Detecting and Characterizing Social Spam Campaigns

Hongyu Gao
Northwestern University
Evanston, IL, USA
hygao@u.northwestern.edu

Jun Hu
HUST
Wuhan, China
junehu1210@gmail.com

Christo Wilson
U. C. Santa Barbara
Santa Barbara, CA USA
bowlin@cs.ucsb.edu

Zhichun Li
Northwestern University
Evanston, IL, USA
lizc@cs.northwestern.edu

Yan Chen
Northwestern University
Evanston, IL, USA
ychen@northwestern.edu

Ben Y. Zhao
U. C. Santa Barbara
Santa Barbara, CA USA
ravenben@cs.ucsb.edu

ABSTRACT

Online social networks (OSNs) are exceptionally useful collaboration and communication tools for millions of users and their friends. Unfortunately, in the wrong hands, they are also extremely effective tools for executing spam campaigns and spreading malware.

In this poster, we present an initial study to detect and quantitatively analyze the coordinated spam campaigns on online social networks in the wild. Our system detected about 200K malicious wall posts with embedded URLs, traced back to roughly 57K accounts. We find that more than 70% of all malicious wall posts are advertising phishing sites.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Human Factors, Measurement, Security

Keywords

Online social networks, Spam, Spam Campaigns

1. INTRODUCTION

Online social networks (OSNs) are exceptionally useful collaboration and communication tools for millions of Internet users. Unfortunately, recent evidence shows that these trusted communities could become highly effective mechanisms for spreading malware and phishing attacks. Popular OSNs have recently become the target of phishing attacks launched from large botnets [2, 3], and account credentials are already being sold online in underground forums [5]. Using compromised or fake accounts, attackers can turn the trusted OSN environment against its users by masquerading phishing attempts and spam messages as communication from friends and family members.

In this project, we present the first attempt to detect and analyze the prevalence of malicious users and spread of malicious content on an OSN. We carry out the study on Facebook, the most popular OSN in the world with over 400 million users. We use the crawled Facebook data between April and June of 2009. We choose 8 regional networks of various sizes (from over 1.6 million users down to ∼14K users) as targets for data collection. For each crawled user we recorded the profile, friend list, and interaction records going back to January 1, 2008. Interaction records include the complete history of status updates and wall posts received by each crawled user within the given time frame. Overall, our complete dataset includes information on over 3.5 million users with more than 187 million wall messages.

We employ the correlation between wall messages, either by the textual content or the contained web address, to identify the spread of potentially malicious content. Our results are confirmed by a number of validation mechanisms. Our subsequent analysis provides insights into the operation of malicious accounts, and has significant implications on the design of future mechanisms to detect malicious behavior.

CCS'10, October 4–8, 2010, Chicago, Illinois, USA.
ACM 978-1-4503-0244-9/10/10.

2. MALICIOUS CAMPAIGN DETECTION AND VALIDATION

2.1 Design Overview

An overview of the system workflow is shown in Figure 1.

The design of our system is guided by intuition about techniques used in spam campaigns. From the recent work [4] and our observation of large number of malicious posts that look similar, we infer that spam wall posts are generated using templates, and posts generated from the same template should contain only small differences. Consequently, we propose to group wall posts with "similar" textual content together. Second, we recognize that the attempt to direct the viewers towards a single destination URL must come from the same spam campaign. Thus we will group together all wall posts that include the same destination URL, including those that have been hidden through URL obfuscation (*e.g.* www dot hack dot com).

We model all wall posts as nodes in a large graph, and build edges when two posts are connected by either similar textual content or same destination URL. During runtime, the system compares the destination URL before computing the approaximate textual similarity between descriptions, since the former is less expensive to compute. Each of the resulting connected subgraphs could represent messages within the same spam campaign. Identifying connected subgraphs is solved by iteratively choosing arbitrary nodes and identifying its transitive closure as a cluster. We summarize the implementation in Algorithm 1. We omit the detail of breadth-first search (BFS) due to the space limitation.

After identifying distinct subgraphs, we use threshold filters on: a) the number of users sending wall posts in the subgraph and b) the

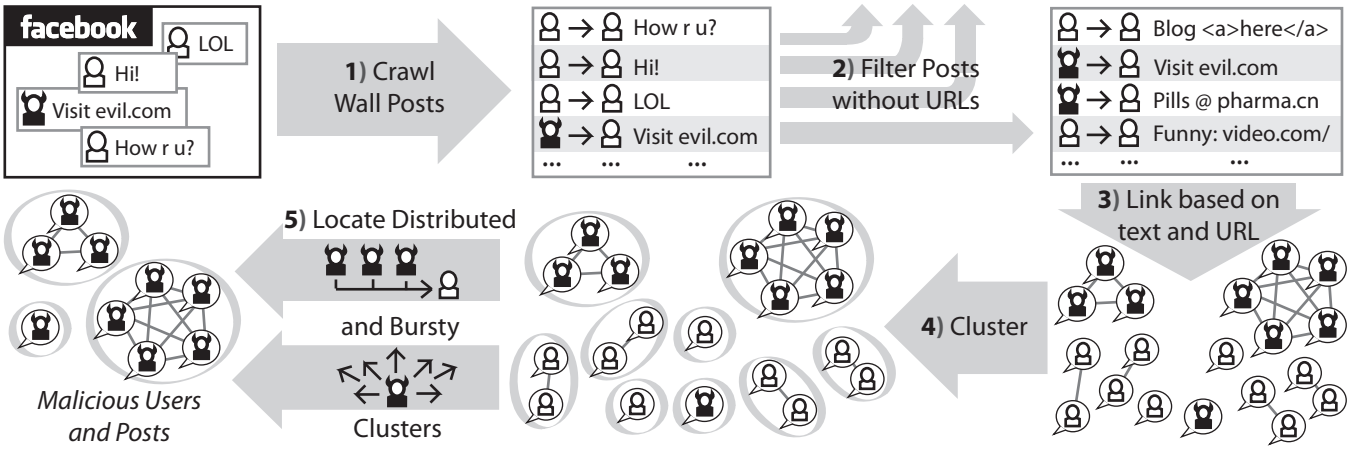

Figure 1: The system design, starting with raw data collection and ending with accurate classification of malicious posts and the corresponding users.

Algorithm 1 **PostSimilarityGraphClustering**($G < V, E >$)

```
traversed ← ∅
clusters ← ∅
Foreach v ∈ V
  If v ∈ traversed
    continue
  EndIf
  one_cluster ← BFS(v)
  traversed ← traversed ∪ one_cluster
  clusters ← clusters ∪ {one_cluster}
EndForeach
return clusters
```

time interval between consecutive wall posts in the subgraphtime to distinguish potentially malicious campaigns.

2.2 Detection Result

The clustering approach produces 1,402,028 clusters. As expected, there are a small number of very large clusters as well as a large number of very small clusters.

The chosen threshold, which is 5 as the minimum number of users that have made wall posts in the cluster and 5400 seconds (1.5 hours) as the maximum median interval between the timestamp of two consecutive wall posts, results in 297 clusters classified as malicious. The total number of wall posts contained in these 297 clusters is 212,863.

2.3 Experimental Validation

We apply a stringent set of heuristic tests to each URL that is contained in the detected malicious posts. Whether the URL is malicious determines whether the wall posts containing it is malicious.

6 steps are adopted. Each step can confirm the malice of a subset of the detection result. For any detection result that cannot be verified as maliciouis, we assume it to be benign, *i.e.* , false positive. These steps and the corresponding validation results are shown in Table 1.

Overall, we observe that our detection methodology results in a very low number of false positives. One additional positive feature of our detection methodology is that it is fully automated and significantly less costly in terms of time than our validation process. As it returns results of roughly equal quality, we believe that our detection methodology represents a step forward for automated detection of spamming activity on OSNs.

Reason for Classification	# of URLs	# of Wall Posts
Obfuscated URL	1003 (6.3%)	45655 (21.4%)
Blacklisted	4485 (28.0%)	55957 (26.3%)
Redirects to a blacklisted URL	4473 (27.9%)	29365 (13.8%)
Contains spam keywords	196 (1.2%)	19018 (8.9%)
Groups with other malicious URLs	5300 (32.5%)	33407 (15.7%)
Manual confirmation	27 (<0.1%)	16380 (7.7%)
Malicious, True Positives	15484 (96.1%)	199782 (93.9%)
Benign, False Positives	616 (3.9%)	13081 (6.1%)

Table 1: ***Validation results. Each row provides the number of confirmed malicious URLs and wall posts in a given validation step. All URLs that remain unvalidated after all steps are assumed to be benign.***

3. CAMPAIGN ANALYSIS

We use “campaign" to refer to a set of malicious posts of a certain type, *e.g.*, pharmaceutical sales. We use the description part of the wall post to distinguish campaigns, without considering the destination URL.

3.1 Campaign identification

We iteratively classify the wall posts by identifying strings characteristic to each campaign with the aid of human knowledge. Malicious posts that cannot be apparently grouped into any campaigns form an additional “other" group. We present all the identified campaigns in addition with a summary of their description in Table 2.

We further associate the campaigns with the clusters produced by the detection mechanism. For most campaigns, their contained clusters form mutually exclusive groups, except for only two instances. More specifically, the “crush" campaign shares one cluster with the “love-calc" campaign and the “PS3" campaign, respectively. These campaigns shares some common embedded URLs. It suggests that there is likely a single authority controlling all these three campaigns, who is using a set of very different templates to generate wall posts.

3.2 Attack categorization

In this subsection, we study the purpose of the attacker to launch the attack. We determine the goal of the attackers based on campaigns. The attacker’s goal is apparent for some campaigns, *e.g.* product selling. For the other campaigns, we rely on McAfee SiteAdvisor’s [1] user review summary of URLs within the cam-

Campaign	Summarized wall post description	Cluster #	Post #
Crush	Someone likes you	21	51082
Ringtone	Invitation for free ringtones	23	31329
Pharma	Pharmaceutical products like viagra	20	17614
Narcotics	Sell drugs	11	16668
Love-calc	Test the love compatibility	5	16354
Macy-gift	Invitation for free macy's giftcard	4	14092
Fake-video	A cool video is provided	114	11464
Pic-misuse	Your photo is misused online	1	10683
Iphone	Invitation for a free iphone	4	6317
Blog	Someone writes you in the blog	2	3948
Fake-fbid	Visit a (fake) facebook profile	1	3556
Fake-news	Visit to read news	1	2707
Is-that-you	Some webpage is about you	4	2620
Ipod-touch	Invitation for a free itouch	1	2125
Denigration	Someone is disparaging you	2	1440
PS3	Invitation for a free Playstation 3	2	1131
Webcam	Video chatting via web camera	4	1127
Luxury	Get cheap luxury product	1	981
Online-job	Work online and earn big money	5	502
Other	No apparent pattern	64	4042

Table 2: Campaigns encountered in the study. The attackers use the description to entice the receiver to visit the contained malicious URL. Cluster # shows how many clusters in the detection result are involved in each campaign. Post # represents the number of malicious posts in each campaign.

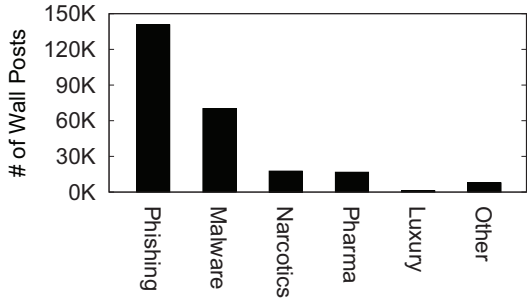

Figure 2: The number of instances in each attack category. Attack instances with multiple malicious intent are counted multiple times.

paign. We identify five different attacker's goals and present them in Figure 2.

The total size of all the categories exceeds the total number of malicious posts, since some malicious posts have multiple goals and are counted under all suitable categories. Phishing is the most common attack (∼70.3%). We encountered two different types of phishing attacks in the study. In the first case the attackers target at confidential information. In the second case, the attackers directly targets at money. Malware propagation is the second most common goal (∼35.1%). Product selling as a whole is still one of the main goals for the attackers to spam the OSNs (∼17.6%).

3.3 Temporal behaviors

We study the temporal features of the identified campaigns and illustrate the result in Figure **??**. The horizontal direction represents the timeline during the period of the data collection. The spam campaigns are represented by different strips. A short, thin vertical line within the strip corresponds to one malicious posts within the campaign. A block in the strip reflects a burst in the campaign, as it is composed of densely distributed vertical lines. The figure clearly shows the bursty nature of all the campaigns. The malicious posts within each campaign are densely distributed in a few relatively short time periods, although the entire campaign may span a much longer time period, like the "crush" campaign.

4. CONCLUSION

In this poster, we describe our work on detecting and characterizing spam campaigns performed using asynchronous wall messages on the Facebook social networks. We use automated techniques to group together wall posts that show strong similarities in advertised URL destination or text description. We identify about 200K malicious wall posts attributable to 57K malicious accounts. Over 70% of these attacks are phishing attacks. More importantly, our work demonstrates that automated detection techniques can be successfully used to detect online social spam.

5. REFERENCES

[1] Mcafee siteadvisor. http://www.siteadvisor.com/.

[2] Users of social networking websites face malware and phishing attacks. Symantec.com Blog.

[3] Zeus botnet targets facebook. http://blog.appriver.com/2009/10/zeus-botnet-targets-facebook.html.

[4] KREIBICH, C., KANICH, C., LEVCHENKO, K., ENRIGHT, B., VOELKER, G., PAXSON, V., AND SAVAGE, S. Spamcraft: An inside look at spam campaign orchestration. In *Proc. of LEET* (2009).

[5] Verisign: 1.5m facebook accounts for sale in web forum. PC Magazine, April 2010.