REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
07-12-2016	Final Report	3-Aug-2012 - 2-Aug-2016

4. TITLE AND SUBTITLE
Final Report: Taming Twitter: Using social media networks to identify deviant behavior

5a. CONTRACT NUMBER
W911NF-12-1-0379

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Tyler McCormick

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Washington
Office of Sponsored Programs
4333 Brooklyn Ave NE Box 359472
Seattle, WA 98195 -9472

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)
62389-NS-YIP.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
We propose a statistical analysis and data collection framework to identify actors in social media networks who are likely to engage in non-normative or deviant behaviors (specifically, driving under the influence of alcohol or use of marijuana) or have a non-normative attribute (specifically, being obese). We focus on Twitter users' reports of behaviors, via the text of their Tweets, and information about their (online) Twitter social network. Our statistical approach uses novel regularization and hierarchical modeling techniques that are informed by sociological theories on stigma and deviance and previous demographic research on the prevalence and correlates of the behaviors and

15. SUBJECT TERMS
Social media data, statistical analysis, deviant behavior

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Tyler McCormick
UU	UU	UU	UU		19b. TELEPHONE NUMBER 206-221-6981

RPPR Final Report
as of 13-Oct-2022

Agency Code: 21XD

Proposal Number: 62389NSYIP **Agreement Number: W911NF-12-1-0379**
INVESTIGATOR(S):

Name: PhD Ali Shojaie
Email: ashojaie@uw.edu
Phone Number: 2066165323
Principal: N

Name: PhD Hedwig Lee
Email: hedylee@u.washington.edu
Phone Number: 2065434572
Principal: N

Name: PhD Tyler McCormick
Email: tylermc@uw.edu
Phone Number: 2062216981
Principal: Y

Organization: **University of Washington**
Address: Office of Sponsored Programs, Seattle, WA 981959472
Country: USA

DUNS Number: 605799469 EIN: 916001537
Report Date: 31-Oct-2016 Date Received: 07-Dec-2016
Final Report for Period Beginning 03-Aug-2012 and Ending 02-Aug-2016
Title: Taming Twitter: Using social media networks to identify deviant behavior
Begin Performance Period: 03-Aug-2012 **End Performance Period:** 31-Jul-2016
Report Term: 0-Other
Submitted By: Email:
Phone:
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:**

Major Goals:

Accomplishments:

Training Opportunities:

Results Dissemination:

Honors and Awards:

Protocol Activity Status:

Technology Transfer:

PARTICIPANTS:

Participant Type: Graduate Student (research assistant)
Participant: Wesley Lee
Person Months Worked: **Funding Support:**
Project Contribution:
National Academy Member:

RPPR Final Report
as of 13-Oct-2022

ARTICLES:

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Sociological Methods and Research
Publication Identifier Type: Publication Identifier:
Volume: 0 Issue: 0 First Page #: 0
Date Submitted: 12/7/16 12:00AM Date Published:
Publication Location:
Article Title: Using Twitter for Demographic and Social Science Research:Tools for Data Collection
Authors: Tyler H. McCormick, Hedwig Lee, Nina Cesare, Ali Shojaie, Emma S. Spiro
Keywords: Amazon Mechanical Turks, data collection
Abstract: Despite recent interest in using Twitter to examine human behavior and attitudes, little work has been done to develop systematic ways of collecting Twitter data for social science research. Further, gleaning key demographic information about Twitter users, a key component of much social science research, remains a challenge. This paper develops a scalable, sustainable toolkit for social science researchers interested in using Twitter data to examine behaviors and attitudes, as well as the demographic characteristics of the populations expressing or engaging in them. We begin by describing how to collect Twitter data on a particular population – in this case, individuals who do not plan to vote in the 2012 U.S. presidential election. We then describe and evaluate a method for processing data to retrieve demographic information reported by users that is not encoded as text (e.g., details of images) and assess the reliability of these techniques. We end by assessing the challenges of thi
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 5-Submitted
Journal: Journal of the American Statistical Association
Publication Identifier Type: DOI Publication Identifier: 10.1080/01621459.2014.991395
Volume: 0 Issue: 0 First Page #: 0
Date Submitted: Date Published:
Publication Location:
Article Title: Latent surface models for networks using Aggregated Relational Data
Authors:
Keywords: social network, Bayesian analysis
Abstract: Despite increased interest across a range of scientific applications in modeling and understanding social network structure, collecting complete network data remains logistically and financially challenging, especially in the social sciences. This paper introduces a latent surface representation of social network structure for partially observed network data. We derive a multivariate measure of expected (latent) distance between an observed actor and unobserved actors with given features. We also draw novel parallels between our work and dependent data in spatial and ecological statistics. We demonstrate the contribution of our model using a random digit-dial telephone survey and a multiyear prospective study of the relationship between network structure and the spread of infectious disease. The model proposed here is related to previous network models which represents high dimensional structure through a projection to a low-dimensional latent geometric surface—encoding dependence as d
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support:

RPPR Final Report
as of 13-Oct-2022

Publication Type: Journal Article Peer Reviewed: N **Publication Status:** 5-Submitted
Journal: Annals of Applied Statistics (accepted)
Publication Identifier Type: Publication Identifier:
Volume: 0 Issue: 0 First Page #: 0
Date Submitted: Date Published:
Publication Location:
Article Title: Estimating Population Size Using the Network Scale UpMethod
Authors:
Keywords: population size estimation, social network, hard-to-reach groups
Abstract: We develop methods for estimating the size of hard-to-reach populations from data collected using network-based questions on standard surveys. Such data arise by asking respondents how many people they know in a specific group (e.g. people named Michael, intravenous drug users). The Network Scale up Method (NSUM) is a tool for producing population size estimates using these indirect measures of respondents' networks. Killworth et al. (1998a,b) proposed maximum likelihood estimators of population size for a fixed e?ects model in which respondents' degrees or personal network sizes are treated as fixed. We extend this by treating personal network sizes as random effects, yielding principled statements of uncertainty. This allows us to generalize the model to account for variation in people's propensity to know people in particular subgroups (barrier e?ects), such as their tendency to know people like themselves, as well as their lack of awareness of or reluctance to acknowledge their con
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:** 2-Awaiting Publicat
Journal: Annals of Applied Statistics
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 12/7/16 12:00AM Date Published: 12/7/16 4:36PM
Publication Location:
Article Title: LATENT SPACE MODELS FOR MULTIVIEW NETWORK DATA
Authors: Michael Salter-Townsend, Tyler McCormick
Keywords: multiview networks
Abstract: Social relationships consist of interactions along multiple dimen- sions. In social networks, this means that individuals form multiple types of relationships with the same person (an individual will not trust all of his/her acquaintances, for example). Statistical models for these data require understanding two related types of dependence structure: (i) structure within each relationship type, or network view, and (ii) the association between views. In this paper we propose a statistical framework that parsimoniously represents dependence be- tween relationship types while also maintaining enough flexibility to allow individuals to serve different roles in different relationship types. Our approach builds on work on latent space models for net- works (see Hoff et al. (2002), for example). These models represent the propensity for two individuals to form edges as conditionally independent given the distance between the individuals in an unobserved social space.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 13-Oct-2022

Partners

,

I certify that the information in the report is complete and accurate:
Signature:
Signature Date:

Report on Scientific Progress

Tyler H. McCormick Hedwig Lee Ali Shojaie

2016

1 Foreword

This report details progress during the third year of proposal number 62389-CS-YIP.

Contents

2 Statement of problem studied

For this project we proposed a statistical analysis and data collection framework to identify actors in social media networks who are likely to engage in non-normative or deviant behavior (specifically, drunk driving or use of marijuana) or have a non-normative attribute (specifically, being obese). More specifically, we use theoretical paradigms to understand the factors associated with disclosure of deviant or stigmatizing behavior. We combine these theoretical paradigms with statistical tools developed for data collected in open, unstructured environments.

In the first period our work focused on developing scalable data collection infrastructure. The second period brought changes in Twitter data access policies which required us to revise some work during the first period. Nonetheless, we were able to build on the work from the first period to begin examining several non-normative outcomes. We also continued exploring the temporal dynamics of the conversation structure on Twitter using newly developed methods for continuous-time network data. We also further explored sampling techniques that are relevant for gathering data in environment, such as Twitter, that do not have a readily available sampling frame. This paper will appear in the *Proceedings of the National Academy of Science*. We further examined a new way to model network structure that is particularly relevant for large, sparse graphs like Twitter.

3 Summary of important results

In the following sections we detail our main results.

- **Continuous-time temporal structure** Relational event data, which consist of events involving pairs of actors over time, are now commonly available at the finest of temporal resolutions. Existing continuous-time methods for modeling such data are based on point processes and directly model interaction "contagion," whereby one interaction increases the propensity of future interactions among actors, often as dictated by some latent variable structure. In this article, we present an alternative approach to using temporal-relational point process models for continuous-time event data. We characterize interactions between a pair of actors as either spurious or that resulting from an underlying, persistent connection in a latent social network. We argue that consistent deviations from expected behavior, rather than solely high frequency counts, are crucial for identifying well-established underlying social relationships. To explore these latent network structures in two contexts: one comprising of college students and another involving barn swallows.

  Analyses of relational event data typically involve representing interactions in data structures congruent with existing discrete-time dynamic social network models. To mirror the form of traditional survey network data, interactions are commonly aggregated into a series of weighted adjacency matrices, also called sociomatrices, whose elements represent the number of interactions between each pair of actors within fixed time intervals. Taken together, the sequence of adjacency matrices can be viewed as a weighted network which evolves in discrete-time. Inference on discrete-time dynamic networks is typically performed using models that characterize the manner in which dyadic interactions change from one time interval to the next, often assuming a Markov process whereby the network at time t only depends on its past states through the network at time $t-1$.

  One issue with the aforementioned aggregation approach is the implicit assumption that unobserved meaningful social (network) relations are easily ascertained from the observed noisy interaction data. The aggregation methods are designed for scenarios in which the data contain complete network information, consisting of the strength of each dyad relation. However, continuous-time interaction data consists only of measurements *when individuals interact*, and the absence of interaction should not be taken an explicit declaration of no relationship. More generally, interaction counts between dyads should not necessarily be taken as a direct measure of the strength of the underlying dyad relationship. Consider call records in which actor A may call actor B multiple times per week and call actor C on Sunday morning once a month. From our perspective, both of these call patterns may indicate strong relations between actor A and the others. Actor B may, for example, represent a roommate, whereas actor C may represent actor A's grandparent. Although communication with the grandparent is relatively infrequent compared to the communication with the roommate, the relationship between the actor A and his/her grandparent should be classified as strong and significant.

  Another key drawback of performing discrete-time data aggregation when modeling continuous-time interaction data is that the time intervals are often chosen arbitrarily and conclusions can be greatly impacted by these choices. In addition, the temporal dynamics of the interactions are possibly lost in the aggregation process depending on the length of the time

interval. Consider Figure 1, for example, where proximity data from a barn swallow study has been aggregated over each day. By aggregating to the level of the day we preclude the

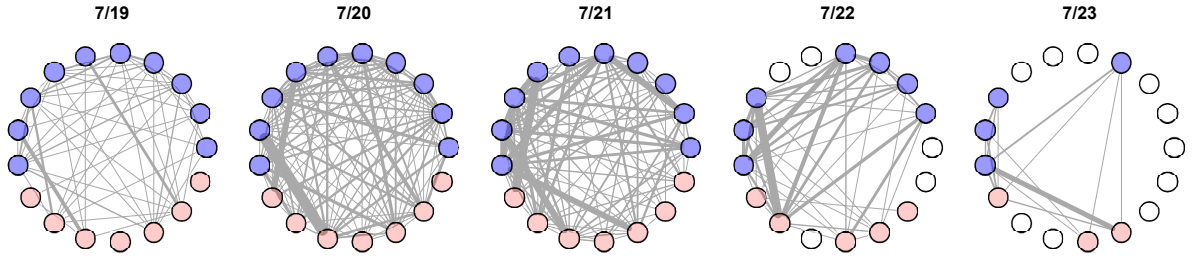

Figure 1: Daily snapshots of aggregated barn swallow interactions. Nodes are colored by sex (females red; males blue) and edges widths are proportional to the number of interactions between each pair. Birds deemed to be unmonitored on a given day are shown in white.

possibility of capturing intricate within-day social dynamics. While smaller time intervals are able to capture more of these dynamics, models with Markov structure may be inappropriate for modeling the resulting series of weighted networks. Conditional dependence among behavior across time would then be more likely to span multiple intervals. Returning to the example of call records, an assumption of Markovian dynamics for summaries of behavior at, for example, an hourly level would be unsuitable since associated individuals are unlikely to call each other on such a frequent basis.

We propose a continuous-time approach to modeling relational event data that explicitly separates interactions from underlying social relations. Specifically, we conceptualize continuous-time interaction data as representing *manifestations of latent network structure*. The model we propose possesses two distinctive properties. First, rather than viewing the data as direct observations of the network of interest, we assume observed interactions arise from a point processes with propensity influenced by the latent network structure. It is the dynamics of this *underlying* social network that we argue is typically most informative about population social structure and of direct interest to researchers. Second, we avoid decisions on temporal resolution by modeling both the observed interactions and dynamics of the latent network structure in continuous-time. The statistical challenge, therefore, is to infer the underlying network structure and its evolution through time. In this paper we assume interactions and connection in the latent network are undirected, although our methods could readily be extended to handle directed interactions and networks.

- **Sampling based on networks.** Respondent-driven sampling (RDS) is a network-based form of chain-referral sampling used to estimate attributes of populations that are difficult to access using standard survey tools. Although it has grown quickly in popularity since its introduction, the statistical properties of RDS estimates remain elusive. In particular, the sampling variability of these estimates has been shown to be much higher than previously acknowledged, and even methods designed to account for RDS result in misleadingly narrow confidence intervals. In this paper, we introduce a tree bootstrap method for estimating uncertainty in RDS estimates based on resampling recruitment trees. We use simulations from known social networks to show that the tree bootstrap method not only outperforms existing methods but also captures the high variability of RDS, even in extreme cases with

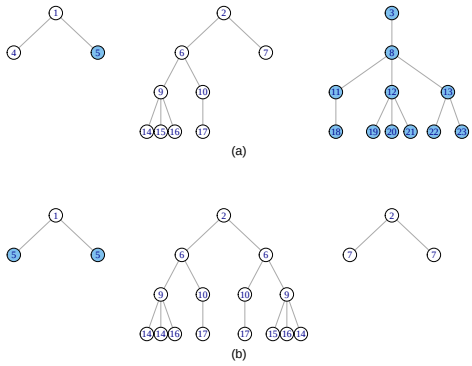

Figure 2: *(a) Example of an RDS recruitment tree, and (b) a resample taken from it by the tree bootstrap method. Individuals are shaded according to their attribute value.*

high design effects. We also apply the method to data from injecting drug users in Ukraine. Unlike other methods, the tree bootstrap depends only on the structure of the sampled recruitment trees, not on the attributes being measured on the respondents, so correlations between attributes can be estimated as well as variability. Our results suggest that it is possible to accurately assess the high level of uncertainty inherent in RDS.

Our method is essentially a multilevel bootstrap procedure applied within the hierarchical framework of the RDS recruitment trees. To draw a bootstrap sample from a set of observed trees, the initial step is to resample with replacement from the seeds of the trees. Next, from each of those seeds we resample with replacement from their recruits, creating the second level of the bootstrap sample trees. From each of these new recruits we then resample with replacement from their recruits to create a third level. This process continues iteratively until no further recruits are available. From the resulting bootstrap sample trees, any statistic of interest, such as the Volz-Heckathorn estimator, can be computed. By taking multiple bootstrap samples, the sampling distribution of the statistic can then be estimated from the observed RDS trees in a way that respects the dependence within the sample. This is similar to other well-known techniques for resampling from correlated data, such as the block bootstrap methods for time series or spatial data, except that in our case the structure of the dependence comes from the RDS recruitment process instead of proximity in time or space. We note that due to the asymmetries of the observed RDS trees, the tree bootstrap produces resamples that vary in size. In order to alleviate any bias that may result from this, any inference based upon the bootstrap distribution of a statistic should be weighted by the effective relative sample sizes (e.g., $\sum_i 1/d_i$ for the Volz-Heckathorn estimator).

Figure 2(b) shows an example of a bootstrap sample drawn from the observed trees in Figure 2(a). Note that seed 2 was selected twice in the initial resampling step while seed 3 was not selected, but the resampled trees resulting from the two copies of seed 2 are quite different due to the further recruits that were selected in later steps. Also note that while the individuals are shaded according to their attribute value, these values do not affect the resampling procedure. However, variablity in the sampling distribution of statistics involving the attribute values will be represented in the changing structure of the resampled bootstrap trees. In fact, we can see from this example that the substantial attribute

homophily observed in Figure 2(a) will result in a higher degree of variability in attribute statistics when using the tree bootstrap method than we would expect if we used a standard bootstrap method.

- **Modeling large, sparse graphs.** Many existing statistical and machine learning tools for social network analysis focus on a single level of analysis. Methods designed for clustering optimize a global partition of the graph, whereas projection based approaches (e.g. the latent space model in the statistics literature) represent in rich detail the roles of individuals. Many pertinent questions in sociology and economics, however, span multiple scales of analysis. Further, many questions involve comparisons across disconnected graphs that will, inevitably be of different sizes, either due to missing data or the inherent heterogeneity in real-world networks. We propose a class of network models that represent network structure on multiple scales and facilitate comparison across graphs with different numbers of individuals. These models differentially invest modeling effort within subgraphs of high density, often termed communities, while maintaining a parsimonious structure between said subgraphs. We show that our model class is projective, highlighting an ongoing discussion in the social network modeling literature on the dependence of inference paradigms on the size of the observed graph.

  we propose a multiresolution model for capturing heterogeneous, complex structure in social networks that exhibit strong community structure. Our modeling framework decomposes network structure into a component that describes between-community relations, i.e. relations between actors belonging to different communities, and another component describing within-community relations. The proposed framework has two distinct advantages over existing methods. First, our framework is that it is able to accommodate a wide variety of models for between- and within-community relations. This feature allows the model to be tailored to reflect different scientific questions that arise when exploring the behavior within and across these communities. The second advantage of our model is that it balances parsimony with model richness by selectively directing modeling efforts towards representing interesting, relevant network structure. Typically, this structure is found within actors' local communities. In such cases, we can exert the most modeling effort (i.e. model complexity and computational effort) within dense pockets, where we expect the most complex dependence structure, and use a parsimonious model to capture between community patterns. A similar approach has been adopted in spatial statistics where locations are partitioned into disjoint dependence neighborhoods. Compared to popular network models that capture global structure, our approach can provide increased resolution on intricate structure within communities. Furthermore, our model is able to apportion little effort to modeling simple structure, resulting in a model that is substantially less complex than existing models focused on local structure for networks, even with only a few hundred actors.

3.1 Student support and engagement

As described in our proposal we hold regular group meetings with all key personnel and paid and unpaid graduate and undergraduate students throughout the duration of the project. We have continued these meetings over the past year, with a notable increase in undergraduate participation. We have established a "mentoring ladder" where we assign undergraduate students

to work on projects in teams with graduate students under our supervision. This provides undergraduates exposure to research while also allowing the graduate students to gain experience managing projects and mentoring junior scholars. We plan to continue these working group meetings along with our additional bi-weekly PI only meetings and our bi-weekly one-on-one meetings with graduate and undergraduate students.

3.2 Dissemination

Along with the papers and conference presentations listed in the report, we would like to highlight a few key dissemination activities. Members of the project team gave presentations at the Population Association of American Annual Meetings, the Joint Statistical Meetings and the American Sociological Association annual meetings.