2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Revealing Censored Information Through Comments and Commenters in Online Social Networks

Giuseppe Cascavilla
Sapienza, University of Rome
Via Salaria, 113, 00198 Roma, Italy
Email: cascavilla@di.uniroma1.it

Mauro Conti
University of Padova
Via Trieste, 63 - 35131, Padua, Italy
Email: conti@math.unipd.it

David G. Schwartz
Bar-Ilan University
Ramat Gan, 5290002 Israel
Email: david.schwartz@biu.ac.il

Inbal Yahav
Bar-Ilan University
Ramat Gan, 5290002 Israel
Email: inbal.yahav@biu.ac.il

***Abstract*—In this work we study information leakage through discussions in online social networks. In particular, we focus on articles published by news pages, in which a person's name is censored, and we examine whether the person is identifiable (de-censored) by analyzing comments and social network graphs of commenters. As a case study for our proposed methodology, in this paper we considered 48 articles (Israeli, military related) with censored content, followed by a threaded discussion. We qualitatively study the set of comments and identify comments (in this case referred as "leakers") and the commenter and the censored person. We denote these commenters as "leakers". We found that such comments are present for some 75% of the articles we considered. Finally, leveraging the social network graphs of the leakers, and specifically the overlap among the graphs of the leakers, we are able to identify the censored person. We show the viability of our methodology through some illustrative use cases.**

I. Introduction

A cyber incident can occur under many circumstances and for many reasons. It can be inadvertent, due to the loss of an electronic device, or deliberate such as from the theft of a device or cyber-based attack from a malicious individual, group, agency, foreign nation, terrorist, or other adversary [1]. Online Social Networks (OSN) present extreme cyber-security challenges. Our ability to detect threats in the form of information breaches and leaks of censored information is hampered by the sheer volume of messages and the growing ability of OSN members to operate anonymously. Addressing the former requires computationally efficient and effective text analysis and classification methods, and addressing the latter requires intelligent network topology, relationship, and activity analysis.

Cybersecurity is concerned with attacks on digital infrastructure and measures taken to protect internet-connected computer systems from unauthorized attack or access, theft of digital assets, and using cyberspace as an infrastructure platform for subversive activity [2]. This article focuses on the latter, studying OSN activities whose presence indicates a threat to events in the real world caused by the revelation of censored information. The field of intelligence gathering is concerned with covert operations, attempts to crack and access protected information assets and supporting infrastructures, and the collection and analysis of Open Source INTelligence (OSINT) [3], [4]. In the intelligence community, the term open refers to overt publicly available sources (as opposed to covert or clandestine sources). Hence, OSINT approaches aim at extracting knowledge from publicly available sources [4].

The cyber environment is becoming increasingly complex and timely actions are often required to counter the threat of attacks that can occur at Internet speed, while 'slow and low' Advanced Persistent Threats (APT) pose threats to both users and national security-the cyber domain being a breeding ground for disorder [5]. Cataldo [6] emphasizes the threat of subversive groups and Cioffi-Revilla [7] discusses how both Social Network Analysis (SNA) and visualization are fundamental to cyber deterrence strategy. Bates and Mooney [8] report on how Cyberspace-Based Psychological Operations (PSYOPS), which includes the recruitment, incitement, and radicalization of target populations, has taken OSN by storm. The above research and analysis point to the growing need to develop advanced detection systems to effectively identify OSINT-based Social Network threats. Our research addresses the technological and behavioral aspects of detection and prevention of subversive activity through social networks, specifically as it relates to the leakage of censored or confidential information.

a) Contribution: Our work contributes to the existing literature on three main dimensions. First, we focus on a underestimated OSINT data source: news comments through online social networks. Second, we demonstrate a methodology to detect hidden or censored information based on news comments. Third, we identify the potential of content-independent and context-independent OSN measures to identify leakage commenters based solely on their OSN characteristics.

b) Organization: The remainder of the paper is organized as follows. In Section II we have an overview about the information leakages, how we analyzed the context and the text of a comment and finally how we visualize a subnetwork of a leaker ID. In Section III we describe how we collected OSN data and the results of our analysis.

II. BACKGROUND

In the following sections, in order to better understand the ideas behind this research paper, we begin by presenting the problem of information leakage from news comments in OSN environment. We then in section II-B explain the analysis of context. Lastly, in section II-C, we explain how we can build a subnetwork of leaker IDs.

ASONAM '15, August 25-28, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2808797.2809290

A. NEWS COMMENT INFORMATION LEAKAGE

Focused on news discussion boards and Facebook-based commenting systems, we combine comment analysis and indicators of knowledge of information [9], [10], with the use of OSINT strategies for participant mapping and identification [3]. As a result, we first **identify** and then **visualize** the **subnet dynamics** of nodes that act as information leakers, and compare subnet characteristics with nodes that do not leak information. Our unique approach has not been attempted in prior research as we will now discuss.

Information leakage includes both the intentional and the accidental or unintentional distribution of private or sensitive data to an unauthorized entity. Sensitive data in organizations include different type of information. Information leakage and data misuse are considered an emerging security threats to organizations, as the number of leakage incidents and the cost they inflict continues to increase, either being caused by malicious intent or by an inadvertent mistake. Such leakage can occur in many forms and in any place [11]. Information leaks are an important concern of many organizations in today's era of online social networks (OSN). Organizations have limited control over their employees' activity in public networks, and even less than that, on the activities of their friends and relatives and of the public at large. The latter two groups, on the other hand, often have limited understanding on what public information sharing is appropriate, and what is not. Many organizations, including companies, the military and the courts, use forms of self-censorship in order to maintain security [12], [13], [14]. Non-identification of individuals, products, or places is commonly seen as a sufficient means of information protection. In the age of social networking and social media, such non-identification has been shown to be ineffective as a security measure [10]. The study of information leakage behavior in human discourse has found that speakers ability to keep information confidential. However sensitive information can be unintentionally compromised despite explicit instructions and high motivation to maintain confidentiality as accepted by the leaking party [15], [16]. This lack of intentionality presents unique cybersecurity risks that have yet to be addressed in research or practice.

B. MINING OSN AND COMMENT ANALYSIS

Mining social media has received much attention recently in different contexts. Examples are Google flu trend[1], a project that aims at detecting epidemic by monitoring users' search content [17], and a series of papers on syndromic surveillance using micro-blogging data from Twitter [18]. Other research groups have focused on information diffusion over social media, offering methods to estimate the fairness of message and its speed [19], [20], and the effect of sampling strategy on estimating information diffusion [21]. Other related research is the pioneering work by AT&T on fraud detection [22]. In their work, the group at AT&T uses social network analysis techniques to monitor contact patterns and identify potential fraudulent users.

The availability of textual data in today's social networks era has kindled an academic interest in text mining. A common text mining application in business related disciplines is the study of opinion and sentiment analysis in blogs and microblogs [23], [24]. Other literature addresses the need to mine emotions and expressions such as sarcasm or fear [25], [26].

The analysis of context is considered a crucial factor in the correct classification of the analyzed text. First, because the content posted on social networks tend to be concise and filled with abbreviations, and therefore popular text analysis techniques are inadequate, especially with learning set being longer, more "official" documents provided by the organization. Second, it is very likely that most cases of information leakage will not include the removal of documents from the organization's network, but rather the creation of new content. This will lead to the content being heavily rephrased, a fact that will make methods that rely on pattern matching far less effective.

Given the difficulty and computational costs of precise comment text analysis, the ability to focus such analysis on specific users to the exclusion of others would provide a cybersecurity advantage. Thus the importance of leak commenter identification based on network characteristics grows in proportion to the computational challenge and cost of comment analysis.

C. SUBNET DETECTION, DYNAMICS AND VISUALIZATION

OSN are web applications that allow users to build connections and establish relationships with other Internet users. If from one side OSN cut distances between users, from the other side OSN are not able to provide an adequate level of security to the data of their users. Studies like the one in [3], [27], [28], [29], [30] show how it is possible to retrieve information that a user assumed to be hidden, highlighting the problem of data leakages that affects most OSN. To connect and visualize all the Facebook identifiers (IDs) into a network we use the *SocialSpy* tool, discussed by Burattin et al. in [3]. *SocialSpy* is able to retrieve a partial list of friends of a given commenter user ID. Unlike related studies such as [31], [32], [33], our system has neither any knowledge about the friendships graph of a commenter user nor knowledge of common friends or shared information with the commenter ID.

III. PROBLEM DESCRIPTION AND METHODOLOGY

In this paper we address the following research question: can a censored reference to an unidentified individual be revealed by studying the subnetworks of leakage commenters? As an example: is it possible to reveal the identity of "Corporal S" that the article in Figure 1 refers to?

Our detection approach consists of four steps, as follows. Given the set of sensitive news articles, the set of comments on those articles, and the subnet of commenter relationships, our detection approach consists of: :

1) Identification of news article comment discourse in which the commenters exhibit knowledge of sensitive information not released in the article.
2) Analysis of commenter subnets to detect implicit relationships between commenters and identify the dynamics of such subnets.

[1] http://www.google.org/flutrends/

3) Visualization of commenting activity and subnet topology to enable quick identification of suspicious activity.
4) Compare visualizations of leaker vs. non-leaker commenter subnets.

More specifically, we augment the results of *SocialSpy* by building the friendships graph of a commenter user and de-anonymize those users that have set up a high level of privacy in a Facebook environment. This is done by using the shared friends between the one-hop friends of a commenter (friends of a commenter ID) and the two-hop friends of a commenter (friends of friends of a commenter ID). On the data retrieved we measure information inference and tie strength.

Once we have the list of friends for each ID, we are able to build the friendships subnetwork. The friendships subnetwork will be then composed by the IDs of our commenters and the IDs of their friends. The network is a graph having as nodes all the IDs (commenters and friends of commenters) and the edges representing the friendships in the OSN. As a result, nodes will be then connected to at least their friends (if any) in the graph. The aim of drawing the friendships network is indeed to have a graphical representation of the subnetwork that shows all the connections (in our case the friendships) between the users.

A. DESCRIPTION OF OSN DATA COLLECTED

We focus on information leakage through comments on Israeli, military-related, news articles published in Facebook, in which a military personnel name in censored. Censorship in our study is the replacement of a name with a supposedly non-identifying initial (e.g., 'Corporal S.'). Information leakage is detected in the comments published by private users.

An example of information leakage of interest is presented in Figure 1. The headline of the news article, as it appears on the Facebook page of a network news service, is: "Karakal combat soldier **Corporal S.** who eliminated a terrorist in the course of an incident on the Egyptian border awarded an honorable citation: "Everyone who was with us deserves it, and of course Nathaniel who fell in the battle". Military information policy dictates that the identity of officers in key positions, or involved in key operations, must not be released to the public. The motivation for this policy is a desire to protect the officer and his or her family from being identified and potentially targeted by hostile persons or forces. In this case, the obfuscated term "Corporal S." is chosen to refer to the censored element of the news item. A particularly verbose comment associated with the news item states: "The brave combatant is the daughter of a good friend of mine. Do you know where the combatant comes from? From Elad of course!". Given the readily available identity of that commenter it only takes a few clicks, to identify Corporal S., who is a Facebook friend of this commenter. We therefore treat this comment and other similar comments as *Leakage* comments. A detailed description of the case is discussed in our previous work [10], [9].

The dataset considered for our study contains 48 articles with censored information in them, with an aggregate total of annotated 3,538 comments, out of which 149 (4.21%) were labeled as *Leakage* comments, spread out through 75% (32) of the articles. A summary of the comments classification is presented in Table I.

Fig. 1: Example of data leakage (translated from Hebrew)

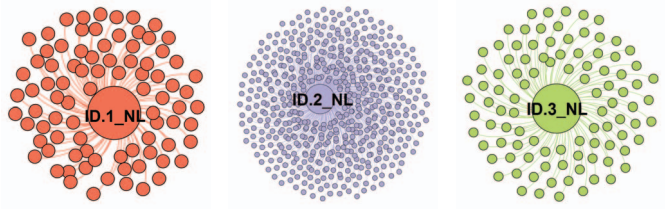
Comment type	f	%
Leakage comments	149	4.21%
Non-leakage comments	3389	95.79%
Total comments	3538	100%

TABLE I: Frequency (f) and proportion (%) of comments' classes.

B. DATA STUDY AND ANALYSIS

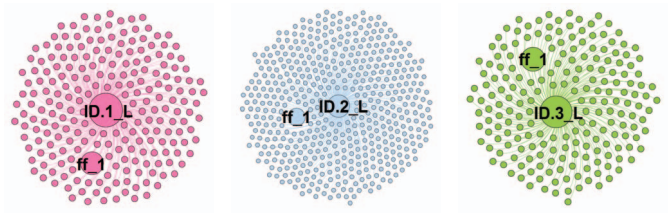
For our study a set of six comments was created from a single censored article that overall received 860 comments. We selected three non-leakage comments and three leakage comments. The Facebook ID of each commenter was then input to the SocialSpy tool [3] for analysis.

As shown in Figure 2 (a-c) and Figure 3 (d-f) our tools produced six individual subnetwork maps of the commenters, where nodes represents users profiles and links friendships among them. Moreover we show a user ID that shares the friendship with all our leaker IDs (mutual friend ff_1).



(a) (b) (c)

Fig. 2: Graphical network representation of the three non-leaker users



(d) (e) (f)

Fig. 3: Graphical network representation of the three Leaker users sharing the friendship with the same user (ff_1)

We collected OSN network parameters for each of the identified commenters and their subnets as follow:

1) We retrieved the friends list from all the IDs of the commenters using SocialSpy [3].
2) On each of the retrieved friends list we ran our new system to retrieve all the common friends between the target IDs and his friends. This new system, first improves the results of SocialSpy retrieving those friend IDs "less" active on Facebook (or with stricter privacy rules). Then it builds all the links between all the IDs. Lastly generates a `.dot` file to visualize the subnetwork.
3) The `.dot` file generated by our system is given to Gephi[2] to create a graphical representation of the subnetwork.

The final aim of this work is to depict the network behind leaker and show if and how they are connected to each other. We decided to have the friends list as target because we believe this might be one of the most important information on Facebook for our purpose. We can exploit the aforementioned list to find even more information since the user profile will no longer be private.

After the data collection step we organized all the retrieved IDs in two different folders: *leaker* and *no-leaker*. On both these folders we apply our tools, SocialSpy and a new approach. In particular, SocialSpy is a tool composed of four different strategies; the first three strategies are based on liked pages, the fourth one exploits likes and comments from the pictures of a target ID. Our new system is based on SocialSpy, and uses the previous results to generate a graphical representation of the retrieved information

We applied SocialSpy on the leaker IDs to retrieve their friends list. Then, we ran our new system that receives as input all the friends list retrieved and: *(i)* finds new IDs missed during the execution of SocialSpy; *(ii)* builds all the edges between all the retrieved IDs; *(iii)* generates the `.dot` file.

Finally, we matched the three graphs of our leaker IDs and found a common friend **ff_1**. The output of SocialSpy tool combined with our new system is depicted in Figure 4, 5 and 6. The three graphs are a graphical representation of the data retrieved by our tools. The pictures below show a subnetwork of all the leaker IDs, all the common friends between each leaker ID (for privacy reasons we renamed them as ID.1_L, ID.2_L and ID.3_L) and the mutual friend **ff_1** and give us the possibility to study the network to may find more information related to friends or leaker IDs.

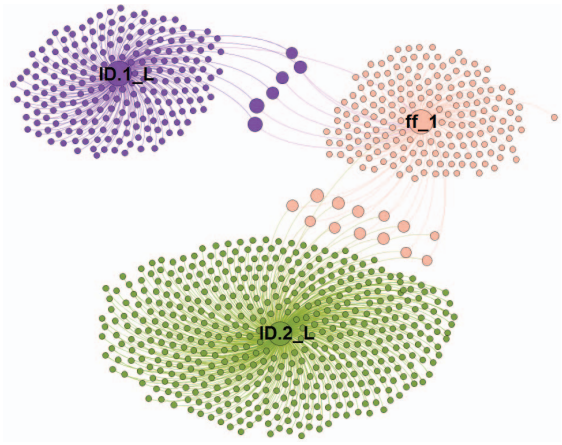


Fig. 4: Subnetwork of leaker IDs ID.1_L and ID.2_L

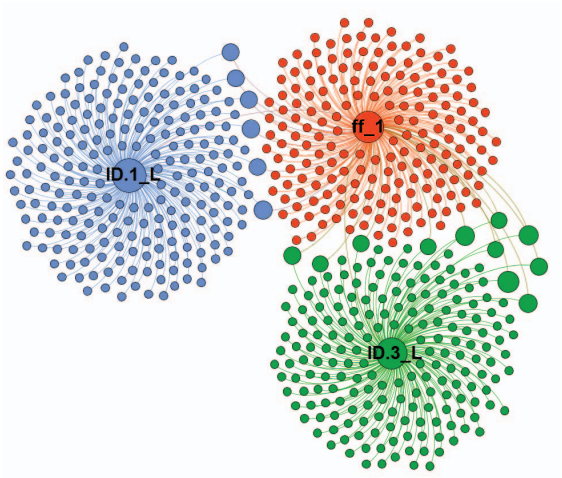


Fig. 5: Subnetwork of leaker IDs ID.1_L and ID.3_L

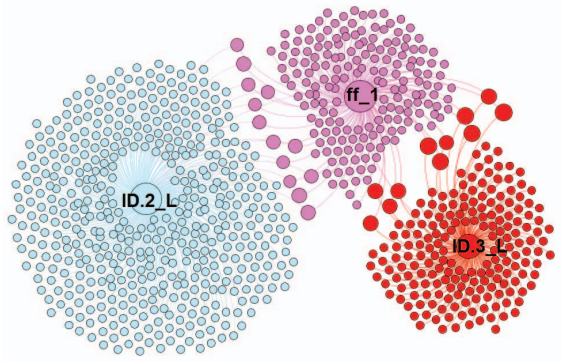


Fig. 6: Subnetwork of leaker IDs ID.2_L and ID.3_L

[2]http://gephi.github.io/

The analysis of the two graphs give evidence of some important information that we can easily summarize in five points below:

1) leakers ID.1_L, ID.2_L, ID.3_L are not friends with each other. Indeed there are no edges that connect the leaker IDs;
2) leakers ID.1_L, ID.2_L, ID.3_L share the friendship only with one ID called **ff_1**;
3) leakers ID.1_L and **ff_1** share the friendship with six mutual IDs;
4) leakers ID.2_L and **ff_1** share the friendship with thirteen mutual IDs;
5) leakers ID.3_L and **ff_1** share the friendship with eleven mutual IDs;
6) starting from three leaker IDs on a comment we have been able to de-anonymize and rebuild part of their subnetwork.

By following our methodology, and based only on three news comments, we were able to positively identify **ff_1** as the censored subject of the article. We independently verified the identification by comparing the blurred photo published with the censored article with an actual profile photo of **ff_1**. Note that no photograph was instead used in the detection algorithm.

IV. Discussion and Conclusion

We proposed an approach to un-censor an unknown and censored military Facebook ID. Censorship in our study is the replacement of a name with a supposedly non-identifying initial (e.g., "Corporal S."). We collected three unintentional information leakage IDs and we ran on them our tools: SocialSpy and a new system. The results of our thorough experiments show the feasibility of our strategies as well as their effectiveness.

Information leaks are an important concern of many organizations in today's era of Online Social Networks. Our test demonstrated how it is possible to de-anonymize a supposed "censored" military ID and his friends list. Moreover we are able to rebuild a subnetwork of friendships showing relationships between the involved IDs. Lastly we draw a friendships subnetwork to better understand the relationships schema behind leaker IDs and how they can compromise other Facebook IDs supposed censored.

In Figure 7 we depicted the final graph to show all the relationships between our leaker IDs. We still avoid using the real IDs because we do not want to expose the safety of those people, but we strongly want to raise a concern about the effectiveness of the censorship methods used by the publishers of similar articles.

There are a number of implications that this methodology may have for future research. Regarding identification of the censored identify in an article, while this study uses comments previously classified [10], manual comment classification is expensive and automated text mining of classification has demonstrated poor performance. The OSN subnetwork analysis of commenters as we have presented here has the potential to reveal the censored person based solely on network structures without considering comment content or classification.

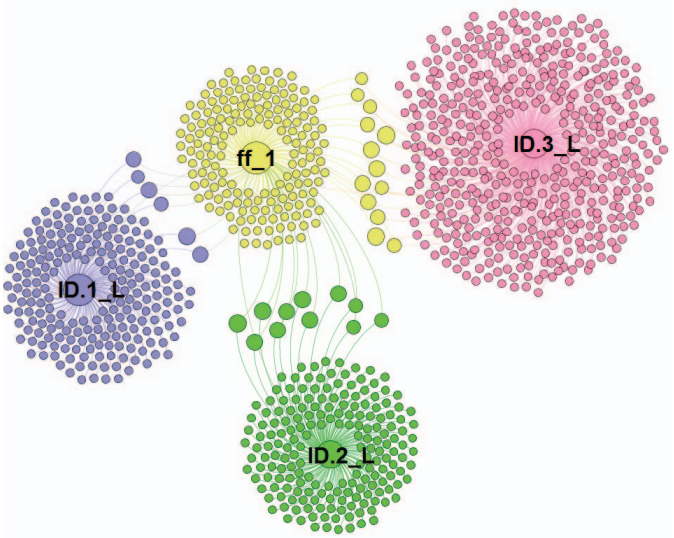


Fig. 7: Graphical representation of the subnetwork of *"leaker"* and *"censored"* IDs.

Regarding the potential for commenter profiling, by cross-referencing comment classification with the subnetworks of commenters who are connected to the censored person, we can potentially create profiles of both leakers and non-leakers. This can be done by determining when the latter who actually know the censored identity have refrained from leaking information in their comments in essence assessing the value of restraint shown by commenters who could have revealed more than they did. Finally, extending the methodology we have described, we can potentially assess the impact of a persons additional undisclosed knowledge on their tendency to share, comment upon, or like a news article.

Acknowledgment

Mauro Conti is supported by a Marie Curie Fellowship funded by the European Commission under the agreement No. PCIG11-GA-2012-321980. This work is also partially supported by the TENACE PRIN Project 20103P34XC funded by the Italian MIUR, and by the Project "Tackling Mobile Malware with Innovative Machine Learning Techniques" funded by the University of Padua.
This research was partially funded by Israel Ministry of Science and Technology research grant 3-9770 Data Leakage in Social Networks: Detection and Prevention.

References

[1] United States Government Accountability Office, "Agencies Need to Improve Cyber Incident Response Practices," Washington, DC, USA, GAO-14-354, Apr. 2014.

[2] A. Liaropoulos, and G.Tsihrintzis, *An Annotated Bibliographical Survey on Cyber Intelligence for Cyber Intelligence Officers*. in Proceedings of the 13th European Conference on Cyber Warfare and Security: ECCWS 2014, 2014, pp. 213220., 2014.

[3] A. Burattin, G. Cascavilla, and M. Conti, "Socialspy: Browsing (supposedly) hidden information in online social networks," in *Risks and Security of Internet and Systems*, 2014, pp. 83–99.

[4] D. Gritzalis, "Open-source intelligence produced from social media: A proactive cyber defense tool," presented at the 13th European Conference on Cyber Warfare and Security, Piraeus, Greece, 2014.

[5] R. Mittu and W. Lawless, "Human factors in cybersecurity and the role for ai," in *Foundations of Autonomy and Its (Cyber) Threats: From Individual to Interdependence, AAAI Spring Symposium Series*, 2015.

[6] G. Cataldo, "Fighting terrorism in cyberspace," in Modelling Cyber Security: Approaches, Methodology, Strategies, U. Gori, Ed. IOS Press, 2009, pp. 160164.

[7] C. Cioffi-Revilla, "Modelling deterence in cyberia," in Modelling Cyber Security: Approaches, Methodology, Strategies, U. Gori, Ed. IOS Press, 2009, pp. 125131.

[8] R. A. Bates and M. Mooney, "Psychological operations and terrorism: The digital domain," J. Public Prof. Sociol., vol. 6, no. 1, p. 2, 2014.

[9] D. G. Schwartz, G. Silverman, and I. Yahav, "News censorship in social networks: A study of circumvention in the commentsphere," Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2604910, 2015.

[10] I. Yahav, D. G. Schwartz, and G. Silverman, "Detecting unintentional information leakage in social media news comments," in 2014 IEEE 15th International Conference on Information Reuse and Integration (IRI), 2014, pp. 7479.

[11] A. Shabtai, Y. Elovici, and L. Rokach, "A survey of data leakage detection and prevention solutions," Springer., 2012.

[12] R. T. Davis, "The us army and the media in the 20th century," vol. 31. Government Printing Office, 2009.

[13] M. Johnson, "Of public interest: How courts handle rape victims privacy suits," Commun. Law Policy, vol. 4, no. 2, 1999, pp. 201242.

[14] J. M. Schumm, "No names, please: The virtual victimization of children, crime victims, the mentally ill, and others in appellate court opinions," Ga. Law Rev., vol. 42, 2008, p. 471.

[15] L. W. Lane, M. Groisman, and V. S. Ferreira, "Dont talk about pink elephants! speakers control over leaking private information during language production," Psychol. Sci., vol. 17, no. 4, Apr. 2006, pp. 273277.

[16] L. W. Lane and M. J. Liersch, "Can you keep a secret? increasing speakers motivation to keep information confidential yields poorer outcomes," Lang. Cogn. Process., vol. 27, no. 3, 2012, pp. 462473.

[17] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," Nature, vol. 457, no. 7232, 2009, pp. 10121014.

[18] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health," in ICWSM, 2011, pp. 265272.

[19] J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter," ICWSM, vol. 10, 2010, pp. 355358.

[20] D. V. Liere, "How far does a tweet travel?: Information brokers in the twitterverse," in Proceedings of the International Workshop on Modeling Social Media, 2010, p. 6.

[21] M. De Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher, "How does the data sampling strategy impact the discovery of information diffusion in social media?" ICWSM, vol. 10, 2010, pp. 3441.

[22] R. A. Becker, C. Volinsky, and A. R. Wilks, "Fraud detection in telecommunications: History and lessons learned," Technometrics, vol. 52, no. 1, 2010.

[23] A. Ghose, P. G. Ipeirotis, and B. Li, "Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content," Mark. Sci., vol. 31, no. 3, 2012, pp. 493520.

[24] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," Knowl. Data Eng. IEEE Trans. On, vol. 23, no. 10, 2011, pp. 14981512.

[25] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, "Mine your own business: Market-structure surveillance through text mining," Mark. Sci., vol. 31, no. 3, 2012, pp. 521543.

[26] O. Tsur, D. Davidov, and A. Rappoport, "ICWSM-A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews," in ICWSM, 2010.

[27] F. Beato, M. Conti, B. Preneel, and D. Vettore, "VirtualFriendship: Hiding interactions on Online Social Networks," in Communications and Network Security (CNS), 2014 IEEE Conference, 2014, pp. 328336.

[28] F. Beato, M. Conti, and B. Preneel, "Friend in the middle (fim): Tackling de-anonymization in social networks," in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference, 2013, pp. 279284.

[29] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, 2005.

[30] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, Apr. 2013.

[31] L. Jin, J. B. Joshi, and M. Anwar, "Mutual-friend based attacks in social network systems," *Computers & Security*, vol. 37, pp. 15 – 30, 2013.

[32] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks." ACM, 2010.

[33] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *IEEE 24th International Conference*, 2008.