Hiding in Plain Sight: Characterizing and Detecting Malicious Facebook *Pages*

Prateek Dewan, Shrey Bagroy, Ponnurangam Kumaraguru
Indraprastha Institute of Information Technology, Delhi (IIIT-D)
Cybersecurity Education and Research Centre (CERC), IIIT-Delhi
Email: {prateekd,shrey14099,pk}@iiitd.ac.in

Abstract—**Facebook is the world's largest Online Social Network, having more than 1 billion users. Like most other social networks, Facebook is home to various categories of hostile entities who abuse the platform by posting malicious content. In this paper, we identify and characterize Facebook *pages* that engage in spreading URLs pointing to malicious domains. We revisit the scope and definition of what is deemed as "malicious" in the modern day Internet, and identify 627 *pages* publishing untrustworthy information, misleading content, adult and child unsafe content, scams, etc. Our findings revealed that at least 8% of all malicious *pages* were dedicated to promote a single malicious domain. Studying the temporal posting activity of *pages* revealed that malicious *pages* were 1.4 times more active daily than benign *pages*. We further identified collusive behavior within a set of malicious *pages* spreading adult and pornographic content. Finally, we attempted to automate the process of detecting malicious Facebook *pages* by training multiple supervised learning algorithms on our dataset. Artificial neural networks trained on a fixed sized bag-of-words performed the best and achieved an accuracy of 84.13%.**

I. Introduction

Internet users around the world use Online Social Networks (OSNs) as primary sources to consume news, updates, and information about events around the world. However, given the enormous volume and veracity, it is hard to moderate all content that is generated and shared on OSNs. This enables hostile entities to generate and promote all sorts of malicious content (including scams, fake information, adult content, etc.) and pollute the information stream for monetary gains, or to compromise system reputation. Such activity degrades user experience, and violates terms of service of OSN platforms.

Researchers in the past have studied and proposed automated techniques to identify malicious user accounts on OSNs [2], [9], [10], [17]. Most of these techniques have a restricted focus of malicious content, which is limited to promotional posts, duplicate bulk messages, campaigns, phishing, and malware. However, with the advent of OSNs and Web 2.0, the scope of what is deemed as "malicious" on the Internet has evolved. Facebook, for example, has established community standards to safeguard users against nudity, hate speech, etc. [7], and considers any entities that confuse, mislead, surprise or defraud people, as abusive [6]. We recently discovered the presence of a similar set of malicious Facebook *pages* accounting for over 30% of malicious posts in our dataset, and have not been studied in detail in the past [4]. Security experts and news sources have also acknowledged the presence of malicious *pages* on Facebook, set up intentionally to spread fraudulent claims, scams, and other types of malicious content. [1] In addition to scams and fake information, researchers have also identified and studied the spread of rumors on Facebook [8]. During events like earthquakes, rumors on OSNs have been observed to contribute to chaos and insecurity in the local population [14]. It is thus, crucial to identify and control the spread of untrustworthy and fake information in order to minimize adverse real world impact.

In this paper, we identify and characterize a set of 627 Facebook *pages* that published one or more malicious URLs in their most recent 100 posts. We focus our analysis on *pages* which spread untrustworthy information, hate speech, nudity, misleading claims, etc., that are deemed as malicious by the community standards [7] and "*Page* Spam" definitions [6] established by Facebook. In an attempt to automate the process of identifying malicious Facebook *pages*, we train and evaluate multiple supervised learning models based on bag-of-words obtained using the textual content published by these *pages*.

II. Background

The definition and scope of what should be labeled as "malicious content" on the Internet has been constantly evolving since the birth of the Internet. With respect to Online Social Networks, state-of-the-art techniques have become efficient in automatically detecting spam campaigns [10], [19], and phishing [1] without human involvement. However, new classes of malicious content pertaining to appropriateness, authenticity, trustworthiness, and credibility of content have emerged in the recent past. Some researchers have studied these classes of malicious content on OSNs and shown their implications in the real world [3], [11], [12], [14]. All of these studies, however, resorted to human expertise to identify untrustworthy and inappropriate content and establish ground truth, due to the absence of efficient automated techniques to identify such content. We aim to study a similar class of malicious content pertaining to trustworthiness and appropriateness in this work, which currently requires human expertise to identify. In particular, we look at Facebook *pages* that generate content deemed as malicious by Facebook's community standards [7] and definitions of "*Page* Spam" [6].

[1] https://grahamcluley.com/2015/09/british-airways-isnt-giving-away-free-flights-year-facebook-scam/

A. Establishing ground truth

In order to obtain ground truth for the malicious content we aim to study, we resort to a crowd sourced approach. Crowdsourcing techniques have been shown to perform well for establishing ground truth for complex and subjective aspects of OSN content such as credibility [3], [11]. For our study, we use the Web of Trust (WOT) API, a crowdsourcing based website reputation service. [2] WOT's crowdsourcing mechanism enables it to spot scams, unreliable web stores, misleading websites, nudity, and questionable content, which largely overlaps with Facebook definitions of spam. To the best of our knowledge, WOT is one of the only services which covers the broader definition of malicious content that is required for our study.

B. Dataset

We collected an initial dataset of 4.4 million public posts published by 390,246 unique *pages* and 2,983,707 unique users on Facebook between April 2013 and July 2014, using Facebook's post search API. These posts were collected by using event related search keywords belonging to 17 real world events that took place in the aforementioned time frame. We queried the WOT API for domain reputations of all URLs present in the 4.4 million posts. URL domains with a low *reputation* score ($<$60) with high *confidence* score ($\geq$10) were marked as malicious. We also marked a URL as malicious if it fell under the *Negative* or *Questionable* category. [2] With this technique, we identified 10,341 posts containing one or more malicious URLs. These 10,341 posts originated from 1,557 *pages* and 5,868 users. Due to space constraints, we omit the detailed description of our dataset. The complete details of this dataset can be found in our prior work [4].

We re-queried the Graph API in August, 2015, and collected the *page* information of the 1,557 *pages* posting malicious URLS. We also collected 100 most recent posts (or all posts, whichever was smaller) published by these *pages* using the Graph API [3] along with all *likes*, *comments*, and *shares* on these posts. We then looked up the WOT API for all URL domains present in the most recent 100 posts, and found that 627 *pages* published one or more malicious URLs. This exercise of rescanning the 1,557 *pages* eliminated those *pages* which had not shown malicious activity in the recent past (recent 100 posts), and could be deemed as non-malicious for our study. For the rest of the paper, we use the remaining 627 *pages* as our dataset of malicious *pages*.

We also drew an equal random sample of 1,557 *pages* from the benign *pages* in our dataset, which had not posted any malicious URLs during our initial data collection phase (April 2013 - July 2014). Similar to our approach for identifying malicious *pages*, we re-queried the Graph API and collected the *page* information along with the most recent 100 posts (including their *likes*, *comments*, and *shares*) published by these *pages*. We found 1,278 *pages* which published no malicious URLs in their most recent 100 posts. These 1,278 *pages* made up our dataset of benign *pages*. Table I shows the descriptive statistics of all Facebook *pages* in our dataset.

TABLE I
DESCRIPTIVE STATISTICS OF OUR DATASET OF FACEBOOK *pages*. NUMBERS IN THE PARENTHESES INDICATE VERIFIED *pages*.

Category	Malicious	Benign
No. of *pages*	627 (31)	1,278 (49)
Recent 100 posts	60,306	120,184
Recent 100 posts with URLs	55,233	92,980
Likes (recent 100 posts)[4]	3,447,669	31,680,263
Comments (recent 100 posts)	354,502	1,245,959
Shares (recent 100 posts)	507,964	1,012,151

III. MALICIOUS *pages* ON FACEBOOK

To understand the differences (and similarities) between malicious and benign *pages*, we studied both the spatial and temporal behavior of these *pages*.

A. Spatial behavior

1) Content: Scanning the most recent 100 posts revealed that almost half the *pages* (49.28%) in our dataset published 10 or less posts containing a malicious URL. Overall, the median number of domains shared by these *pages* was 24.5. On the contrary, the median number of domains shared by the other half of the *pages* posting more than 10 posts containing a malicious URL (50.72%) was 5. We found a weak declining trend in the number of domains as the number of malicious posts increased (r = -0.223, *p-value*$<$0.01). This declining trend (and negative correlation) indicated that *pages* posting a large number of malicious URLs tend to do so from a small subset of domains. In fact, 53 *pages* (8.45%) shared URLs from only 1 domain, and published more than 90 posts containing a malicious URL. Most certainly, such *pages* exist for the sole purpose of promoting a single (malicious) domain, and are successful in engaging thousands of Facebook users. This sort of activity resembles social spam campaigns studied widely in the past [10], [19]. However, since most past research has focused on more obvious threats like unsolicited and targeted spam, advertising, and bulk messaging, other types of malicious content concerned with trustworthiness and child safety has largely remained unaddressed.

Note that there also exist multiple legitimate *pages* on Facebook dedicated to promote a single domain, for example, the BBC News *page* (exclusively posting bbc.com URLs). We found 118 *pages* in our benign dataset (9.23%) which were dedicated to promote a particular domain. Such behavior cannot therefore be associated exclusively with malicious activity. Malicious *pages* seem to take advantage of this fact and continue their activity, hiding in plain sight. However, the vocabulary used in the content published by these *pages* can be used to differentiate between the malicious and benign classes using a bag-of-words.

[2] https://www.mywot.com/wiki/API

[3] https://developers.facebook.com/docs/graph-api/reference/page/feed

[4] Due to API rate limitations, we had to restrict our data collection to 50,000 *likes* per post. 2 malicious and 291 benign posts exceeded this limit.

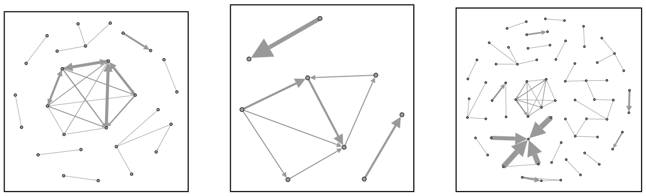
2) Network: Past research has shown that decentralized networks are prone to *sybil attacks*, wherein malicious entities tend to collude together and attempt to infiltrate the legitimate part of the network [5]. Such attacks have also been studied in context of OSNs [18]. To investigate if such phenomenon exists for Facebook *pages* too, we analyzed the inter, and intra page *like, comment*, and *share* networks for both malicious and benign *pages* in our dataset. In particular, we analyzed networks consisting of *pages* and users *liking, commenting on,* or *sharing* posts from two or more *pages* in our dataset (inter-*page* networks), and networks of *pages liking, commenting on,* or *sharing* posts from *pages* within our dataset (intra-*page* networks) for malicious and benign pages separately. To keep the network size comparable, we averaged out the results for 10 random samples of 627 benign *pages* each (same size as malicious *pages* dataset) drawn from the full 1,278 benign *pages* dataset.

We found that the inter-likes network for benign *pages* (83,799 nodes) was much larger and stronger (avg. weighted degree: 41.695) than the inter-likes network for malicious *pages* (21,947 nodes, avg. weighted degree: 24.273), indicating that more users *liked* posts from two or more benign *pages* as compared to the number of users who *liked* posts from two or more malicious *pages* in our dataset. More interestingly, we found stronger ties (avg. weighted degree) within malicious *pages* in all aspects (*likes, comments*, and *shares*) as compared to benign *pages*, indicating collusion and sybil behavior within malicious *pages*. We also found a much larger number of connected components in all inter-*page* networks for benign *pages* as compared to inter-*page* networks for malicious *pages*, indicating a larger and more diverse audience for benign *pages* as compared to malicious *pages*.

Stronger ties within malicious *pages* prompted us to further investigate the components we detected from intra-*page likes, comments*, and *shares* networks. Figure 1 shows the network graphs of the detected components. We observed that post *sharing* was the most prominent intra-*page* activity, followed by *liking* and *commenting*. The network graphs also revealed a distinct component of six Facebook *pages* fully connected to each other in terms of *likes* (Figure 1(a)) and *shares* (Figure 1(c)). Five out of these six *pages* also formed a component in the intra-comments graph (Fig 1(b)). We manually inspected and observed that all *pages* in this component belonged to adult stars and promoted pornographic content. This behavior closely resembled a sybil network, and indicated that all these *pages* were controlled by / belong to the same real-world entity (person or organization). We also found multiple two-*page* components involving politically polarized *pages*, where one *page* heavily engaged in *liking, commenting on*, and *sharing* the other page's content.

B. Temporal behavior

We explored the temporal activity of all *pages* in our dataset to determine how active the *pages* were. To be able to quantitatively compare the activity of malicious and benign *pages*, we calculated a *daily activity ratio* for each page, defined by the ratio of "number of days a *page* was active (published one or more posts)" versus the "total number of days between the first and hundredth post" by the page.



(a) *Pages liking* each others' posts. (b) *Pages commenting* on each others' posts. (c) *Pages sharing* each others' posts.

Fig. 1. Network graphs capturing intra-*page* activity of malicious *pages* in our dataset. We found multiple two-node components.

We observed that 27.43% of all malicious *pages* were active daily as compared to only 8.60% daily active benign *pages*. On average, malicious *pages* were 1.4 times more active daily as compared to benign *pages* in our dataset. We also calculated activity ratio in terms on number of hours and number of weeks, and observed similar results. All activity ratio values were compared using Mann-Whitney U test and the differences were found to be statistically significant (*p-value*<0.01 for all experiments) [13]. These difference confirmed that malicious *pages* in our dataset were more active as compared to benign *pages*, and published more frequently.

IV. Automatic detection of malicious pages

Past research has shown that URL blacklists and reputation services are ineffective initially, and take time to update [16]. Moreover, lack of blacklists and reputation services for malicious content other than phishing, and malware demand the need for an automated solution to analyze and detect malicious Facebook *pages*. To fulfil this need, we trained supervised learning algorithms on our labeled dataset to create an effective model for automatic detection of malicious Facebook *pages*, independent of third party reputation services.

We used a bag-of-words model to automatically identify malicious Facebook *pages* in our dataset. We collected textual content from three sources (wherever present), viz. status message in the post, name and description of the link present in the post (if any). We performed experiments by calculating term frequencies of unigrams, bigrams, and trigrams, and limited our vocabulary size to the top 10,000 features. We used balanced training and test sets containing equal numbers of positive and negative examples (627 malicious *pages*, and 627 benign *pages* randomly picked from our dataset).

A bag-of-words with 10,000 features produced a sparse feature vector. This sparse data prompted us to explore more state-of-the-art learning techniques for fast and effective classification. We chose Sparsenn for this task. [5] Sparsenn is a C implementation of artificial neural networks based on stochastic gradient descent, designed for learning neural networks from high dimensional sparse data. Table II presents the results of our experiments.

[5] http://lowrank.net/nikos//sparsenn/

Neural networks on trigrams (hidden units: 64, learning rate: 0.07, determined experimentally) performed the best, achieving an accuracy of 84.13% (ROC area under curve: 0.9). Thus artificial neural networks trained on the top 10,000 trigrams outperformed all the other learning models.

TABLE II
RESULTS FOR SUPERVISED LEARNING EXPERIMENTS. ARTIFICIAL NEURAL NETWORKS PERFORMED THE BEST.

Classifier	Feature set	Acc. (%)	ROC AUC
Naive Bayesian	Unigrams	68.27	0.682
	Bigrams	69.06	0.690
	Trigrams	69.77	0.697
Logistic Regression	Unigrams	74.18	0.795
	Bigrams	74.34	0.791
	Trigrams	73.93	0.789
Random Forest	Unigrams	72.26	0.794
	Bigrams	71.80	0.802
	Trigrams	72.18	0.794
Neural Networks	Unigrams	81.74	0.862
	Bigrams	84.12	0.872
	Trigrams	**84.13**	**0.900**

V. RELATED WORK

Gao et al. characterized spam campaigns launched using accounts on Facebook using a dataset of 187 million posts [10]. Authors relied on URL blacklists to detect spam, phishing and malware. Following up their work, Gao et al. presented an online spam filtering system to inspect messages generated by users in real time [9]. In an attempt to protect Facebook users from malicious posts, Rahman et al. designed a social malware detection method which took advantage of the social context of posts [15]. Authors were able to achieve a maximum true positive rate of 97%, using a SVM based classifier trained on 6 features. Stringhini et al. [17] utilized a honeypot model to collect information about spammers on Facebook. Ahmed et al. presented a Markov Clustering (MCL) based approach for the detection of spam profiles on Facebook. Authors crawled the public content posted by 320 handpicked Facebook users, out of which, 165 were manually identified as spammers, and extracted 3 features from these profiles, which served as input to the Markov Clustering model [2].

Most aforementioned research relied on URL blacklists to identify ground truth spam, phishing, and malware, and tried to identify patterns which could be used to design effective measures to curb the spread of spam on OSN platforms. However, fewer attempts have been made to go beyond the traditional spam, phishing, and malware, and address other classes of malicious content on OSNs which include untrustworthy content, hate and discrimination, etc. that are non-trivial to identify through automated means. There has been some research in the space of identifying credible content on Twitter [3], [11], but state-of-the-art techniques proposed by researchers to detect content credibility have not been able to achieve the degree of efficiency that has been achieved in detecting traditional spam, phishing and malware.

VI. DISCUSSION

Pages on Facebook have a lot in common with Facebook groups and events. Groups and events can also be used to target large audiences at once. Our analysis can thus be easily extended to study malicious groups and events as well.

Our bag-of-words model is based on a limited history (100 posts) of *page* activity. Although it is possible to collect and analyze the entire history for all *pages*, doing so would be time consuming and computationally expensive. Moreover, *pages* can change behavior over time; malicious *pages* may stop spreading malicious content, while benign *pages* may start engaging in posting malicious content over time. To accommodate such changes in behavior, we recommend a self-adaptive model which relies on the most recent activity by the page. The history (number of posts) to consider can be decided experimentally. Such a model would be accommodative of the changing behavior of *pages* over time.

VII. ACKNOWLEDGEMENTS

We would like to thank the members of Precog and CERC research groups at IIIT-Delhi for their support and feedback.

REFERENCES

[1] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru. Phishari: Automatic Realtime Phishing Detection on Twitter. In *eCrime Researchers Summit (eCrime), 2012*, pages 1–12. IEEE, 2012.

[2] F. Ahmed and M. Abulaish. An MCL-based Approach for Spam Profile Detection in Online Social Networks. In *IEEE TrustCom*, pages 602–608. IEEE, 2012.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In *WWW*, pages 675–684. ACM, 2011.

[4] P. Dewan and P. Kumaraguru. Towards Automatic Real Time Identification of Malicious Posts on Facebook. In *13th Annual Conference on Privacy, Security and Trust (PST)*, pages 85–92. IEEE, 2015.

[5] J. R. Douceur. The Sybil Attack. In *Peer-to-peer Systems*, pages 251–260. Springer, 2002.

[6] Facebook. What is page spam? *https://www.facebook.com/help/116053525145846*, 2015.

[7] Facebook.com. Facebook Community Standards. *https://www.facebook.com/communitystandards*, 2015.

[8] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor Cascades. In *ICWSM*, 2014.

[9] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. N. Choudhary. Towards Online Spam Filtering in Social Networks. In *NDSS*, 2012.

[10] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and Characterizing Social Spam Campaigns. In *Internet Measurement Conference*, pages 35–47. ACM, 2010.

[11] A. Gupta and P. Kumaraguru. Credibility Ranking of Tweets During High Impact Events. In *PSOSM*. ACM, 2012.

[12] M. Gupta, P. Zhao, and J. Han. Evaluating Event Credibility on Twitter. In *SDM*, pages 153–164. SIAM, 2012.

[13] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[14] M. Mendoza, B. Poblete, and C. Castillo. Twitter Under Crisis: Can We Trust What We RT? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.

[15] M. S. Rahman, T.-K. Huang, H. V. Madhyastha, and M. Faloutsos. Efficient and Scalable Socware Detection in Online Social Networks. In *USENIX Security Symposium*, pages 663–678, 2012.

[16] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang. An Empirical Analysis of Phishing Blacklists. In *Sixth Conference on Email and Anti-Spam (CEAS)*, 2009.

[17] G. Stringhini, C. Kruegel, and G. Vigna. Detecting Spammers on Social Networks. In *ACSAC*, pages 1–9. ACM, 2010.

[18] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai. Uncovering Social Network Sybils in the Wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(1):2, 2014.

[19] X. Zhang, S. Zhu, and W. Liang. Detecting Spam and Promoting Campaigns in the Twitter Social Network. In *IEEE 12th International Conference on Data Mining (ICDM)*, pages 1194–1199. IEEE, 2012.