Detecting and Mitigating the Effect of Manipulated Reputation on Online Social Networks

Anupama Aggarwal
« supervised by Dr. Ponnurangam Kumaraguru »
Indraprastha Institute of Information Technology, Delhi
New Delhi, India
anupamaa@iiitd.ac.in

ABSTRACT

In recent times, online social networks (OSNs) are being used not only to communicate but to also create a public/social image. Artists, celebrities and even common people are using social networks to build their brand value and gain more visibility either amongst a restricted set of people or public. In order to enable user to connect to other users in the OSN and gain following and appreciation from them, various OSNs provide different social metrics to the user such as Facebook `likes`, Twitter `followers` and Tumblr `reblogs`. Hence, these metrics give a sense of social reputation to the OSN user. As more users are trying to leverage social media to create a brand value and become more influential, spammers are luring such users to help manipulate their social reputation with the help of paid service (black markets) or collusion networks.

In this work, we aim to build a robust alternate social reputation system and detect users with manipulated social reputation. In order to do so, we first start by understanding the underlying structure of various sources of crowdsourced social reputation manipulation like blackmarkets, supply-driven microtask websites and collusion networks. We then build a mechanism for an early detection of users with manipulated social reputation. Our initial results are encouraging and substantiate the possibility of a robust social reputation system.

Keywords

social reputation; online social networks; user behaviour; crowdsourced manipulation

1. PROBLEM

The social network revolution has led to the rise of several online social networks (OSNs) like Facebook, Twitter and Yelp [1] which have rapidly acquired millions of users. OSN users use these services to communicate with other users, spread and consume information, and even build and strengthen social connections. Due to the growing rise of OSNs as a platform for mass communication, brands, celebrities and political parties have started using them extensively to engage with users. Some of the most popular and influential celebrities have garnered millions of users following their activities on OSNs. [2] The user following and crowdsourced ratings give an OSN user a sense of social reputation which she tries to maintain and boost to be more influential in the network and attract more following. Various OSNs have a different measure of user following and social reputation, like `followers` on Twitter and Instagram, `likes` on Facebook and `ratings` on Yelp. However, these reputation metrics can be manipulated in several ways.

One of the most prevalent methods to alter social reputation is online blackmarket that not only generates a misleading reputation but also injects the network with false identities. There exist several online services from where an OSN user can purchase bulk `followers` and `likes`. Such services are very cheap and let users choose from different packages like *"1000 Followers for $3"*. [3] Security researchers have recently estimated that the revenue generated by blackmarket for Twitter followers is between $40 million to $360 million. Another host for fake accounts peddlers is supply-driven microtask markets like *Fiverr* and *SEOClerks*. Account peddlers exploit these services to cater to OSN users who inflate their social media metrics such as – followers, likes and shares in the hope to become more influential and popular on the network. These marketplaces often provide newly created or stockpiled, fake and inactive accounts to the users. Such services are a big threat to the credibility of the social networks that rely on crowdsourced ratings and reviews for product recommendation and build user trust. Recently Amazon sued 1,114 sellers on Fiverr for posting fake reviews of Amazon products [15]. Infiltration of fake accounts and metrics also has a damaging effect on the advertisement revenue framework of OSNs. An advertisement might be shown to certain users with seemingly high popularity and reputation, but will most likely not get clicked by the expected number of real users due to the accounts being fake or bots. To address such issues and mitigate the damaging effect on the OSN ecosystem, there is a dire need to build alternate social reputation systems and weed out users and entities on social networks with manipulated followers, likes and ratings.

[1] www.facebook.com www.twitter.com www.yelp.com

[2] Barack Obama has 67.2M Twitter `followers`, Shakira has 103.8M Facebook `likes`

[3] `http://www.buycheapfollowersfast.com/twitter/`

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2888601.

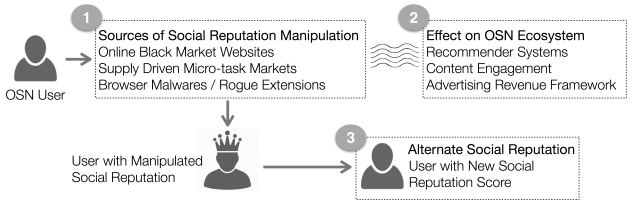

Figure 1: Illustration of three step study objectives to mitigate manipulated social reputation on online social networks

Recent studies show that there exist a thriving blackmarket to create fake identities [12, 21]. There have also been strong evidence of browser malwares that attack social networking sites and trigger activities such as liking and sharing posts without user's knowledge [10, 11]. The attackers behind blackmarket or browser malwares either create fake identities or compromise existing users to manipulate social reputation in bulk. These sybil identities are used to manipulate social reputation, such as Twitter `followers` [17], Facebook `likes` [18] and Yelp `reviews` and `ratings` [13]. Although there is a strong evidence of manipulated social reputation on various networks and its damaging effects [19], there does not exist alternate social reputation systems to compute the difference between percieved and real reputation of an OSN user. Coupled with this gap, there exist a lack of understanding of how manipulated social reputation affects OSN's ecosystem. An artificially inflated social reputation can adversely affect OSN's recommender system, post engagement and advertising revenue framework, which we plan to study. Therefore, my dissertation explores the following –

> Building a robust alternate Social Reputation System which does not get affected by manipulated social metrics of the OSN user.

To address the above, this work is broadly divided into three parts, also illustrated in Figure 1 –

- Landscape the sources of manipulation of social reputation on OSNs - online blackmarkets, collusion networks, browser malwares
- Detect and measure the effect of sybil nodes in OSNs contributing to the fake social reputation; and an early detection mechanism for users with suspicious social reputation
- Use meta-information of an entity on OSN to draw out differences between perceived and real social reputation and build an alternate social reputation system for OSNs which is robust

2. STATE OF THE ART

Identifying sources of social reputation manipulation. Researchers have lately discovered several sources of social reputation manipulation on OSNs which adversely affect the economy and tarnish the network's credibility. Stringhini et al. detected several thousand blackmarket websites which either create collusion networks of participating users or boost Twitter follower count of paying user with the help of sybil nodes (fake or compromised) [16]. Other studies have also uncovered ways to detect various sources of crowdsourced manipulation of social reputation involving microtask supply-driven markets or cheap online labour markets [8, 12]. As first part of this work, we want to exhaustively landscape the sources of social reputation manipulation. We have landscaped blackmarket websites to understand who the sellers, customers and the sybil nodes are and explain them in more detail in section 5.1.

Detect sybil nodes on online social networks. There has been significant research focusing on identification of sybil nodes in a network. Researchers have developed techniques to distinguish between sybil and legitimate identities and effectively identify inorganic behaviour on social networks using supervised, as well as unsupervised machine learning methods [2, 9, 20]. These techniques largely rely on the network and behavioral features of each sybil entity itself. In this work, instead of focusing on detection of each sybil node, we focus on the damage caused by a group of such sybil nodes and hence aim to detect a crowdsourced manipulation of social reputation done by these sybil nodes. Bimal et al. have looked at detection of crowdsourced tampering of a social media metric like Twitter follower count or Yelp ratings by sybil entities by finding an anomalous pattern in the joining date of all the OSN nodes involved in computing the like of rating of each entity under suspicion [19]. While this methodology proves to be effective to some extent, it needs details of each node participating in contributing to the social reputation score of the suspicious user. As the second part in this work, we want to build an effective mechanism to detect OSN users with manipulated reputation score. Our preliminary results based on a heuristic approach and only the attributes of the suspicious user seem promising, and we describe them in more detail in section 5.3. There exist few alternate reputation systems like Klout and PeerIndex, but they are not robust and yield poor results. The next step based on these results would be to build an alternate social reputation system for each OSN, which is robust and is not affected by crowdsourced manipulation of number of followers, likes or shares. Such an alternate social reputation system can be helpful to mitigate the adverse effects on OSN's ecosystem.

Effect of crowdsourced manipulation of social reputation. There have been several reports stating the adverse effects of manipulated social reputation on popular OSNs like generating fake following by politicians [5], misleading product reviews [15] and fake hits on Facebook advertisements [4]. Researchers have shown that blackmarkets for crowdsourced social reputation on OSNs have a potential to generate revenue between $40 to $360 million [14]. However, the effect of social reputation manipulation of users on OSN's friend recommender system, prioritization of search results and advertising revenue of OSN is largely unexplored and not quantified. As the final part of this work, we would measure the adverse effect of manipulated social reputation on the OSN's ecosystem.

3. BACKGROUND & APPROACH

Every online social network provides mechanisms to its users to become socially visible and influential. This work explores the sources of crowdsourced manipulation of social reputation and detection of the same. Table 1 gives a brief overview of popular OSNs and correspondingly, which met-

ric gives the concerned OSN entity a sense of social reputation. The larger goal of this work is to be able to effectively detect and measure the effects of manipulation of such social reputation metrics.

Table 1: Metrics on online social networks which give a sense of social reputation

OSN	OSN Entity	Perceived Reputation Score
Twitter, Instagram, Pinterest	User	Follower Count
Facebook	Page	Likes
Yelp, Foursquare	Business/ Venue	Rating
Amazon, Flipkart, Ebay	Product	Rating
Tumblr	Post	Reblogs
YouTube	Video	Likes

To achieve our goal as described in section 3, we divide our work into a three phases. We describe our current and future approach to complete each phase in more detail as followed –

3.1 Landscaping the Source of Social Reputation Manipulation

We identify three main sources of bulk social reputation manipulation viz. (i) Online Blackmarket, (ii) Scratch-Back Services, (iii) Browser Malwares. There has been some work done to identify ***Online Blackmarkets*** by finding websites on search engines using keywords like *buy cheap followers* and *purchase Facebook likes* [16]. Online blackmarkets are however not hard to find since they advertise themselves to get more hits. We focus on the more challenging task of profiling and measuring the impact of these websites. We identify market leaders that inject the network with most of the sybil nodes and hence have a more damaging effect on the network.

Another source of social reputation manipulation, ***Scratch-Back*** services are the ones where users do not need to pay money to boost their social reputation. Instead, they can do favors to other interested users to get back the same and hence become part of a collusion network. Since each individual node involved in such a collusion network can be a genuine user trying to manipulate its social reputation, such manipulation is very hard to detect. There do not exist any studies with tangible results to uncover large-scale collusion networks involved in manipulating social reputation. However, we plan to take cues from seminal research been done on social spam campaigns [6, 3] to detect social reputation manipulation by scratch-back services.

Browser Malwares is one of the other primary sources of social reputation manipulation which we plan to study. The infected users are unaware of the likes, shares and comments triggered from their accounts and unknowingly become a proxy for a manipulated social reputation of the attacker. This area is largely unexplored with only evidence of browser malwares attacking OSNs [7]. We plan to collect a dataset of rogue browser extensions that trigger such attacks and study the propagation of infection in the network.

3.2 Early Detection of Users with Manipulated Social Reputation

While there have been studies to detect individual sybil nodes involved in crowdsourced manipulation, there has been very less research to effectively identify OSN users with manipulated social reputation. Bimal et al. proposed to identify crowdsourced manipulation by investigating the timestamp of activities by involved sybil nodes [19]. Since it is not feasible to fetch the timestamp in almost every OSN of the past activity, they rely on the joining dates distribution. Their assumption is that the distribution of joining dates of nodes involved in a crowdsourced manipulation will have significant divergence from that of a reliable reference set of high-quality nodes. However, the disadvantage of this approach is that it will fail in case the sybil nodes are stockpiled and hence there won't be any anomaly in their joining dates distribution.

We propose a robust approach to detect crowdsourced social reputation manipulation by leveraging the basics of each OSN used to gain true attention and reputation in a network. For instance, a Twitter user will genuinely garner followers when she posts interesting content (or is a celebrity in the real world). Therefore, to validate her follower count, we build a heuristic to determine her real reputation based on various other factors like tweet-follower ratio, evidence of being a topic expert and overlap of her interests with followers. The initial results of our approach seem promising and based on our findings we later plan to propose an alternate social reputation score.

3.3 Measuring the Effects of Manipulated Social Reputation on OSN's Ecosystem

Measuring and quantifying the adverse effects of crowdsourced manipulation remains an unexplored area. We plan to conduct the following experiments to understand the effect on OSN's ecosystem –

- Effect on Recommender Systems: Do more people follow a user after her social reputation is artificially boosted? Does she have a higher chance to get suggested by friend/follow recommender system of the social network?
- Effect on Social Capital of User: Does she become more impactful and visible? Does a manipulated social reputation also bump up her existing alternate social reputation scores like Klout and PeerIndex?
- Effect during Events: Does the user's posts about a trending event get higher priority in search results after her social reputation is manipulated? Does it trigger a chain reaction garnering higher engagement and attention from other users following the event?

4. EVALUATION

Our three-fold study will be evaluated as followed –

- To effectively profile the sources of crowdsourced manipulation of social reputation, we will collect a true positive large-scale dataset of online blackmarket websites, scratch-back services and browser malwares. To understand how these services operate, we plan to create dummy accounts and use these services. As initial part of our experiments, we created dummy ac-

counts and purchased Twitter followers from blackmarket websites. [4] We plan to similarly collect large scale data of scratch-back services and browser malwares for an effective evaluation

- To evaluate our approach to detect users with crowdsourced manipulation of social reputation, we plan to compare our technique with (i) existing social reputation systems like Klout and PeerIndex, and (ii) methodology proposed in existing work by Bimal et al. [19].

- We plan to conduct in-the-wild online experiments to study the effect of crowdsourced social reputation manipulation. Collecting temporal data with multiple snapshots before and after the manipulation will help us evaluate how effectively we can measure and detect the changes in effects on OSN's ecosystem.

5. RESULTS

We have following results for the experiments described in the previous sections –

5.1 Landscaping the Online Blackmarket

The first part involves landscaping the sources of crowdsourced social reputation manipulation. As a preliminary study, we have been able to effectively study the underlying structure of online blackmarket which sells Twitter followers. We purchased over fake followers from over 60 most popular (by Alexa ranking) merchant websites and discovered the following –

- We defined Quality of Service (QoS) metric for blackmarket services and discovered that 95% of these services have a QoS score of 0.28 (on a scale of 0 to 1, with 1 being highest) or lesser. While this is not surprising since such services are not reliable and may not deliver as promised, they still attract heavy traffic as shown by Alexa rankings.

- We use several parameters like spread on social networks, promotion by affected users and Alexa rankings to determine the market leaders. We discovered that only a few merchant websites contribute towards more than 70% popularity and usage for manipulation of Twitter follower count. This shows that there is an underlying *oligopolistic* structure, and we found evidence of collusion among these leading merchant websites. Figure 2 shows this phenomenon in more detail.

More details on landscaping the blackmarket services can be found in our recently published work [1].

5.2 Detection of Fake Twitter Followers

To understand how the sybil nodes created by the blackmarket services operate, we studied their behaviour in initial part of our work. To conduct the experiment, we used the same data purchased for the experiment in the previous section. With respect to these sybil nodes, in this case, fake (purchased) Twitter followers, we discovered the following –

[4]We ensured that all money we paid to underground merchants to acquire fake followers was exclusively for the dummy Twitter accounts we created. The dummy accounts were fully controlled by us and were for the sole purpose of conducting experiments.

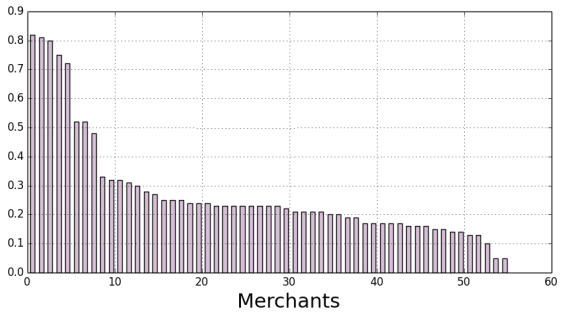

Figure 2: Popularity and spread of blackmarket websites selling fake Twitter followers. Few sellers are market leaders

- Fake followers have low social engagement. These users do not indulge in conversations with other users. We also found that they do not post original content and most of the tweets consist of RTs (Retweets). One of the other striking features of fake followers was that they have followers and followees tweeting in several languages. We found that 13% percent users had followees tweeting in 5 or more languages.

- We can detect fake following behaviour from legitimate with an accuracy of 89.2% by using temporal behavioral and network features of the suspicious users. Some of the most important attributes we discovered were number of unfollows by suspicious user over time, language overlap with the users she follows and social engagement like RTs and @-mentions. Table 2 elaborates results of the evaluation metrics for detection of suspicious following behaviour.

Table 2: Confusion Matrix – Classification Results of distinguishing legitimate users from those exhibiting suspicious following behaviour.

		Predicted	
		Suspicious	Legitimate
True	Suspicious	88.5	11.5
	Legitimate	9.7	89.9

Detailed explanation of characteristics of fake followers can be seen in our recent work [1].

5.3 Detection of Users with Manipulated Social Reputation

In this section we move towards the more challenging task of effective detection of users with manipulated social reputation. We conduct our initial experiment on Twitter. The perceived social reputation is defined by the follower count of a user. To assess the credibility of this follower count, we use the following parameters: tweet-follower ratio, evidence of user being a topic expert and overlap of her interests with followers. To build our baseline, we collected a random sample of 1.6 million Twitter users and used a multiple linear regression model to fit the three parameters. We define our model as –

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

For a particular user under suspicion, we find the deviation of its distribution based on the same three parameters from that of the random sample and define *suspect_ratio* as the inverse of the deviation. Therefore, smaller *suspect_ratio* indicates higher suspicion towards the perceived social reputation (follower count) of the user. We then conducted an in-the-wild experiment over 1% Twitter stream data and labeled a user as suspicious if her *suspect_ratio* was significantly low below a certain threshold. So far we have been able to label over 56,000 users as suspicious. A small sample of the users detected by our proposed methodology can be seen at `http://bit.ly/FakeFollowProj`.

6. CONCLUSION AND FUTURE WORK

This work aims to detect and measure the deviation from perceived social reputation of an OSN user. We start by landscaping the sources of such manipulation like blackmarkets and scratch-back services. Preliminary results bring out the underlying structure of blackmarket which can be helpful to uncover the market leaders. Eliminating or hindering their operations can significantly bring down crowdsourced manipulation of social reputation. Initial results also show that a robust and adaptive technique can be built to detect social reputation manipulation. However, our proposed framework is at a very nascent stage and needs much more improvement and rigorous evaluation. Much work yet remains to leverage this framework to build an alternate social reputation system and measure the effects of social reputation manipulation on OSN's ecosystem.

7. REFERENCES

[1] A. Aggarwal and P. Kumaraguru. What they do in shadows: Twitter underground follower market. In *Privacy, Security and Trust (PST)*, 2015.

[2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12, 2010.

[3] Z. Chu, I. Widjaja, and H. Wang. Detecting social spam campaigns on twitter. In *Applied Cryptography and Network Security*, pages 455–472. Springer, 2012.

[4] D. B. Clark. The bot bubble. `https://newrepublic.com/article/121551/bot-bubble-click-farms-have-inflated-social-media`, April 2015.

[5] DailyMail. More than 2 million of hillary clinton's twitter followers are fake or never tweet. `http://www.dailymail.co.uk/news/article-3038621/More-2-MILLION-Hillary-Clinton-s-Twitter-followers`, April 2015.

[6] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 35–47. ACM, 2010.

[7] A. Kapravelos, C. Grier, N. Chachra, C. Kruegel, G. Vigna, and V. Paxson. Hulk: Eliciting malicious behavior in browser extensions. In *Proceedings of the 23rd Usenix Security Symposium*, 2014.

[8] K. Lee, S. Webb, and H. Ge. Characterizing and automatically detecting crowdturfing in fiverr and twitter. *Social Network Analysis and Mining*, 5(1):1–16, 2015.

[9] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948. ACM, 2010.

[10] Microsoft. Trojan:js/febipos.a. `www.microsoft.com/security/portal/threat/encyclopedia/entry.aspx?Name=Trojan:JS/Febipos.A`, August 2013.

[11] Microsoft. Trojan:js/kilim.a. `https://www.microsoft.com/security/portal/threat/encyclopedia/entry.aspx?Name=JS/Kilim`, June 2013.

[12] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. Dirty jobs: The role of freelance labor in web service abuse. In *Proceedings of the 20th USENIX conference on Security*, pages 14–14. USENIX Association, 2011.

[13] NYTimes. A rave, a pan, or just a fake? `http://www.nytimes.com/2011/05/22/your-money/22haggler.html`, May 2011.

[14] NYTimes. Fake twitter followers become multimillion-dollar business. `http://bits.blogs.nytimes.com/2013/04/05/fake-twitter-followers-becomes-multimillion-dollar-business/`, April 2013.

[15] NYTimes. All the product reviews money can buy. http://www.nytimes.com/2015/12/06/your-money/all-the-product-reviews-money-can-buy.html, December 2015.

[16] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao. Follow the green: growth and dynamics in twitter follower markets. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 163–176. ACM, 2013.

[17] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson. Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In *USENIX Security*, pages 195–210. Citeseer, 2013.

[18] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In *Proceedings of the 23rd USENIX Security Symposium (USENIX Security)*, 2014.

[19] B. Viswanath, M. A. Bashir, M. B. Zafar, S. Bouget, S. Guha, K. P. Gummadi, A. Kate, and A. Mislove. Strength in numbers: Robust tamper detection in crowd computations. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, pages 113–124. ACM, 2015.

[20] G. Wang, T. Wang, H. Zheng, and B. Y. Zhao. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *23rd USENIX Security Symposium, USENIX Association, CA*, 2014.

[21] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao. Serf and turf: crowdturfing for fun and profit. In *Proceedings of the 21st international conference on World Wide Web*, pages 679–688. ACM, 2012.