AFRL-AFOSR-VA-TR-2018-0349

Collective Attention Threats: Models and Deterrence

James Caverlee
TEXAS ENGINEERING EXPERIMENT STATION COLLEGE STATION

07/17/2018
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
15-07-2018	Final Performance Report	May 2015 - April 2018

4. TITLE AND SUBTITLE

Collective Attention Threats: Models and Deterrence

5a. CONTRACT NUMBER

5b. GRANT NUMBER

FA9550-15-1-0149

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

PI: James Caverlee, Associate Professor
Department of Computer Science and Engineering
Texas A&M University
Co-PI: Anna Squicciarini
Pennsylvania State University

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Texas Engineering Experiment Station
1470 William D Fitch Pkwy
College Station, TX 77845-4645

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Scientific Research
875 North Randolph Street
Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A - Approved for Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This project investigates threats to collective attention and develops both models of these phenomena as well as methods to deter their impact and reach. These collective attention threats manipulate opinion, rapidly spread malware, and disseminate misinformation, all amplified by collective attention. Unlike traditional threats, users themselves are unwitting accomplices to the spread, infection rate, and success of these new threats. This project aims to develop new models, algorithms, and systems for defending against emergent collective attention threats in large-scale systems.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(Include area code)*

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATE COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33315-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report. e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Title: Collective Attention Threats: Models and Deterrence (FA9550-15-1-0149)

Primary Investigator: James Caverlee, Department of Computer Science and Engineering, Texas A&M University, caverlee@cse.tamu.edu

Co-Investigator: Anna Squicciarini, College of Information Sciences and Technology Pennsylvania State University, asquicciarini@ist.psu.edu

Problem / Objective: This project investigates threats to collective attention and develops both models of these phenomena as well as methods to deter their impact and reach. These collective attention threats manipulate opinion, rapidly spread malware, and disseminate misinformation, all amplified by collective attention. Unlike traditional threats, users themselves are unwitting accomplices to the spread, infection rate, and success of these new threats. Toward tackling this challenge, this project aims to develop new models, algorithms, and systems for defending against emergent collective attention threats in large-scale systems.

Results & Impact: We have made a number of key advances in this project. First, we have developed methods to uncover hidden campaigns. Compared to traditional spam bots that typically leave identifiable footprints, these human-powered deceptive campaigns are inherently distinct, linked only by their common theme and not in common keywords, phrases, or other easily identifiable signals. And since deceptive campaigns are generated by humans rather than bots, their ongoing detection is even more challenging since crowds can actively circumvent detection methods. We have pioneered methods to extract signatures of crowdsourced manipulators, exploit graph-based locality to uncover hidden manipulators, and uncover two-faced manipulators who engage in a mix of legitimate and malicious behaviors. Second, we have developed algorithmic approaches toward controlling or at least managing some instances of collective malicious behavior. We have started with crowdsourcing platforms, where users' contribution can be measured and controlled before it is actually used for task completion or knowledge extraction. We also developed a user reputation system that aims to balance the estimation accuracy of users' reliability and its cost by combining and controlling different types of users' performance measures (i.e., including peer consistency evaluations and gold standard tests). We have formulated integer-programming solutions that identify early evidence of deviance, and in a recent study, how to control possible attempts of deviance through a dynamic rewarding system (i.e. a pricing policy in case of paid contributions). Finally, we have investigated the strategies of propagandists like ISIS where we find through an analysis of all Arabic tweets in 2015 (more than 9 billion) that ISIS is relatively ineffective. Although hailed by the FBI as the "most adept terrorist group" on social media, we find that their collective influence has been steadily waning and was relatively small to begin with (mainly since their interactions online tend to be insular or augmented by bots rather than organic social media users).

Publications / Accomplishments:

C. Qiu, A. Squicciarini, C. Griffin, P. Umar. Combating Behavioral Deviance via User Behavior Control. International Conference on Autonomous Agents and Multiagent Systems, 2018. To appear.

C. Qiu, A. Squicciarini, D. Khare, B. Carminati, and J. Caverlee. CrowdEval: A Cost-Efficient Strategy to Evaluate Crowdsourced Worker's Reliability. International Conference on Autonomous Agents and Multiagent Systems, 2018. To appear.

P. Kaghazgaran and J. Caverlee. Combating Crowdsourced Review Manipulators: A Neighborhood-Based Approach. 11th ACM Conference on Web Search and Data Mining, 2018. DOI: 10.1145/3159652.3159726

W. Yao, Z. Dai, R. Huang, and J. Caverlee. Online Deception Detection Refueled by Real World Data Collection. Recent Advances in Natural Language Processing, 2017. DOI: 10.26615/978-954-452-049-6_102

M. Alfifi and J. Caverlee. Badly Evolved? Exploring Long-Surviving Suspicious Users on Twitter, 9th International Conference on Social Informatics, 2017. DOI: 10.1007/978-3-319-67217-5_14

A. Squicciarini, S. Rajtmajer and C. Griffin. Positive and Negative Behavioral Analysis in Social Networks, *WIREs Data Mining and Knowledge Discovery*, 7(3), 2017.

C. Qiu, A. Squicciarini, S. Rajtmajer and J. Caverlee. Dynamic Contract Design for Heterogeneous Workers in Crowdsourcing for Quality Control. IEEE ICDCS 2017. DOI: 10.1109/ICDCS.2017.187

S. Li, J. Caverlee, W. Niu, and J. Caverlee. Crowdsourced App Review Manipulation (short paper). ACM SIGIR 2017. DOI: 10.1145/3077136.3080741

P. Kaghazgaran, J. Caverlee, and M. Alfifi. Behavioral Analysis of Review Fraud: Linking Malicious Crowdsourcing to Amazon and Beyond (short paper). AAAI ICWSM 2017.

C. Qiu, A. Squicciarini, B. Carminati, and J. Caverlee. CrowdSelect: Increasing Accuracy of Crowdsourcing Tasks through Behavior Prediction and User Selection. ACM CIKM 2016. DOI: 10.1145/2983323.2983830

H. Ge, J. Caverlee, N. Zhang, and A. Squicciarini. Uncovering the Spatio-Temporal Dynamics of Memes in the Presence of Incomplete Information. ACM CIKM 2016. DOI: 10.1145/2983323.2983782

C. Liao, A. Squicciarini, C. Griffin, and S. Rajtmajer. A hybrid epidemic model for Deindividuation and Antinormative Behavior in Online Social Networks. Social Network Analysis and Mining Journal, Springer (2016) 6:13. DOI: 10.1007/s13278-016-0321-5

Invited Addresses:

- Access Control and Information Disclosure in Social Networks, Boise State University, Guest Seminar, April 3, 2018.
- Invited expert, ISAT/DARPA Reality Jamming: Technology-Enabled Misinformation at Scale Workshop at Columbia University, October 26-27, 2017.
- A Peak into the Future Present of Strategic Manipulation, University of New Mexico, Distinguished Lecture, October 11, 2017.
- A Peak into the Future Present of Strategic Manipulation, U.S. Department of Defense Strategic Multi-Layer Assessment (SMA), Invited Speaker, August 21, 2017.
- How to define and detect misinformation? Invited speaker at the ICWSM Digital Misinformation Workshop, May 15, 2017.
- Automating Image Privacy Controls, National Academy of Science/Kavli Symposium, Ambon, Indonesia. July 20, 2017
- Weaponized Crowdsourcing, NSA/DHS National Centers for Academic Excellence CAE Forum, April 27, 2017.
- Caverlee Lab Overview, National Security Agency (NSA) San Antonio, April 10, 2017.
- Invited expert, ARO-Sponsored Workshop on Trustworthy Human-Centric Social Networking, May 12-13, 2016.
- Geo-Social Media Analytics: Opportunities and Challenges, Lab 41 Technical Outreach Day invited speaker, January 13, 2016.
- Caverlee Lab Overview, National Security Agency (NSA), November 16, 2015.