbour, 2006).

These constructions may rely on preexisting vocabulary for a subject, but clarity of communication and a demonstration of competence is not the goal. In fact, ambiguity in meaning can be an advantage when pursuing a rhetorical goal (Eisenberg, 1984). Instead, authors have incentives to influence their audience by manipulating the flow of the audience's attention through the sequence of constructions they use (Carley & Kaufer, 1993; Danowski, 2011). For example, authors may find ways to deploy "slogans," sequences of words that are relevant to many topics but which dull associations with additional ideas (Carley & Kaufer, 1993). By guiding the audience to think of a slogan, the audience then finds it hard to reason about the problem at hand. For example, on July 29, 2011, 78 Democrats and 123 Republicans released statements on the issue of whether Congress would raise the "debt ceiling" limit on government spending. On that day, 23 Republicans (19%) used the phrase "cut, cap, and balance" in promoting their alternative plan for the budget. This phrase, used by only two (3%) of Democrats, does not appear to refer to any concrete plan described in these statements.

The ability to emphasize particular aspects of an idea can also depend on the way individual words are juxtaposed (Lakoff, 2004). In their reference to the ACA as "the president's health care law," Republicans introduce a concept, "the president," that likely activated negative associations in the mind of many of those who might be persuaded to agree with them about the law (Hardisty, Johnson, & Weber, 2010). By placing this term first in the sequence, it influences their interpretation of the combination of words as a whole (Wisniewski, 1996).

Because slogans and juxtapositions depend on sequence, but do not need to be widely recognized as stable ideas with agreed meaning, they can be deployed at substantially longer *n*-gram lengths than vocabulary terms. Furthermore, because *n*-gram sharing across authors becomes generically less likely as *n*-grams become longer, the fact that rhetorically motivated sequences can produce similarity in longer *n*-grams should make these shared textual sequences more apparent at longer *n*-gram lengths. For example, the phrase "I will continue to" is used by 33 (19%) of the Republicans who explicitly state they will attempt to "repeal" the ACA, and used by only two (5%) of the Republicans who do not make this commitment. Another *n*-gram, "and replace it with" is also common among those advocating repeal. It turns out that more members say the 1-gram "repeal" than either of these longer phrases. However, these 4-gram sequences are more salient in relative terms. Each is in the top 10 of most shared 4-grams (5^{th} and 8^{th}, respectively) while "repeal" is only the 15^{th} most shared 1-gram. "Repeal" is harder to detect because it is buried among other short *n*-grams, such as "health" and "care" that are chosen as subject matter words. In the analysis of longer *n*-grams, there are fewer vocabulary terms and the rhetorical constructions are more salient.

The length of the sequences for which an exact match should be expected to emerge from shared rhetorical goals is also limited, however. Not all differences in word sequence will make a material impact on interpretation. Two individuals can convey the same frame with similar, but nonidentical sequences where key, framing concepts are in the same sequence but less important placeholders and stop-words are not. Although 49 Republicans say "the president's health care" and 47 say "president's health care law," only 16 say "of the president's health care law." At some point the rhetorical motivations are not strong enough to constrain discursive choices of such distinction. From this we derive:

> **H2 (Shared rhetorical goals):** To the extent to which two authors share a rhetorical goal, they will be more similar in their use of moderate *n*-grams than in their use of short *n*-grams or long *n*-grams.

Shared sources. In mimetic processes, where authors copy one another or a shared third party, there is no constraint on *n*-gram length. However, copies may fit contexts less frequently than shorter stock vocabulary and framing

DOI: 10.1002/asi

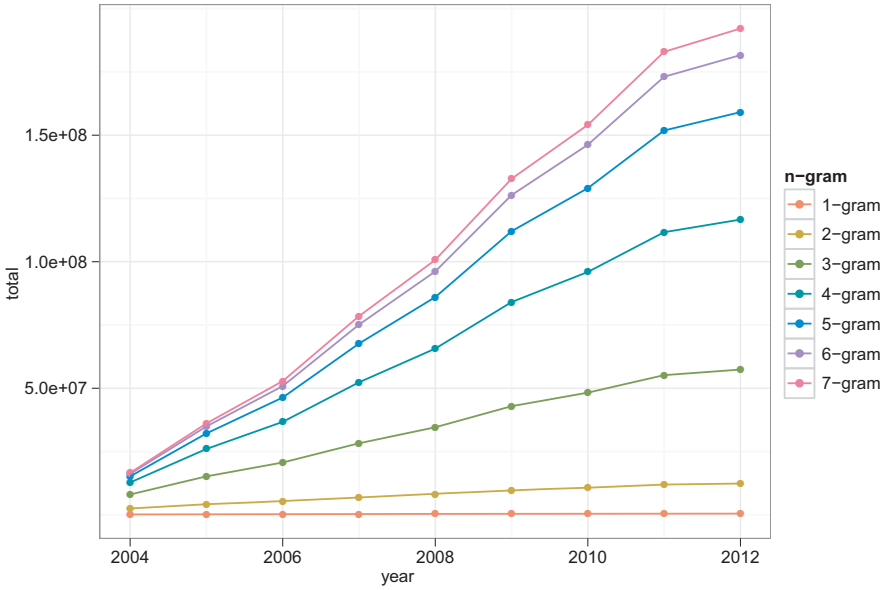


FIG. 3. Growth of *n*-gram. The total number of unique *n*-grams increases much faster for longer *n*-gams. On average, Members of the U.S. Congress introduced 43 K new words, 6 M new trigrams, and 19.5 M 5-grams per year. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

constructions, so their absolute frequency across texts may be likely to be limited when only short *n*-grams are taken into account. We thus anticipate that absolute sharing of very long *n*-grams will be low, but even small numbers of long shared *n*-grams will be evidence of sharing sources, as it is very difficult to explain the highly improbable convergence of authors around very long constructions in the absence of a common source (Barrón-Cedeño & Rosso, 2009).

H3 (Shared texts): To the extent to which two authors share a textual source, they will be more similar in their use of long *n*-grams than in their use of shorter *n*-grams.

Method

Data

We gathered 0.4 million documents from the public statements of members of the U.S. Congress from the Vote Smart Project website.[2] According to Vote Smart, these public statements include any press releases, statements, newspaper articles, interviews, blog entries, newsletters, legislative committee websites, campaign websites, and cable news show websites ("Meet the Press", "This Week", etc.) that contain direct quotes from the member. We include all of these in our data and focus on the statements made by the members of the 112th Senate and House, during the period between January 2011 through August 2012.

The individual attributes of the members of Congress, such as party and district, were retrieved using Sunlight Congress API.[3] We use the DW-NOMINATE scores for the U.S. Congress (Carroll, Lewis, Lo, Poole, & Rosenthal, 2009) as measures of legislators' ideological locations.[4] The estimates are based on the history of roll call votes by the members of Congress and have been widely used in political science studies and related fields. Based on their method, each member's ideological point is estimated along two dimensions. Previous research has shown that the first dimension reveals standard left–right or economic cleavages and the second dimension reflects social and sectional divisions.

Randomized N-*Gram Extraction*

Computational challenge: n-*gram explosion.* A naive method to observe authors using overlapping *n*-grams would be to scan all documents and count each author's use of every *n*-gram that appears in the corpus. The problem with this approach is that the number of unique *n*-grams increases drastically as *n* increases.

Figure 3 shows the growth of *n*-grams in our data set. The total number of unique *n*-grams increases much faster for longer *n*-grams. On average, members of the U.S. Congress introduced 43 K new words, 6 M new trigrams, and 19.5 M 5-grams per year. The growth curves of *n*-grams are roughly linear, but curves of longer *n*-grams have larger slopes. Table 1 shows the number of total and unique *n*-grams in our data set for $n \in \{2, \ldots, 7\}$. The total number of *n*-grams is roughly the same for different *n*'s, but the total number of unique *n*-grams for larger *n*'s is much higher. The fraction of

[2]http://www.votesmart.org
[3]http://services.sunlightlabs.com/docs/Sunlight_Congress_API/
[4]Estimates for the 112th Congress are available at: http://www.voteview.com/housevotes.asp

DOI: 10.1002/asi

unique n-grams quickly approaches 1 when n increases. Moreover, as shown in Table 2, the average fraction of unique 1-grams (words) in a document is around 50%, the fractions of unique 3-grams and 5-grams per document are around 95% and 98–99% (similar statistics appear in documents in other years). This indicates that the fraction of unique n-grams in a document quickly approaches 1 when n increases, and is greater than 95% when $n > 3$. As a consequence, the total memory needed for counting the n-gram frequency approaches the length of corpus when n increases, and most of the counters are used to store unique n-grams.

Extraction technique. We developed a randomized dictionary algorithm to tackle this computational issue. The idea is to iteratively sample (with replacement) a pair of documents from the corpus and only count n-grams that appear in both documents in each iteration. The algorithm will always find n-grams that appear in more than one document, and thus no memory is wasted storing unique n-grams. The performance improvement is significant—for example, in the case of 7-grams, the nonunique fraction is less than 18%.

The question remaining for using this randomized dictionary algorithm is *how many samples are needed to find all sufficiently frequent n-grams*. The number of samples can be derived as follows. Suppose our goal is to find all n-grams that appear in at least p fraction of documents. Let w be one of such n-grams. In our algorithm, the probability of missing w in one iteration is $1 - p^2$. After k iterations, the probability of missing w is $(1 - p^2)^k$. Thus, the probability of finding w in k samples is $1 - (1 - p^2)^k$. Based on the inequality $1 - x \le e^{-x}$, we have:

$$\Pr[\text{find } w \text{ in } k \text{ samples}] = 1 - \left(1 - p^2\right)^k \le 1 - e^{-kp2}$$

By making k larger, the probability of successfully finding w can be made arbitrarily close to 1. Hence, the number of samples needed to find w with success rate r is $k > (1/p^2)ln(1/(1 - r))$. The time complexity for sampling is $O(1/p^2)$, which means the sampling cost depends only on the document fraction of targeting n-grams, and is independent of the size of the corpus or the length of n-grams. For example, to find all n-grams that appear in 1% of total documents, approximately 20,000 samples are required in order to guarantee a 90% success rate.

We summarize the randomized dictionary algorithm in Table 3. For comparing documents from different authors, Step 3 can be further restricted as: randomly draw a document pair (*Da, Db*) from *D* with the constraint *author*(*Da*) ≠ *author*(*Db*). Step 8 is used to remove n-grams without sufficient frequency, but for the purpose of our study this step can be omitted.

The problem formulation falls within the setting of the frequency items problem—of which the goal is to find all items or sets of items whose frequency exceeds a specified fraction of the total amount of input data (Cormode & Hadjieleftheriou, 2009). In particular, our algorithm has a connection to the two-pass randomized sampling-based approach (Toivonen, 1996). The important difference between our method and the prior work is that we sample a document pair and add the candidate n-grams in each sampling iteration, rather than sample a document per sampling iteration and determine the candidates after the sampling pass. This is based on the key observation that the number of n-grams is enormous but most of them are unique. Our setting considers each n-gram as a single item, not a set of unigrams, which is also different from the frequent itemsets setting that usually breaks the relationship between items in the extracted itemsets.

Validation. To evaluate the performance of the randomized dictionary algorithm we use the naïve method to extract all 3-grams and 5-grams in the corpus to create ground truth from which to measure the success rate of the algorithm. Then we evaluate the success rate of the randomized dictionary with respect to different targeting document fractions (p). Figure 4 shows the results of this computational experiment. The gray (solid and dashed) curves are theoretical lower bounds in terms of number of samples. The colored curves are experimental results where a point (x, y) indicates success rates y for n-grams with document fraction $p = x$ (i.e., those that appear in at least x fraction of documents). The computational experiments are repeated 30 times for each k and the means and standard deviations are marked with error bars. The figure shows that, for all $k = 200, \ldots, 200{,}000$ samples, the empirical success rates are bounded below by the

TABLE 1. Fraction of unique n-grams.

n	Total	Unique	Fraction
1	236,071,504	463,101	0.002
2	235,704,167	12,349,893	0.052
3	235,336,830	57,379,878	0.244
4	234,969,494	116,653,788	0.496
5	234,602,158	159,103,419	0.678
6	234,234,830	181,644,125	0.775
7	233,867,506	192,138,833	0.822

TABLE 2. Average fraction of unique n-gram per document.

Year	Max. doc size	Avg. doc size	Avg. 1-gram	Avg. 3-gram	Avg. 5-gram
2008	11,512	680.841	0.505	0.959	0.989
2009	11,819	790.346	0.506	0.959	0.990
2010	11,538	560.536	0.530	0.963	0.990

DOI: 10.1002/asi

corresponding theoretical curves (in gray). This indicates that the algorithm performs well, as expected.

Author-Author Similarity Networks

Our hypotheses are at the level of relationships between authors. Below we describe the author–author networks used in this study.

N-*gram similarity.* *N*-gram similarity was calculated for each pair of members of Congress for all documents in the data set that were not exact copies or "joint press releases" where multiple authors issued completely identical documents. Each member received an *n*-gram similarity score with each other member at each of the six lengths of *n*-gram for which matches were tracked. The number for this pair represents the extent to which these two members issued statements that tended to share *n*-grams of that length. This score was calculated as the total number of *n*-gram overlaps normalized by the number of statements made by the member. Specifically, for each author i we construct an *n*-gram vector $\overline{w_i}$ where each entry w_{ij} represents the number of occurrences of an *n*-gram j appear in the documents authored by i. The normalized vector is computed by $\overline{w_i} \Big/ \sum_j w_{ij}$, and at each *n*-gram length, the pairwise *n*-gram similarity is computed as the cosine similarity of the pair of normalized *n*-gram vectors. The results of these calculations are six directed 534×534 networks reflecting author similarity, one for each length of *n*-gram analyzed. The aggregated text similarities scores showed a highly skewed distribution, and thus scores were log-transformed to create a more appropriate distribution for a correlation analysis.

TABLE 3. Randomized dictionary algorithm.

Input: a set of document D, a document fraction threshold p and targeting success rate r
Output: the collection W consisting of frequent *n*-grams and their frequencies
Method:

```
W←∅
/* sampling*/
Repeat k times with k > (1/p^2)ln(1/(1 − r))
randomly draw a document pair (D_a, D_b) from D
for each n-gram w in the overlapping text D_a ∩ D_b
W ← {w} ∪ W
/* counting */
For each d ∈ D
  for each w ∈ W, increment count(w) by 1
W ← W − {w ∈ W : count(w) < p|D|} /* optional */
Output W and {count(w) : w ∈ W}
```

Topic similarity. Hypothesis 1 proposes that when authors share subject matter their statements will overlap more in short *n*-grams than in long *n*-grams. To test this hypothesis we use standard topic models (LDA; Blei et al., 2003) to assign topical distribution to a randomly sampled 20% of documents in our corpus. The topic model was based on bag-of-unigram features. For each document pair we compute similarity based on topic distribution. In the topic

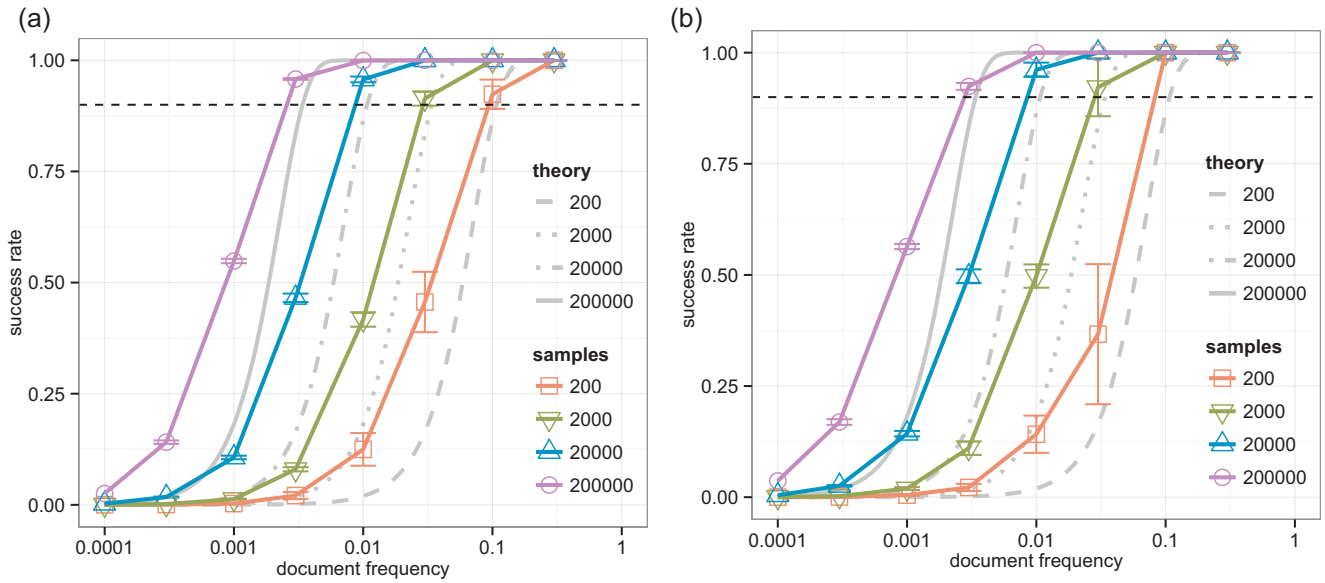


FIG. 4. Sampling success rates. The success rates (y-axis) over *n*-gram document frequencies (x-axis), by using different number of samples (k = 200, . . . , 200,000). The gray curves are theoretical lower bounds for $1 - exp(-kx^2)$ different samples k. The colored lines are experimental results on success rates y for (a) 3-grams and (b) 5-grams with document frequency x. The empirical success rates for samples are bounded below by the theoretical lower bound for k. The experiments were conducted 30 times for each k and the means and standard deviations are reported. Error bars show the standard deviations. The horizontal dashed line indicates a 90% success rate can be guaranteed at all levels of document frequency. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

modeling, the number of topics K can be determined by predictive likelihood (i.e., perplexity; Blei et al., 2003) based on cross-validation. We explored the topics with $K = 50$, 100, and 150 and found $K = 100$ has the highest perplexity. Note that in our experiment we are not interested in finding the exact optimal number of topics because the topics are used to build a length-K vector for the posterior topic proportions of each document. The topical similarity is defined as the cosine similarity between the topic proportion vectors of the pairs of documents. These similarities of document pairs are then converted into edge weights in the author-to-author networks based on the document authorship.

DW-Nominate similarity. Hypothesis 2 stated that when the authors of documents shared a rhetorical goal their documents would share moderate length n-grams. It is difficult to obtain ground truth for an author's intentions on a particular document. Nonetheless, it is possible to measure the overall rhetorical goals of U.S. Congress members through the policy preferences expressed in their votes as captured by the first dimension of their DW-Nominate scores (Carroll et al., 2009). That is, the more similar two members are in their voting record, the more similar we assume are the persuasive goals of their discourse. The maximum score for DW-Nominate is 1, indicating most conservative, and the minimum score is −1, indicating most liberal. For each pair of members the difference of their DW-Nominate scores was subtracted from the maximum difference (2) to transform it into a similarity score. The DW-Nominate similarity network is thus a 534×534 undirected network where scores range from 2 (*most similar*) to 0 (*least similar*).

Chamber similarity. Hypothesis 3 stated that when the authors rely on common sources from which they may copy text this similarity will be most visible long n-gram lengths. To test Hypothesis 3 we require a relation that does not map directly onto ideology, such that individuals could share ideology (but be less likely to share the source) or share the source (but be less likely to share ideology). We elected to use the chamber structure of the Congress. Both the House and the Senate contain a range of member DW-Nominate scores. There is also an existing mechanism through which members from the same chamber can share sources. A common way for members to solicit sponsors for legislation they propose is through "Dear Colleague" letters (Krutz, 2005; Petersen, 2005), in which they write several paragraphs justifying why their bill should become law. Although these letters can be circulated to all members, they are often sent only within one chamber (Petersen, 2005), and, since most bills never make it to the next chamber for consideration, most sponsors are gathered from within a member's own chamber only (Krutz, 2005).

Analysis

Correlations between n-gram networks and other author–author networks were tested using the quadratic assignment procedure (QAP) (Krackhardt, 1987). The QAP uses the Pearson correlation coefficient between the dyadic scores in each network but calculates the significance based on permutations to control for network interdependency. Since both DW-Nominate similarity and chamber membership are compared to the same networks using the same technique, their partial correlation with n-gram networks was also tested using network regression based on the QAP (Dekker, Krackhardt, & Snijders, 2007). To assess the significance of the differences in the n-gram correlations, we generate bootstrap samples with 1,000 replicates from each of the n-gram networks and calculate the pairwise Tukey's Honest Significant Differences test.

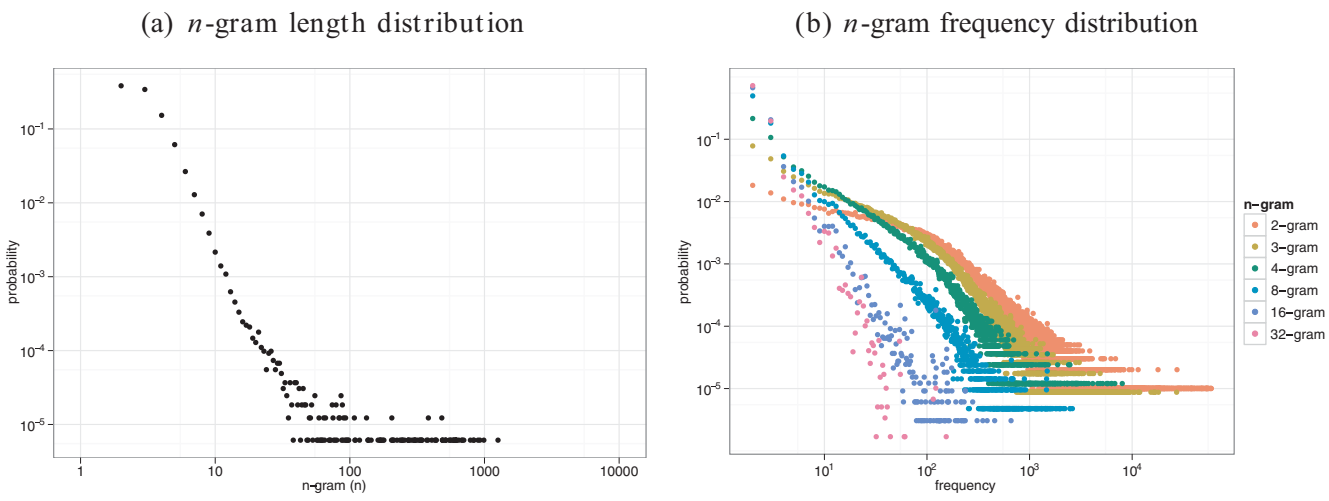


FIG. 5. *N*-grams found by randomized dictionary. The n-grams found by randomized dictionary exhibit heavy-tailed length and frequency distributions. (A) The length distribution indicates the existence of very long n-grams. (B) The frequency distribution for n-grams of different lengths. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

Results

Descriptive Analysis

Figure 5 shows the length and frequency distributions of the *n*-grams found by a randomized dictionary algorithm. The length distribution is heavy-tailed, indicating the existence of very long *n*-grams. The frequency distribution for *n*-grams of different lengths is also heavy-tailed. This means there are a few *n*-grams occurring very frequently, and many *n*-grams occurring rarely. In uni-grams, the most frequent ones such as "and" and "the" are often referred to as "stopwords" and removed before more sophisticated text analysis as they do not carry useful semantic information. Figure 5 indicates that, even in the set of very long *n*-grams, there exist "stop-*n*-grams" that occur extremely frequently compared to the rest of the set. Inspection of these "stop-*n*-grams" showed that they tend to be procedural language used within the congressional forum, such as "I yield back the remainder of my time," that appear in a very large number of documents. Like stopwords, these stop-*n*-grams need to be removed before further analysis. We describe the procedure to filter such stop-*n*-grams in the Appendix.

Table 4 shows additional examples of sampled short (bi-grams), moderate (4-grams) and long *n*-grams (for $n \geq 16$) in three broad categories: health, taxes, and jobs.

To examine the evidence for the theorized "bumps" in the distribution of observed *n*-gram similarity we randomly sample 20% of the documents from the corpus and then compute Jaccard similarity of the *n*-gram sets for $n \in [1, 2^{10}]$ in each pair of documents (excluding those authored by the same individual). Figure 6 shows the distribution of similarities by *n*-gram length. A multimodal distribution appears in both (a) similarity calculated based on the Jaccard similarity measure of the *n*-gram sets, and (b) number of matched *n*-grams in the pairs of documents by distinct authors. Consistent with expectations, we observe three distinct regions of similarity, one for short *n*-grams, one for moderate *n*-grams, and one for long *n*-grams. This provides evidence that there are distinct similarity generating processes happening at different *n*-gram lengths and suggests that any analysis that looks at only one length of *n*-gram ignores important information.

TABLE 4. Example of short and long *n*-gram (numbers in parentheses indicate the sampled document frequency of the particular *n*-grams).

Length	*N*-gram examples
	Health
2	Health care (10,334); health insurance (2,933); public health (1,934); health service (1,054); affordable health (881); mental health (868); health reform (750); health benefit (696); women's health (597)
4	Health and human services (1,796); the health care law (1,169); our health care system (668); access to health care (620); the president's health care (587)
≥16	Who have health insurance through their employer or the market for private insurance eliminating health care tax credits for up to (11); bans insurance companies from imposing lifetime dollar limits on health benefits freeing cancer patients and individuals suffering from other chronic (10); on health benefits freeing cancer patients and individuals suffering from other chronic diseases from having to worry about going without (9); health care coverage to 47 million americans including 39 million seniors and 8 million people under 65 (9); to community health centers hospitals doctors and other healthcare providers in the district to improve the community's health (9)
	Tax
2	Tax cut (3,208); taxpayer dollar (2,757); tax increase (2,381); american tax (1,994); tax reform (1,611); income tax (1,224); tax relief (1,052); tax hike (1,014); corporate tax (802); tax loophole (782)
4	The payroll tax cut (783); the bush tax cuts (448); to raise taxes on (297); the middle class tax (285); the payroll tax holiday (257)
≥16	Tax credit of up to $5 600 for hiring veterans who have been looking for a job for more than six months (20); tax credit of up to $9 600 for hiring veterans with service-connected disabilities who have been looking for a job for more than six (17); than 500 employees to take a tax deduction equal to 20% of their active business income (13); included in the payroll tax cut extension bill passed in december didn't give him enough time to review the project in fact (12); expands much needed lending to millions of small businesses and offers tax incentives to help small businesses grow hire and fuel (12)
	Job
2	Create job (8,554); american job (4,063); job creator (3,761); new job (3,222); job growth (2,830); their jobs (1,851); more job (1,802); jobs bill (1,726); million job (1,679); jobs act (1,608)
4	To create jobs and (1,430); of thousands of jobs (757); economic growth and job (668); jobs here at home (417); that will create jobs (385)
≥16	$5,600 for hiring veterans who have been looking for a job for more than six months (28); $9,600 for hiring veterans with service-connected disabilities who have been looking for a job for more than (24); to begin the federal employment process prior to separation in order to facilitate a truly seamless transition from the military to jobs (23); been looking for a job for more than six months as well as a $2 400 credit (23)

Hypothesis Tests

Hypothesis 1 stated that authors who shared subject matter would share short *n*-grams but not long *n*-grams. Figure 7 and Table 5 show the first-order dyadic correlations between the author–author similarity at different *n*-gram lengths and Topic similarity, DW-Nominate similarity and Chamber similarity, respectively. The three types of similarities over different *n*-gram lengths are shown as three curves in Figure 7. Table 6a–c shows the results of our statistical tests for different lengths of *n*-gram correlations with Topic similarity, DW-Nominate, and Same Chamber, respectively. Each table corresponds to a curve shown in Figure 7. In each of the three tables (a–c), each row indicates which pair of correlations is being compared. For example, the first row in Table 6a is the (QAP) correlation of (3-gram, topic) minus the correlation of (2-gram, topic), and the columns show the difference in the observed means of the

DOI: 10.1002/asi

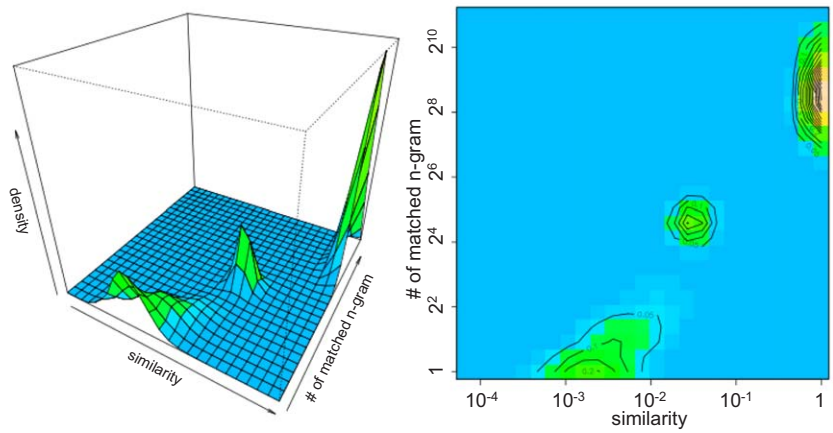


FIG. 6. Empirical evidence that different social processes compete to produce "bumps" in the distribution of observed *n*-gram similarity. Consistent with expectations, we observe three distinct regions of similarity, one for short *n*-grams, one for moderate *n*-grams, and one for long *n*-grams. This provides evidence that there are distinct data-generating processes happening at different levels of *n*-gram similarity. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

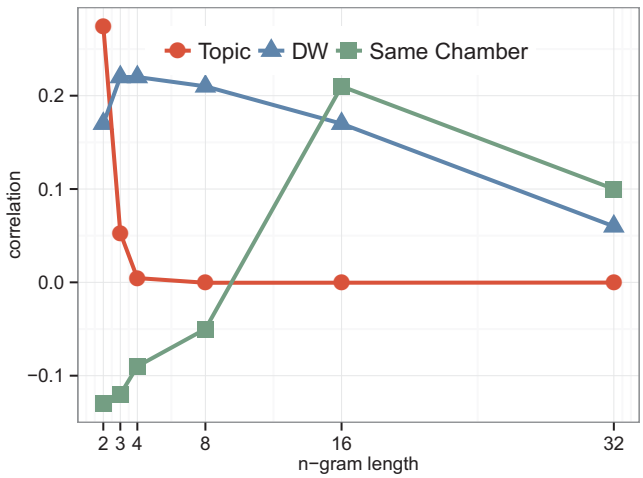


FIG. 7. First-order dyadic correlation over different *n*-gram lengths. The three curves show the first-order dyadic correlations between the author–author similarity at different *n*-gram lengths and Topic similarity, DW-Nominate similarity, and Chamber similarity, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

pairwise correlations, the lower and upper end points of the interval, and the *p*-value after adjustment for the multiple comparisons. The difference is significant if the *p*-value is small. The results indicate that (a) the highest correlation between topic and *n*-gram similarities appears at $n = 2$, and (b) the correlations drop gradually at $n = 3, 4, 8$ (no significant difference after $n > 8$). This is consistent with Hypothesis 1. Authors who share topics are similar in shorter *n*-grams but are not particularly likely to be similar in longer *n*-grams.

Hypothesis 2 stated that when the authors shared a rhetorical goal, their documents would be similar moderate-length *n*-grams. Table 5 shows the results of the QAP tests. The relationship between shared rhetorical goals and semantic similarity is statistically significant at each length of *n*-gram. Table 6b shows the highest correlation between DW and *n*-gram similarities appears at $n = 8$, and gradually tapers off on both ends.

Hypothesis 3 stated that when the authors construct their documents from common sources, their documents can be similar even at long *n*-gram lengths. Table 5 shows the results of the QAP tests. The relationship between working in the same chamber and text similarity is statistically significant at each length of *n*-gram. However, at short and moderate *n*-gram lengths ($n \leq 8$) the correlation is negative. At $n = 16$ and $n = 32$, the correlation is positive. The highest correlation between same-chamber and *n*-gram similarities appears at $n = 16$ and gradually tapers off on both ends. In other words, the test results provide statistical significance for the peaks in each of the three curves in Figure 7.

To compare the relative association between DW-Nominate similarity and chamber membership and *n*-gram similarity at different levels of *n*, we run a network regression that yields the partial correlation coefficient for each of these networks relative to the *n*-gram similarity network (Table 7). The results are consistent with the first-order correlation analyses. The DW-similarity coefficient is significant and positive in each regression. The shared chamber membership is negative for $n = 2$ thru 8, and is not significant for $n = 3$ or $n = 8$. At $n = 16$, however, the coefficient is positive and significant. At $n = 32$, the coefficient is positive, significant, and larger than the coefficient for DW-nominate similarity (DW-nominate similarity is on a 0–2 scale, but 2× this coefficient is less than the coefficient for shared chamber at $n = 32$). In other words, at the long *n*-gram lengths of 16 and 32, the mimetic processes begins to compete with and then overtake shared rhetorical goals as the process producing similarity.

DOI: 10.1002/asi

TABLE 5. First-order correlation (QAP) with *n*-gram networks. (*$p < .01$, **$p < .01$, ***$p < .001$, all significance estimates calculated using QAP.)

	N-gram length					
	2	3	4	8	16	32
Topic	0.27***	.05***	0.005	0.000	0.000	0.000
DW	0.17***	0.22***	0.22***	0.21***	0.17***	0.06***
Chamber	–0.13***	–0.12***	–0.09***	–0.05***	0.21***	0.1***

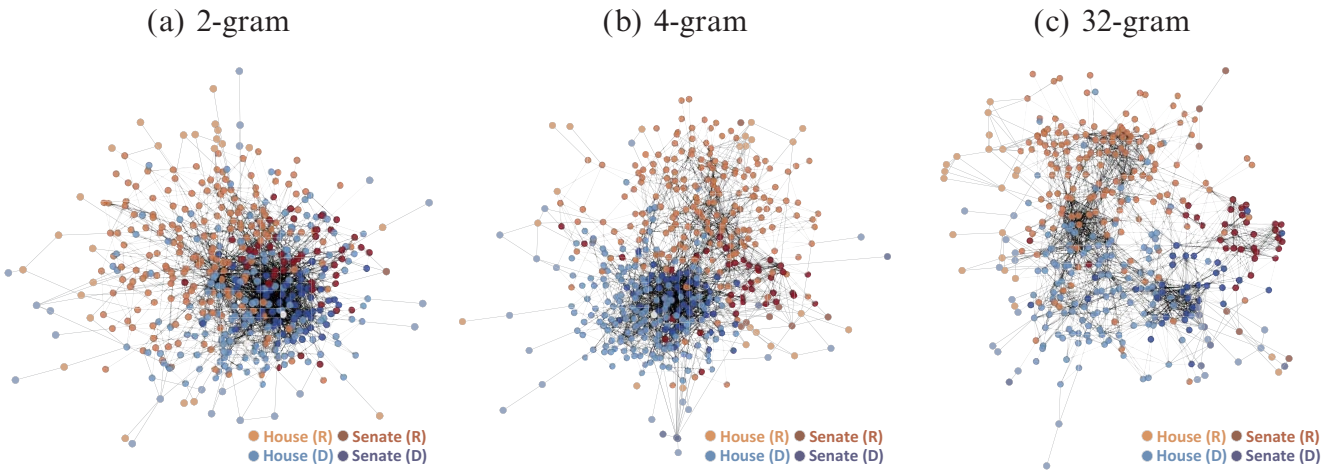


FIG. 8. Congressional social networks constructed based on co-mentioning (a) 2-grams, (b) 4-grams, and (c) 32-grams in their public statements. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

In Figure 8, we visualize the congressional social networks based on the *n*-gram similarity of different lengths. In the networks, each node represents a member of Congress and each edge represents *n*-gram similarity based on length *n*. For ease of visual comparison, all three networks are constructed to have approximately the same number of edges (no edges are removed in the statistical analyses). Specifically, we keep removing low-similarity edges in each network until the final edge density reaches 0.1 (this threshold is empirically determined in order to balance between data filtering and visual clarity). The figure shows that ideological differences are apparent at moderate n-gram similarity but the chamber structure of the Congress begins to emerge at longer *n*-gram similarity.

Discussion

Applied and Theoretical Implications

We provide evidence that different shared processes of authorship, reflecting different kinds of social relationships between authors, can be detected by examining authors' semantic similarity at different *n*-gram lengths. We first show that although most existing approaches assume only one authorship process and assume that they have chosen an appropriate *n*-gram length, there is theory (Figure 2) and evidence (Figure 6) that the full spectrum of text similarities contains systematic information about authors and their relationships. We also show that similarities at specific *n*-gram lengths have detectable correspondence to specific nontextual attributes and contexts such as voting ideology and the social context (chamber) in which authors work.

These findings suggest a number of applications for designing information systems to support social network analysis and related social science investigations (Karanasios et al., 2013). One application of particular importance is in the parsing of strategic language use in text. Strategic (rhetorical) language is largely ignored by, and therefore confounds, many existing techniques. Our analysis suggests that *n*-gram-based similarity measures can be used to identify and potentially "control out" the influence of this process, returning discourse that meets the assumptions of other techniques.

Distinguishing subject sharing and goal sharing is particularly important to studies in media and politics (e.g., Glazier & Boydstun, 2012) and organizational change (e.g., Rice, 2005) where changes in semantic similarity are used to measure changes in consensus. The results of our study suggest that consensus in texts can be both conceptualized and measured in two distinct ways: as consensus about what to talk about versus consensus about how to think about what is being discussed. These might be measured independently and tracked over time.

DOI: 10.1002/asi

TABLE 6. Pairwise comparisons using Tukey's Honest Significant Differences test.

	Diff	Lower	Upper	Adj *p*-value
(a) Topic				
3-2	–0.281	–0.283	–0.280	0.000
4-2	–0.349	–0.350	–0.348	0.000
8-2	–0.358	–0.359	–0.356	0.000
16-2	–0.358	–0.359	–0.356	0.000
32-2	–0.358	–0.359	–0.356	0.000
4-3	–0.068	–0.069	–0.066	0.000
8-3	–0.076	–0.078	–0.075	0.000
16-3	–0.076	–0.078	–0.075	0.000
32-3	–0.076	–0.077	–0.075	0.000
8-4	–0.009	–0.010	–0.007	0.000
16-4	–0.009	–0.010	–0.007	0.000
32-4	–0.009	–0.010	–0.007	0.000
16-8	0.000	–0.001	0.001	1.000
32-8	0.000	–0.001	0.001	1.000
32-16	0.000	–0.001	0.001	1.000
(b) DW				
3-2	0.033	0.028	0.038	0.000
4-2	0.043	0.038	0.048	0.000
8-2	0.048	0.043	0.053	0.000
16-2	–0.014	–0.019	–0.009	0.000
32-2	–0.098	–0.103	–0.093	0.000
4-3	0.010	0.005	0.014	0.000
8-3	0.015	0.010	0.020	0.000
16-3	–0.047	–0.052	–0.042	0.000
32-3	–0.131	–0.136	–0.126	0.000
8-4	0.005	0.000	0.010	0.030
16-4	–0.057	–0.062	–0.052	0.000
32-4	–0.141	–0.146	–0.136	0.000
16-8	–0.062	–0.067	–0.057	0.000
32-8	–0.146	–0.151	–0.141	0.000
32-16	–0.084	–0.089	–0.079	0.000
(c) Same chamber				
3-2	0.008	0.002	0.013	0.002
4-2	0.043	0.037	0.048	0.000
8-2	0.088	0.082	0.093	0.000
16-2	0.328	0.322	0.333	0.000
32-2	0.261	0.256	0.267	0.000
4-3	0.035	0.030	0.041	0.000
8-3	0.080	0.075	0.086	0.000
16-3	0.320	0.315	0.326	0.000
32-3	0.254	0.248	0.259	0.000
8-4	0.045	0.040	0.051	0.000
16-4	0.285	0.279	0.290	0.000
32-4	0.219	0.213	0.224	0.000
16-8	0.240	0.234	0.245	0.000
32-8	0.174	0.168	0.179	0.000
32-16	–0.066	–0.072	–0.061	0.000

Our emphasis on and evidence for the importance of distinguishing mimetic processes in such systems is also important. To date, work on mimetic processes has largely taken place in a parallel but separate track from other studies of textual similarity. It may be possible to use *n*-gram similarity to answer basic questions about mimetic processes such as whether individuals tend to copy others who use similar, but not copied, words and phrases.

More broadly, our results suggest new lines of theoretical and methodological investigation. Theoretically, more attention is needed to distinguish different authorship processes. For example, shared subject matter and shared goals appear to be a process of structural equivalence, in which authors independently choose similar constructions, whereas shared sources appear to be a relationship of social influence.[5] Our study also suggests the value of testing generic text analysis techniques on corpuses produced by authors with known attributes. Our technique focuses on features—shared *n*-grams—that can be calculated across any set of documents without knowledge of the native language. Rather than relying on dictionaries or the interpretation of coders, we test our hypotheses through the use of other author behavioral data. Future work may consider adapting this approach to other settings where texts are difficult to interpret and social relationships of interest are difficult to observe, such as discovering terrorist networks (Danowski, 2011).

Limitations

Our analysis is a first attempt to distinguish disparate social relationships from text archives. In general, the ground truth for connecting texts with possible social processes is difficult to obtain. In our case, because our analysis relies on the known attributes of political actors, there might be concerns about generalizability to nonpolitical contexts. One way to address this is to test these hypotheses on corpuses from other domains. Sports reporting in local newspapers may provide a good case, as sports writing describes events for which there is a well-known vocabulary but does so from a particular point of view. Sportswriters also share sources in the form of press conference interviews. Essays written by undergraduates for different assignments and under different group conditions might also be a useful place to run tests.

We also do not intend to suggest that text similarity as a function of *n*-gram is the only dimension for distinguishing social processes. Further investigation of possible dimensions may prove fruitful for areas including document retrieval, natural language processing, author attribution, and social network analysis.

Conclusion and Future Work

In this paper we present an integrated approach for analyzing text similarity in which multiple processes are considered. Drawing on communication theory we identify three broad processes by which two documents might be detectably similar in their textual features based on the length of the *n*-grams in which document texts overlap. We employ a novel technique for detecting *n*-gram overlaps of

[5]We are grateful to the anonymous reviewer for this suggestion.

DOI: 10.1002/asi

TABLE 7. Partial correlation (estimated using random sample of 100 members; significance level: *$p < .01$, **$p < .01$, ***$p < .001$).

	n-gram length					
	2	3	4	8	16	32
DW	0.555***	0.864***	0.859***	1.306***	2.668***	0.359
Same chamber	0.428***	0.122	0.679***	0.198	3.145***	0.010***
R-squared	0.057	0.057	0.062	0.068	0.073	0.06
Coefficient ratio (Abs)	1.295	7.058	1.264	6.580	0.848	0.445

arbitrary length to show that there exist distinct text signatures reflecting the operation of three different social processes within a corpus where information about the authors' subject interests, goals, and exposure to one another is available. This work takes the first step in computationally distinguishing textual similarity due to known theoretical processes and has important implications for modeling text and underlying social processes. Our future work includes probing the generalizability of our hypotheses in additional corpuses and investigating different features that may also distinguish social processes.

Acknowledgments

The authors gratefully acknowledge the support of MURI grant #504026, DTRA grant #509475, and ARO #504033. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors alone.

References

Adamic, L.A., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. ACM.

Adamic, L.A., Lento, T.M., Adar, E., & Ng, P.C. (2014). Information evolution in social networks. arXiv Preprint arXiv:1402.6792.

Aerts, D., & Gabora, L. (2005). A theory of concepts and their combinations i. Kybernetes, 34(1/2), 151–175.

Atteveldt, W.V., Kleinnijenhuis, J., & Ruigrok, N. (2008). Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. Political Analysis, 16(4), 428–446.

Barnett, G.A., Wigand, R.T., Harrison, R.P., Woelfel, J., & Cohen, A.A. (1981). Communication and cultural development: A multidimensional analysis. Human Organization, 40(4), 330–337.

Barrón-Cedeño, A., & Rosso, P. (2009). On automatic plagiarism detection based on *n*-grams comparison. In Advances in information retrieval (Vol. 5478, pp. 696–700). Springer Berlin Heidelberg.

Benford, R.D., & Snow, D.A. (2000). Framing processes and social movements: An overview and assessment. Annual Review of Sociology, 26(1), 611–639.

Blei, D. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. The Journal of Machine Learning Research, 3, 993–1022.

Boydstun, A.E., Glazier, R.A., & Pietryka, M.T. (2013). Playing to the crowd: Agenda control in presidential debates. Political Communication, 30(2), 254–277. doi: 10.1080/10584609.2012.737423

Bramsen, P., Escobar-Molano, M., Patel, A., & Alonso, R. (2011). Extracting social power relationships from natural language. Proceedings of ACL HLT, 773–782.

Bryden, J., Funk, S., & Jansen, V.A. (2013). Word usage mirrors community structure in the online social network Twitter. EPJ Data Science, 2(1), 1–9.

Carley, K., & Kaufer, D. (1993). Semantic connectivity: An approach for analyzing symbols in semantic networks. Communication Theory, 3, 183–213.

Carley, K.M. (1997). Extracting team mental models through textual analysis. Journal of Organizational Behavior, 18, 533–558.

Carroll, R., Lewis, J., Lo, J., Poole, K., & Rosenthal, H. (2009). Measuring bias and uncertainty in DW-nominate ideal point estimates via the parametric bootstrap. Political Analysis, 17(3), 261–275.

Chapel, G. (1976). Speechwriting in the Nixon administration. Journal of Communication, 26(2), 65–72.

Chong, D., & Druckman, J.N. (2007). Framing theory. Annual Review of Political Science, 10(1), 103–126.

Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22, 1–39.

Cordell, R. (in press). "Taken possession of": The reprinting and re-authorship of Hawthorne's "the celestial railroad" in the antebellum religious press. *Digital Humanities Quarterly*.

Corman, S.R., Kuhn, T., McPhee, R.D., & Dooley, K.J. (2002). Studying complex discursive systems. Human Communication Research, 28(2), 157–206.

Cormode, G., & Hadjieleftheriou, M. (2009). Finding the frequent items in streams of data. Communications of the ACM, 52(10), 97–105.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In Proceedings of the 21st International Conference on World Wide Web (pp. 699–708).

Danowski, J.A. (1982). Computer-mediated communication: A network-based content analysis using a CBBS conference. In M. Burgoon (Ed.), Communication Yearbook (Vol. 6, pp. 874–904). Beverly Hills, CA: Sage.

Danowski, J.A. (2011). Counterterrorism mining for individuals semantically-similar to watchlist members. In U.K. Wiil (Ed.), Counterterrorism and open source intelligence (Vol. 2, pp. 223–247). New York: Spring-Verlag/Wien.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

Dekker, D., Krackhardt, D., & Snijders, T. (2007). Sensitivity of mrqap tests to collinearity and autocorrelation conditions. Psychometrika, 72(4), 563–581.

DiMaggio, P.J., & Powell, W.W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American Sociological Review, 48, 147–160.

Doerfel, M.L., & Barnett, G.A. (1999). A semantic network analysis of the International Communication Association. Human Communication Research, 25(4), 589–603.

Doerfel, M.L., & Connaughton, S.L. (2009). Semantic networks and competition: Election year winners and losers in U.S. televised presidential debates, 1960–2004. Journal of the American Society for Information Science and Technology, 60(1), 201–218.

Eisenberg, E.M. (1984). Ambiguity as strategy in organizational communication. Communication Monographs, 51(3), 227–242.

Entman, R. (1993). Framing: Toward clarification of a fractured paradigm. Journal of Communication, 43, 51–58.
Gandy, O. (1982). Beyond agenda setting: Information subsidies and public policy. Norwood, NJ: Ablex.
Gentzkow, M., & Shapiro, J.M. (2010). What drives media slant? Evidence from U.S. daily newspapers. Econometrica: Journal of the Econometric Society, 78(1), 35–71.
Glazier, R.A., & Boydstun, A.E. (2012). The president, the press, and the war: A tale of two framing agendas. Political Communication, 29(4), 428–446.
Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2005). Integrating topics and syntax. Advances in Neural Information Processing Systems, 17, 537–544.
Grimmer, J. (2009). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. Political Analysis, 18(1), 1–35.
Grozea, C., Gehl, C., & Popescu, M. (2009). Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection. In 3rd Pan Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse (p. 10).
Hardisty, D.J., Johnson, E.J., & Weber, E.U. (2010). A dirty word or a dirty world?: Attribute framing, political affiliation, and query theory. Psychological Science, 21(1), 86–92. doi: 10.1177/0956797609355572
Hart, R.P., Childers, J.P., & Lind, C.J. (2013). Political tone: How leaders talk and why Chicago: University of Chicago.
Hassan, A., Fader, A., Crespin, M., Quinn, K., Monroe, B., Colaresi, M., & Radev, D. (2008). Tracking the dynamic evolution of participant salience in a discussion. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp. 313–320).
Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42(1), 177–196.
Jerit, J. (2008). Issue framing and engagement: Rhetorical strategy in public policy debates. Political Behavior, 30(1), 1–24.
Karanasios, S., Thakker, D., Lau, L., Allen, D., Dimitrova, V., & Norman, A. (2013). Making sense of digital traces: An activity theory driven ontological approach. Journal of the American Society for Information Science and Technology, 64(12), 2452–2467.
Kauffman, S. (1993). The origins of life: Self-organization and selection in evolution. New York: Oxford University Press.
Krackhardt, D. (1987). QAP partially as a test of spuriousness. Social Networks, 9(2), 171–186.
Krauss, R.M., & Fussell, S.R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. Social Cognition, 9(1), 2–24.
Krutz, G.S. (2005). Issues and institutions: "Winnowing" in the U.S. Congress. American Journal of Political Science, 49(2), 313–326.
Lakoff, G. (2004). Don't think of an elephant!: Know your values and frame the debate. White River Junction, VT : Chelsea Green.
Lammers, J.C., & Barbour, J.B. (2006). An institutional theory of organizational communication. Communication Theory, 16, 356–377.
Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211–240.
Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In Proceedings of the 15th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining (pp. 497–506).
Lin, Y.-R., Margolin, D., & Lazer, D. (2014). Tracing coordination and cooperation structures via semantic burst detection. EAI Endorsed Transactions on Collaborative Computing, 1(2), e7.
Loewenstein, J., Ocasio, W., & Jones, C. (2012). Vocabularies and vocabulary structure: A new approach linking categories, practices, and institutions. The Academy of Management Annals, 6(1), 41–86.
Lynch, E.B., Coley, J.D., & Medin, D.L. (2000). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. Memory and Cognition, 28(1), 41–50.
Lyon, C., Malcolm, J., & Dickerson, B. (2001). Detecting short passages of similar text in large document collections. In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (pp. 118–125).
Lyon, C., Barrett, R., & Malcolm, J. (2004). A theoretical basis to the automated detection of copying between texts, and its practical implementation in the ferret plagiarism and collusion detector. Plagiarism: Prevention, Practice and Policies.
March, J.G., & Simon, H.A. (1993). Organizations (2nd ed.). Cambridge, MA: Blackwell.
Margolin, D.B., & Monge, P. (2013). Conceptual retention in epistemic communities. In P. Moy (Ed.), Communication and Community (pp. 1–22). New York: Hampton Press.
McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. Journal of Artificial Intelligence Research, 30(1), 249–272.
McCombs, M., & Shaw, D. (1972). The agenda-setting function of mass media. Public Opinion Quarterly, 36, 176–187.
Meyer, J.W., & Rowan, B. (1991). Institutionalized Organizations: Formal structure as myth and ceremony. In W.W. Powell & P.J. DiMaggio (Eds.), The new institutionalism in organizational analysis (pp. 41–62). Chicago: University of Chicago.
Monge, P., & Poole, M. (2008). The evolution of organizational communication. Journal of Communication, 58, 679–692.
O'Connor, B. (2013). Learning frames from text with an unsupervised latent variable model. arXiv preprint arXiv:1307.7382.
Petersen, R.E. (2005). "Dear Colleague" letters: A brief overview. Washington, DC: Congressional Research Service. Retrieved from http://digital.library.unt.edu/ark:/67531/metacrs6161/m1/1/high_res_d/RS21667_2005Jan04.pdf
Quinn, K., Monroe, B., Colaresi, M., Crespin, M., & Radev, D. (2006). An automated method of topic-coding legislative speech over time with application to the 105th–108th US Senate. In Midwest Political Science Association Meeting.
Rice, R.E. (2005). New media/internet research topics of the association of internet researchers. The Information Society, 21(4), 285–299.
Rice, R.E., & Danowski, J.A. (1993). Is it really just like a fancy answering machine? Comparing semantic networks of different types of voice mail users. Journal of Business Communication, 30(4), 369–397.
Rogers, E.M., & Kincaid, D.L. (1981). Communication networks: Toward a new paradigm of research. New York: Free Press.
Rosch, E. (1978). Principles of categorization. In E. Rosch & B.B. Lloyd (Eds.), Cognition and categorization (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.
Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. Cognitive Psychology, 8, 382–439.
Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (pp. 487–494).
Simmons, M., Adamic, L., & Adar, E. (2011). Memes online: Extracted, subtracted, injected, and recollected. ICWSM 2011.
Tamburrini, N., Cinnirella, M., Jansen, V.A.A., & Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. Social Networks, 40, 84–89.
Toivonen, H. (1996). Sampling large databases for association rules. In VLDB (Vol. 96, pp. 134–145).
Wallach, H. (2006). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd International Conference on Machine Learning (pp. 977–984).
Wang, X., McCallum, A., & Wei, X. (2007). Topical *n*-grams: Phrase and topic discovery, with an application to information retrieval. In Seventh IEEE International Conference on Data Mining, 2007 (pp. 697–702).
Wisniewski, E.J. (1996). Construal and similarity in conceptual combination. Journal of Memory and Language, 35, 434–453.
Yuan, E.J., Feng, M., & Danowski, J.A. (2013). "Privacy" in semantic networks on Chinese social media: The case of Sina weibo: Privacy in semantic networks on Sina weibo. Journal of Communication, 63(6), 1011–1031.

Appendix: Stop *N*-grams

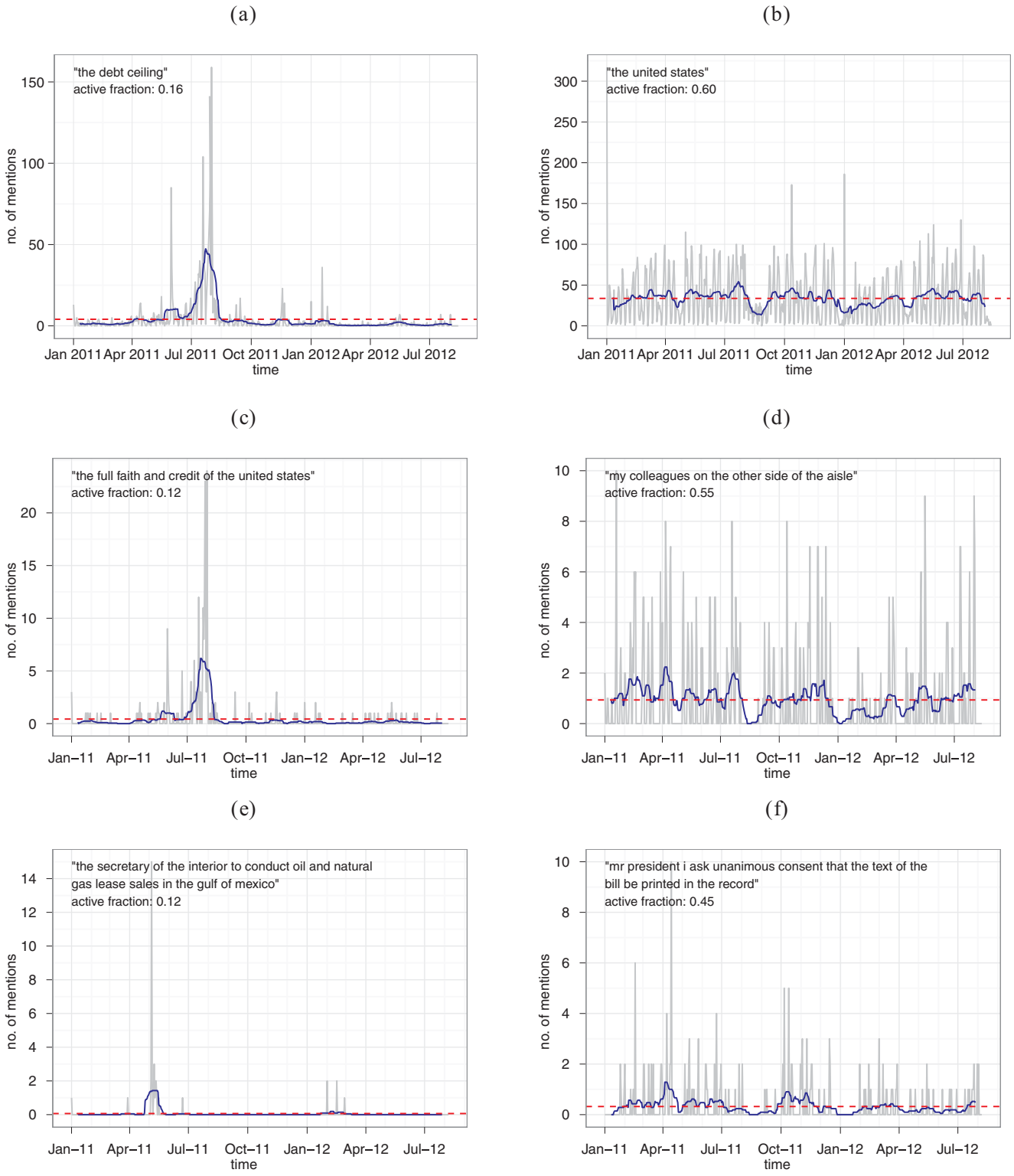


FIG. A1. *N*-gram active phase. The figure shows six examples of *n*-grams with active phase (A,C,E) and without active phase (B,D,F). Gray, blue, and red lines indicate number of documents mentioning the *n*-gram per days, the smoothed numbers, and the mean of the smoothed curves, respectively. The active fraction is computed based on the portion of above-mean in the smoothed curves. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

The idea is that such stop-*n*-grams are used for reasons independent of the document context, and hence can be differentiated by how they are used in the corpus over time. Figure A1 shows the time series of six example *n*-grams. For each *n*-gram, we show the number of documents mentioning the *n*-gram over days in a gray curve. We can see the *n*-grams such as "the united states" and "my colleagues on the other side of the aisle" are used almost all the time and are not necessarily determined by the context of the documents, while the *n*-grams such as "the debt ceiling" and "the full faith and credit of the united states" are used in relating to particular political issues. Note that the 3-gram "the united states" occurs frequently but the 9-gram "the full faith and credit of the united states" does not.

We differentiate those stop-*n*-grams by computing the active fraction of each *n*-gram. The active fraction α is defined as the above-mean ratio in the smoothed *n*-gram mentioning time series. Here we empirically chose a 21-day sliding window to smooth the time series. As can be seen in Figure A1, those with larger active fractions are less relevant to specific document context.

DOI: 10.1002/asi

PLOS ONE

RESEARCH ARTICLE

Timescales of Massive Human Entrainment

Riccardo Fusaroli[1,2]*, Marcus Perlman[3], Alan Mislove[4], Alexandra Paxton[5], Teenie Matlock[5], Rick Dale[5]

1 Interacting Minds Centre, Aarhus University, Aarhus, Denmark, **2** Center for Semiotics, Aarhus University, Aarhus, Denmark, **3** Department of Psychology, University of Wisconsin, Madison, Wisconsin, United States of America, **4** College of Computer and Information Science, Northeastern University, Boston, Massachusetts, United States of America, **5** Cognitive and Information Sciences, University of California Merced, Merced, California, United States of America

* fusaroli@gmail.com

CrossMark
click for updates

OPEN ACCESS

Citation: Fusaroli R, Perlman M, Mislove A, Paxton A, Matlock T, Dale R (2015) Timescales of Massive Human Entrainment. PLoS ONE 10(4): e0122742. doi:10.1371/journal.pone.0122742

Academic Editor: Christopher M. Danforth, University of Vermont, UNITED STATES

Received: October 17, 2014

Accepted: February 12, 2015

Published: April 16, 2015

Copyright: © 2015 Fusaroli et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the Danish Council for Independent Research (RF), the EUROCORES EuroUnderstanding program (RF), the Interacting Minds Center (RF) and NSF BCS-0826825 (RD). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

The past two decades have seen an upsurge of interest in the collective behaviors of complex systems composed of many agents entrained to each other and to external events. In this paper, we extend the concept of entrainment to the dynamics of human collective attention. We conducted a detailed investigation of the unfolding of human entrainment—as expressed by the content and patterns of hundreds of thousands of messages on Twitter—during the 2012 US presidential debates. By time-locking these data sources, we quantify the impact of the unfolding debate on human attention at three time scales. We show that collective social behavior covaries second-by-second to the interactional dynamics of the debates: A candidate speaking induces rapid increases in mentions of his name on social media and decreases in mentions of the other candidate. Moreover, interruptions by an interlocutor increase the attention received. We also highlight a distinct time scale for the impact of salient content during the debates: Across well-known remarks in each debate, mentions in social media start within 5–10 seconds after it occurs; peak at approximately one minute; and slowly decay in a consistent fashion across well-known events during the debates. Finally, we show that public attention after an initial burst slowly decays through the course of the debates. Thus we demonstrate that large-scale human entrainment may hold across a number of distinct scales, in an exquisitely time-locked fashion. The methods and results pave the way for careful study of the dynamics and mechanisms of large-scale human entrainment.

Introduction

Interest in the collective behaviors of complex systems composed of many agents has dramatically increased over the past couple of decades. This interest may stem in no small part from a new ability to measure and model collective behaviors. In a canonical case, Strogatz and Stewart [1] highlight firefly behavior as illustrative of fundamental principles underlying entrained systems [2, 3]. In parts of Southeast Asia, one may happen upon a sea of fireflies, in which each firefly's intrinsic oscillatory dynamics have become entrained to others around it. The result is

a large-scale collective behavior: The fireflies fire in sync in an impressive display brought on by subtle mutual influences. They are *entrained* in that they match their behavior to the temporal structure of events in the environment [4–6]. This process might involve elements of reciprocal influence between individual agents as in the case of the fireflies, or it might depend predominantly on external environmental events. The firefly model has inspired the investigation of entrainment across many physiological and technological phenomena, from neuronal firing to electric power networks [7]. However, it is still unclear how complex cognitive agents, such as human beings, might also exhibit patterns of large-scale entrainment.

In this paper we employ a series of massively shared media events to examine the entrainment of human collective attentional behavior at several time scales. We analyzed the three 2012 US presidential debates between Barack Obama and Mitt Romney—altogether watched by 192 million viewers—and the associated use of Twitter, a popular social media service. These events were thus (a) *shared at a massive scale*, and, via Twitter, (b) *induced the rapid spread of social behavior across a network of agents*. We time-locked the corresponding Twitter data with video of each debate to match precise behaviors in the debates with the second-by-second rate of tweets involving mentions of the candidates. With these two time series in hand, we examined whether human behavior is entrained at three different time scales: i) short-term entrainment to conversational dynamics; ii) slower entrainment to salient content of the debates; and iii) long-term entrainment to the duration of the debates. We define statistical models that can capture the aggregate tendencies of human behavior at these different scales, and test these on each debate to assess whether the effects generalize across them. The findings show massive behavioral entrainment in humans, which is intrinsically multi-scale and reproduces across events (the three debates).

A massively shared event: US presidential debates

There are good reasons to choose the US presidential debates as our arena for exploring large-scale human entrainment. Since the televised debates of Kennedy and Nixon in 1960, they have attracted the attention of a hundred million or more television viewers each election cycle. The enormous magnitude of public attention has turned the debates into major events in the US presidential elections, as candidates have the chance to sway millions of voters through the discussion of controversial issues and planned policies [8–10]. In addition to their massive television viewership, the most recent 2012 US presidential debates—between candidates Barack Obama and Mitt Romney—were notable in the extent to which viewers were not just passive spectators isolated in front of a television set. Through the use of social media like Twitter and Facebook, millions of viewers participated in a global dialogue in which they generated tens of millions of interactive messages in real-time response to the debates.

The presidential debates present many salient aspects to public attention. Commentary on the debates emphasizes the highly competitive conversational interactions, dense with retorts, reciprocal interruptions and struggles for keeping or taking the floor [11–14], with much space devoted to assessing which candidate acted most presidentially [15–20]. Other studies have emphasized the content of the debates and how candidates frame the issues that are discussed [10, 21, 22], not least indicating the role of debates in creating widespread memes [23]. Finally, the debates, as any other large event, have a natural development as they warm up, reach their peak and then fade as they lose their novelty [24].

A massively social behavior: the Twitter "gardenhose" stream

The recent development of massive social media networks yields a prime forum in which to examine the phenomenon of human collective entrainment. The use of social media technologies

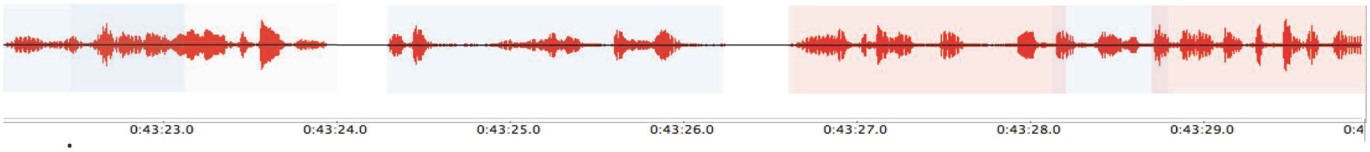
enables people to extend the existing constraints on the distance, timing, and connectivity of communication, facilitating the rapid cascade of information across the digital networks [25]. To investigate the impact of the presidential debates on human behavioral entrainment, we employed Twitter, a popular micro-blogging platform that launched in 2006. Twitter is widely used by marketers, public authorities, and the general public and has become a major mechanism for the rapid spread of information. As such it offers an unprecedented window into how large populations collectively experience and respond to a wide range of real-world events [26]. Researchers have used social media to describe—and sometimes anticipate—epidemics, earthquakes, stock options, the effect of time and weather on mood, reality show outcomes, and political elections [27–36]. Little is known, however, about the precise temporal dynamics through which the use of online social media reflects and interacts with the unfolding action of massively shared events. We chose to investigate these dynamics with Twitter because of the near-instantaneous nature of its message: Its short format (140 characters per message) and widespread integration with mobile devices facilitates fast messaging and reactions. Twitter provides a grasp of the precise temporal dynamics of how real-world events drive and resonate with human social behavior.

The dynamics of human collective entrainment: Three time scales

The purpose of this study was to explore human entrainment to the presidential debates through Twitter. Human social entrainment is arguably more complex than that of other species; events that reflect the sophisticated format of human interaction may shape entrainment in distinct ways. We thus hypothesized that the fine-grained conversational dynamics of the debates would directly drive and constrain Twitter discourse concerning the events at (at least) three time-scales of interest.

i) Interactional entrainment. We hypothesized that assertive behaviors—keeping the ground, interrupting the adversary, and so on—would strongly impact Twitter mentions and lead to higher rates of tweeting about the respective candidate. Thus candidates would generate tweets as they interrupted their opponent and asserted their turn, and they would continue to generate tweets for as long as they maintained the floor. This hypothesis was motivated by political and media studies suggesting that presidential debates are employed as heuristic or judgmental shortcut for viewers to assess future presidential performance [15, 16]. Both experimental settings and real life analyses showed that human beings tend to perceive and support leadership in individuals with extroverted personalities [37, 38] and relatedly in those who display assertiveness, boldness, initiative, proactivity, and risk-taking [39–42]. Corroborating this view is extensive coverage by the news media of the interactional style of the candidates—who behaved more presidentially, who was being defensive—with victory often defined in terms of the level of interruptions and direct confrontation [19, 20]. For the current paper, we did not consider emotional valence of attention. Instead, we hypothesized that display of assertiveness would capture the attention of viewers, irrespective of whether that attention was positive or negative. We leave to future studies the evaluation of judgments, emotional valence and more sophisticated clustering in response to assertiveness.

ii) Content entrainment. Besides this ebb-and-flow dynamics of interaction, debates are also rife with pointed or "salient" remarks that propagate through social media—often as "memes" that cascade through communications in forums like Twitter [43]. Indeed, viewers pay attention to the content of the debates, focusing their attention on particularly salient, amusing, or controversial elements [23]. We hypothesized that viewers would react to these salient events, however, in different ways than to the candidates' conversational dynamics. Content entrainment is likely to require more intensive cognitive processing and therefore happen



Obama:	(0:43:22.09)	His running mate –	(0:43:23.18)
Lehrer:	(0:43:22.37)	– And you and you don't support that?	(0:43:24.06)
Obama:	(0:43:24.42)	I don't. And and let me explain why	(0:43:26.17)
Romney:	(0:43:26.83)	Again, that's for future people –	(0:43:28.26)
Obama:	(0:43:28.08)	– I understand –	(0:43:28.91)
Romney:	(0:43:28.69)	– right, not for current retirees	(0:43:30.03)

Fig 1. Excerpt of the waveform and related transcript from the first presidential debate. Blue highlighting indicates Obama speech turns, red Romney's and grey Lehrer's (the moderator). The transcripts were retrieved from the National Public Radio website, cleaned and edited to better reflect the audio files. Start and end time of each speech turn as well as interruptions are aligned to the debate by combination of careful coding and automated processing.

doi:10.1371/journal.pone.0122742.g001

at longer time scales. Moreover, interest in salient events is expected to be partially self-sustaining: Once a high level of attention has been raised, the tweets produced will help maintain the attention on the topic, although the debate might have moved on.

iii) Long-term attention decay. Finally, despite the relatively longer scale of content entrainment, attention and interest are unlikely to be sustained for a long period, being subject to bursts and decays [25]. Therefore, we expected the general interest in the debate to decay after an initial burst, thus showing long-term attentional dynamics.

Below we demonstrate how the entrainment of Twitter behavior to the presidential debates is aptly characterized by these three time scales, both individually and in a multi-scale model.

Materials and Methods

Analysis of the debate

There were three 2012 US presidential debates between former Massachusetts Governor Mitt Romney and incumbent US President Barack Obama. The first took place on October 3rd at University of Denver, Denver, Colorado; the second on October 16th at Hofstra University, Hempstead, New York; and the third on October 22nd at Lynn University, Boca Raton, Florida. Each debate lasted about 90 minutes.

The audio recordings and transcripts of the three debates between President Barack Obama and Governor Mitt Romney were collected from National Public Radio (www.npr.org). The transcripts were cleaned and edited to better reflect the audio files. Through careful listening supplemented by an in-depth examination of the waveform and automated analysis of variations in pitch and intensity using Praat [44] and MATLAB (Mathworks Inc.), we individuated start and end time at a 10-millisecond scale for each speech turn as well as interruptions and a selection of salient moments discussed in popular media after the debate events (see Fig. 1). This was performed blind from any inspection of the Twitter data (see below). By identifying the precise timestamp of the debate onset, we time-aligned the Twitter data and the debate data (see Fig. 2).

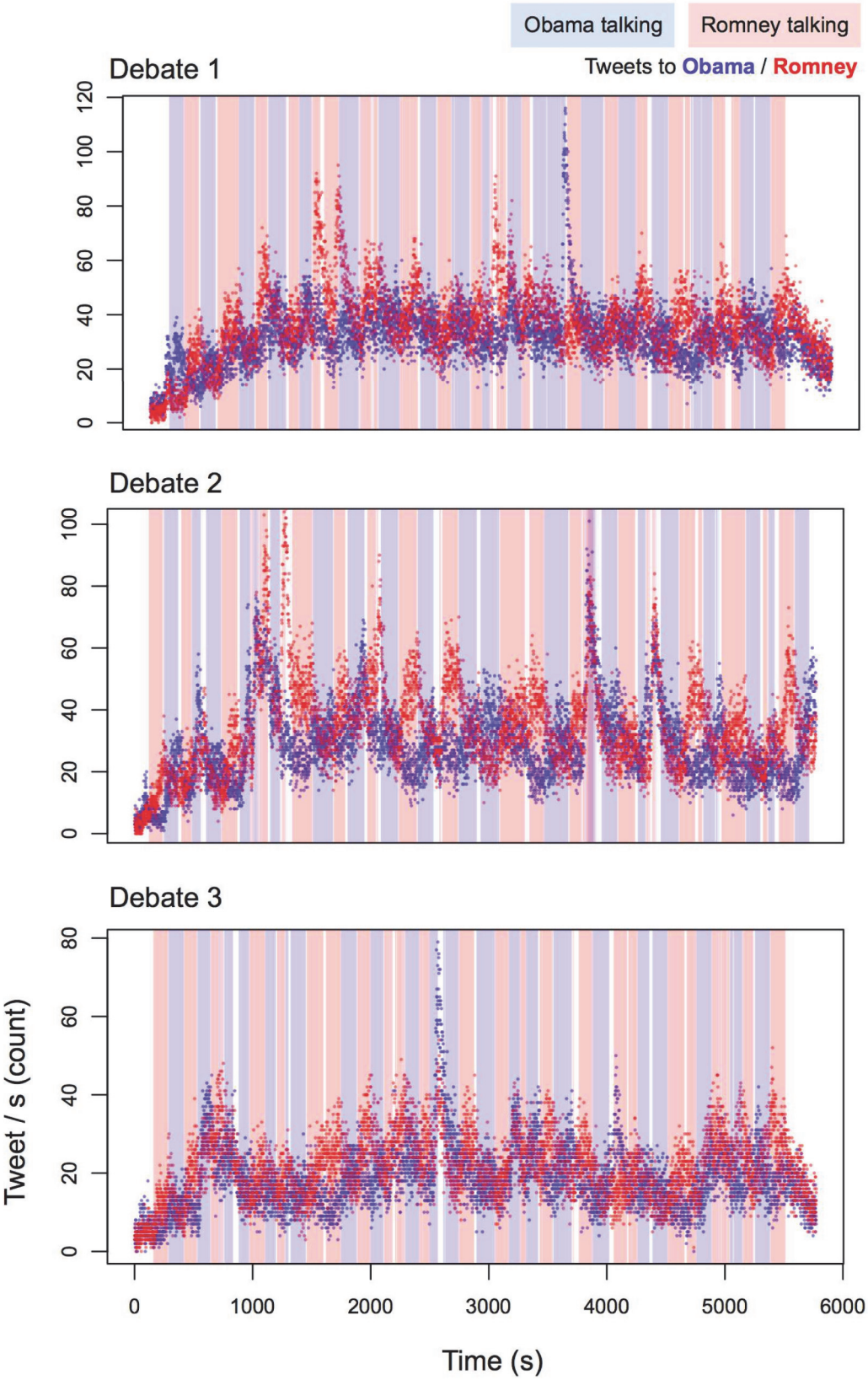

Obama talking
Romney talking
Debate 1
Tweets to Obama / Romney
Debate 2
Debate 3
Tweet / s (count)
Time (s)
0
20
40
60
80
100
120
1000
2000
3000
4000
5000
6000

Fig 2. Tweet rate and turn-taking during the presidential debates. Light red and blue rectangles are periods of time during which candidates were speaking during the debates. Darker red and blue dots represent per-second tweet rate mentioning the corresponding candidates. Visual inspect reveals relatively periodic patterns of Twitter mentions that seem to be cued by turn onset. Plots include both tweets and retweets in the tweet / s rate.

doi:10.1371/journal.pone.0122742.g002

Analysis of the tweets

The Twitter data consisted of a random sample of approximately 10% of all public tweets ("gardenhose" stream), collected during each 90-minute presidential debate. The Twitter data collected as part of this study currently resides on and is archived by co-author Mislove's research cluster at Northeastern University. While the data source (Twitter's streaming service) is publicly available, Twitter's Terms of Service prevent making the raw tweets available. Instead, we make the list of unique tweet identifiers (tweetIDs) publicly available (See S1 to S3 Datasets, also on http://www.ccs.neu.edu/home/amislove/obama-romney/), similar to previous studies of Twitter and Twitter-based benchmarks.

We filtered tweets to select only those that mentioned "Obama" or "Romney," either in the text or in their hashtag, and we excluded those containing URLs (to exclude spambot-generated tweets). This resulted in 713,642; 686,805; and 406,242 tweets for the first, second, and third debates, respectively. Each set of tweets was generated by a large number of unique user accounts: 442,368; 413,537; and 255,644 accounts respectively for each debate (see Table 1). "Retweets" (i.e., when another Twitter user merely reposted the original message) were omitted from the analysis, which ensured these patterns were not simply generated by repetitions of the same messages [45]. However, analyses including retweets show similar robust patterns (see S1 to S4 Figs.).

Statistical analysis of combined debate and Twitter data

We assessed the impact of debate events on human entrainment as measured in tweet rate per second at three key time scales. An overlay of tweet rate per second and turn-taking for each debate is shown in Fig. 2. We first modeled each scale individually. We then built a multiple regression model including all three time scales to assess their relative and overall predictive power for public attention. We hypothesized the three debates to display the same trends, with statistically significant attentional entrainment at the three time scales. To ensure effects were not driven by one debate only, we fitted each model to each single debate and report them separately. To further ensure the generality of our results after fitting the full multiple regression model to the first debate, we employed it to predict attentional entrainment in the other two debates. Full details of the analyses are reported in the following paragraphs. All models were developed with the lme4 and MuMIn libraries in R, and the R code is available in S1 Code.

Interaction: Turn-taking and interruptions. The first time scale was modeled on the turn-taking dynamics, using number of tweets per second (measured at a 1-second scale) as the dependent variable and "speaker", "speaking time", and "interruption" as independent variables. Speaker was a dichotomous factor indicating which speaker held the floor. Speaking time

Table 1. Basic descriptive statistics of Twitter data collected for the debates.

Debate	Total tweets	Retweets	Mean tweets / sec (SD)	"Obama"	"Romney"
1	713642	381797	110.4 (47.2)	411391	468583
2	686805	368010	104.5 (47.9)	375506	462159
3	406368	212262	63.0 (27.8)	231778	266801

Sum of "Obama" and "Romney" may exceed total tweet count because tweets can mention both of them.

doi:10.1371/journal.pone.0122742.t001

was a measure of how long the speaker had been speaking in the current speech turn. Interruption was a dichotomous factor indicating whether the current speaker had interrupted his interlocutors to gain the floor. Linear mixed effects models were used to test these patterns for each debate. The first model included a main effect for speaker, duration of speaking in each speech turn, and an interaction between these two fixed factors. The models included a random effect for turn number, along with nested slopes for both candidate identity and time within turn number. The second model built on the first model by including interruption as an additional fixed factor. Goodness of fit of the models was calculated using R^2: in the context of mixed effects models, marginal R^2 (R^2m) indicates the variance explained by fixed effects alone, while conditional R^2 (R^2c) indicates the variance explained by the full model, including random effects.

Content: Momentary salient events. To investigate the second time scale, the impact of content, we chose three distinct salient events that took place in the interaction. These events, which quickly evolved into Internet "memes," were identified based on popularly discussed comments by the candidates. We chose one salient remark per debate: Romney declaring "I love Big Bird" in the first debate, Romney mentioning that he received "whole binders full of women" in the second debate, and Obama noting that the army had fewer "bayonets" in the third debate. Each of these events spread rapidly on the Internet, becoming the dominant topics of debate-related Twitter conversations and online searches for each of them totaled hundreds of thousands of mentions [23].

We expected attention to salient events to have partially self-sustaining dynamics. When enough tweets are produced on a given topic, they should keep public attention focused on that topic, although the debate might have moved on. To estimate how long a salient event can be expected to influence overall tweet counts, collective attentional entrainment at this scale was modeled as an exponential decay function coupled to a sigmoid. This serves as a simple mathematical model for a meme. The decay component relates to the fall from a burst of mentions due to novelty or salience of the event, $N(t) = e^{-\lambda t}$, with λ reflecting the decay rate. If that saliency achieves a particular prominence, or threshold, then the continuing attention to the event may sustain it as a meme, which could be characterized as a rapid-onset sigmoid function, $M(t) = 1 / (1+e^{-m(t-s)})$, where s is the point (in seconds) at which tweet rate is increasing maximally for the "meme," and m reflects the slope of that rate. The following product of these two functions captures the general patterns seen in the tweets: $M(t)$ $[N(t)\text{-}b]$, where b is the mean base tweet rate observed in the final 100s of the data, reflecting the stable sustained tweet rate after the initial rapid decay. The model was fit to the three events by performing a simple parameter search within reasonable ranges of λ, s, and m, and choosing parameter values that maximized the correlation between the model and the observed data.

Long-term attention: The whole debate. The longest timescale was represented as a quadratic time term that rises from the onset of the debate, and drops at its end. This is motivated by the notion that human social responsiveness to the debate will itself be driven by the onset and offset of the massively shared event. We tested for the impact of long-term attention by employing a linear multiple regression model with tweets per second (measured at a 1-second scale) as dependent variable and a second-order polynomial as independent variable to account for linear and quadratic temporal development. The presence of decay in the second half of the debate was further tested by assessing the fit of the quadratic term alone (which involves only predicted decay in the second half of the inverse quadratic function). Goodness of fit of the models was assessed using R^2.

Multi-scale dynamics: Predicting public attention. We combined the three time scales variables into one regression model per debate that predicted overall rate of tweets. Thus, we employed number of tweets per second as the dependent variable, and "speaker duration",

"interruption", "salient moment", and "quadratic time" as independent variables. As shown below, each time scale contributes uniquely to the model, suggesting that entrainment of large-scale social attention is complex and driven by several time-varying factors. Finally, we tested whether the model generated from the first debate would generalize to predict tweet rates in the second and third debates. We chose not to include salient events in this last test as their analysis relied on post-hoc assessment of which events went viral and therefore would not be easily generalizable to new debates.

Results

Interactional Entrainment 1: Tweet mentions co-vary with speaker

Twitter activity was tightly time-locked with turn-taking exchanges in each of the debates (Fig. 3). When one candidate started to speak, tweet rate increased for that candidate within seconds of the turn switch. The models for debates 1 to 3 explained at least 10% of the variance, with the tweet rate of debate 2 being the best explained by the model, at over 30% of its variance, for both Obama- and Romney-centered attention (all marginal R^2's $> .10$). The models revealed main effects of speaker and duration, with a significant interaction of the two (see Table 2). The positive main effect of speaker indicates that when a candidate spoke he received proportionally more attention, $ß$'s $<-.21$, t's < -2.7, p's $< .001$. The negative main effect of duration, $ß$'s $> .45$, t's > 3.3, p's $< .0001$, might seem less intuitive, until one considers the significant positive interaction with speaker, $ß$'s $> .40$, t's > 4, p's $< .0001$. Thus, on average tweets about the candidates decreased the longer the current speech turn, however, the tweets about the speaking candidate himself increased. In other words, attention follows mostly the one who is speaking at the moment, neglecting the other candidate. The results suggest that entrainment to the turn-taking structure of the debate is rapid, requiring only a few seconds to exert an observable influence on massive social attention. All three debates display the same significant factors, with analogous effect size and direction.

Estimates of $ß$ were calculated by standardizing all continuous variables. Across all three debates, speaker mention substantially drives attention (tweet mention). $ß$ and t's are reported as Obama / Romney, as a model was devised to test the effect of each speaker's turns.

Interactional Entrainment 2: Tweet rate increases with conversational interruptions

Tweet rate was also influenced by interruptions, which significantly increased Twitter mentions of both candidates. Fig. 4 shows the tweet rate for both candidates and moderator together when their turns were interruptions or not. Numerous interruptions took place in the debates and were of varying lengths (Table 3). Results revealed a general increase in the mention of both candidates during interrupting events. Using a mixed effects model similar to the prior analysis, all debates show a reliable contribution of interruption, with marginal R^2's = .07, .02, and .12, for debates 1–3, respectively. Though the effect of interruptions is much smaller, all three debates show a significant coefficient for the interruption term, $ß$'s $> .50$, t's > 1.9, p's $< .05$ (see Table 4). All three debates display the same significant patterns, with analogous effect size and direction.

Content entrainment: Twitter bursts to "memes"

Twitter behavior was influenced by the occurrence of salient remarks that took place during the debates. Focusing on tweets containing the root terms "big bird" (10,076 mentions), "binder" (2,889), or "bayonet" (5,458), we analyzed the temporal development of Twitter behavior

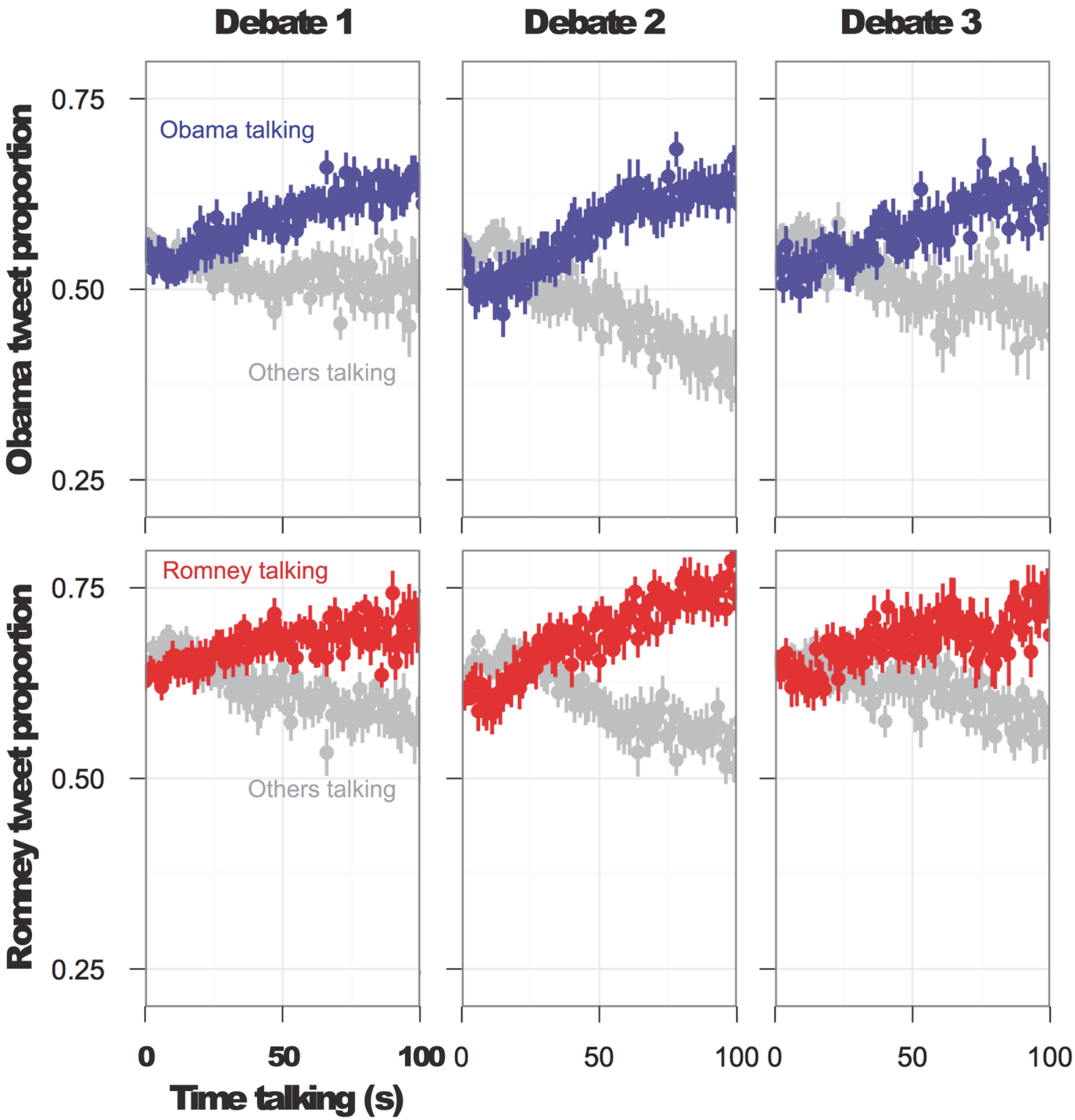


Fig 3. Effects of taking and holding the ground on Twitter mentions. Starting from the onset of each turn per candidate, plots show relative proportion of Twitter mention rises during that candidate's turn. While others are speaking, proportion mentions drops. Proportions are based on, for example, dividing mention to "Obama" divided by the sum of mentions to "Obama" and "Romney" together. Importantly, these plots only include original tweets, showing the anticipated effect is independent of retweets.

doi:10.1371/journal.pone.0122742.g003

Table 2. Speaking by candidates strongly invokes Twitter attention.

Debate	*Speaker ß, t*	*Speaking time ß, t*	*Speaker x Speaking ß, t*	R^2m R^2c
1	Obama: 0.75, 4.6	Obama: –0.30, –4.3	Obama. 0.70, 6.3	Obama: 0.20 0.73
	Romney: 0.67. 4.8	Romney: –0.37, –5.1	Romney: 0.69, 7.4	Romney: 0.16 0.75
2	Obama: 1.2, 8.6	Obama: –0.50, –6.8	Obama: 0.98, 9.1	Obama: 0.39 0.82
	Romney: 1.0, 8.2	Romney: –0.42, –5.4	Romney: 0.98, 9.1	Romney: 0.36 0.78
3	Obama: 0.47, 3.4	Obama: –0.21, –2.7	Obama: 0.50, 4.3	Obama: 0.11 0.58
	Romney: 0.60, 4.5	Romney: –0.22, –3.1	Romney: 0.43, 4.0	Romney: 0.12 0.57

doi:10.1371/journal.pone.0122742.t002

following the precise onset of each event. Our analysis shows that Twitter behavior displayed a remarkably similar temporal profile for each of these events. The first mention of the terms occurred within 11 seconds, and tweet rates peaked at about one minute after its onset, followed by a slow decay over the next few minutes (Fig. 5). Using the model of meme initiation and propagation we described in the previous section (Eq. 1), we model these temporal profiles in Fig. 6. It can be observed that distinct meme-like events can be modeled with the same functional form, and model parameters may serve to characterize subtle distinctions among them as further shown in the discussion.

Long-term attentional decay

We assessed the longer time scale of the debate itself, where we would expect both a gradual increase in attention, but one that trails off as the end of the debate approaches. Such a pattern is

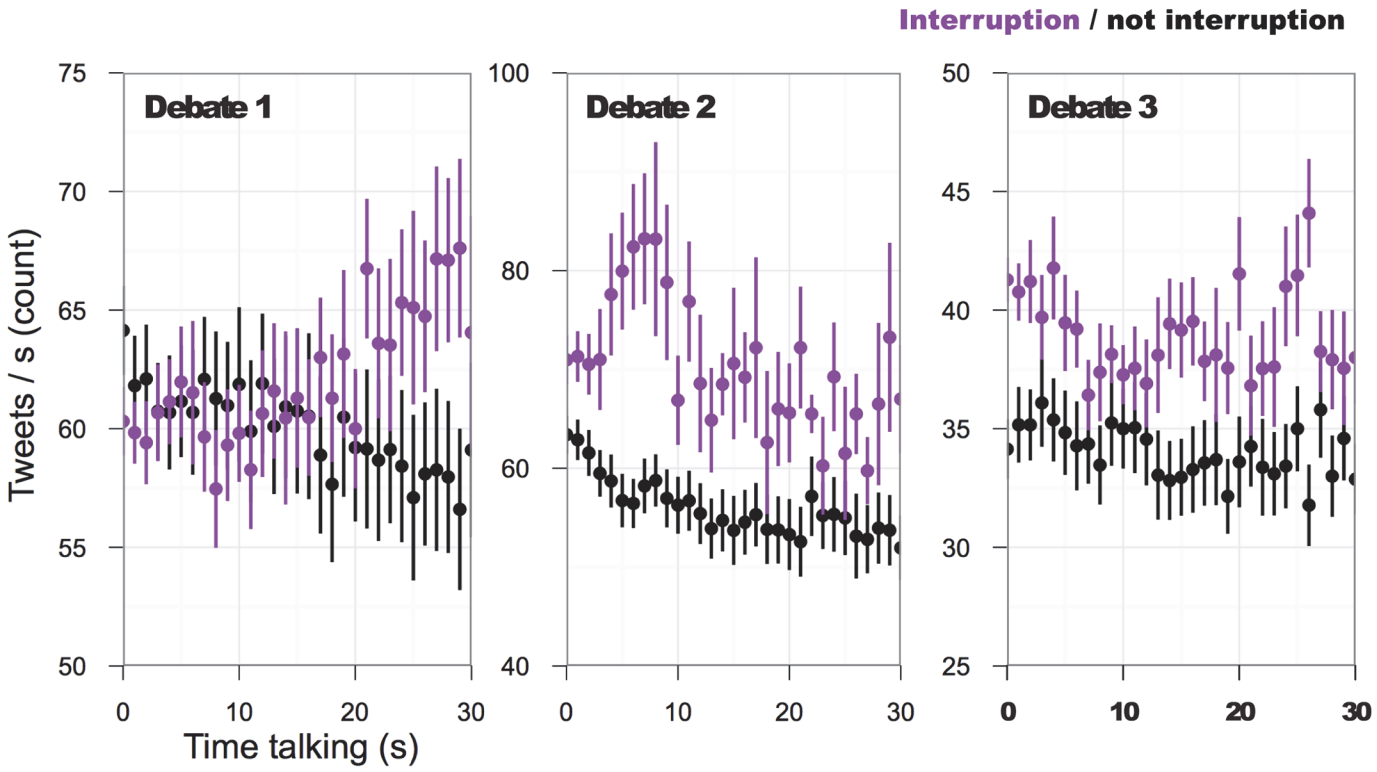


Fig 4. Effects of interruptions on Twitter mentions. At the onset of speaking, results show that the volume of tweets increases when that spoken turn is in the form of an interruption. Each panel represents the results from one of the debates. Importantly this figure only shows original tweets, omitting retweets.

doi:10.1371/journal.pone.0122742.g004

Table 3. The number of interruptions identified in each debate.

Debate	Turn count	Interruptions	Duration range (s)	Obama	Romney	Moderator
1	214	115	0.1–130.6	23	45	47
2	266	105	0–208.7	39	37	29
3	190	117	0.1–117.7	41	45	31

Duration range is the minimum / maximum length of turn identified as an interruption. The final three columns identify interruption counts within speaker.

doi:10.1371/journal.pone.0122742.t003

evident in Fig. 2. To test this quantitatively, we used a second-order polynomial regression model, with first- and second-order time terms predicting overall tweet rate. Both are highly significant, and account for over 20% of the variance from the two terms alone, for each debate. The linearly increasing term is strongly significant, β's $> .28$, t's > 20.0, p's $< .0001$. However there appears to be a larger effect magnitude for the quadratic term, which specifies both a relative increase at the beginning of the debate and a *decrease* by the end of the debate, β's $> .34$, t's > 25.0, p's $< .0001$. This larger effect for the quadratic term holds for all three debates (see Table 5). Importantly, this was not driven just by the beginning of the debate, for which the nonlinear second-order term may be considered to fit better; the last half of the debate, which only includes the decay portion of the quadratic term, still shows a significant contribution of the decay term when included alone, p's $< .001$ for all debates. All three debates display the same significant patterns, with analogous effect size and direction.

Regression model to test entrainment timescales

The prior analyses demonstrated each time scale's relevance separately, and we wished to test in a simple way whether all of these factors contribute simultaneously to tweet rate. To do so, we developed a multiple regression model with all time scales as variables accounting for tweet rate. We factored in salient events, modeled as a decay function along with temporal variables for speaker, whether interruptions were taking place, and at the broadest scale, a quadratic term representing the start and end of the debate. In each debate, the full regression model accounted for almost 50% of the variance in tweet rate (see Fig. 7). All variables also significantly ($p < 0.05$) and uniquely contributed to this variance (see Table 6). This regression analysis suggests that all time-varying properties that we have analyzed above contribute to the ebb and flow of public attention as reflected in tweet mentions. Put another way, the temporal variation in tweet rate may contain signatures of various time-scales of attentional processes taking place simultaneously in these massively shared experiences. These processes are influenced by broad exposure to the debate itself, by more local events, such as conversational interruptions and by the salient remarks that give rise to memes.

Table 4. Interruption by candidates increases Twitter activity.

Debate	*Interruption ß, t*	*Speaking time ß, t*	*Interruption x Speaking time ß, t*	R^2m R^2c
1	0.72, 2.0	–0.20, –3.8	0.80, 2.5	0.08 0.85
2	0.55, 2.5	–0.07, –0.64	0.11, 0.72	0.02 0.93
3	0.94, 4.2	–0.08, –1.34	0.29, 1.6	0.12 0.71

Estimates of ***ß*** were calculated by standardizing all continuous variables. Across all three debates, interruptions drive debate attention (overall tweet rate).

doi:10.1371/journal.pone.0122742.t004

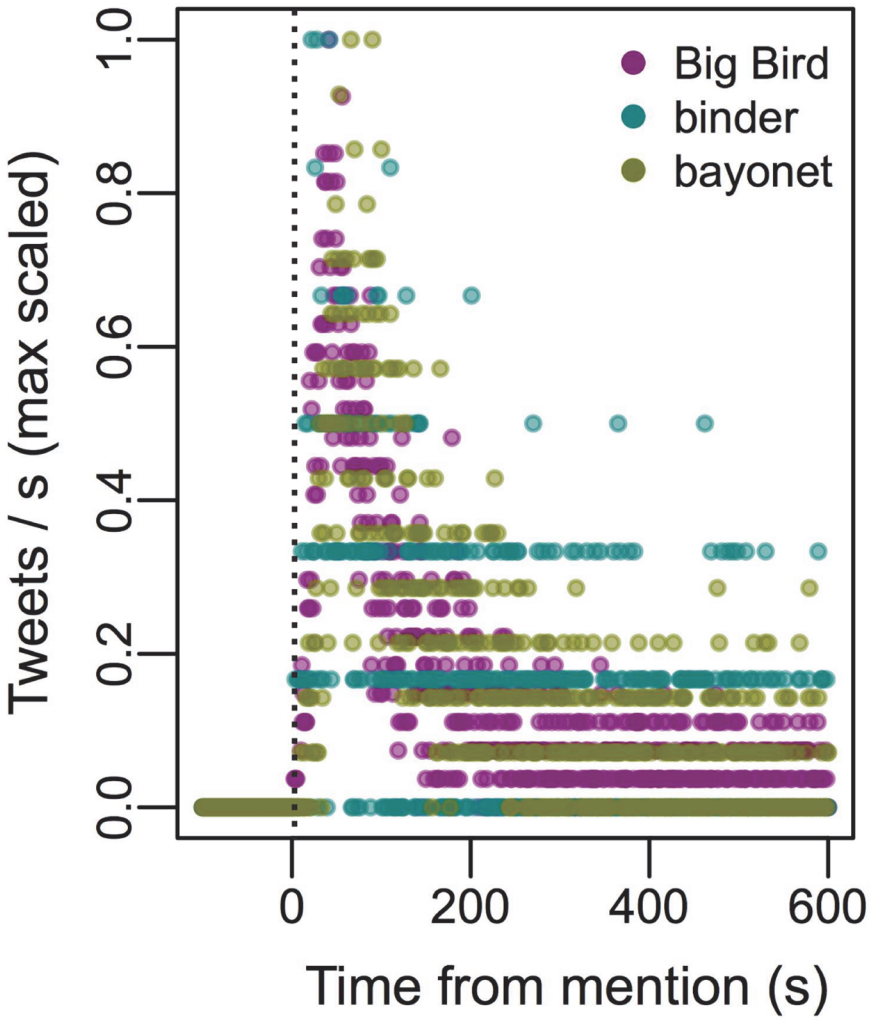


Fig 5. The temporal profile of public attention to salient events. At the onset of a salient event, mention of the word (in the context of either "Obama" or "Romney") rapidly rises within 10 seconds (left panel). Mentions are max scaled to facilitate comparison. Right panel shows retweets separately from original tweets, showing the expected delay. Interestingly, these salient events show distinct temporal signatures in their onset and rise to maximum, both in the profile of tweets and retweets. For original tweets, first mention for Big Bird, binder, and bayonet respectively is 4, 5, and 11 seconds; their maximum is achieved at 42, 23, and 67 seconds. In the retweet data, this is lagged, with first retweets at 31, 14, and 17 seconds; maximum achieved at 99, 80, and 78 seconds, respectively for Big Bird, binder, and bayonet.

doi:10.1371/journal.pone.0122742.g005

Lastly, we used the model from the first debate to predict the tweet rate from the subsequent two debates. Can basic information about a debate—knowing the time point of the debate, whether one of the candidates is speaking, and whether one is interrupting the other—predict tweet rate from one debate to the next? Even with just these two timescales (speaker duration/interruption, the duration of the debate), the model from the first debate can capture about 10% or more of the variance in the second and third, r's = .41, .32, respectively, p's < 0001. A simple and efficient representation of basic conversational processes (speaker, interruption) and time terms (second-order polynomial) can significantly predict large-scale social attention across debates.

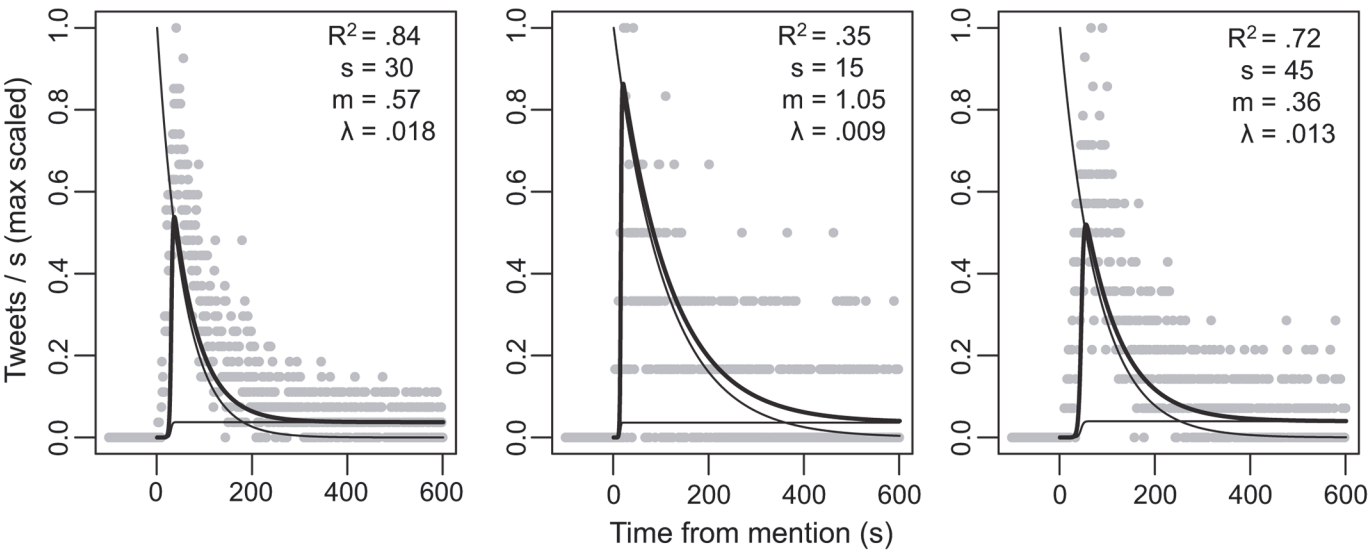


Fig 6. A model of public attention to salient events. The model of public attention reactions to salient events as fit to the three case studies: "Big bird", "binder" and "bayonet," from left to right. Note two interlocked timescales: a saliency/novelty followed by the establishment of a meme that sustains a base-level of continued attention. In each panel R^2 indicates the fit of the model, *s* is the point (in seconds) at which tweet rate is increasing maximally for the "meme," *m* reflects the slope of that rate, and λ reflects the decay rate. For more details on the equation cf. methods section.

doi:10.1371/journal.pone.0122742.g006

Discussion

We hypothesized that the dynamics of a massively shared event—such as the 2012 US presidential debates—would be reflected in the second-by-second, larger-scale dynamics of public attention. Specifically, the generation of Twitter messages would exhibit entrainment to the debates at (at least) three time scales: short-term conversational dynamics, mid-term content and long-term attentional entrainment.

i) Conversational entrainment

Public attention and response are time-locked to the conversational dynamics (e.g. turn-taking, interruptions) of the debates. Within seconds of initiating their conversational turn, speakers generate increased Twitter mentions to themselves, with correspondingly fewer mentions to their opponent. Moreover, the longer the speaker holds the ground the greater the increase in attention he receives from the tweeting audience. Interrupting the adversary emphasizes this effect and increases attention on one's speech turn. In other words, collective attention is time-

Table 5. Long-term trends in the Twitter activity.

Debate	R^2	*Linear term ß, t*	*Quadratic term ß, t*
1	0.29	0.34, 28.5	–0.46, –38.1
2	0.23	0.36, 24.0	–0.53, –35.2
3	0.16	0.28, 20.9	–0.34, –26.0

An OLS regression was used to predict tweet rate using a linear term representing the increasing time of the debate, and a quadratic term over the same time frame, which reflects an increase and then decrease. Both are highly significant, with the quadratic term in general having the larger effect size. That last column shows that the decay portion of the quadratic term still significantly predicts tweet rate when included alone. There is thus a longer timescale process of height activity then decay.

doi:10.1371/journal.pone.0122742.t005

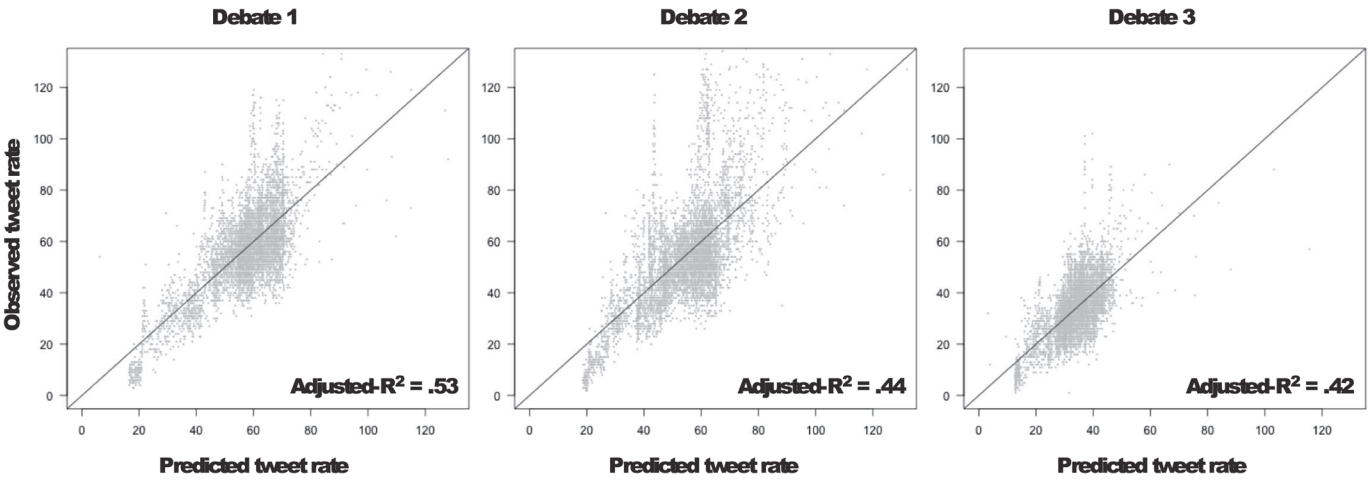


Fig 7. Multiple regression fits for all 3 debates. Variance accounted for by salient events, a quadratic time term, who is talking, and whether interruption is taking place accounts for between 42% and 53% of the variance in observed tweet rate. See text and Supporting Information for further details on the model.

doi:10.1371/journal.pone.0122742.g007

locked to cues of assertiveness and maybe even "presidentiality." It has to be restated that our findings are limited to how assertiveness display entrains viewers' attention and do not include a more nuanced perspective on the emotional valence of the tweets, or even the way tweeters and tweets cluster in response to different candidates.

ii) Content entrainment

In addition to a more immediate entrainment, we have shown slower dynamics as the public tunes its attention and elaborates on salient events. The first mention occurs within 11 seconds, overall mentions peak at 1 minute, then gradually fade over about 10 minutes' time. The dynamics of this profile can be modeled as an interaction between the decrease of saliency of the event itself over time and the sustained interest generated by new mentions of the event on Twitter. This highlights the more demanding cognitive processing of actual semantic content, and the importance of intrinsic dynamics in the social media, which can keep a salient event alive beyond its instantiation in the debate. Interestingly, the model parameters individuated can be used to characterize subtle distinctions in the memes. For example, our results suggest that some memes may resonate more strongly in the social media sphere: the salient event "binders," despite having a lower raw tweet rate relative to the other two salient events, had both the slowest decaying and the most rapidly rising meme formation. This resonates with analyses by Lin et al. [23] showing that the "staying power" of a meme is not only related to the

Table 6. Performance of simultaneous multiple regression models.

Debate	R^2	Quadratic	Speaker duration	Salient	Interruptions
1	0.53	0.20	0.06	0.11	0.02
2	0.44	0.12	0.04	0.11	0.04
3	0.42	0.09	0.03	0.14	0.06

The variables developed in prior analyses were used in one multiple regression model, predicting tweet rate by a variety of factors. All contribute significantly. Proportion variance *uniquely* associated with each variable in the model is shown, by entering it last into the regression. Note that models include all interaction terms among our primary variables analyzed above (speaker, interruptions, etc.). See Supplementary Information for the full model specification.

doi:10.1371/journal.pone.0122742.t006

raw quantity of mention, but also other social factors like conversational vibrancy (i.e., the prominence of the tweeters involved) and the interactivity of their audience.

iii) Long-term attention

Not least, collective entrainment displays long-term dynamics with an initial increase as the debate unfolds, followed by a decrease as it nears its conclusion.

Taken together, our findings suggest that human collective entrainment is multi-scale. Each of these three scales contributes uniquely and significantly to a multiple regression model predicting public attention in the form of Tweet rate. The debates generally present slight difference in the strength of the different predictors and the goodness of fit of the models. This is unavoidable as the debates are complex social phenomena taking place in an evolving political and communicational context. Amongst the differences between them was the overall structuring of the debate: the first and the second debate with six thematic segments, the second with eleven questions from the public. That all our effects were in the same directions and statistical significant anyway is a cue to their robustness.

While these results strongly indicate the presence of collective entrainment, they do not fully describe the complexity of human collective entrainment as many additional factors could and should be explored in future studies. Three dimensions in particular seem to be crucial for the current case study: i) emotional valence; ii) networks of political affiliation and pre-existing beliefs; and iii) impact on public opinion. Assertiveness and interruptions may generate positive appraisal as showing presidential qualities or may be more negatively assessed, and these reactions are likely to be mediated by political affiliation and pre-existing beliefs. Just as blogs cluster around political orientation [30], politically active Twitter users might primarily respond to their preferred candidate only, or may modulate their appraisal of assertiveness and salient events so as to cast a good light on his behavior. Promising work has been done on automatically segmenting Twitter users according to their political orientation [46–48], on automated assessment of conceptual and emotional dimensions of political discourse and tweets [29, 33, 49, 50], and on the impact of conversational dynamics between tweets [23, 25]. Future work will also have to investigate the details of conversational entrainment through Twitter and its impact on public opinion.

We live in an age in which local events can be broadcast in real-time to hundreds of millions of people around the world, and in response, people can interact instantaneously with each other through the use of online social media. This qualitatively new capacity for communication is changing the nature of large-scale politics and coordinated action across the globe. The situation calls for the development of large-scale analysis and models that both characterize and predict these emerging social dynamics. A growing number of studies are dedicated to identifying and categorizing events, including earthquakes, and even successful and unsuccessful political speeches, according to the public attention dedicated to them [25, 45, 51–53]. Yet little is known about the dynamics of this local-global interaction. How does the unfolding action of debates and other broadcasted events impact real-time public attention and response in social media? By combining quantitative assessments of conversational dynamics [54–58] with the analysis of hundreds of thousands of Twitter messages, this study is the first to assess the unfolding impact of a single event on the large-scale dynamics of public attention. Our results highlight how the dynamics of a local conversation can entrain the communicative behavior of massive populations of spectators. They also demonstrate the value of fine-grained temporal analyses at different time scales in uncovering the powerful relationship between social media and public events.

Conclusion

Collective and self-organizing behaviors are endemic to many social species, at many scales [59]. Entrainment is one frequently cited collective behavioral pattern, famous in fireflies [2] but found across numerous species, including murmurations of starlings, schooling in various fish species and more (see [60] for a review). Human communication might seem a smaller-scale phenomenon, likely built on a foundation of dialogical and spatially limited interactive dynamics [61]. Recent studies, however, argue for the existence of large-scale human entrainment dynamics, with local dialogical exchanges combining at a societal level and over time [62–64]. The advent of social media and information technologies allows humans to scale and speed up these dynamics to showcase massive and rapidly self-organizing patterns of entrainment. Indeed, our findings highlight that the massively shared experience of a political event induces complex patterns of collective attentional entrainment: an exquisite time-locking of observed behavior with the structure of the political event itself, content entrainment with partially self-sustaining dynamics, and large-scale attention bursts and decays. Put simply, like "congregating fireflies," humans show massive sustained entrainment across hundreds of thousands of individuals, in matters of seconds and minutes.

Supporting Information

S1 Code. Commented R code employed to run the analyses in the paper. The file is a pdf containing the code used to run the analyses, commented for understandability, and the output of the code.
(PDF)

S1 Dataset. Unique TweetIDs for the first debate. This file is a text file containing the unique TweetIDs for the first debate.
(TXT)

S2 Dataset. Unique TweetIDs for the second debate. This file is a text file containing the unique TweetIDs for the second debate.
(TXT)

S3 Dataset. Unique TweetIDs for the third debate. This file is a text file containing the unique TweetIDs for the third debate.
(TXT)

S1 Fig. Tweet rate and turn-taking the presidential debate: tweets vs. tweets+retweets. A comparison of tweet vs. total (tweet and retweet) rate (per second) of mentions to "Obama" and "Romney" across the debate. Patterns are highly similar, and retweets appear to happen very promptly following the volume of initial tweets.
(EPS)

S2 Fig. Effects of taking and holding the ground on Twitter mentions: tweets vs. tweets+-retweets. At the onset of speaking, results show that both the volume of tweets (on the left) and the total volume (tweets plus retweets, on the right) increase when that spoken turn is in the form of an interruption; this effect appears to be stronger in the original tweets.
(EPS)

S3 Fig. Effects of interruptions on Twitter mentions: tweets vs. tweets+retweets. Original tweets (above) and total data (tweets plus retweets, below) both exhibit the interruption effect: during interruption by individuals during the debate, the raw tweet rate (per second) increases.
(EPS)

S4 Fig. The temporal profile of public attention to salient events: tweets vs. tweets+-retweets. The left panel shows the original tweet data as displayed in the main paper. The retweets show a distinct time delay that is still nevertheless highly similar in structure across all three pointed moments during the debates.
(EPS)

Acknowledgments

The authors wish to thank Chris Kello, Chris Frith, Andrea Baronchelli and Peer Christensen for useful suggestions, Skye Smith and Benjamin Riis for help in time-coding the transcripts.

Author Contributions

Conceived and designed the experiments: RF RD MP. Performed the experiments: RD MP RF AM. Analyzed the data: RD MP RF. Contributed reagents/materials/analysis tools: RD MP RF AM TM AP. Wrote the paper: RF RD MP TM AP.

References

1. Strogatz SH, Stewart I. Coupled oscillators and biological synchronization. Sci Am. 1993; 269(6):102–9. Epub 1993/12/01. PMID: 8266056.
2. Mirollo RE, Strogatz SH. Synchronization of Pulse-Coupled Biological Oscillators. Siam J Appl Math. 1990; 50(6):1645–62. doi: 10.1137/0150098 PMID: WOS:A1990EK97600010.
3. Pikovsky A, Rosenblum M, Kurths J. Synchronization: a universal concept in nonlinear sciences. Cambridge: Cambridge University Press; 2001. xix, 411 p. p.
4. Ancona D, Chong C-L. Entrainment: Pace, cycle, and rhythm in organizational behavior. Research in organizational behavior. 1996; 18:251–84.
5. Schmidt RC, Richardson MJ, Arsenault C, Galantucci B. Visual tracking and entrainment to an environmental rhythm. Journal of Experimental Psychology Human Perception and Performance. 2007; 33 (4):860–70. Epub 2007/08/09. doi: 2007-11560-008 [pii] doi: 10.1037/0096-1523.33.4.860 PMID: 17683233.
6. Patel AD, Iversen JR, Bregman MR, Schulz I. Experimental evidence for synchronization to a musical beat in a nonhuman animal. Current Biology. 2009; 19(10):827–30. doi: 10.1016/j.cub.2009.03.038 PMID: 19409790
7. Dörfler F, Bullo F. Synchronization in complex networks of phase oscillators: A survey. Automatica. 2014.
8. Jamieson KH, Birdsell DS. Presidential debates: The challenge of creating an informed electorate. New York: Oxford University Press.; 1988.
9. Carlin DP. A rationale for a focus group study. In: Carlin DB, McKinney MS, editors. The 1992 presidential debates in focus. Westport, CT: Praeger; 1994. p. 3–19.
10. Benoit WL, Hansen GJ, Verser RM. A Meta-Analysis of the effects of viewing U.S. Presidential Debates. Communication Monographs. 2003; 70(4):335–50. PMID: 12740504
11. Grimshaw AD. Conflict talk: Sociolinguistic investigations of arguments in conversations: Cambridge University Press; 1990.
12. Morello JT. The "look" and language of clash: Visual structuring of argument in the 1988 Bush-Dukakis debates. Southern Communication Journal. 1992; 57(205–218).
13. Benoit WL, McKinney MS, Lance Holber R. Beyond Learning and Persona: Extending the Scope of Presidential Debate Effects. Communication Monographs. 2001; 68(2):259–73.
14. Pfau M. The subtle nature of presidential debate influence. Argumentation and Advocacy. 2002; 38:251–62.
15. Patterson ML, Churchill ME, Burger GK, Powell JL. Verbal and nonverbal modality effects on impressions of political candidates. Analysis from the 1984 presidential debates. Communication Monographs. 1992; 1992(59):231–42.
16. Masters RD, Sullivan D. Nonverbal behavior and leadership. In: Iyengar S, McGuire WJ, editors. Explorations in Political Psychology. Durham (NC): Duke University Press; 1993.

17. Sullivan DG, Masters RD. Biopolitics, the media, and leadership: Nonverbal cues, emotions, and trait attributions in the evaluation of leaders. In: Somit A, Peterson SA, editors. Research in Biopolitics. 2. Greenwich, CT: JAI Press; 1994. p. 237–73.
18. Bucy EP, Grabe ME. "Happy warriors" revisited: hedonic and agonic display repertoires of presidential candidates on the evening news. Politics and the life sciences: the journal of the Association for Politics and the Life Sciences. 2008; 27(1):78–98. Epub 2009/02/14. PMID: 19213305.
19. Hamby P, Preston M, Steinhauser P. 5 things we learned from the presidential debate. CNN.com; 2012.
20. Ward J. Mitt Romney Versus Obama: 4 Key Moments From First Presidential Debate. The Huffington Post. 2012 October 4th, 2012.
21. Clayman SE. Defining moments, presidential debates, and the dynamics of quotability. Journal of Communication. 1995; 45(3):118–47.
22. Matlock T. Framing Political Messages with Grammar and Metaphor-How something is said may be as important as what is said. American Scientist. 2012; 100(6):478.
23. Lin YL, Margolin D, Keegan B, Baronchelli A, Lazer D. #Bigbirds never die: understaing social dynamics of emergent hashtags. Proceedings of 7th International AAAI Conference on Weblogs and Social Media. 2013:370–9.
24. Schroeder A. The Presidential Debates: Fifty Years of High-Risk TV: Columbia University Press; 2008.
25. Lin YR, Keegan B, Margolin D, Lazer D. Rising Tides or Rising Stars?: Dynamics of Shared Attention on Twitter during Media Events. Plos One. 2014; 9(5):e94093. Epub 2014/05/24. doi: 10.1371/journal.pone.0094093 PMID: 24854030; PubMed Central PMCID: PMC4031071.
26. Miller G. Social Scientists wade into the Tweet stream. Science. 2012; 333:1814–5.
27. Culotta A, editor. Towards detecting influenza epidemics by analyzing Twitter messages. First workshop on social media analytics ACM; 2010; New York.
28. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci. 2011; 2:1–8.
29. Golder SA, Macy MW. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science. 2011; 333(6051):1878–81. Epub 2011/10/01. doi: 10.1126/science.1202775 PMID: 21960633.
30. Livne A, Simmons M, Adar E, Adamic L. The party is over here: structure and content in the 2010 election. ICWSM 11. 2011:17–21.
31. Ciulla F, Mocanu D, Baronchelli A, Goncalves B, Perra N, Vespignani A. Beating the news using social media: the case study of American Idol. EPJ Data Science. 2012; 1(8). doi: 10.1186/1687-4153-2012-8 PMID: 22793672
32. Conover M, Goncalves B, Flammini A, Menczer F. Partisan asymmetries in online political activity. EPJ Data Science. 2012; 1(6).
33. Hannak A, Anderson E, Barrett LF, Lehmann S, Mislove A, Riedewald M, editors. Tweetin' in the rain: Exploring societal-scale effects of weather on mood. 6th International AAAI Conference on Weblogs and Social Media; 2012 June 2012; Dublin, Ireland.
34. Metaxas PT, Mustafaraj E. Social Media and the Elections. Science. 2012; 338(6106):472–3. doi: 10.1126/science.1230456 PMID: 23112315
35. Yu S, Kak S. A Survey of Prediction Using Social Media. arXiv preprint arXiv:12031647. 2012.
36. Mitchell L, Harris KD, Frank MR, Dodds PS, Danforth CM. The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. arXiv preprint arXiv:13023299. 2013.
37. Hogan J, Holland B. Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. Journal of Applied Psychology. 2003; 88:100–12. PMID: 12675398
38. Judge TA, Ilies R, Colbert AE. Intelligence and leadership: A quantitative review and test of theoretical propositions. Journal of Applied Psychology. 2004; 89(542–552). PMID: 15161411
39. House RJ, Aditya RN. The social scientific study of leadership: Quo vadis?. Journal of Management. 1997; 23:409–73.
40. Ames DR, Flynn FJ. What breaks a leader: The curvilinear relation between assertiveness and leadership. J Pers Soc Psychol. 2007; 92:307–24. PMID: 17279851
41. Bass BM, Bass R, Bass BM. The Bass handbook of leadership: theory, research, and managerial applications. 4th ed. New York: Free Press; 2008. xix, 1516 p. p.
42. King AJ, Johnson DD, Van Vugt M. The origins and evolution of leadership. Current biology: CB. 2009; 19(19):R911–6. Epub 2009/10/15. doi: 10.1016/j.cub.2009.07.027 PMID: 19825357.

43. Weng L, Flammini A, Vespignani A, Menczer F. Competition among memes in a world with limited attention. Scientific Reports. 2012;2.
44. Boersma P. Praat, a system for doing phonetics by computer. Glottology International. 2001; 5:341–5.
45. Sasahara K, Hirata Y, Toyoda M, Kitsuregawa M, Aihara K. Quantifying collective attention from tweet stream. Plos One. 2013; 8(4):e61823. doi: 10.1371/journal.pone.0061823 PMID: 23637913; PubMed Central PMCID: PMC3640043.
46. Lin Y-R, Margolin D, Keegan B, Lazer D, editors. Voices of victory: A computational focus group framework for tracking opinion shift in real time. Proceedings of the 22nd international conference on World Wide Web; 2013: International World Wide Web Conferences Steering Committee.
47. Pennacchiotti M, Popescu A-M, editors. Democrats, republicans and starbucks afficionados: user classification in twitter. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining; 2011: ACM.
48. Boutet A, Kim H, Yoneki E, editors. What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election. ICWSM; 2012.
49. Slatcher RB, Chung CK, Pennebaker JW, Stone LD. Winning words: Individual differences in linguistic style among US presidential and vice presidential candidates. Journal of Research in Personality. 2007; 41(1):63–75.
50. Angus D, Smith A, Wiles J. Human communication as coupled time series: Quantifying multi-participant recurrence. IEEE Transactions on Audio, Speech and Language Processing. 2012; 20:1795–807.
51. Wu F, Huberman BA. Novelty and collective attention. Proc Natl Acad Sci U S A. 2007; 104(45):17599–601. doi: 10.1073/pnas.0704916104 PMID: 17962416; PubMed Central PMCID: PMC2077036.
52. Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system. Proc Natl Acad Sci U S A. 2008; 105(41):15649–53. doi: 10.1073/pnas.0803685105 PMID: 18824681; PubMed Central PMCID: PMC2572957.
53. Lehmann J, Gonçalves B, Ramasco JJ, Cattuto C, editors. Dynamical classes of collective attention in twitter. Proceedings of the 21st international conference on World Wide Web; 2012: ACM.
54. Shockley K, Richardson DC, Dale R. Conversation and coordinative structures. Topics in Cognitive Science. 2009; 1(2):305–19. doi: 10.1111/j.1756-8765.2009.01021.x PMID: 25164935
55. Dale R, Fusaroli R, Duran N, Richardson DC. The self-organization of human interaction. Psychology of Learning and Motivation. 2013; 59:43–95.
56. Fusaroli R, Konvalinka I, Wallot S. Analyzing social interactions: Promises and challenges of cross recurrence quantification analysis. Springer Proceedings in Mathematics & Statistics. 2014; 103:137–55. PMID: 25484995
57. Fusaroli R, Raczaszek-Leonardi J, Tylén K. Dialog as interpersonal synergy. New Ideas in Psychology. 2014; 32:147–57.
58. Fusaroli R, Tylén K. Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance. Cognitive Science. in press.
59. Strogatz SH. Sync: the emerging science of spontaneous order. 1st ed. New York: Hyperion; 2003. viii, 338 p. p.
60. Kauffman S. At Home in the Universe: The Search for the Laws of Self-Organization and Complexity. Oxford: Oxford University Press; 1996 1996.
61. MacWhinney B. The emergence of language. Mahwah, NJ: Taylor & Francis; 1999.
62. Loreto V, Steels L. Social dynamics—Emergence of language. Nature Physics. 2007; 3(11):758–60. PMID: 17955017
63. Galantucci B, Garrod S. Experimental semiotics: A new approach for studying the emergence and the evolution of human communication. Interaction Studies. 2010; 11(1):1–13.
64. Tylén K, Fusaroli R, Bundgaard P, Østergaard S. Making sense together: A dynamical account of linguistic meaning making. Semiotica. 2013; 194:39–62.

Of Pins and Tweets: Investigating How Users Behave Across Image- and Text-Based Social Networks

Raphael Ottoni †
rapha@dcc.ufmg.br

Diego Las Casas †
diegolascasas@ufmg.br

João Paulo Pesce †
jpesce@dcc.ufmg.br

Wagner Meira Jr. †
meira@dcc.ufmg.br

Christo Wilson ‡
cbw@ccs.neu.edu

Alan Mislove ‡
amislove@ccs.neu.edu

Virgilio Almeida †
virgilio@dcc.ufmg.br

† Universidade Federal de Minas Gerais, Brazil
‡ Northeastern University, USA

Abstract

Today, it is the norm for online social (OSN) users to have accounts on multiple services. For example, a recent study showed that 34% of all Twitter users also use Pinterest. This situation leads to interesting questions such as: Are the activities that users perform on each site disjoint? Alternatively, if users perform the same actions on multiple sites, where does the information originate? Given the interlinking between social networks, failure to understand activity across multiple sites may obfuscate the true information dissemination dynamics of the social web.

In this study, we take the first steps towards a more complete understanding of user behavior across multiple OSNs. We collect a sample of over 30,000 users that have accounts on both Twitter and Pinterest, crawling their profile information and activity on a daily basis for a period of almost three months. We develop a novel methodology for comparing activity across these two sites. We find that the global patterns of use across the two sites differ significantly, and that users tend to post items to Pinterest before posting them on Twitter. Our findings can inform the understanding of the behavior of users on individual sites, as well as the dynamics of sharing across the social web.

1 Introduction

Online social networks (OSNs) are now a popular way for individuals to connect, communicate, and share content; many now serve as the de-facto Internet portal for millions of users. Because of their massive popularity, data about the users and their communication offer unprecedented opportunities to examine how human society functions at scale.

In recent years, the number and diversity of social networks have both increased beyond traditional OSNs like Facebook; popular examples include sites like Tumblr (short-form blogging), Pinterest (sharing image collections), and FourSquare (sharing user locations). In fact, today, many users have active accounts on multiple such sites (*e.g.,* it has recently been shown that over 34% of Twitter users also have a Pinterest account (Pew Research Center 2013)).

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

While there have been many studies that have examined user behavior on these sites, it remains unknown how users distribute their time and activity *across* multiple sites. For example, do users share the same kinds of content on multiple sites? Are there sites where content tends to originate?

In this paper, we take a closer look at *cross-OSN user behavior*, focusing on the popular sites Twitter and Pinterest. Pinterest is a photo-sharing web site that allows users to share collections ("boards") of photos with others. Pinterest is primarily designed for users to share ("pin") images from other Web sites (unlike sites like Flickr, which are centered around having users upload their own photos). Like Twitter, Pinterest users can "follow" others, meaning the other user's pinned photos will show up in a feed when the user logs in.

Our goals are to understand how user activity is correlated across these two sites. First, we aim to understand whether users share a common *identity* across these sites. Second, we investigate the interests of users on these two different platforms. In Pinterest, user interest is conveyed by the images a user pins and repins in different boards, each of which is tagged with a category. In Twitter, user interest can be deduced from the words in their tweets. Questions that we aim to answer include: where does user interest manifest first, Twitter or Pinterest? Does a user share the same interests in both networks? How does one compare the interests represented by images versus the interests found in tweets?

We face a number of challenges in comparing activity on Twitter and Pinterest. Unlike Twitter, Pinterest is challenging to study, as the content shared is images (as opposed to text) and there is no official API for gathering data. We develop techniques for gather Pinterest data at scale, and collect a set of 30,000 users who have both Pinterest and Twitter accounts. We download all pinned photos and tweets of these users, resulting in a data set of over 2.9 million pins and 7.1 million tweets. Additionally, we develop a novel methodology that is able to compare image-based content (Pinterest) with text-based content (Twitter). We do so using *categories* of content, based on the categories provided by Pinterest; we demonstrate that our approach to categorization on Twitter shows high accuracy on many different types

of tweets.

Overall, our analysis makes three contributions: First, we aim to understand the macro-scale patterns of activity across these sites. We find that, despite considering the same set of users on both sites, we see remarkably different global patterns of activity. Second, we use our categorization approach to study how users distribute their activity across the sites; we find that users engage in more categories of content on Pinterest than Twitter, but Twitter categories have greater predictive power. Finally, we explore the pollination of content from one site to the other and find that new content tends to originate on Pinterest before spreading to Twitter. This result suggests that while Twitter is an incredibly popular global communication platform, sites like Pinterest play a crucial role in the generation of new ideas and content.

2 Background

2.1 Pinterest

Pinterest is a pinboard-style image sharing social network designed to let users collect and share images and videos in an organized, categorized way. Pinterest was founded in 2010, and boasts a user population of 70 million as of July 2013.[1] It is currently the world's 27th most popular website.[2] and is the fastest growing OSN in both unique visitors and clicks via search engines (Walker 2012) Pinterest users are 80% female (Chafkin 2012; Ottoni et al. 2013) and the average monthly usage time per visitor is 98 minutes, which makes Pinterest the second most used OSN behind Facebook.[3] Pinterest has become especially important as a driver for e-commerce traffic: a recent study showed that Pinterest users are worth almost twice as much as Facebook users.[4]

On Pinterest, users create a personal profile as well as one or more *boards*. Each board is given a name and a description (both freeform text fields) as well as a category. Pinterest pre-defines 33 categories, varying from "Women's Fashion" and "Hair Beauty" to "Geek" and "Tattoos". The basic units of data on Pinterest are the images and videos users *pin* to their boards. Each pin is characterized by a relative timestamp (*e.g.,* "posted 22 hours ago", "posted 5 days ago"), a description (freeform text), and a link to the source of the content (if it originated from a third-party website). The only interactions supported by Pinterest are *repins*, comments on pins, and *likes*. By default, all data on Pinterest is public.

The organization and functions of the Pinterest social network are similar to Twitter. Personal profiles on Pinterest include a profile image, a brief self-description, and lists of the user's boards, pins, likes, followers, and friends (*i.e.,* those who the user follows). Figure 1 shows a typical user profile on Pinterest. Like Twitter, Pinterest users are presented with a chronologically-ordered *timeline* of pins from their friends upon logging in. Unlike Twitter, Pinterest users may follow other users, or follow specific boards. In the former case, all of the friend's pins appear in the timeline; in the latter case only pins from the specific board appear.

[1] http://bit.ly/1jfjRHL
[2] http://bit.ly/1ksdYHv
[3] http://bit.ly/1dxEJSX
[4] http://bit.ly/OFmPrj

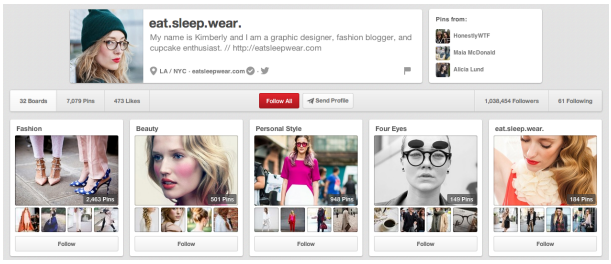


Figure 1: Pinterest profile of a famous designer/blogger.

2.2 Twitter

Twitter is the de-facto standard for micro-blogging in most of the world. Founded in 2006, Twitter currently has 645 million users who generate over 500 million 140-character *tweets* per day.[5] Twitter is now the 11th most popular website in the world.[2] On Twitter, users create personal profiles that include a profile image, a brief self-description, and lists of the user's tweets, followers, and friends. Each tweet is characterized by an exact timestamp, and may include embedded images, URLs, #hashtags, @mentions, or geotagged coordinates. Users can *retweet* messages from other users to their own followers. Although Twitter allows users to set their profiles as "protected" (in which case only approved people may follow the account and view its tweets), the vast majority of users accept the default policy in which all data are public (Wasserman 2012).

3 Dataset and Methodology

In this section, we present the dataset we will be using in our analysis. First, we introduce our raw data and describe how it was collected. Second, we briefly compare the users in our dataset against random samples of Pinterest and Twitter users, to quantify any bias in our target population. Third, we describe our methodology for labeling tweets with high-level *categories* (*e.g.,* design, fashion, politics). This categorization facilitates our comparison of Pinterest and Twitter in § 4, since all pins are categorized by default. Finally, we describe how we group tweets and pins into *sessions*, which represent a sequential period of user activity.

3.1 Dataset and Collection Methodology

The goal of our work is to compare and contrast the behavior of users across Pinterest and Twitter. In order to perform this comparison, we need to locate users that have accounts on *both* OSNs. Fortunately, there is a straightforward way to identify such users: when Pinterest first began accepting users, the only way to create an account was log-in with a pre-existing Facebook or Twitter account. Between August 21th and October 9th, 2012, we crawled 2 million users from Pinterest, of which 210,426 had signed-up with their Twitter account (Ottoni et al. 2013). Of these users, 76% have either protected or closed their Twitter account, leaving us with 50,702 users for our study. For clarity, we will refer to these users as the *selected population*.

[5] http://bit.ly/18KIwd2

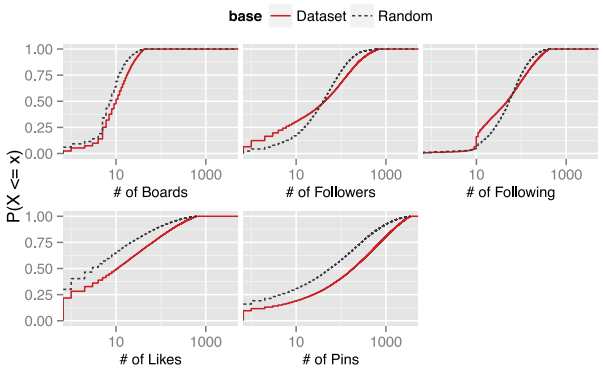


Figure 2: Comparison of our target Pinterest users to a random sample of Pinterest users.

To analyze user behavior, we need to collect pins and tweets generated by users in the selected population. To gather tweets, we can simply use the Twitter REST API v1.1. However, collecting pins requires addressing two technical challenges: 1) each pin must be gathered within 24-hours after it was generated, and 2) each pin must be gathered individually. The 24-hour requirement stems from the fact that Pinterest displays relative timestamps with decreasing resolution, *i.e.,* a pin from an hour ago will display "posted 1 hour ago," whereas a pin from yesterday will display "posted 1 day ago." Our study necessitates high-resolution timestamps, thus we need to crawl pins while the timestamps are displayed in hours. The second challenge occurs because Pinterest does not have a public API for gathering pins with timestamps in bulk.

To address these challenges, we built a distributed crawler that collected pins from the selected users every day. The Pinterest website is a complex Javascript application, so we leveraged PhantomJS to crawl the site. PhantomJS is a headless version of the WebKit browser engine that can be automated using Javascript. Our crawler recorded the daily pin activity from each user. Unfortunately, it was impossible to perform daily crawls on all 50K users in the selected population due to Pinterest's rate limits. To work around this issue, we focused our crawls on random sample of 30K users from the selected population. We refer to these users as the *sampled population*.

Final Dataset Our final dataset consists of just under three months of pins and tweets from the sampled population. We actively crawled Pinterest between September 12 and December 3, 2013; since we only have accurate timestamps for pins collected during this period, we constrain our focus to tweets that were also generated during this time period. Table 1 shows the number of users from the sampled population that performed specific actions (*i.e.,* pinning or tweeting) during the time window; we observed 23,313 users who had at least one activity. We refer to these 23K users as the *target population*, and we focus on them for the remainder of this study.

	Action	Users	Pins	Tweets
Ignored	No activity	6,687	—	—
Target population	Pinned	15,329	1,778,798	—
	Tweeted	19,107	—	4,427,040
	Pinned & Tweeted	11,123	1,299,266	2,746,472

Table 1: Our final 82-day (09/12/13 to 12/03/13) dataset, comprising our *target population*.

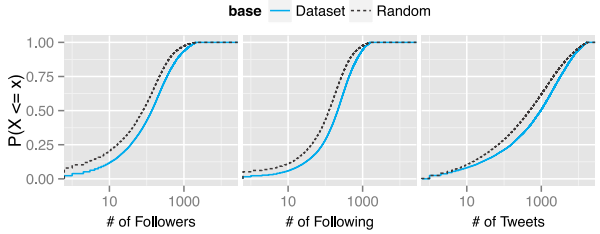


Figure 3: Comparison of our target Twitter users to a random sample of Twitter users.

3.2 Dataset Validation

We now present a brief comparison of the characteristics of users in the target population versus random samples of Pinterest and Twitter users. The purpose of this comparison is to quantify any bias in our target population. To enable this comparison, we selected 30K Pinterest users uniformly at random from our original sample of 2 million Pinterest users. Similarly, we selected 30K Twitter users uniformly at random from the "gardenhose" feed (10K each from September, October, and November 2013).

Figure 2 compares the characteristics of our target and random Pinterest users. The target users are slightly more active: they have more boards, likes, and pins. This suggests that the target users are more active social networkers, which is not surprising given that the target users have accounts on Pinterest and Twitter. Despite this, both populations have similar average numbers of friends and followers.

Figure 3 compares the characteristics of our target and random Twitter users. Much like on Pinterest, we observe that the target users generate slightly more tweets than the random sample. In this case, the target users also have slightly more friends and followers, further confirming that the target users are more active than random OSN users.

Overall, Figures 2 and 3 reveal that there are minor differences between the target and random user populations. However, despite this divergence, we believe that the target population is broadly representative of the active user community on Pinterest and Twitter.

3.3 Labeling Tweets with Categories

The next step in our methodology is to classify tweets from our target users with category labels. This step is necessary in order to make pins and tweets directly comparable: on Pinterest, all pins fall into 1 of 33 pre-defined categories,

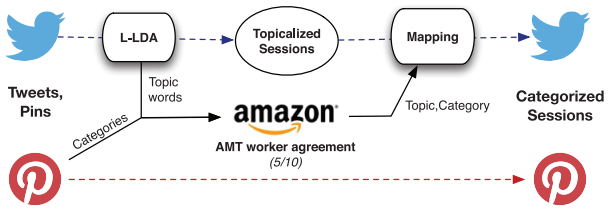


Figure 4: Diagram of our tweet categorization methodology.

while tweets are freeform text. Figure 4 depicts the novel process we developed to categorize tweets.

Selecting Categories The first step in our classification process is to select relevant categories. We use the 33 categories provided by Pinterest as a natural starting point, and added the following five additional categories, since they have been commonly identified as important in Twitter-related studies (Duh et al. 2011; Hong and Davison 2010; Lee et al. 2011; Zhao et al. 2011): "Politics", "Religion", "Charity", "Twitter-Specific", and "Business". "Twitter-specific" refers to tweets that discuss Twitter culture, *e.g.,* #followfriday.

To determine whether these categories were sufficient, we conducted a brief pilot study on Amazon Mechanical Turk (AMT) where we asked AMT workers to label tweets with one of the aforementioned 38 categories. The results of this pilot study revealed some problems with the categories. For example, the AMT workers had difficulty labeling tweets as "Men's Fashion" or "Women's Fashion", so we merged these two categories together as "Fashion". Similarly, we merged "Health & Fitness" with "Sports", and "Home Decoration" with "DIY". We also generalized "Kids" to "Kids & Family". After this process, we were left with 35 categories.

Identifying Topics on Twitter The second step in our classification process is to extract *topics* from our corpus of tweets. The key idea here is to leverage well-known topic extraction tools as a stepping stone towards categorization: if we can extract topics from tweets, and map the topics into our 35 categories, then we can use the topic labels to also apply category labels to the tweets.

To extract topics from our tweets, we leverage Labeled Latent Dirichlet Allocation (L-LDA) (Ramage et al. 2009). L-LDA uses the same underlying mechanisms as LDA, however each topic is seeded with a label (*i.e.,* a word chosen by the researcher), to help anchor the topic extraction process. We chose L-LDA because prior work has shown it to be more effective than LDA at extracting topics from microblogs (Quercia, Askham, and Crowcroft 2012; Ramage, Dumais, and Liebling 2010).

We ran L-LDA on a random sample of 2 million tweets from our corpus. This was the maximum number of tweets we could process on a machine with 70GB of RAM. We preprocessed the tweets by removing the top 300 words (*i.e.,* the stop words), URLs, and all words that appeared less than 10 times. As shown in Figure 4, we parameterized the L-LDA algorithm with $\alpha = 0.167$, $\beta = 0.001$, and $k = 300$. We selected the 300 most common hashtags from our tweets as topic labels. The final output of L-LDA is 300 topics, each containing ≈50 words ranked by frequency in that topic.

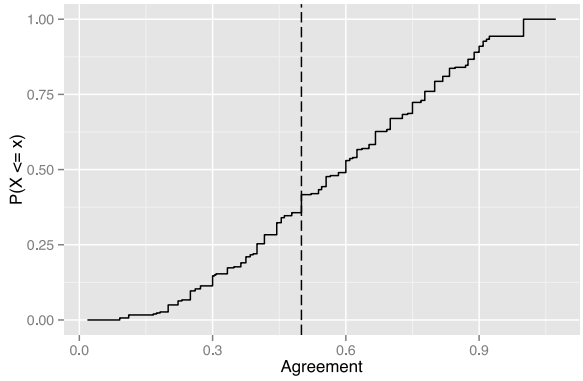


Figure 5: Agreement between AMT workers on the category label for each of the 300 Twitter topics.

Mapping Topics to Categories The third step in our classification process is to map the topics extracted from Twitter to our 35 categories. As shown in Figure 4, we leveraged AMT for this task. We presented each Turker with 20 of the topics extracted by L-LDA, and asked them to select the 3 most applicable categories for each topic. AMT workers were shown the top 25 words associated with each topic. We divided our 300/20 topics into 15 tasks, and had 10 AMT workers complete each task.[6] AMT workers were paid $0.25 per task, and this process took 2 days to complete.

After collecting the results from the AMT workers[7], we analyzed the data to determine which of the 300 L-LDA topics could be mapped with high confidence onto our 35 categories. Figure 5 shows the percentage of agreement between the AMT workers on the category for each topic, *i.e.,* whether they listed a particular category as one of their three options. 60% of the topics have >50% agreement between the 10 AMT workers, meaning we can say with high confidence that we know the best mapping of those topics to categories. In the remainder of our categorization methodology, we only use the topics that have >50% agreement; we discard the remaining low-agreement topics.

Tweet Categorization The fourth step of our classification process is to calculate the most likely category for each tweet based on its topic distribution. Initially, we attempted to use LDA to generate the topic distribution for each tweet (the output of LDA labeling is a topic distribution, representing the likelihood that a given document contains each topic). However, we had three independent raters examine a random sample of the labeled tweets, and discovered that the accuracy of the topic labels was ≈30%. This low accuracy stems from the fact that many tweets are very short,

[6] An example task can be found here: http://dcc.ufmg.br/~rapha/ofpinsandtweets/categorization/?s=1

[7] The labeling dataset can be found here: http://dcc.ufmg.br/~rapha/ofpinsandtweets/labelingdataset

and some cannot be interpreted when removed from their conversation context.

To remedy this situation, we group tweets into *sessions*. We define a session as sequentially-generated tweets from a single user, where the time-gap between any two successive tweets is $\leq$2 hours. Intuitively, sessions group tweets that are generated in rapid succession, and are therefore likely to fall into the same topic. We chose 2 hours as the time delimiter between sessions on Twitter because 2 hours is the smallest unit of time that we can accurately measure on Pinterest (recall that timestamps on Pinterest are shown in increments of hours, *e.g.,* "posted 4 hours ago"). In § 3.4, we will also group pins into sessions, so that our categorized data from Pinterest and Twitter are directly comparable.

We can now assign a category to each Twitter session by generating a topic distribution for the session. We treat each Twitter session as a document and use LDA to generate the topic distributions. The results from the AMT workers give us a mapping between L-LDA topics and categories, however this is a many-to-one mapping (*i.e.,* many topics may map to the same category). Thus, to determine the category of a session, we subtract $1/numberoftopics$ (*i.e.,* the baseline probability) from the probability of each topic, and then sum the positive results of all the topics that map to each category. We then assign the category with the highest aggregate probability to each session.

As a final step, we manually validated the categories assigned to 100 random Twitter sessions. As before, three independent raters examined each session to determine the appropriateness of the category. Of the 35 categories, 11 were judged as having accuracy $>$60% (*i.e.,* the majority of the time, the category accurately described the tweets in the session). These 11 categories are: "Art," "Charity," "Design," "DIY & Crafts," "Fashion," "Food & Drink," "Hair & Beauty," "Health, Fitness & Sports," "Politics," "Technology," and "Weddings". Thus, in § 4, we focus on Twitter sessions from these 11 categories, and ignore all sessions from other categories.

We make our code and data for classifying Twitter sessions into topics available to the research community.[8]

3.4 Sessions on Pinterest and Twitter

At this point, we have grouped tweets from our target users into sessions and categorized each session. After filtering out the sessions with low-accuracy categories, we are left with 100,212 sessions (out of 665,980 total sessions). To ensure that our Pinterest and Twitter data remain comparable, we also grouped pins from the target users into 2 hour sessions. The category for a Pinterest session is simply the most frequent category among the pins in that session.

Figure 6 shows the number of categorized sessions per user in our dataset. Our users tend to have fewer categorized Twitter sessions than Pinterest sessions; this is understandable, given that we have filtered out Twitter sessions that belong to low-accuracy categories. The most important line in Figure 6 is "Both," which plots the number of users with at least k categorized sessions on *both* OSNs. Since our goal

[8]http://dcc.ufmg.br/~rapha/ofpinsandtweets

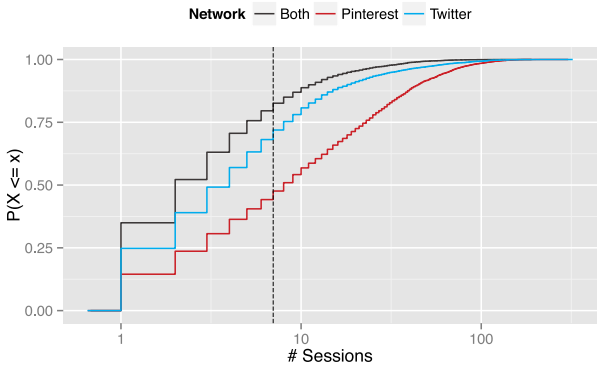


Figure 6: Number of categorized sessions per user by network.

is to examine user behavior across Pinterest and Twitter, we must focus on the subset of users who are active on both OSNs. Thus, in § 4, we focus on the users who have $\geq$7 categorized sessions on both Pinterest and Twitter. In total, this dataset includes 41,928 Pinterest sessions and 33,452 Twitter sessions generated by 1,227 users.

4 Analysis

We now present our analysis on cross-user behavior on Twitter and Pinterest. We first explore global patterns of user behavior (§ 4.1) on the two sites, followed by an examination of the similarity in user attributes across the sites (§ 4.2). We then turn to investigate the content shared on these two sites, based on the categories we discussed in the previous section. We first look at global category popularity (§ 4.3) followed by pollination of content categories across sites (§ 4.4).

4.1 General and Temporal Analysis

We begin our analysis by examining the high-level characteristics of our target users on Pinterest and Twitter. In this section, we examine all 23K users, 1.8M pins, and 4.4M tweets in our 82 day dataset (see Table 1 for details).

First, we examine overall pinning and tweeting behavior. Figure 7 is a scatterplot showing the number of pins and tweets per user in our dataset. Black regions are occupied by $\approx$50 users, while light grey regions are occupied by $<$10 users.

Figure 7 reveals that most users in our target population tend to fall in one of two regions of the plot. There is a group of users along the axes who are active on either Pinterest or Twitter. Conversely, there is a group of users in the center of the plot that generate on the order of 100 pins and 100 tweets during our 82 day sample. This demonstrates that our target population does include many users who are equally active on both OSNs.

Next, we examine the temporal dynamics of Pinterest and Twitter users. Figure 8 plots a moving average of the number of pins and tweets per day between October 15 and November 15, 2013. Two conclusions can be drawn from Figure 8: first, our target users generate around two times as many tweets per day as pins. Second, Pinterest and Twitter

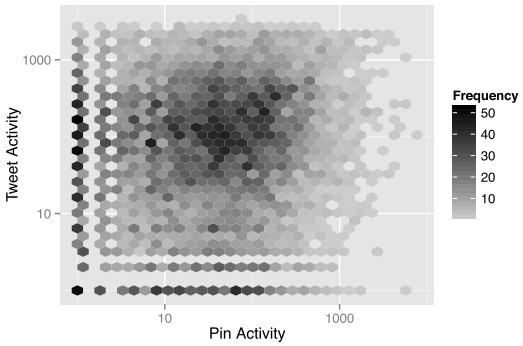


Figure 7: Scatter plot of activity by user in 82 days.

exhibit different long-term temporal dynamics. Twitter use peaks during the week, while Pinterest use peaks later in the week, heading into the weekend. One deviation occurs the week of October 28: we hypothesize that Pinterest saw increased activity this week due Halloween-related pins.

Figure 9 plots the daily activity patterns of our target users, broken down by individual days of the week. Our target users are primarily located in the U.S., Figure 9 is normalized to Eastern Standard Time. Unsurprisingly, the usage of both OSNs troughs in the early-morning. However, Twitter users frequently tweet late at night, and there is a pronounced reduction in tweeting on Saturday and Sunday (see Figure 8). In contrast, Pinterest use tends to slowly rise over the course of each day, especially on Sundays.

4.2 User and Linguistic Identity

In this section, we examine the personal profiles and linguistic features of our 23k target users. We pose the question: *do users have unique* identities *on Pinterest and Twitter, or do they share one identity across both OSNs*? To quantify user identity, we examine three textual features: usernames, descriptions in personal profiles, and words used in pin descriptions and tweets. We focus on textual features because prior work has been able to successfully measure individual emotional and cognitive factors using automated text analysis (Tausczik and Pennebaker 2009; Kahn et al. 2007; Veltman 2006).

First, we compare the usernames chosen by individuals on Pinterest and Twitter. Studies have shown that the usernames chosen by people on online systems are often reflective of their individual personalities, the way they want to be perceived by others, and their cultural environment (Suler 2002; Bechar-Israeli 1995). Thus, we hypothesize that if users choose the same username on Pinterest and Twitter, this indicates a homogeneous identity across both social platforms.

Figure 10(a) plots the Levenshtein Ratio (LR) between each user's username on Pinterest and Twitter. LR is defined as $LR(s_i, s_j) = 1 - ED(s_i, s_j)/\max(|s_i|, |s_j|)$, where s_i and s_j are strings, $ED(s_i, s_j)$ is the Levenshtein edit-distance between s_i and s_j, and $|s|$ is the length of string s. Intuitively, $ED(s_i, s_j) = 1$ means the strings are identical, while $ED(s_i, s_j) = 0$ means they are completely different. As shown in Figure 10(a), a significant number of

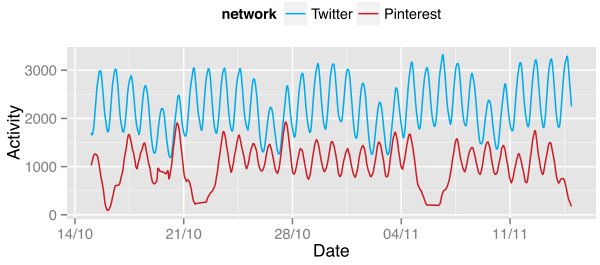


Figure 8: Average pins and tweets per day over one month.

our target users use the same, or very similar, usernames on both OSNs. This suggests that users share a common identity across Pinterest and Twitter.

Second, we compare the textual descriptions that user's write in their personal profiles. We analyze each description with LIWC (Pennebaker, Francis, and Booth 1999; Pennebaker et al. 2007), which produces a vector of linguistic attributes for each description. We then plot the cosine similarity between users' LIWC vectors on Pinterest and Twitter in Figure 10(b). Similar to Figure 10(a), the majority of users in our target population have the same, or very similar, linguistic attributes in descriptions across both OSNs. This reaffirms our finding that many users share the same personal identity on Pinterest and Twitter. However, unlike Figure 10(a), there are a significant fraction of users with zero similarity, indicating that some users may use Pinterest and Twitter for different purposes, or perhaps desire to separate their identities on these two social platforms.

Third, we compare the language used in pin descriptions and tweets. Although Pinterest is primarily on image-based OSN, users may write (typically brief) freeform textual descriptions for each pin. This leads to the question: *do users share the same linguistic style across pin descriptions and tweets?* To answer this question, we concatenated all of each users pin descriptions into a single document, and analyzed the document with LIWC. We also performed the same process for each user's tweets. This produces a pin-LIWC vector and a tweet-LIWC vector for each user. In order to compare each user's vectors to the language usage of the general population, we also created *network* LIWC vectors by gathering a random sample of pin descriptions/tweets, concate-

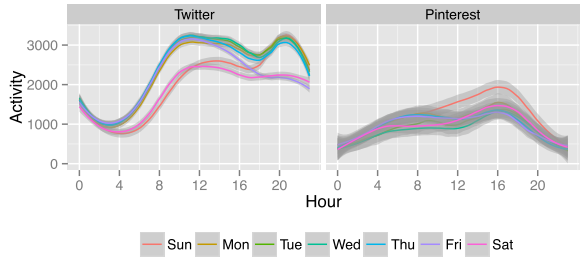


Figure 9: Daily activity on Pinterest and Twitter.

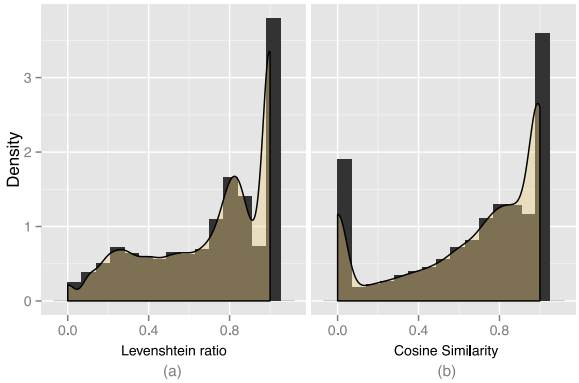


Figure 10: (a) The Levenshtein Ratio between each user's username on Pinterest and Twitter. (b) The cosine similarity between the LIWC vectors derived from each user's profile description on Pinterest and Twitter.

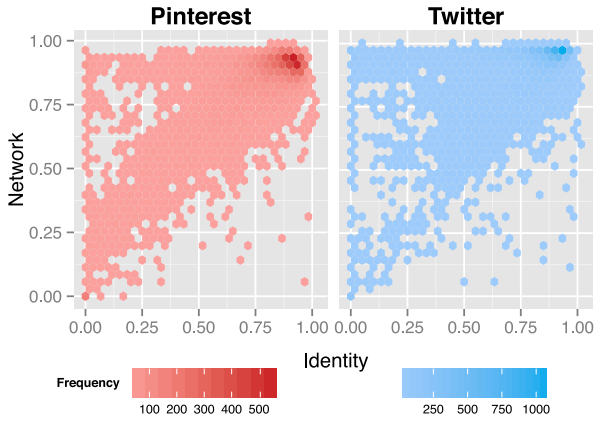


Figure 11: Linguistic identity of users across Pinterest and Twitter, versus similarity when compared to the linguistic norms of each respective OSN.

nating them into two respective documents, and analyzing these two documents with LIWC.

Figure 11 present the results of our pin and tweet LIWC analysis. Each of the two x-axes captures the same quantity: the cosine similarity between each user's Pinterest and Twitter LIWC-vector. We refer to this as the "Identity" axis. The y-axis of the Pinterest scatter plot captures the cosine similarity between each user's Pinterest LIWC-vector and the network LIWC-vector for Pinterest. The y-axis for the Twitter scatter plot is the same, except it substitutes the individual and network LIWC-vectors for Twitter. We refer to these as the "Network" axes. For example, a user at position (0, 0) in the Pinterest plot would have 1) zero similarity between the LIWC of their own pins and tweets, and 2) zero similarity between the LIWC of their pins and the pins of other Pinterest users. Conversely, a user at position (1,1) in the Pinterest plot would have 1) exact similarity between the LIWC of their own pins and tweets, and 2) exact similarity between the LIWC of their pins and the pins of other Pinterest users. The intensity of color in Figure 11 captures the number of users with a given set of similarity scores.

There are several takeaways from Figure 11. First, the majority of users cluster in the top-right of both plots, meaning that these individuals use similar language across both OSNs, and this language is similar to other users. This suggests that both OSNs have similar linguistic norms, and that most users adopt these conventions. Second, there are many users in the upper-left half of each graph, but not in the lower-right. These users tend to use different language constructs on Pinterest and Twitter. However, although these users diverge from themselves, they still tend to conform to the linguistic norms of the platform. This suggests that although an individual's linguistic identity may vary across OSNs, they still tend to adopt the conventions of each platform. Prior work has observed similar linguistic adaption on online forums (Danescu-Niculescu-Mizil et al. 2013).

4.3 Categories on Pinterest and Twitter

In this section, we analyze the characteristics of categories on Pinterest and Twitter. For this analysis, we leverage the 41,928 Pinterest and 33,452 Twitter categorized sessions introduced in § 3.4. For clarity, we organize this section around four high-level questions.

The first question we ask is very simple: *which categories are the most popular overall on Pinterest and Twitter?* To answer this, Figure 12 plots the number of sessions classified amongst the top 10 categories on Pinterest on Twitter. The top 2 categories ("Fashion" and "Food & Drink") are shared across both sites, while "Design, Decorations, & Crafts" and "Hair & Beauty" are more popular on Pinterest, and "Health & Sports" is more popular on Twitter. Although other categories, such as "Technology" and "Politics," are very popular amongst the general Twitter population, it is unsurprising that these categories are not as popular with our target users. Our target population intentionally contains many avid Pinterest users, which is reflected in the popularity of the design, food, and beauty categories.

The second question we ask is: *how do users distribute their content across categories?* Are users highly focused (*i.e.,* most of their content is in a few categories), or are users more varied? To answer this question, we calculate the *Shannon Entropy* of each user from the distribution of categories across their sessions. Shannon Entropy is defined as

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i) \quad (1)$$

where p is the user's distribution of sessions, n is the number of categories present in the user's sessions, and p_i is the probability that a given session from u will be of category i. A user specializing in a single category will have $H(p) = 0$, while a user whose sessions are evenly distributed across all categories will have maximum entropy.

Figure 13(a) shows the CDF distribution of *session entropy* for our target users. Figure 13(a) reveals that even though the Pinterest platform has a wider range of popular

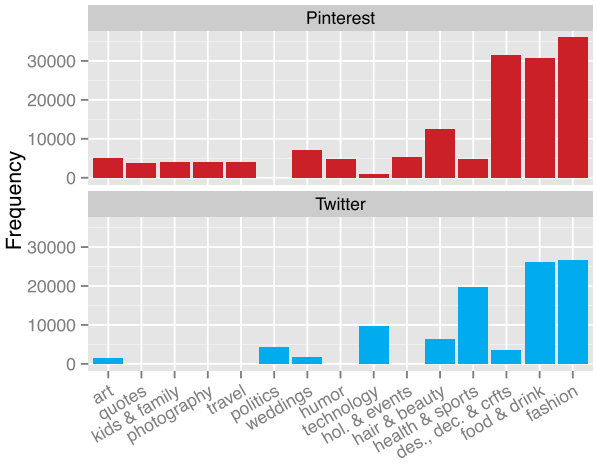


Figure 12: Popularity of the top 10 most frequent categories on Pinterest and Twitter.

categories (see Figure 12), users tend to specialize more on Pinterest than on Twitter. This result seems to agree with the prevailing understanding of Twitter as a general communication platform, versus Pinterest which is organized around particular interest groups (*e.g.,* fashion, food, photography).

The third question we ask stems naturally from our second question: *can you predict the categories of a user's tweets based on their pins, or vice-versa?* To answer this question, we use the metric *Cross Entropy*, defined as:

$$H(p,m) = -\sum_{i=1}^{n} p_i \log(m_i), \qquad (2)$$

where p and m are the user's distributions of sessions either in Pinterest or Twitter. *Cross Entropy* is used to compare two probability distributions: if $p = m$ then $H(p) = H(p,m)$, therefore the closer the $H(p,m)$ is to the true entropy $H(p)$ (eq. 1), the better m is as an approximation of p. Intuitively, *Cross Entropy* scores closer to zero (*i.e.,* low entropy) indicate that the model distribution m is a good predictor of the observed distribution p.

Figure 13(b) plots the CDF of *Cross Entropy* for our target users. The red line represents $H(Pinterest, Twitter)$ (*i.e.,* Twitter is the model distribution), while the blue line represents $H(Twitter, Pinterest)$. Figure 13(b) reveals that for our users, Twitter categories are a better predictor of Pinterest categories than vice-versa. This finding may stem from the fact that the set of popular Twitter categories amongst our target users is more constrained than the set of Pinterest categories (see Figure 12), thus giving the Twitter categories more predictive power.

The final question we ask is: *does each user's session stream exhibit* locality of interest*?* A user with high locality engages in the same categories over many sequential sessions, while a user with low locality spreads their sessions over many different categories with random interleaving. To quantify locality of interest, we use the *stack distance* metric (Cascaval and Padua 2003; Almeida et al. 1996) to analyze each user's session timeline. A stack distance of zero

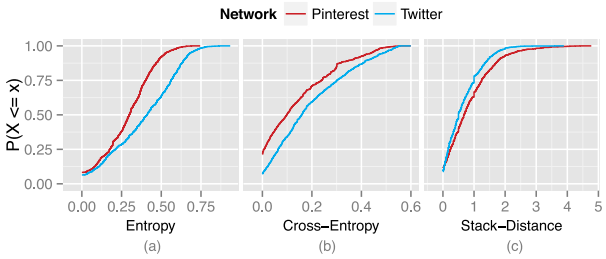


Figure 13: Distributions of *Entropy*, *Cross-entropy* and *Stack-distance* for each user's categorized sessions on Pinterest and Twitter.

indicates that sessions of each category appear close to each other in the timeline.

Figure 13(c) plots the distribution of *stack distances* for our target users. Both OSNs show similar a similar trend, with $>$60% of users having stack distance $\leq$1. This shows that users on both OSNs tend to have many subsequent sessions about the same category of content. This finding suggests that not only do user's interests cross-over between OSNs, but that those interests tend to remain stable for relatively long periods of time.

4.4 Pollination of Categories

In this section, we examine the temporal flow of ideas from Pinterest to Twitter, and vice-versa. Specifically, we pose the question: *where do new ideas first appear: on Pinterest, or on Twitter?* This question is fundamental to the study of information dissemination across OSN boundaries. To answer this question, we again focus on the categorized session introduced in § 3.4.

To begin to address this question, Figure 14 plots a heatmap of transitions from category to category across all of our target users. For each user, we take all of their session on Pinterest *and* Twitter and put them into a single, chronological timeline. We then examine the transitions between categories and OSNs over time.

The red region of Figure 14 represents intra-Pinterest transitions, while blue is intra-Twitter, and grey is inter-OSN. The brighter the color, the more frequently we observe that transition. The transitions happen from column to row, *e.g.,* if the user's current session is about technology in Pinterest, the probability that the next session will be about technology in Twitter is $\approx$15%.

There are three noteworthy features of Figure 14. First, there are many self-transitions (*i.e.,* category c to category c) on both OSNs, as shown on the diagonal of the heatmap. This reinforces the finding from the *stack distance* plot (Figure 13(c)). Second, we observe many transitions from all categories into the most popular categories on Pinterest and Twitter, which manifest as the horizontal bands of color in the heatmap. Although transitions from one OSN into a popular category on the other OSN do occur, they are less frequent than intra-OSN transitions.

Third, there are specific inter-OSN self-transitions that occur more frequently than others. Examples include "De-

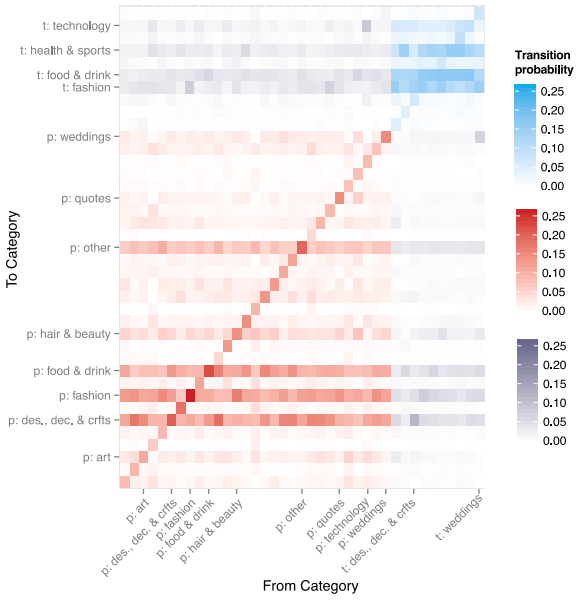


Figure 14: Heatmap of inter and intra-OSN category transitions.

sign, Decorations, & Crafts" and "Weddings" from Twitter to Pinterest, and "Fashion" and "Technology" from Pinterest to Twitter. These cases epitomize *cross-pollination*, or information transfer, from one OSN to another. One question that arises from Figure 14 is: *why are inter-OSN transitions less frequent than intra-OSN transitions?* In future work, we plan to examine whether the proliferation of mobile social apps plays are role in this, since it is more difficult to multi-task between apps on a smartphone than on a desktop.

Network Leaning Although Figure 14 shows which inter-OSN category transitions are common between Pinterest and Twitter, it does not tell us the strength of information transfer in either direction. In other words, *which OSN drives innovation: do ideas tend to originate on Pinterest and then move to Twitter, or vice-versa?*

To answer this question, we define a metric called *Cross Network Precedence* (CNP) as the number of times a session of category c on OSN X at start time t precedes another session of category c on OSN Y at time $t' > t$ for a specific user. Simply put, CNP counts the number inter-OSN self-transitions for each user. Now, we define the *Network Leaning* (NL) score of a user as the *relative difference* between the sum of all CNP of his/her sessions across the OSNs:

$$NL(u) = \frac{\sum_{i=1}^{S_P} CNP(i) - \sum_{i=1}^{S_T} CNP(i)}{\sum_{i=1}^{S_P} CNP(i) + \sum_{i=1}^{S_T} CNP(i)}, \qquad (3)$$

where S_P and S_T are the user's sessions in *Pinterest* and *Twitter*, respectively. NL varies from -1 to 1, with values close to -1 indicating stronger influence of Twitter over Pinterest, while values close to 1 indicate the contrary.

Figure 15 plots a histogram of the NL scores for our target population. The majority of users fall somewhere in the middle of the range (*i.e.,* mild preference for one OSN over the other), while a defiant subset of users have -1 and 1 scores (*i.e.,* strong preference for one OSN over the other). However, the mean of the NL scores is 0.15 and the median is 0.16 (with skew of -0.2), demonstrating that overall, users tend to start generating content on Pinterest, and then transition to Twitter. This finding demonstrates that although Twitter may have a large, vibrant user community, smaller OSNs like Pinterest play a key role in fostering new ideas, which can then transition onto and trend on Twitter.

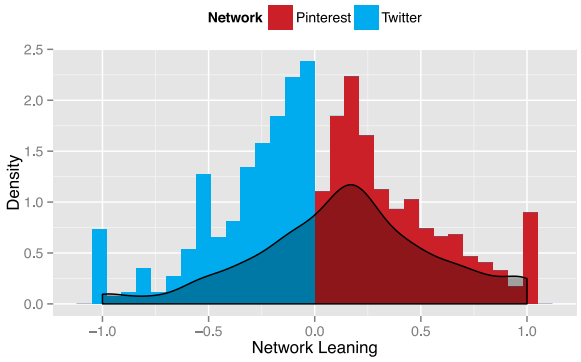


Figure 15: Histogram of Network Leaning scores for our target users. (*mean* = 0.15, *median* = 0.16, *skew* = -0.2)

5 Related work

We now briefly summarize existing studies of cross-OSN user behavior, as well as studies of user activity on Twitter and Pinterest. The most related work to ours explores how users curate content in Pinterest and Last.fm (Zhong et al. 2013). While the authors focus on the popularity of items that users choose, we focus on how users post across multiple sites; thus, the two studies are highly complementary.

There have been a number of approaches to categorizing Twitter content, *e.g.,* (Duh et al. 2011; Hong and Davison 2010). Twitter itself categorizes its trending hashtags/topics as *trends* in reviews published yearly. For example, the 2012 Review[9] labels classes of hashtags as politics, sport, tech, and food. Unfortunately, Twitter's classification methodology is not public. Other work (Lee et al. 2011) proposes a method to map Twitter's *trending topics* to higher level categories. The result is similar to our methodology, with most of the categories overlapping with our Pinterest-based categories. In the domain of news, Zhao et al. (Zhao et al. 2011) topicalize tweets based on news labels from the New York Times, with some additions of their own. Finally, other work (Hong and Davison 2010) uses categories from Twitter Suggestions, the former official user recommendation tool; almost all of these categories have an equivalent in Pinterest.

As Pinterest has grown, there have been a number recent studies (*e.g.,* (Feng et al. 2013)) that focus on quantifying and analyzing Pinterest user behavior. Recent work (Zoghbi, Vulić, and Moens 2013) verifies that the images posted to

[9]https://2012.twitter.com/en/trends.html

sites like Pinterest are an accurate reflection of a user's interest, and can be used to recommend relevant products. Other work (Ottoni et al. 2013) has shown that behavior on Pinterest differs significantly by gender.

Finally, the use of AMT to obtain ground truth for experiments has been examined by a number of studies in the past. For example, recent work (Komarov, Reinecke, and Gajos 2013) studies different setting designs in AMT and compares the results with lab-based settings. They find no significant differences between lab and AMT experiments, suggesting that using AMT for ground truth is likely to yield high accuracy.

6 Concluding Discussion

Today, online social networks (OSNs) are extremely popular; many users actively maintain accounts of a variety of sites. While these sites have enabled significant research into the functioning of society at scale, most prior work has focused on user activity within a single site. It remains unclear how the OSN ecosystem fits together, what roles the various sites play, and how users are choosing to distribute their activities across the wide variety of sites today.

We took a first step towards answering these questions in this paper, focusing on users who have accounts on both Twitter and Pinterest. We observe that even though many users seem to maintain a single identity and interests across the two sites, they show markedly different global patterns of activity. Using a novel tweet categorization methodology, we investigate how users distribute their content across the two sites; we find that users tend to engage in more categories of content on Pinterest than on Twitter, and (crucially) that new content tends to germinate on Pinterest, then transfer to Twitter. This underscores the notion that while Twitter may be an incredibly popular communication platform, smaller topic-specific sites like Pinterest play a key role in the generation of new ideas and content. However, because users' activities are more narrowly focused on Twitter, it can actually serve as a good predictor of activity on smaller OSNs like Pinterest.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported in part by NSF grants CNS-1054233 and CNS-1319019, ARO grant W911NF-12-1-0556, and an Amazon Web Services in Education Grant.

References

Almeida, V.; Bestavros, A.; Crovella, M.; and De Oliveira, A. 1996. Characterizing reference locality in the www. In *TPDS*.

Bechar-Israeli, H. 1995. From <bonehead> to <clonehead>: Nicknames, play, and identity on internet relay chat. *Journal of Computer-Mediated Communication* 1(2).

Cascaval, C., and Padua, D. A. 2003. Estimating cache misses and locality using stack distances. ACM.

Chafkin, M. 2012. Can ben silbermann turn pinterest into the world's greatest shopfront? http://bit.ly/OgmHYl.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *WWW*.

Duh, K.; Hirao, T.; Kimura, A.; and Ishiguro, K. 2011. Creating Stories: Social Curation of Twitter Messages. In *ICWSM*.

Feng, Z.; Cong, F.; Chen, K.; and Yu, Y. 2013. An empirical study of user behaviors on pinterest social network. In *WI-IAT*.

Hong, L., and Davison, B. D. 2010. Empirical study of topic modeling in Twitter. In *SOMA*.

Kahn, J. H.; Tobin, R. M.; Massey, A. E.; and Anderson, J. A. 2007. Measuring emotional expression with the linguistic inquiry and word count. *The American Journal of Psychology* 120(2).

Komarov, S.; Reinecke, K.; and Gajos, K. Z. 2013. Crowdsourcing performance evaluations of user interfaces. In *CHI*.

Lee, K.; Palsetia, D.; Narayanan, R.; Patwary, M. M. A.; Agrawal, A.; and Choudhary, A. 2011. Twitter Trending Topic Classification. In *ICDM Workshop*.

Ottoni, R.; Pesce, J. P.; Casas, D. L.; Franciscani Jr., G.; Meira Jr., W.; Kumaraguru, P.; and Almeida, V. 2013. Ladies first: Analyzing gender roles and behaviors in pinterest. In *ICWSM*.

Pennebaker, J. W.; Chung, C. K.; Ireland, M.; Gonzales, A.; and Booth, R. J. 2007. The development and psychometric properties of liwc2007 the university of texas at austin. *Development* 1(2).

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 1999. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum.

Pew Research Center. 2013. Social media update. http://bit.ly/1fXokJy.

Quercia, D.; Askham, H.; and Crowcroft, J. 2012. TweetLDA: Supervised Topic Classification and Link Prediction in Twitter. In *WebSci*.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing Microblogs with Topic Models. In *ICWSM*.

Suler, J. R. 2002. Identity management in cyberspace. *Journal of Applied Psychoanalytic Studies* 4(4).

Tausczik, Y. R., and Pennebaker, J. W. 2009. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1).

Veltman, B. R. 2006. *Linguistic analysis of the semantic content of the Rorschach Inkblot Test.* Ph.D. Dissertation.

Walker, T. 2012. State of us internet in q1 2012. http://bit.ly/V9L10Y.

Wasserman, T. 2012. Twitter user id numbers cross into the billions. http://on.mash.to/1c5ifeM.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; and Lim, E.-p. 2011. Comparing Twitter and Traditional Media Using Topic Models. In *ECIR*, 338–349.

Zhong, C.; Shah, S.; Sundaravadivelan, K.; and Sastry, N. 2013. Sharing the Loves: Understanding the How and Why of Online Content Curation. In *ICWSM*.

Zoghbi, S.; Vulić, I.; and Moens, M.-F. 2013. I pinned it. where can i buy one like it?: Automatically linking pinterest pins to online webshops. In *DUBMOD*, 9–12. New York, NY, USA: ACM.