Advancing the Science of Social Cyber-Security

Kathleen M. Carley

In 1672, Charles II issued a proclamation "to restrain the spreading of false news" that was helping "to nourish an universal jealousie and dissatisfaction in the minds of all His Majesties good subjects". Today, legislative bodies around the world are asking the same thing; although, reframed as "how can social media platforms restrain the spreading of false news?" Scientists, journalists, policy makers, the general public and the social media corporations are looking at posts in the cyber-mediated information environment and realizing that there is high variance in the credibility of posts. Further, it is generally recognized that in this information environment disinformation, misinformation and ok information are being spread at an unprecedented rate to groups of unprecedented size. While research is underway to measure the credibility of a post and to identify who is spreading it; less is understood about how this unprecedented scale of false information and its spread will impact human activity, and the mechanisms that enable or contain that spread.

What is Proposed

To address this gap a two-day workshop is proposed that will focus on these two questions: How does the unprecedented scale of false information and its spread impact human activity? What are the mechanisms that enable or contain the spread of false information? This workshop will consider what is needed to characterize, measures, and effect the impact of information with different levels of credibility on, and its spread to, the diverse consumers of that information. This workshop begins with the assumption that even if the credibility of information could be measured and the networks through which the information was spread were identified; we would still not be able to understand how the spread of false information in the cyber-mediated information environment impacts human activity, nor how to enable or contain its spread. This workshop is intentionally framed from the broader perspective of social cyber-security so as to reduce the tendency of participants and observers to over-focus on issues of identification of disinformation and those who spread it, and instead to focus more on questions related to who is being impacted, how great is that impact, and how can the potential impact be changed.

Objective

The purpose of this workshop is to refine the proposed statement of the problem space, to identify the immediate research challenges, and to characterize the potential scientific paths forward that appear to hold the greatest promise for understanding the socio-cultural impact of information diffusion at scale for information varying in credibility and the mechanisms that enable or contain its spread. In doing this work we expect to forge a shared understanding of the major influencers, factors, and the interactions among those factors at a data, theory, and method level, that must be considered to advance our ability to understand, measure, predict and affect how the unprecedented scale of false information and its spread will impact human activity and the mechanism that enable or contain this spread. Specifically, the objective is to identify what basic science is needed to advance out understanding of the spread of, susceptibility to, and impact on groups of disinformation and misinformation in contrast to ok information in the cyber-mediated environment. The goal is to identify the theoretical basis of relevance in addressing this issue, gaps in our understanding, and the types of scientific approaches that may

be productive. This workshop is aimed at defining the way ahead for advancing the science of social cyber-security as it concerns the spread and impact of dis/mis and ok information.

Background

The issue of concern is: how will the unprecedented scale of false information and its spread impact human behavior, and what are the mechanisms that are enabling or containing that spread. Scientists have approached this issue in many ways. Indeed indepth literature reviews in this area point to thousands of researchers producing hundreds of papers (Carley et al, 2018) from multiple theoretical perspectives (Paletz et al, 2018). Before continuing to describe the research on this issue it is important to note that this issue is one of the dominant issues in the emerging area of social cyber-security. According to Carley et al (2018) "Social Cyber-security is an emerging scientific area focused on the science to characterize, understand, and forecast **cyber-mediated** changes in human behavior, social, cultural and political outcomes, and to build the cyber-infrastructure needed for **society** to persist in its essential character in a **cyber-mediated** information environment under changing conditions, and actual or imminent social cyber-threats. … Fundamental to this area is the perspective that we need to maintain and preserve a free and open information environment in which ideas can be exchanged freely, the information source is known, disinformation and false data are identifiable and minimized, and technology is not used to distort public opinion." The other major issues being participatory democracy and to a much less extent insider threat. Of the over 800 papers identified in this study, over 500 were concerned with misinformation, disinformation, propaganda, who was spreading it and the role of bots in its spread, its impact on the population, and the way information could be used to influence and manipulate individuals and groups.

This broader context is valuable for a number of reasons. First, it moves raises the discussion of the spread of false information from the form of a problem dujour to the form of an emerging science. Second, it stresses the critical nature of moving from characterization of a phenomena to understanding, predicting and impacting. That is it moves the discussion beyond what is false information to why do we care that it is spreading. This provides a background against which to develop a research agenda. Third, it paves the way for reasoning about the unprecedented quantity and spread of false information as just one of the activities that can destroy a free and open information environment.

The study by Carley et al (2018) places social cyber-security as a discipline, and the sub-issues related to the spread of false information, at the intersection of computational social science, media and marketing, and policy and law – see Figure 1. It is argued that this is an emergent computational social science field where computer science approaches to social science problems or social science utilization of existing computer science techniques are insufficient. Rather, new transdisciplinary methods are needed that blend the computational and the social. Moreover, it is an applied field in which social change, policy/law, and new technologies must co-evolve in a synergistic fashion.

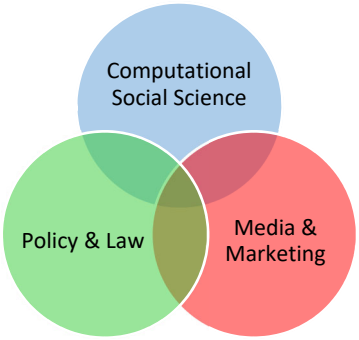

Figure 1. A new field emerging at the intersection.

Key relevant theories are related to persuasion (Grass and Seiter, 2015), social influence (Benigni et al, 2017), individualized collective action (Bennett, 2012), information diffusion (Wu and Lin, 2018), manipulation (Colliander and Dahlén, 2011), identity creation (Josph et al, 2016), strategic messaging (Benigni et al, 2017), information warfare (Cordesman and Cordesman, 2002), digital forensics (Al-Khateeb et al, 2017) and power (Entman, 2007). Researchers in this area employ multi-technology computational social science tool chains (Benigni and Carley, 2016) employing network analysis and visualization (Carley et al, 2016), language technologies (Hu and Liu, 2012), data-mining and statistics (Agarwal et al, 2012), spatial analytics (Cervone et al, 2016), and machine learning (Wei et al, 2016). Finally, the theoretical results and analytics are often multi-level focusing simultaneously on change at the community and conversation level, change at the individual and group level, and so forth

Research in social cyber-security, and on our specific issue, has grown exponentially – see Figure 2. Note when collecting this data the authors did a multi-start snowball, combined with multiple key word searchers, and then removed all papers that were pure machine learning or privacy related. In the final analysis there are over 60 fields involved, and those that currently dominate are shown in Figure 3. What is key about this results is that while there are papers from many social science disciplines and many branches of computer science, it tends to be the more interdisciplinary/transdisciplinary variants that stand out. This interdisciplinary/transdisciplinary approach and the growth in the amount of science in this area is also reflected in the review by Paletz. The review by Paletz covers less of the political science, computer science, data-mining and social networks. However, approximately two-thirds of the papers covered Paletz that actually dealt with the cyber-mediated information environment are part of the Carley et al study. These reviews lead to a number of key insights. Four of these are: Credibility assessment is insufficient. Stationarity cannot be assumed. Structure, cognition and social cognition impact spread. A deep understanding of the media technology is required.

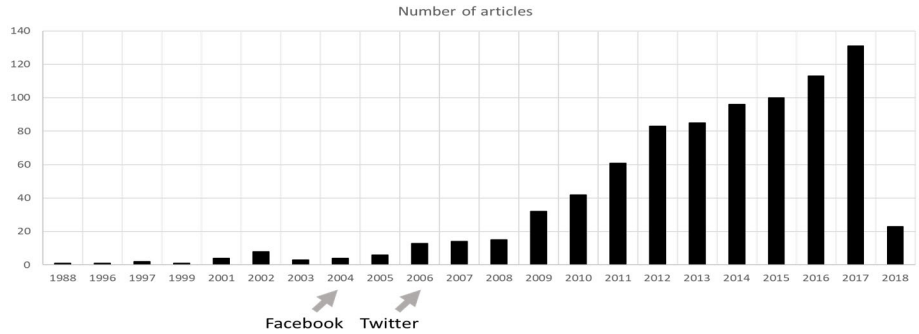

Figure 2. Growth of research in social cyber-security.

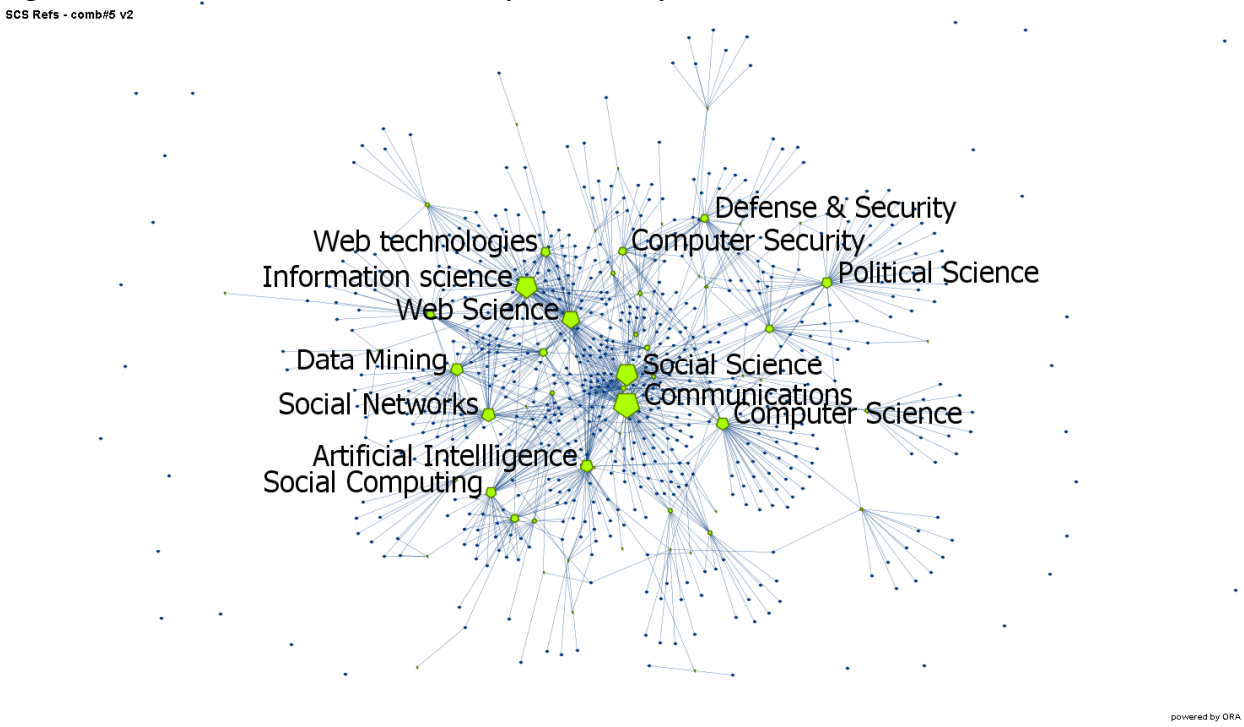

Figure 3. Network depiction showing links between articles and disciplines, such that the disciplines are sized by total degree centrality (number of articles in it).

Credibility assessment is insufficient.

Initially, information credibility was thought to be easily identifiable leading to numerous computer science machine learning solutions and some human fact checking solutions. The upshot, however, is that the scale makes human fact checking impossible. Moreover, the literature shows that there are many types of "false" information including, but not limited to, complete ridiculous lies (Hilary Clinton's alien baby), logical errors (anti-vax campaign), confusion of correlation and causation (the anti-vax campaign), misinterpretation of statistics (climate change), satire, and sarcasm (e.g., Tandock etal, 2018; Babcock et al, 2018). This means detection is unlikely to ever be perfectly accurate. Further, while much research to detect false information and measure the credibility of posts is underway, it may not affect how false information spreads. People spread false information knowing its false (Gupta et al, 2013) and

being told something is false doesn't make people believe it to be so (Polange, 2012). Multiple approaches for countering false information have been suggested (Alemanno, 2018). Moreover, the mere spread may set agendas that persist independent of the veracity of the claims (Vargo et al, 2018).

Nevertheless, much of the work on information credibility and social media is framed as a technical problem of detecting and filtering "fake news," "misinformation," or "disinformation" in social media networks. This framing tends to draw researchers primarily from computer science and information science, and to a lesser extent from sociology, communications, and social networks. Many of these researchers tend to approach the problem by attempting to develop new tools employing statistical, machine learning, or language technology approaches to identify and classify the messages or the messengers. Issues of social influence, cognitive biases, marketing and so forth are often relegated to the backseat. In this case, defining the problem as detecting and filtering implies that if we just removed the inaccurate information and nefarious actors from the network, information accuracy would be improved and information would be more credible. Machine learning-oriented methods often make assumptions that there is a ground truth, and that the truth value of a piece, collection, or source of information is binary and absolute; none of which may be true.

.

Stationarity cannot be assumed.

A great deal of research in this area has considered the issue – "who is spreading the information?". For example, in these literature reviews 78 papers were identified that focused on the role of bots. The bots described, however, are evolving (Oentrayo et al, 2016). Early bots spread spam or simply retweeted (Messias et al, 2013). Then bots that participated socially appeared (Maréchal, 2016). Then networks of emerged that engaged in repeating each others activities (Ferrara et al, 2016). More recently more sophisticated forms of bot coordination are emerging, and bots are acting to exploit both features of the social media technology and social cognition (Benigni et al, 2018). Social media companies have responded in knee jerk reactions by removing actors that spread malicious or false information, changing the rules of services, or altering what data is available through the APIS. See for example – "This is How Facebook Has Changed Over the Past 12 Years." This continual flux impact the repeatability of the scientific studies (Wei et al, 2016). Moreover the flux is so fast that training sets cannot be constructed so many standard machine learning techniques won't work.

Structure, cognition and social cognition impact spread.

Another dominant approach is traditional social network analysis which focuses on issues of social influence. In this case the structure of the network is examined to identify structures and actors that facilitate spread. Typical networks examined include the retweet network in Twitter (e.g., Chatfiel and Brajawidagda, 2012), and the friend network in Facebook (Lewis et al, 2008). Collectively this research is suggesting that traditional methods of influence may be inappropriate as actors move between media (Qasem et al, 2015)), ties are often more representative of passive listening than influence (Romero et al, 2011), and influence even in the same media is a multiple tie phenomenon (Cha et al, 2010). By defining the problem as structural, this body of research is implying that if we just fixed the structure and intervened at the right point the potential influence of non-credible information would be diminished, and/or its flow could be stopped. This approach however, treats the actors as not having agency, and

ignores the cultural, cognitive, and social biases that impact how information is processed and the social organizational and media factors that impact the structure of the network. Whereas, non-network based research points to the criticality of cognitive biases (Bandura, 2009), delivery features (Scott, 2015), and organizational position in social media influence (Segerberg, and Bennett, 2011).

A deep understanding of the media technology is required.

Advances in this area are increasingly requiring a deep understanding of the social media technology itself. Early work on Twitter made the erroneous assumption that the retweet network reflected who retweeted whom; whereas, Twitter assigns the C's retweet of B's retweet of A to being a retweet of A. This error renders invalid some of these early findings. Other examples abound. For example, understanding how the Twitter geo-location spheres operate and the APIS is critical for understanding differences in the results of different types of data samples (Carley et al, 2016). As another example, understanding how API calls are filled is critical for understanding the biases in the data (Morstatter et al, 2013). As the media technology is changing rapidly, researchers in this area need to stay abreast of all changes. Since the changes are often not published, understanding the event or policy change that resulted in the technology change can be crucial in ferreting out what might have changed and building the right test.

This deep understanding of the problem space, indepth assessments of the ongoing research, and experience editing special issues related to this area forms the basis for the design of the proposed workshop.

References

Agarwal, N., Kumar, S., Gao, H., Zafarani, R., Liu, H.: Analyzing behavior of the influentials across social media. In Behavior Computing, pp. 3-19. Springer, London, (2012).

Alemanno, Alberto. "How to Counter Fake News? A Taxonomy of Anti-fake News Approaches." European Journal of Risk Regulation 9, no. 1 (2018): 1-5.

Al-khateeb, S., Hussain, M.N., Agarwal, N.: Social Cyber Forensics Approach to Study Twitter's and Blogs' Influence on Propaganda Campaigns. In: Lee, D., Lin, Y.R., Osgood, N., Thomson, R. (eds.) Proceedings of the International Conference on Social Computing, Behavioral-Cultural and Prediction and Behavior Representation in Modeling and Simulation, pp. 108-113. Springer, Switzerland (2017).

Bandura, Albert. "Social cognitive theory of mass communication." In Media effects, pp. 110-140. Routledge, 2009.

Benigni, M. Joseph, K., Carley, K.M.: Online Extremism and the Communities that Sustain It: Detecting the ISIS Supporting Community on Twitter. PLOS ONE, 12(12), e0181405 (2017).

Benigni, M., Carley, K.M.: From Tweets to Intelligence: Understanding the Islamic Jihad Supporting Community on Twitter. In **Xu**, K.S., **Reitter**, D., **Lee**, D., **Osgood**, N. (eds.) Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pp. 346-355. Springer, Switzerland (2016).

Benigni, Mathew, Kenneth Joseph, and Kathleen M. Carley, 2018-forthcoming, Bot-ivistm: Assessing Information Manipulation in Social Media Using Network Analytics , Emerging

Research Challenges and Opportunities in Social Network Analysis and Mining. Agarwal, Nitin & Dokoohaki, Nima (Eds.). Springer.

Benigni,M., Joseph, K., Carley, K.M.: Mining Online Communities to Inform Strategic Messaging: practical methods to identify community-level insights. Computational and Mathematical Organization Theory, pp. 1-19 (2017).

Bennett, W.L.: The personalization of politics: Political identity, social media, and changing patterns of participation. The ANNALS of the American Academy of Political and Social Science 644(1), 20-39 (2012).

Carley, K. M., Guido Cervone, Nitin Agarwal, Huan Liu, 2018, "Social Cyber-Security," In Proceedings of the International Conference SBP-BRiMS 2018, Halil Bisgin, Ayaz Hyder, Chris Dancy, and Robert Thomson (Eds.) July 10-13, 2018 Washington DC, Springer.

Carley, K.M., Momin, M., Landwehr, P/M., Pfeffer, J. Kowalchuck, M.: 2016, Crowd Sourcing Disaster Management: The Complex Nature of Twitter Usage in Padang Indonesia. Safety Science, 90, 48-61 (2016).

Carley, K.M., Wei, W., Joseph, K.: High Dimensional Network Analytics: Mapping Topic Networks in Twitter Data During the Arab Spring. In Shuguan Cui, Alfred Hero, Zhi-Quan Luo and Jose Moura (eds) Big Data Over Networks, Cambridge University Press, Boston MA (2016).

Carley, Kathleen M., Momin Malik, Pater M. Landwehr, **Jürgen** Pfeffer, and Michael Kowalchuck, 2016, "Crowd Sourcing Disaster Management: The Complex Nature of Twitter Usage in Padang Indonesia." Safety Science, 90: 48-61.

Cervone, G., Sava, E., Huang, Q., Schnebele, E., Harrison, J., Waters, N.: Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. International Journal of Remote Sensing 37(1), 100-124 (2016).

Cha, Meeyoung, Hamed Haddadi, Fabricio Benevenuto, and P. Krishna Gummadi. "Measuring user influence in twitter: The million follower fallacy." Icwsm 10, no. 10-17 (2010): 30.

Chatfield, Akemi, and Uuf Brajawidagda. "Twitter tsunami early warning network: a social network analysis of Twitter information flows." (2012).

Colliander, J., Dahlén, M.: Following the Fashionable Friend: The Power of Social Media: Weighing Publicity Effectiveness of Blogs versus Online Magazines. Journal of advertising research 51(1), 313-320 (2011).

Cordesman, A.H., Cordesman, J.G.: Cyber-threats, information warfare, and critical infrastructure protection: defending the US homeland. Greenwood Publishing Group, Westport, CT (2002).

Entman, R.M.: Framing bias: Media in the distribution of power. Journal of communication 57(1), 163-173 (2007).

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. "The rise of social bots." Communications of the ACM 59, no. 7 (2016): 96-104.

Gass, R.H., Seiter, J.S.: Persuasion: Social influence and compliance gaining. Routledge, UK (2015).

Gupta, Aditi, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy."

In Proceedings of the 22nd international conference on World Wide Web, pp. 729-736. ACM, 2013.

Hu, X., Liu, H.: Text analytics in social media. In Mining text data, pp. 385-414. Springer US (2012).

Joseph, K., Wei , W., Benigni, M., Carley, K.M.: A Social-event Based Approach to Sentiment Analysis of Identities and Behaviors in Text. Journal of Mathematical Sociology. 40(3), 137-166 (2016).

Landwehr, Peter M., Wei Wei, Michael Kowalchuck and Kathleen M. Carley, 2016, Using Tweets to Support Disaster Planning, Warning and Response, Safety Science, 90: 33-47.

Lewis, Kevin, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. "Tastes, ties, and time: A new social network dataset using Facebook. com." Social networks 30, no. 4 (2008): 330-342.

Maréchal, Nathalie. "Automation, algorithms, and politics| when bots tweet: Toward a normative framework for bots on social networking sites (feature)." International Journal of Communication 10 (2016): 10.

Messias, Johnnatan, Lucas Schmidt, Ricardo Augusto Rabelo de Oliveira, and Fabrício Rodrigues Benevenuto. "You followed my bot! Transforming robots into influential users in Twitter." (2013).

Morstatter, F., Pfeffer, J., Liu, H. Carley, K.M.: “Is the Sample Good Enough? Comparing Data (2013).

Mullens, Jenna, 2016, https://www.eonline.com/news/736769/this-is-how-facebook-has-changed-over-the-past-12-years

Oentaryo, Richard J., Arinto Murdopo, Philips K. Prasetyo, and Ee-Peng Lim. "On profiling bots in social media." In International Conference on Social Informatics, pp. 92-109. Springer, Cham, 2016.

Paletz, S. B. F., Auxier, B. E., & Golonka, E. M. (2018). A multidisciplinary, theoretical model of information propagation: Why do people share information and narratives on social media? College Park, MD: University of Maryland Center for Advanced Study of Language.

Polage, Danielle C. "Making up history: False memories of fake news stories." Europe’s Journal of Psychology 8, no. 2 (2012): 245-250.

Qasem, Ziyaad, Marc Jansen, Tobias Hecking, and H. Ulrich Hoppe. "On the detection of influential actors in social media." In Signal-Image Technology & Internet-Based Systems (SITIS), 2015 11th International Conference on, pp. 421-427. IEEE, 2015.

Romero, Daniel M., Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. "Influence and passivity in social media." In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 18-33. Springer, Berlin, Heidelberg, 2011.

Scott, David Meerman. The new rules of marketing and PR: How to use social media, online video, mobile applications, blogs, news releases, and viral marketing to reach buyers directly. John Wiley & Sons, 2015.

Segerberg, Alexandra, and W. Lance Bennett. "Social media and the organization of collective action: Using Twitter to explore the ecologies of two climate change protests." The Communication Review 14, no. 3 (2011): 197-215.

Tandoc Jr, Edson C., Zheng Wei Lim, and Richard Ling. "Defining “fake news” A typology of scholarly definitions." Digital Journalism 6, no. 2 (2018): 137-153.

Vargo, Chris J., Lei Guo, and Michelle A. Amazeen. "The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016." new media & society 20, no. 5 (2018): 2028-2049.

Wei Wei, Kenneth Joseph, Huan Liu and Kathleen M. Carley, 2016, “Exploring Characteristics of Suspended Users and Network Stability on Twitter.” Social network analysis and mining, 6:51.

Wei, W., Joseph, K., Liu, H., Carley, K.M.: Exploring Characteristics of Suspended Users and Network Stability on Twitter. Social network analysis and mining, 6(1), 51. (2016).

Wu, L., Liu, H.: Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In the Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM2018), ACM, NY NY (2018).

Yang, Zi, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang, and Zhong Su. "Understanding retweeting behaviors in social networks." In Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1633-1636. ACM, 2010.