Antisocial Behavior on the Web: Characterization and Detection

Srijan Kumar
University of Maryland
srijan@cs.umd.edu

Justin Cheng
Stanford University
jcccf@cs.stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Web platforms enable unprecedented breadth and speed in transmission of knowledge, and allow users to communicate and shape opinions. However, the safety, usability and reliability of these platforms are compromised by the prevalence of online antisocial behavior, for e.g., 40% of users have experienced online harassment [3]. Antisocial behavior is present in the form of antisocial users, such as trolls, sockpuppets and vandals, and misinformation, such as hoaxes, rumors and fraudulent reviews [37]. This tutorial presents the state-of-the-art research spanning two aspects of antisocial behavior: characterization of their behavioral properties, and development of algorithms for identifying and predicting them.

The tutorial first discusses antisocial users — trolls, sockpuppets and vandals. We present the causes, community effects, and linguistic, social and temporal characteristics of trolls. Then we discuss the types of sockpuppets, i.e. multiple accounts of the same user, and their behavioral characteristics in Wikipedia and online discussion forums. Vandals make destructive edits on Wikipedia and we discuss the properties of vandals and vandalism edits. In each case, detection and prediction algorithms of the antisocial user are also discussed.

The second part of the tutorial discusses misinformation — hoaxes, rumors and fraudulent reviews. We present the characteristics and impact of hoaxes on Wikipedia, followed by the spread and evolution of rumors on social media. Then, we discuss the algorithms to identify fake reviews and reviewers from their characteristics, and the camouflage and coordination among sophisticated fraudsters. Again, in each case, we present the detection algorithms, using textual, temporal, sentiment, network structure and rating patterns. Finally, the tutorial concludes with future research avenues.

Keywords

Malicious users; False information; Antisocial behavior; Vandals; Trolls; Sockpuppets; Shills; Hoax; Rumor; Fake review; Fraud; Spam; Cyberbully; Bots;

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04. http://dx.doi.org/10.1145/3041021.3051106

1. INTRODUCTION

The web is a space for all, where everybody can read, publish and share information. The interconnectedness of the web enables dissemination of information, ideas and opinions to a large audience at an unprecedented speed, which has had revolutionary effects on the lives of billions of people. While benign users try to keep the web safe and usable, online antisocial behavior threatens the usability and safety of web platforms [44, 52]. This exists in the form of antisocial users (e.g., trolls [11, 53], sockpuppets [21, 59] and vandals [22]) and misinformation (e.g., rumors and hoaxes [9, 44, 49], and fake reviews [45]). Their extent is widespread, for instance, 40% of web users have experienced online harassment [3], 8–10% social network accounts are fake [2, 32], and roughly 16% Yelp reviews are fake [42]. Dissociative anonymity in online interactions further encourages antisocial behavior [58]. The effect of antisocial behavior on people's lives have been detrimental, ranging from experiencing distress [6], offline harassment [60] and in some cases, has even led to fatalities [26]. Therefore, it is primarily essential to maintain the quality and safety of web platforms.

Antisocial users are present on the web in several forms, such as trolls, vandals, and sockpuppets. *Trolls* are antisocial users that harass others [48]. They misbehave when participating in online discussions, make irrelevant posts and are more abusive [15]. As a result, they are treated more harshly by the community, by being reported and banned more often [15]. Trolling behavior is also affected by both the users' mood and discussion context [13]. Another type of antisocial users are *sockpuppets*, i.e., multiple accounts operated by the same user. These accounts are often benign, for instance, to diversify interests and keep separate accounts for separate activities [23]. But quite frequently, they are used to deceive and manipulate others [24], as in case of online discussions [33], Wikipedia [54] and social networks [21]. Deceptive sockpuppets tend to be collusive and behave similar to each other, i.e. they are mostly active during similar times and write similarly [33, 54, 10]. The third type of antisocial users we study are *vandals*, which are users who make destructive edits, prominently on collaborative spaces such as Wikipedia [5, 36], WikiMapia [7] and OpenStreetMaps [47]. Vandals tend to be faster in editing, get involved in arguments and make incoherent edits [36]. Similarly, vandalism edits are shorter, have lower quality and are reverted more often [5, 4].

Antisocial behavior also manifests in the form of misinformation, such as rumors, hoaxes, and fake reviews. *Rumors and hoaxes* are forms of false information purposely created to masquerade as truth. Their occurrence tends to increase around popular and widely covered events, such as hurricanes [25] and plane crashes [61]. They are bursty in nature [16], spread quickly across social networks [18], and evolve over time [20]. Wikipedia hoaxes tend

to be less related to other articles and have less supporting references [38]. Another type of misinformation is *fraudulent review* that is given to artificially boost or reduce ratings of products on e-commerce platforms [42, 43]. Such reviews substantially increase the profits of targeted services (e.g. products, hotels, restaurants, etc.) [55, 41]. These reviews are given in a short period of time [34, 46, 27], have extreme scores [46], and are shorter in length [45, 51].

Several computational approaches have been developed to predict antisocial behavior (see Section 2 for details). In general, any antisocial behavior prediction algorithm should satisfy two basic criteria: (i) it needs to be highly accurate, i.e., it should not predict benign users as antisocial, and vice-versa; and (ii) it needs to reasonably predict the antisocial activity as soon as possible, so that harm is pre-emptively avoided.

However, there are several challenges that lie in understanding and predicting antisocial behavior [57]. Antisocial activities forms only a small part of all activities, leading to information imbalance. Often limited ground-truth data is available about antisocial behavior. Additionally, antisocial entities camouflage themselves to masquerade as benign entities. Finally, antisocial behavior is ever-evolving, and dynamically adaptive to any predictive system.

This tutorial will cover two important aspects of online antisocial behavior: provide better understanding of their characteristics, and discuss the state-of-the-art algorithms to identify and predict them.

2. OUTLINE

In the first part of the tutorial, we focus on antisocial users. We start our discussion with the causes of trolling, and their textual, social, and temporal characteristics. We further discuss the effect of community on trolling behavior. Then we describe algorithms to identify and predict trolls from non-trolls. Moving on to the topic of sockpuppets, we illustrate the different types of sockpuppetry — deceptive vs non-deceptive, supportive vs dissenting — and their linguistic, activity, and temporal properties. We then explain the algorithms to identify sockpuppets in web discussion forums and social networks. Finally, we present the properties of vandals and vandalism edits, and then explain the algorithms to detect them.

In the next part of the tutorial, we focus on online antisocial behavior in the form of misinformation. First, we explore the spread and evolution of rumors in social networks, the impact and textual characteristics of hoaxes and the users who create them. Then we outline algorithms to identify hoaxes and rumors from genuine information. Next, we specify several properties of fraudulent reviews — textual, temporal, rating, and social characteristics. We further discuss coordination and camouflage of fake reviews and reviewers. Lastly, we present the state-of-the-art algorithms to detect them.

We conclude the tutorial with the future research directions and unsolved challenges.

An outline of the tutorial is given below:

1. Introduction
   (a) Definitions of antisocial behavior. Overview of antisocial behavior on the web - trolls, sockpuppets, vandals, hoaxes, fraudulent reviews, cyberbullies, bots, and spammers.
   (b) Requirements of antisocial behavior detection system: high accuracy, early prediction.
   (c) Challenges: unbalanced data, lack of ground truth, dynamically evolving behavior.
2. Antisocial Users on the Web
   (a) Trolls
      i. Causes of trolling [13]
      ii. Behavioral characteristics of trolls [15]
      iii. Effects of the community on behavior of trolls [15, 14]
      iv. Detecting and predicting trolls: linguistic, temporal and network structure [15, 35, 39]
   (b) Sockpuppets
      i. Types of sockpuppets: deceptive vs non-deceptive, supportive vs dissenting [33]
      ii. Behavioral characteristics of sockpuppets on social platforms and Wikipedia: linguistic [33, 62], social [63], sentiment [10]
      iii. Detecting sockpuppets [33, 10, 62, 63]
   (c) Vandals
      i. Behavioral characteristics of vandals: temporal and activity [36]
      ii. Properties of vandalism: textual and temporal [5]
      iii. Detection and prediction of vandals and vandalism [5, 36, 1]
   (d) Brief discussion of other antisocial users - cyberbullies [29], bots [17, 19, 56], fake accounts [12, 45], spammers [45, 30, 40]
3. Misinformation on the Web
   (a) Hoaxes and Rumors
      i. Characteristics and impact of hoaxes in Wikipedia [38]
      ii. Spread of rumors: patterns, dynamics and evolution [20]
      iii. Detecting hoaxes and rumors with network structure, temporal, linguistic and activity properties [20, 38, 50]
   (b) Fraudulent reviews
      i. Characteristics of fraudulent reviews: temporal [34], ratings network [34, 27], sentiment [42] and linguistic [46]
      ii. Camouflage and coordination in fraudulent reviews [8, 31, 28]
      iii. Detection of fraudulent reviewers and reviews [34, 27, 42, 46]
   (c) Conclusion and Open Research Avenues

3. AUDIENCE AND PREREQUISITES

This tutorial targets academic, industry and government researchers and practitioners with interests in social network anomaly detection, user behavior modeling, graph mining, cybersecurity, and community policy design. Beginners in the area will learn the basics of these algorithms. Experts in the area will learn in-depth algorithms and case-studies to detect online antisocial behavior that are both platform-specific techniques and platform-independent. This tutorial should appeal to researchers of several disciplines.

There are no prerequisites for attending the tutorial. We cover basics as well as advanced techniques.

4. TUTORIAL MATERIAL

The tutorial slides, links to relevant papers, datasets and codes are available at `http://snap.stanford.edu/www2017tutorial/`.

5. PRESENTERS

Srijan Kumar. is a Ph.D. candidate in the Computer Science department at the University of Maryland, College Park. His research broadly lies in data mining and social network analysis, and focuses on malicious user and information detection. He is a recipient of WorldQuant PhD Fellowship, University of Maryland Outstanding Graduate Student Dean's Fellowship, and has been awarded Dr. Bidhan Chandra Roy Gold Medal by Indian Institute of Technology (IIT), Kharagpur. More details can be found at `http://cs.umd.edu/~srijan/`

Justin Cheng. is a Ph.D. candidate in the Computer Science department at Stanford University. His research lies at the intersection of data science and human-computer interaction, and focuses on cascading behavior in social networks. This work has received several best paper nominations at CHI, CSCW, and ICWSM. He is also a recipient of a Microsoft Research PhD Fellowship and a Stanford Graduate Fellowship. More details can be found at `http://www.clr3.com`

Jure Leskovec. is an associate professor of Computer Science at Stanford University and chief scientist at Pinterest. Computation over massive data is at the heart of his research and has applications in computer science, social sciences, economics, marketing, and healthcare. This research has won several awards including a Lagrange Prize, Microsoft Research Faculty Fellowship, the Alfred P. Sloan Fellowship, and numerous best paper awards. Leskovec received his bachelor's degree in computer science from University of Ljubljana, Slovenia, and his PhD in in machine learning from the Carnegie Mellon University and postdoctoral training at Cornell University. More details can be found at `https://cs.stanford.edu/~jure`

6. ACKNOWLEDGEMENT

Parts of this work were supported by US Army Research Office under Grant Number W911NF1610342, NSF IIS-1149837, ARO MURI, DARPA NGS2, Stanford Data Science Initiative, and Microsoft Research PhD fellowship.

7. REFERENCES

[1] Cluebot ng. `https://en.wikipedia.org/wiki/User:ClueBot_NG`, 2010.

[2] How many of the internet's users are fake. `http://www.dailyinfographic.com/how-many-of-the-internets-users-are-fake`, 2014.

[3] Online harassment, pew research center. `http://www.pewinternet.org/2014/10/22/online-harassment`, 2014.

[4] B. Adler, L. De Alfaro, and I. Pye. Detecting wikipedia vandalism using wikitrust. *Notebook papers of CLEF*, 1:22–23, 2010.

[5] B. T. Adler, L. De Alfaro, S. M. Mola-Velasco, P. Rosso, and A. G. West. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, 2011.

[6] Y. Akbulut, Y. L. Sahin, and B. Eristi. Cyberbullying victimization among turkish online social utility members. *Educational Technology & Society*, 13(4):192–201, 2010.

[7] A. Ballatore. Defacing the map: Cartographic vandalism in the digital commons. *The Cartographic Journal*, 51(3):214–224, 2014.

[8] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd international conference on World Wide Web*, 2013.

[9] P. Bordia and N. DiFonzo. Problem solving in social interactions on the internet: Rumor as social cognition. *Social Psychology Quarterly*, 67(1):33–49, 2004.

[10] Z. Bu, Z. Xia, and J. Wang. A sock puppet detection algorithm on virtual spaces. *Knowledge-Based Systems*, 37:366–377, 2013.

[11] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus. Trolls just want to have fun. *Personality and individual Differences*, 67:97–102, 2014.

[12] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, 2012.

[13] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2017.

[14] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. How community feedback shapes user behavior. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[15] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 2015.

[16] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi. The anatomy of a scientific rumor. *Scientific Reports*, 3, 2013.

[17] J. P. Dickerson, V. Kagan, and V. Subrahmanian. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2014.

[18] B. Doerr, M. Fouz, and T. Friedrich. Why rumors spread so quickly in social networks. *Communications of the ACM*, 55(6):70–75, 2012.

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

[20] A. Friggeri, L. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[21] K. Gani, H. Hacid, and R. Skraba. Towards multiple identity detection in social networks. In *Proceedings of the 21st International Conference on World Wide Web*, 2012.

[22] R. S. Geiger and D. Ribes. The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 2010.

[23] R. Gilbert, V. Thadani, C. Handy, H. Andrews, T. Sguigna, A. Sasso, and S. Payne. The psychological functions of avatars and alt (s): A qualitative study. *Computers in Human Behavior*, 32:1–8, 2014.

[24] R. L. Gilbert, J. A. Foss, and N. A. Murphy. Multiple personality order: Physical and personality characteristics of the self, primary avatar and alt. In *Reinventing ourselves: Contemporary concepts of identity in virtual worlds*, pages 213–234. Springer, 2011.

[25] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, 2013.

[26] S. Hinduja and J. W. Patchin. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221, 2010.

[27] B. Hooi, N. Shah, A. Beutel, S. Gunneman, L. Akoglu, M. Kumar, D. Makhija, and C. Faloutsos. Birdnest: Bayesian inference for ratings-fraud detection. *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016.

[28] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[29] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2014.

[30] X. Hu, J. Tang, H. Gao, and H. Liu. Social spammer detection with sentiment information. In *Proceedings of the IEEE International Conference on Data Mining*, 2014.

[31] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Catchsync: catching synchronized behavior in large directed graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[32] K. Krombholz, D. Merkl, and E. Weippl. Fake identities in social media: A case study on the sustainability of the facebook business model. *Journal of Service Science Research*, 4(2):175–212, 2012.

[33] S. Kumar, J. Cheng, J. Leskovec, and V. Subrahmanian. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[34] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian. Fairjudge: Trustworthy user prediction in rating platforms. 2017.

[35] S. Kumar, F. Spezzano, and V. Subrahmanian. Accurately detecting trolls in slashdot zoo via decluttering. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2014.

[36] S. Kumar, F. Spezzano, and V. Subrahmanian. Vews: A wikipedia vandal early warning system. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[37] S. Kumar, F. Spezzano, and V. Subrahmanian. Identifying malicious actors on social media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016.

[38] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[39] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, 2009.

[40] Y. Li, O. Martinez, X. Chen, Y. Li, and J. E. Hopcroft. In a world that counts: Clustering and detecting fake social engagement at scale. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[41] M. Luca. Reviews, reputation, and revenue: The case of yelp. com. *Harvard Business School NOM Unit Working Paper*, 2011.

[42] M. Luca and G. Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427, 2016.

[43] J. Malbon. Taking fake online consumer reviews seriously. *Journal of Consumer Policy*, 36(2):139–157, 2013.

[44] A. P. Mintz. *Web of deception: Misinformation on the Internet*. Information Today, Inc., 2002.

[45] A. Mukherjee, B. Liu, and N. Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, 2012.

[46] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance. What yelp fake review filter might be doing? In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[47] P. Neis, M. Goetz, and A. Zipf. Towards automatic vandalism detection in openstreetmap. *ISPRS International Journal of Geo-Information*, 1(3):315–332, 2012.

[48] W. Phillips. Loling at tragedy: Facebook trolls, memorial pages and resistance to grief online. *First Monday*, 16(12), 2011.

[49] P. S. Piper. Better read that again: Web hoaxes and misinformation. *Searcher*, 8(8), 2000.

[50] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[51] S. Rayana and L. Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[52] C. Seife. *Virtual Unreality: Just Because the Internet Told You, how Do You Know It's True?* Penguin, 2014.

[53] P. Shachaf and N. Hara. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3):357–370, 2010.

[54] T. Solorio, R. Hasan, and M. Mizan. A case study of sockpuppet detection in wikipedia. In *Workshop on Language Analysis in Social Media*, 2013.

[55] D. Streitfeld. Fake reviews, real problem. *New York Times. http://query. nytimes. com/gst/fullpage. html*, 2012.

[56] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016.

[57] V. Subrahmanian and S. Kumar. Predicting human behavior: The next frontiers. *Science*, 355(6324):489–489, 2017.

[58] J. Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.

[59] M. Tsikerdekis and S. Zeadally. Multiple account identity deception detection in social media using nonverbal behavior. *IEEE Transactions on Information Forensics and Security*, 9(8):1311–1321, 2014.

[60] D. Wiener. Negligent publication of statements posted on electronic bulletin boards: Is there any liability left after zeran. *Santa Clara L. Rev.*, 39:905, 1998.

[61] S. Wu, Q. Liu, Y. Liu, L. Wang, and T. Tan. Information credibility evaluation on social media. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[62] Z. Yamak, J. Saunier, and L. Vercouter. Detection of multiple identity manipulation in collaborative projects. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016.

[63] X. Zheng, Y. M. Lai, K.-P. Chow, L. C. Hui, and S.-M. Yiu. Sockpuppet detection in online discussion forums. In *Proceedings of the Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2011.