Hiding in Plain Sight: The Anatomy of Malicious *Pages* on Facebook*

Prateek Dewan, Shrey Bagroy, and Ponnurangam Kumaraguru

IIIT-Delhi, India
{prateekd, shrey14099, pk}@iiitd.ac.in

Abstract. Facebook is the world's largest Online Social Network, having more than 1 billion users. Like most social networks, Facebook is home to various categories of hostile entities who abuse the platform by posting malicious content. In this chapter, 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. We perform in-depth characterization of *pages* through spatial and temporal analysis. Upon analyzing these *pages*, our findings reveal dominant presence of politically polarized entities engaging in spreading content from untrustworthy web domains. Studying the temporal posting activity of *pages* reveal that malicious *pages* are 1.4 times more active daily than benign *pages*. We further identify collusive behavior within a set of malicious *pages* spreading adult and pornographic content. Finally, we attempt to automate the process of detecting malicious Facebook *pages* by extensively experimenting with multiple supervised learning algorithms and multiple feature sets. Artificial neural networks trained on a fixed sized bag-of-words perform the best and achieve a maximum ROC area under curve value of 0.931.

1 Introduction

Online Social Networks (OSNs) are an integral part of the modern Internet. Users around the world use OSNs as primary sources to consume news, updates, and information about events around the world. However, given the enormous volume and veracity of content on social networks, 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 not only degrades user experience, but also violates the terms of service of OSN platforms.

* This chapter is an extended version of the paper titled *"Hiding in Plain Sight: Characterizing and Detecting Malicious Facebook Pages"* previously accepted at ASONAM 2016.

Researchers have studied and proposed automated techniques to identify malicious user accounts on OSNs [2, 16, 17, 26, 38]. 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. [13], and considers any pages, groups or events that confuse, mislead, surprise or defraud people, as abusive [12]. In a recent study, we 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 [9]. 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. A group of scammers, for example, set up a fake British Airways *page*, offering free flights to customers for a year (Figure 1). The *page* asked users to *share* a photo, *like* the *page* and leave a *comment* to claim their free flights[1]. In another similar incident, an international football player's name was used as bait to set up a Facebook *page*, and users were asked to sign a fake petition[2].

In addition to scams and fake information, researchers have also identified and studied the spread of rumors (which are a class of untrustworthy information) on Facebook [15]. In case of events like earthquakes, rumors on OSNs have been observed to contribute to chaos and insecurity in the local population [29]. Facebook also faced criticism because of presence of fake news and polarized politics on the platform during the recent presidential elections in the USA [18]. Such instances highlight a new and emerging class of malicious content, which is much harder to identify using automated means, and hasn't been widely explored in literature. It is thus crucial to identify and control the spread of untrustworthy and fake information, and minimize adverse real world impact.

In this chapter, we identify and characterize a set of 627 Facebook *pages* that published one or more malicious URLs (URLs that point to webpages comprising malicious content) in their most recent 100 posts. We expand on our characterization study, feature sets, and supervised learning experiments from the previous version [8]. 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 [13] and "*Page* Spam" definitions [12] established by Facebook. We use our labeled dataset to train and evaluate multiple supervised learning models to automate the process of identifying malicious Facebook *pages*. We extract a total of 96 features from *page* information, and posts published by the *pages*. Further, we train and evaluate supervised learning models using a bag-of-words obtained using the textual content published by these *pages*. Our broad findings are as follows:

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

[2] http://www.marca.com/2014/07/18/en/football/barcelona/1405709402.html

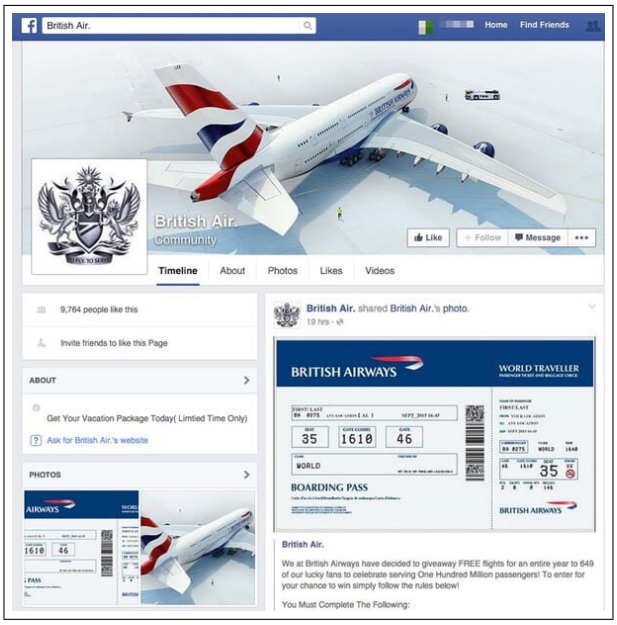


Fig. 1. Fake British Air Facebook page offering free flights for a year in return for liking, commenting on, and sharing their post.

- **Politically polarized malicious entities**: We identified and manually verified the presence of numerous politically polarized entities, which dominated our dataset of malicious *pages*, and published URLs from untrustworthy web domains.
- **Malicious *pages* were more active**: We found that malicious *pages* were more active (in terms of posting) than benign *pages*; the number of malicious *pages* that were active daily was over three times the number of benign *pages* that were active daily.
- **Malicious *pages* showed collusive behavior**: We found presence of collusive behavior within malicious *pages* in our dataset; malicious *pages* engaged in promoting (*liking, commenting on*, and *sharing*) each others' content.
- **Malicious and benign *pages* had similar temporal behavior**: We performed a longitudinal study over a period of one year, by capturing daily snapshots of malicious and benign *pages* in our dataset, and found minimal statistically significant difference between the two types of *pages* in terms of temporal behavior.
- **Artificial neural networks outperformed all other algorithms**: Artificial neural networks trained on a bag-of-words outperformed all other supervised learning algorithms for automatic detection of malicious *pages*, achieving an area under the ROC curve value of 0.9. Grid search experi-

ments help improve the performance further, attaining a maximum ROC AUC value of 0.931.

The rest of the chapter is structured as follows. Related work is discussed in Section 2. Section 3 gives the background and scope of our research, and explains the data collection process. Characterization and analysis of malicious *pages* make up Section 4. We present the results of our supervised learning experiments in Section 5. Section 6 discusses the implications and limitations of our results. We conclude and discuss the future directions of our work in Section 7.

2 Related Work

Detecting malicious content beyond spam and phishing, like fake reviews, link farming, social spam, etc. has received some attention by the research community [23]. These pieces of research have highlighted the surge of non-traditional spam on OSNs in general, and have also proposed network based techniques to combat suspicious behavior [22]. These techniques are however, restricted by the scarcity of data available in practice.

Social spam detection: Fake reviews and opinionated spam has been prevalent and well studied in the ecommerce domain. Mukherjee et al. employed human labelers to identify groups of fake reviewers on Amazon. Authors proposed several behavior features derived from collusion among fake reviewers, and proposed a relation-based ranking model for automatic detection of spammer groups. Due to the human labeling required for generating ground truth, this method was limited in terms of scalability and generalizability [31]. Owing to this drawback, authors proposed an unsupervised Bayesian inference model to identify spammers [30]. Although this technique was novel and shown to be effective, the features used by authors (for example, extreme rating, rating deviation) cannot be ported directly to the social network domain.

Similar attempts were made by Jindal et al. [24], and Lim et al. [27], both of which exploited review ratings and feedbacks from reviews to identify fraudulent activity using ranking and classification techniques under supervised settings. Manual effort needed for these techniques, in terms of ground truth generation and evaluation respectively, was a limitation in both of these attempts.

Akoglu et al. [3] and Fei et al. [14] proposed network based approaches to identify fraudulent reviewers by modeling the reviewer network as a graph. Although these approaches performed reasonably well, the lack of network level data in practice on OSNs like Facebook makes it difficult for porting and evaluating such approaches for identifying malicious entities on Facebook. Also, network level approaches help identify fraudulent users on the network, while the focus of our study is more on the content.

Ratkiewicz et al. studied astroturf political campaigns on Twitter using supervised learning. Similar to previous approaches, this work also required human labeling for ground truth generation [35]. Authors followed up on this approach

and presented Truthy, a web service that tracked political memes in Twitter and helped detect astroturfing, smear campaigns, and other misinformation in the context of U.S. political elections [34]. This research is closely aligned to our work. We overcome the drawbacks and limitations introduced because of human labeling by leveraging crowdsourced services like Web of Trust to extract ground truth for our study.

Malicious content on Facebook: Multiple researchers have studied and proposed techniques to detect malicious content on Facebook. Gao et al. presented an initial study to quantify and characterize spam campaigns launched using accounts on Facebook [17]. They studied a large anonymized dataset of 187 million asynchronous "wall" messages between Facebook users, and used a set of automated techniques to detect and characterize coordinated spam campaigns. Authors of this work relied on URL blacklists to detect spam URLs and concentrated on spam, phishing and malware. Following up their work, Gao et al. presented an online spam filtering system that could be deployed as a component of the OSN platform to inspect messages generated by users in real time [16].

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 [33]. Authors were able to achieve a maximum true positive rate of 97%, using a SVM based classifier trained on seven features. These features included no. of likes, no. of comments, no. of wall posts, no. of news feed posts, message similarity score, spam keyword score, and a boolean feature capturing whether a URL present in a post was obfuscated. Similar to Gao et al's work [17], this work was also targeted at detecting spam campaigns.

Stringhini et al. [38] utilized a honeypot model to collect information about spammers on Facebook. They created and monitored a honey profile for over one year, and manually identified 173 spam profiles among a total of 3,831 friendship requests they received. Studying the spam profiles, authors developed six features, viz. follower-friend ratio, message similarity, URL ratio, no. of friends, no. of messages sent, and friend choice. Authors trained a Random Forest classifier on 173 spam and 827 legitimate profiles, and reported a false positive rate of less than 3% for both Facebook and Twitter networks. 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. Authors then extracted 3 features from these profiles, viz. Active friends, *Page* Likes, and URLs to generate a weighted graph, 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 [5, 20], 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 spam, phishing and malware.

3 Scope and data collection

Facebook (unlike Twitter, Instagram, etc.) poses a restriction on the number of connections a user can have (max. 5,000 friends), and provides *pages* to enable large following for celebrities, groups, businesses, etc. A Facebook *page* can have multiple administrators managing the *page* under the same name, without the audience knowing. This allows *pages* to have a higher degree of interaction with its audience and keeping it more active as compared to a single user profile. Facebook *pages* are an important and integral part of the Facebook ecosystem, that offer a free platform for promotion of businesses, brands and organizations[3]. From an attacker's perspective, Facebook *pages* are potentially lucrative tools to gather large audiences and target all of them at once. Our past research has shown greater participation of Facebook *pages* in posting malicious URLs as compared to posting benign URLs [9]. Such inflated malicious activity and reach of Facebook *pages* make them a vital aspect to study in detail.

3.1 Scope

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. Researchers have been studying malicious content in the form of spam and phishing for over two decades. With respect to Online Social Networks, state-of-the-art techniques have become efficient in automatically detecting spam campaigns [17, 45], 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 [5, 19, 21, 29]. 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 and definitions of "*Page* Spam". Facebook defines "*Page* Spam" as *pages* that *confuse, mislead, surprise or defraud people* [12]. Additionally, *pages* that are misleading, deceptive, or otherwise misrepresent the prize or any other aspect of promotion are considered as "*Page* Spam". Facebook has also established

[3] https://www.facebook.com/help/174987089221178

community standards to protect users against nudity, hate speech, violence and graphic content, fraud, sexual violence etc. [13].

3.2 Establishing ground truth

Given the complex definition of malicious content for the scope of our study, there exist no accurate detectors for establishing ground truth. Detectors such as URL blacklists (Google Safebrowsing, PhishTank, SURBL, SpamHaus, etc.) used for identifying malicious content, are restricted to identifying classical threats like phishing, malware, etc. In order to obtain ground truth for the malicious content we aim to study, we resorted 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 [5, 19]. For our study, we used the Web of Trust (WOT) API [43]. WOT leverages crowdsourcing to collect ratings and reviews from millions of users who rate and comment about websites, based on their personal experiences. This crowdsourced, community based mechanism enables WOT to protect users against threats that only the human eye can spot such as 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.

We understand that WOT ratings are obtained through crowd sourcing, and may seem to suffer from biases. However, WOT states that in order to keep ratings more reliable, the system tracks each user's rating behavior before deciding how much it trusts the user. In addition, the meritocratic nature of WOT makes it far more difficult for spammers to abuse. This approach is similar to other crowdsourcing services like Amazon's Mechanical Turk[4] and CrowdFlower[5], which have been widely used in OSN research (as discussed above). To further increase the confidence in the ratings, we used conservative thresholds for confidence values associated with the reputation scores. We discuss these thresholds in more detail in Section 3.3.

3.3 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. All events we picked for our analysis, made headlines in international news. To maintain diversity, we selected events covering various domains of news events like political, sport, natural hazards,

[4] http://mturk.com/
[5] http://crowdflower.com/

terror strikes and entertainment news. For all 17 events, we started data collection from the time the event took place, and stopped about two weeks after the event ended. Table 1 shows the detailed description of this data.

We queried the WOT API for domain reputations of all URLs present in the 4.4 million posts (480,407 unique URLs), and identified 10,341 posts containing one or more malicious URLs (4,622 unique URLs). We will discuss the exact definition of "malicious" later in the section. These 10,341 posts containing malicious URLs originated from 1,557 *pages* and 5,868 users.

To capture a more recent view of the 1,557 *pages* posting malicious URLs, we re-queried the Graph API and collected their *page* information in August 2015. We also collected 100 most recent posts (or all posts, whichever was smaller) published by these *pages* using the */page-id/posts* endpoint of the Graph API[6] along with all *likes*, *comments*, and *shares* on these posts. Collecting posts along with their likes, comments, and shares is a computationally expensive and time consuming task. Due to limited time and resources, we had to restrict our dataset to 100 most recent 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 chapter, we use the remaining 627 *pages* as our dataset of malicious *pages*.

According to its documentation[7], the WOT API returns a reputation score for a given domain. Reputations are measured for domains in several *components*, for example, trustworthiness. For each {`domain, component`} pair, the system computes two values: a *reputation* estimate and the *confidence* in the reputation. Together, these indicate the amount of trust in the domain in the given component. A *reputation* estimate of below 60 indicates *unsatisfactory*. The WOT browser add-on requires a confidence value of ≥ 10 before it presents a warning about a website. We tested the domain of each URL in our dataset for *Trustworthiness* and *Child Safety* components. For our analysis, a URL was marked as malicious if both the aforementioned conditions were satisfied (*reputation*$<$60; *confidence*$\geq$10). In addition to reputations, the WOT rating system also computes categories for websites based on votes from users and third parties. We marked a URL as malicious if it fell under the *Negative* or *Questionable* category (Table 2). We used the same approach previously to develop techniques for automatic identification of individual malicious posts on Facebook [9].

We also drew an equal random sample of 1,557 *pages* from the benign *pages* in our dataset of 4.4 million posts, 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

[6] https://developers.facebook.com/docs/graph-api/reference/page/feed
[7] https://www.mywot.com/wiki/API

Table 1. Event name, keywords used as search queries, number of posts, and description for the 17 events in our dataset.

Event *(keywords)*	# Posts	Description
Missing Air Algerie Flight AH5017 *(ah5017; air algerie)*	6,767	Air Algerie flight 5017 disappeared from radar 50 minutes after take off on July 24, 2014. Found crashed near Mali; no survivors.
Boston Marathon Blasts *(prayforboston; marathon blasts; boston marathon)*	1,480,467	Two pressure cooker bombs exploded during the Boston Marathon at 2:49 pm EDT, April 15, 2013, killing 3 and injuring 264.
Cyclone Phailin *(phailin; cyclonephailin)*	60,016	Phailin was the second-strongest tropical cyclone ever to make landfall in India on October 11, 2013.
FIFA World Cup 2014 *(worldcup; fifaworldcup)*	67,406	20th edition of FIFA world cup, began on June 12, 2014. Germany beat Argentina in the final to win the tournament.
Unrest in Gaza *(gaza)*	31,302	Israel launched Operation Protective Edge in the Hamas-ruled Gaza Strip on July 8, 2014.
Heartbleed bug in OpenSSL *(heartbleed)*	8,362	Security bug in OpenSSL disclosed on April 1, 2014. About 17% of the world's web servers found to be at risk.
IPL 2013 *(ipl; ipl6; ipl2013)*	708,483	Edition 6 of IPL cricket tournament hosted in India, April-May 2013.
IPL 2014 *(ipl; ipl7)*	59,126	Edition 7 of IPL cricket tournament jointly hosted by United Arab Emirates and India, April-May 2013.
Lee Rigby's murder in Woolwich *(woolwich; londonattack)*	86,083	British soldier Lee Rigby attacked and murdered by Michael Adebolajo and Michael Adebowale in Woolwich, London on May 22, 2013.
Malaysian Airlines Flight MH17 shot down *(mh17)*	27,624	Malaysia Airlines Flight 17 crashed on 17 July 2014, presumed to have been shot down, killing all 298 on board.
Metro-North Train Derailment *(bronx derailment; metro north derailment; metronorth)*	1,165	A Metro-North Railroad Hudson Line passenger train derailed near the Spuyten Duyvil station in the New York City borough of the Bronx on December 1, 2013. Four killed, 59 injured.
Washington Navy Yard Shootings *(washington navy yard; navy yard shooting; NavyYardShooting)*	4,562	Lone gunman Aaron Alexis killed 12 and injured 3 in a mass shooting at the Naval Sea Systems Command (NAVSEA) headquarters inside the Washington Navy Yard in Washington, D.C. on Sept. 16, 2013.
Death of Nelson Mandela *(nelson; mandela; nelsonmandela; madiba)*	1,319,745	Nelson Mandela, the first elected President of South Africa, died on December 5, 2013. He was 95.
Birth of the fist Royal Baby *(RoyalBabyWatch; kate middleton; royalbaby)*	90,096	Prince George of Cambridge, first son of Prince William, and Catherine (Kate Middleton), was born on July 22, 2013.
Typhoon Haiyan *(haiyan; yolanda; typhoon philippines)*	486,325	Typhoon Haiyan (Yolanda), one of the strongest tropical cyclones ever recorded, devastated parts of Southeast Asia on Nov. 8, 2013.
T20 Cricket World Cup *(wt20; wt2014)*	25,209	Fifth ICC World Twenty20 cricket competition, hosted in Bangladesh during March-April, 2014. Sri Lanka won the tournament.
Wimbledon Tennis 2014 *(wimbledon)*	2,633	128th Wimbledon Tennis championship held between June 23, and July 6, 2014. Novak Djokovic from Serbia won the championship.

Table 2. Category labels and descriptions returned by WOT API. Source: WOT API Wiki (https://www.mywot.com/wiki/API).

Category	Description
Negative	Malware, viruses, poor customer experience, phishing, scam, potentially illegal, adult content
Questionable	Misleading claims, unethical, privacy risks, suspicious, hate, discrimination, spam, potentially unwanted programs, ads, pop-ups, incidental nudity, gruesome / shocking

1,278 *pages* made up our dataset of benign *pages*. The remaining *pages* were either deleted, migrated, or had been merged with other *pages*. Table 3 shows the descriptive statistics of all Facebook *pages* in our dataset.

Table 4 provides a detailed description of the number of posts and *pages* along with their WOT categories in our dataset of 627 malicious *pages*. We found a total of 20,999 posts which contained one or more malicious URLs. These posts engaged a total of 675,162 unique users who *liked*, *commented*, or *shared* these posts. Interestingly, we found that spam and phishing (two of the most common types of malicious content studied in literature) were least common in our dataset. Child unsafe content was the most common, followed by untrustworthy content.

Table 3. 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)[8]	3,447,669	31,680,263
Comments (recent 100 posts)	354,502	1,245,959
Shares (recent 100 posts)	507,964	1,012,151

We understand that our sample size of 627 malicious *pages* is not a large dataset as compared to some of the other studies done on OSNs. However, gathering Facebook data is a challenging task now. To the best of our knowledge, our dataset of 4.4 million public Facebook posts (from which we identified 627 malicious *pages*) is one of the biggest samples of Facebook data studied in literature, in terms of user generated content. The only dataset of Facebook posts (user gen-

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

Table 4. Number of malicious posts and *pages* in each category in our dataset. Number of *pages* posting phishing and spam URLs was the lowest.

WOT Response	No. of *pages*	No. of posts
Child unsafe	387	10,891
Untrustworthy	317	8,057
Questionable	312	8,859
Negative	266	5,863
Adult content	162	3,290
Spam	124	4,985
Phishing	39	495
Total	627	20,999

erated content) larger than ours was collected by Gao et al. [17]. This dataset was gathered in 2009 by performing large scale crawls on 8 regional Facebook networks over 3 months. Authors gathered 187 million posts which originated from roughly 3.5 million users (almost equal to the 3.3 million users + *pages* in our dataset). In contrast, we gathered all our data through authenticated requests made to the Graph API over a much larger time span of 16 months. All other studies on user generated content on Facebook have used much smaller datasets [2, 33, 38]. We discussed these studies in more detail in Section 2.

4 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*. We present our findings in detail in this section.

4.1 Spatial behavior

Most OSNs can be divided into three basic components that make up the social network; *the entity* (user / *page*), *the content* it posts, and its *network* (friends / followers / subscribers). We study all these three components separately.

Entities We performed term-frequency analysis on unigrams, bigrams, and trigrams obtained from *page* names in our dataset to identify the most prominent entities generating malicious and benign content. Table 5 lists the top 30 unigrams appearing in *page* names in our dataset after removing English stopwords. Manual analysis revealed dominant presence of politically polarized entities and religious groups with keywords like *american, british, english, league, patriot, defense, etc.* in malicious *pages*. Bigram and trigram analysis confirmed wide presence of entities like *British National Party (BNP)*, *The Tea Party*, *English Defense League*, *American Defense League*, *American conservatives*, *Geert Wilders supporters*, etc. We also found some malicious *pages* dedicated to pop

bands (One Direction), radio channels (Kiss FM), *pages* using *anonymous* in their names, etc. We manually inspected all the aforementioned *pages* and validated that the *page* names were aligned with the content they published, and were not misleading.

Table 5. Word frequency of the top 30 terms appearing in page names in our dataset. We found substantial presence of politically polarized entities among malicious pages.

Malicious page names				Benign page names			
Keyword	Freq.	Keyword	Freq.	Keyword	Freq.	Keyword	Freq.
news	11	group	5	church	20	county	8
league	11	one	5	center	15	one	8
defense	10	world	5	llc	14	services	8
online	8	videos	5	love	14	fans	8
american	8	national	5	photography	14	south	8
party	8	cricket	5	inc.	13	national	8
english	8	new	4	news	12	life	7
free	7	network	4	united	12	get	7
media	7	bnp	4	school	11	arts	7
truth	7	division	4	team	11	confessions	7
british	6	says	4	community	10	world	7
direction	6	club	4	club	9	health	7
edl	6	tea	4	st.	9	fire	7
forum	5	patriot	4	page	9	dr.	6
radio	5	united	4	cricket	9	beauty	6

Facebook has been shown to play a significant role in the political context, especially during elections [18, 41, 4]. Researchers have found that knowledge gained by youngsters from Facebook about electoral candidates influenced their evaluation of the candidates [11]. Such impactful role of Facebook on the users prompted us to study and understand the sentiment and emotion of content generated by politically polarized entities in our dataset.

Using the bigram and trigram analysis, we divided *pages* belonging to politically polarized entities into four broad groups based on *page* names to help us study them better. These groups were i) America (9 *pages*), containing *pages* with "america" or "american" in their *page* name, ii) British National Party (7 *pages*), containing *pages* mentioning British National Party or BNP in their *page* name, iii) Conservative (6 *pages*), containing *pages* with the term "conservative" in the *page* name, and iv) Defence League (11 *pages*), containing *pages* using the phrase "defence league" in the *page* name. We manually verified each *page* to ensure that they fit in the group they were assigned. To maintain anonymity, we do not reveal the exact *page* names. We performed linguistic analysis on the content published by these 4 categories of *pages* separately using LIWC2007 [32]. LIWC is a text analysis software to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. It determines the rate at which certain cognitions and emotions (for example,

personal concerns like religion, death, and positive or negative emotions) are present in the text. LIWC has been widely used to study social media content related to politics [37, 39, 40].

We focused our analysis on 12 dimensions in order to profile the linguistic patterns of content published by these groups of *pages*: Positive emotion, negative emotion, anxiety, anger, sadness, money, religion, death, sexuality, past orientation, future orientation, and swear words. Figure 2 shows the results of our analysis. We found high degree of anger in content from all categories. We also observed that only one category of *pages* (British National Party) had more positive emotions that negative emotions. The Defence League *pages* had much higher negative emotions as compared to positive emotions, followed by America *pages*. Conservative *pages* were almost equal in terms of positive and negative emotions. These findings contradicted prior results where researchers found that positive emotions outweighed negative emotions by 2 to 1 for profiles of all German political candidates [40]. We also found substantial presence of content related to religion. These observations are indicative of the kind of influence that politically polarized *pages* in our dataset can have on their audience.

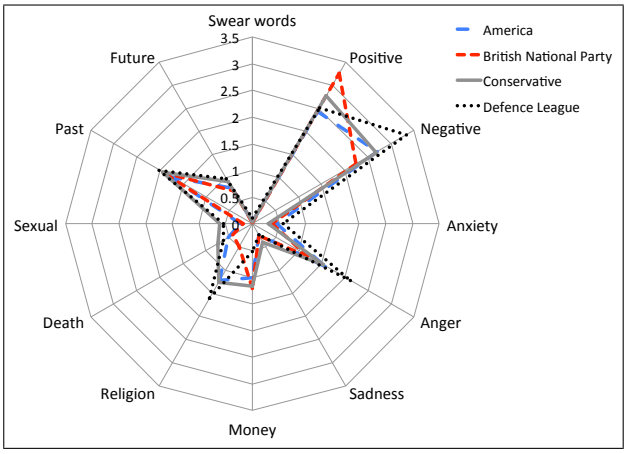


Fig. 2. Linguistic analysis of content produced by politically polarized groups of pages in our dataset. We found high presence of negative emotion, anger, and religion related content.

Benign *page* names were found to represent a variety of categories and interests like *photography, school, love, news, confessions*, etc. Bigram and trigram analysis revealed presence of a set of *methodist church pages*. We also found some overlap between malicious and benign *page* names, for example, *One Direction* fan *pages*, and radio channel *pages*. Unlike malicious *pages*, we did not find any fixed category dominating in benign *pages*.

From the above findings, it is evident that politically polarized entities that exist in the real world, also have a strong online presence. These results can be used to identify such entities on other social networks as well, and control (if not eliminate) the spread of polar political views online.

Content Scanning the most recent 100 posts (Section 3.3) revealed that almost half of 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. Figure 3 shows the distribution of the number of malicious posts versus the total number of domains shared by all malicious *pages* in our dataset. 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).

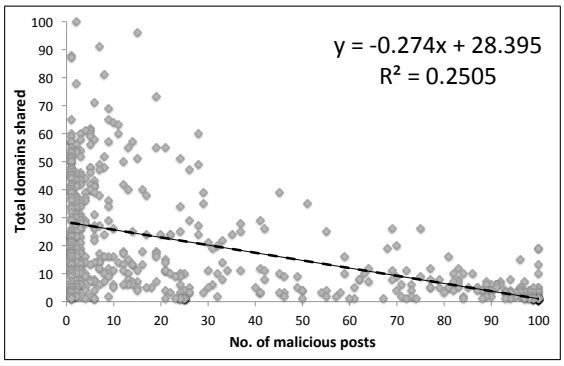


Fig. 3. Number of malicious posts versus all domains published by all 627 *pages* in our dataset. We observed a weak declining trend in the total number of domains when the number of malicious posts published by a *page* increased. Three outliers (sharing 1,257, 202 and 140 domains) have been removed in this graph.

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, 84 *pages* in our dataset (13.39%) shared URLs from only 1 domain. Out of these 84 *pages*, 53 *pages* (8.45%) published more than 90 posts containing a malicious URL, gathering *likes* and *comments* from 55,171 and 31,390 distinct users respectively. 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 closely resembles social spam campaigns, which have been studied by multiple researchers [17, 45]. 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, FIFA World Cup *page* (exclusively posting fifa.com URLs), BBC News *page* (exclusively posting bbc.com URLs), etc. 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. We explore this possibility, and report our findings in Section 5.2.

Top domains: Table 6 lists the 10 most frequently occurring domains in our dataset of malicious *pages*, along with their WOT classification, Facebook audience, and Alexa world ranking[9]. For each domain, we calculated the number of posts the domain appeared in, the sum of *likes*, *comments*, and *shares* on all these posts, the number of *pages* the domain appeared in, and the sum of *likes* on all these *pages*. It was interesting to observe that 3 out of the top 10 domains were very famous, and were ranked within the top 3,000 domains worldwide on the Alexa ranking. Two of these domains were reported for being unsafe for children and spreading adult content. Although the Internet does not restrict the creation and promotion of adult and child unsafe content, most OSNs including Facebook have established community standards which restrict the display of adult and explicit content [13]. All of the other domains had low Alexa ranking worldwide. Only 3 of the top 10 domains were marked as spam, and none of the domains in the top 10 were reported for phishing or malware, signifying that untrustworthy and child unsafe content is much more prominent on Facebook than traditional forms of malicious content like spam and phishing.

The number of posts and *pages* associated with each of the top 10 domains revealed that there existed multiple Facebook *pages* dedicated to promoting all of these domains. We observed that all of the top 10 domains appeared in 2 or more *pages*, and two of the domains appeared in over 10 *pages* (ridichegratis.com in 21 *pages*; 9cric.com in 13 *pages*). At least 4 of the top 10 domains (ammboi.com, ghanafilla.net, pulsd.com, and 970wfla.com) had two or more Facebook *pages* each (3 for ghanafilla.net, 5 for ammboi.com), heavily promoting their respective domains (over 90 out of the 100 posts containing the domain, per *page*). Pages set up for these domains also had a substantial audience, with 6 out of the 10 domains collectively having over 100,000 *likes* on their *pages*. Two of the top 10 domains had over 1 million *likes* (collectively) on *pages* promoting them. The collective number of *likes*, *comments*, and *shares* on posts was however, considerably low as compared to collective *likes* on the *pages*. Only 3 out of the top 10 domains managed 1,000 or more *likes* on the posts associated with them. This indicated that while malicious domains in our dataset were successful in gathering a substantial audience in the form of *page likes*, they were not as

[9] http://www.alexa.com/

Table 6. Top 10 malicious domains in our dataset with their Web of Trust classification, Facebook audience, and Alexa world rank.

Domain	WOT class, categories	Posts	Likes — comments — shares	Pages	Page likes	Alexa rank
ammboi.com	Untrustworthy, suspicious, spam, privacy risks	456	666 — 61 — 195	5	109,012	352,191
ridichegratis.com	Untrustworthy	424	428 — 14 — 252	21	2,650,802	-
blesk.cz	Child unsafe, adult content	402	3,674 — 2,103 — 1,494	8	864,554	2,924
says.com	Child unsafe	386	387 — 15 — 62	5	97,784	27,684
ghanafilla.net	Untrustworthy, scam, spam, suspicious	296	192 — 8 — 6	3	54,246	1,360,634
9cric.com	Child unsafe	281	1,189 — 121 — 177	13	193,348	923,243
perezhilton.com	Child unsafe, adult content	274	26,088 — 3,516 — 1,128	8	1,701,834	2,192
nairaland.com	Untrustworthy, misleading claims or unethical	201	238 — 89 — 31	3	116,131	1,329
pulsd.com	Untrustworthy, child unsafe	199	2 — 0 — 0	2	19,020	247,480
970wfla.com	Spam	194	700 — 448 — 280	2	22,486	277,467

successful in engaging the audience with their content. We also observed that 2 of the 3 domains with high Alexa rank (blesk.cz and perezhilton.com) also had high number of *page likes* and high number of *likes*, *comments*, and *shares* on posts. This signified that domains which were popular (more visited) on the Internet were also more famous on Facebook.

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 [10]. Such attacks have also been studied in context of OSNs [44]. To investigate if such a phenomenon exists for Facebook *pages* too, we analyzed the *like, comment*, and *share* networks for both malicious and benign *pages* in our dataset. Facebook does not provide an API endpoint to gather the list of users who have *liked* (subscribed to) a *page*. However, it is possible to collect the list of users who have *liked, commented* on, or *shared* posts published by a *page*. As described in Section 3.3, we collected all *likes, comments*, and *shares* on the most recent 100 posts of all *pages* in our dataset, and analyzed the inter and intra-*page* networks. In particular,

we analyzed networks consisting of *pages* and users *liking, commenting on,* or *sharing* posts from two or more *pages* in our dataset (malicious and benign separately) (inter-*page* networks), and networks of *pages liking, commenting on,* or *sharing* posts from *pages* within our dataset (malicious and benign separately) (intra-*page* networks). 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.

Table 7 shows the details of the network analysis. We found that the Inter-page *likes* network for benign *pages* (83,799 nodes) was much larger and stronger (avg. weighted degree: 41.695) than the Inter-page *likes* network for malicious *pages* (21,947 nodes, avg. weighted degree: 24.273), indicating that a larger number of users *liked* two or more benign *pages* as compared to the number of users who *liked* two or more malicious *pages* in our dataset. More interestingly, we found stronger ties (avg. weighted degree for Intra-*page* networks) 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 communities 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 communities we detected from Intra-*page likes, comments*, and *shares* networks. Figure 5 shows the network graphs of the detected communities. We observed that post *sharing* was the most prominent intra-*page* activity, followed by *liking* and *commenting*. The network graphs also revealed a distinct community of six Facebook *pages* completely connected to each other in terms of *likes* (Figure 5(a)) and *shares* (Figure 5(c)). Five out of these six *pages* also formed a community in the intra-page comments graph (Fig 5(b)). We manually inspected and observed that all *pages* in this community 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* communities involving politically polarized *pages*, where one *page* heavily engaged in *liking, commenting on*, and *sharing* the other *page*'s content.

Metadata Analyzing the metadata of posts in our dataset revealed some significant differences in the type of content published by malicious and benign *pages*. Figure 4 shows the distribution of the content type of posts published by all *pages* in our dataset. We observed that more than half of the content published by benign *pages* were photos and videos (50.16%). This percentage went down to 32.42% for malicious *pages*. The metadata also revealed that over half of the posts published by malicious *pages* were links (54.69%), where as less than a quarter of all posts published by benign *pages* were links (24.45%). These numbers indicate that malicious *pages* are inclined towards posting links, and directing user traffic to external websites. On the other hand, benign *pages*

tend to post more pictures, which can be consumed by users without leaving the OSN. In addition to content types, we looked at the status types of posts and found that benign *pages* published almost double the amount of content through mobile devices (23.80%) as compared to malicious *pages* (12.33%).

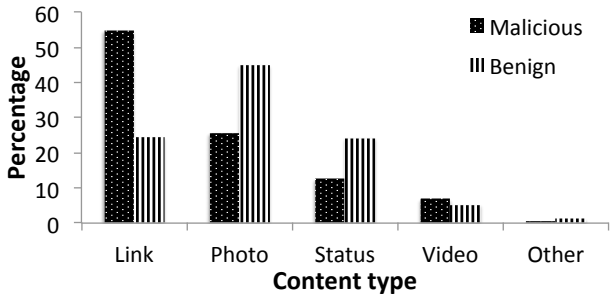


Fig. 4. Types of content published by malicious and benign pages in our dataset. Malicious pages published more links, while benign pages published more pictures.

All *pages* on Facebook have a *category* associated with them, for example *Community, Company, Personal Website*, etc. This category is assigned to the *page* by the *page* administrator(s) at the time of *page* creation, according to the person / organization represented by the *page*, and content that the *page* generates. To see if any subset of categories was more popular among a particular class of *pages* (malicious or benign), we compared category ranks and found strong correlation between category ranks across malicious and benign *pages* (Spearman's $\rho = 0.67$, $p\text{-}value<0.01$). This indicated that the distribution of malicious and benign *pages* across various categories was fairly similar, and that categories more popular among malicious *pages* were also more popular among benign *pages*. We also compared the *page likes* and *page* mentions (*talking_about_count* field) of malicious and benign *pages*, and did not find any significant differences.

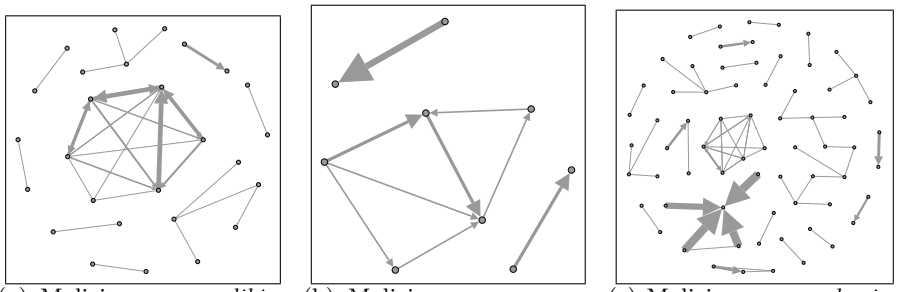
These observations indicated that apart from the type and source of published content, there were no significant differences in the meta information between malicious and benign *pages* in our dataset. Metrics like popularity (*likes*) and user mentions (*talking_about_count*) associated with OSN entities can be used to identify spammers, since they capture the notion of influence of entities in the network [6]. However, similarities in such metrics across malicious and benign *pages* can aid malicious *pages* to continue operating regularly and go undetected for long periods of time, hiding in plain sight.

4.2 Temporal behavior

We explored the temporal posting activity of all *pages* in our dataset to determine how active the *pages* were, in terms of publishing posts. We also monitored the

Table 7. Network analysis of *likes*, *comments* and *shares* networks within and between *pages* in our dataset. We observed that malicious *pages* had stronger intra-network ties as compared to benign *pages*.

Network type	Total nodes	Total edges	Avg. weighted degree	Density	No. of communities
Malicious (All 627 *pages*)					
Inter-*page likes* network	21,947	103,683	24.273	0	18
Inter-*page comments* network	3,901	13,957	11.255	0.001	19
Inter-*page shares* network	14,318	67,513	15.796	0	14
Intra-*page likes* network	27	35	8.333	0.05	9
Intra-*page comments* network	9	9	1.667	0.125	3
Intra-*page shares* network	68	65	6.309	0.014	21
Benign (Results averaged across 10 random samples of 627 benign *pages* each)					
Inter-*page likes* network	83,799	390,854	41.695	0	3070
Inter-*page comments* network	2,958	7,722	8.919	0.001	142
Inter-*page shares* network	3,406	10,234	9.920	0.001	30
Intra-*page likes* network	4.3	3.6	0.408	0.075	0.7
Intra-*page comments* network	0	0	0	0	0
Intra-*page shares* network	7.8	6.9	1.168	0.072	1.1



(a) Malicious *pages liking* each others' posts; intra-page likes network.

(b) Malicious *pages commenting* on each others' posts; intra-page comments network.

(c) Malicious *pages sharing* each others' posts; intra-page shares network.

Fig. 5. Network graphs capturing intra-*page* activity of malicious *pages* in our dataset. We found multiple two-node communities and a few bigger communities.

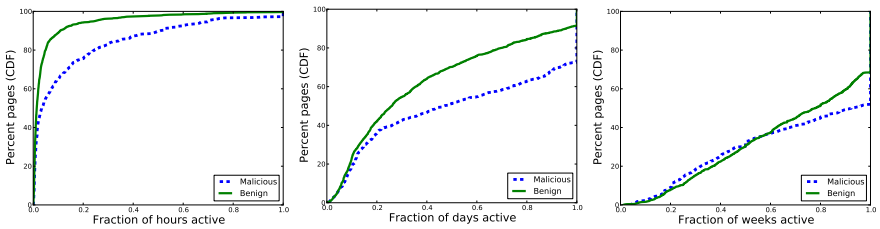
status of these *pages* daily, for a period of over one year to observe any changes in the *pages*' behavior and attributes over time.

Posting Activity 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*.

$$daily\ activity\ ratio = \frac{no.\ of\ days\ active}{no.\ of\ days\ between\ first\ and\ last\ post}$$

Figure 6(b) shows the *daily activity ratio* plots of all malicious and benign *pages* in our dataset. 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) [28]. These difference confirmed that malicious *pages* in our dataset were more active as compared to benign *pages*, and published more frequently.



(a) Hourly activity ratio. (b) Daily activity ratio. (c) Weekly activity ratio.

Fig. 6. Daily, hourly, and weekly temporal activity of *pages* in our dataset. We found that malicious *pages* were more active than benign *pages*.

Attributes over time We studied the temporal behavior of all *pages* in our dataset over the period of an year, between October 2015 and October 2016. During this period, we captured daily snapshots of the *page* information for all the *pages* through the Graph API. The aim of this study was to observe the change in attributes of malicious *pages* over time, and to identify if these changes were significantly different from the changes in attributes for benign *pages*. In particular, we studied changes in two types of attributes over time; a) popularity, and b) description.

Popularity To study the change in popularity over time, we computed a *gain factor* corresponding to the change in the number of *likes* on all *pages* in our dataset as follows:

$$gainFactor_P = \frac{likesOnDayLast_P - likesOnDayOne_P}{likesOnDayOne_P} \times 100$$

where likesOnDayLast$_P$ = no. of *likes* on *page* P on the last day (October 15, 2016), and likesOnDayOne$_P$ = no. of likes on *page* P on the first day (October 16, 2015) of the study. A positive value of the *gain factor* for a *page* P indicates an increase in the number of *likes*, while a negative value depicts a drop in the number of *likes* for P over the span of one year.

Figure 7 shows the *gain factor* between malicious and benign *pages* for all *pages* in our dataset. We observed that a larger proportion of malicious *pages* (28.54%) lost *likes* as compared to benign *pages* (20.26%). Contrarily, while computing the average *gain* over all *pages*, we found that malicious *pages* had a larger *gain factor* (32.52%) as compared to benign *pages* (24.03%). This difference, however, was statistically insignificant (*p-value*>0.1). Prior statistics show that the average growth rate for a Facebook *page* is 0.64% per week (approximately 33.28% per year) [25]. Interestingly, this number is much closer to malicious *pages* in our dataset. However, given the statistical insignificance of our results, we cannot conclude that the growth rate of malicious *pages* is more similar to an average Facebook *page* as compared to benign *pages* in our dataset.

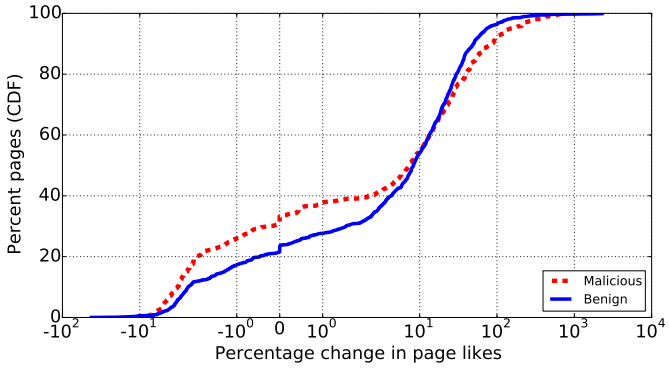


Fig. 7. Percentage change in *page likes* (*gain factor*) over one year for all *pages* in our dataset. We observed that a larger proportion of malicious *pages* lost likes over time as compared to benign *pages*.

We investigated the change in popularity over time in more detail, by computing the rate of change of the number of likes per day for all *pages* in our dataset, to see if there was a statistically significant difference between malicious and benign *pages* with respect to this metric. We modelled the growth rate of *likes* on a *page* as a linear function over time and studied the distribution of the gradient for *page likes* across malicious and benign classes. This technique has been used previously to study popularity on OSNs over time [7]. We observed low values for standard error of the estimated gradient, and significant *p-values* for both classes, signifying good fit (see Table 8).

Table 8. Mean values for standard error of estimated gradient and correlation p-values for linear model. We obtained low error rates and p-values signifying a good fit.

Metric	Class	Malicious	Benign
Standard error of estimated gradient	μ_{err}	0.8273	0.5583
	σ_{err}	4.6282	4.3492
p-value for correlation	μ_p	0.016	0.009
	σ_p	0.121	0.082

Figure 8 shows the distribution of the gradients ($\tan^{-1} m$, where $m = \frac{y-c}{x}$; y = no. of likes, x = days, c = intercept) we obtained for malicious and benign classes. We observed that gradients for malicious *pages* were more evenly distributed as compared to benign *pages*. The median gradient value for the malicious class (7.85) was lower than the median gradient value for the benign class (9.11), but the difference in the two distributions was statistically insignificant (p-*value*=0.38).

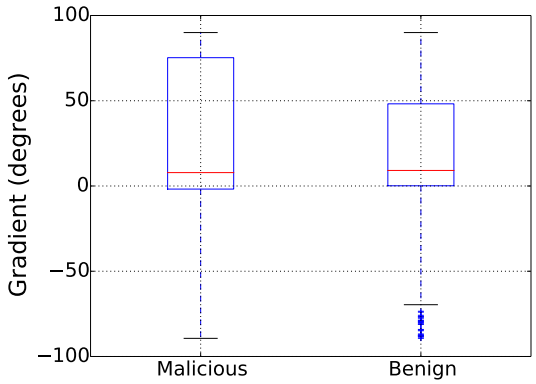


Fig. 8. Distribution of the popularity gradients (in degrees) for malicious and benign pages in our dataset. Although the distributions look different, we did not find the difference to be statistically significant.

Description Each Facebook *page* has multiple attributes that make up its description, for example, *username, description, general_info, personal_info, category, location, phone_number, mission, bio,* etc. While some attributes (like *username, category*) are available for all *pages*, the presence of other attributes (like *general_info, mission, bio, etc.*) is dependent on the category of the *page*. We examined changes in all such attributes (wherever available) for both, malicious and benign *pages* in our dataset. Figure 9 shows the top 20 attributes in

which we observed at least one change during the one year period of our study. We came across a total of 44 attributes that were changed once or more[10]. The remaining 24 attributes were changed by less than 1% of all *pages* in our dataset.

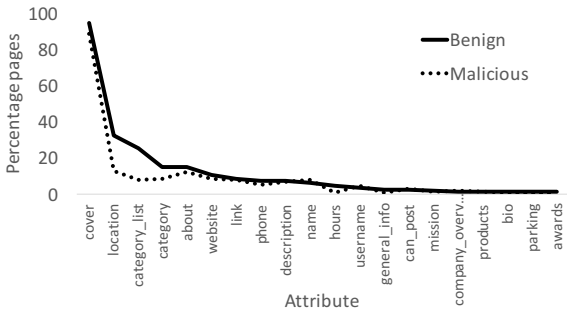


Fig. 9. Top 20 attributes across malicious and benign pages that were changed at least once during one year of observation. We identified a total of 44 such attributes, but all remaining attributes underwent one or more changes in less than 1% pages.

We observed a strong correlation between malicious and benign *pages* in terms of the proportion of *pages* changing each attribute, for all 44 attributes ($r = 0.967$, $p - value < 0.01$). This correlation depicted that an attribute that was changed by a large proportion of benign *pages*, was changed by a large percentage of malicious *pages* too, and vice versa. For example, the *cover* attribute (cover picture) was changed by 94.44% of benign *pages* and 88.51% of malicious *pages*, while *name* was changed by 6.25% of benign *pages* and 7.49% of malicious *pages*.

We further investigated each of the top 20 attributes individually to see if the changing behavior of any of these attributes could help distinguish between malicious and benign *pages*. We applied the Kolmogorov-Smirnov (KS) 2 sample test to compare the distributions of the number of times each of these attributes were changed by malicious and benign *pages* in our dataset, and found that 19 out of the 20 attributes were not informative (p-*value*>0.1). The only statistically significant distribution corresponded to changing behavior of the *category_list* attribute (p-*value*<0.05).

The above results corroborate with our previous findings, suggesting minimal presence of a significant difference between malicious and benign *pages*, even in terms of popularity and attribute change behavior over time. These findings further suggest that distinguishing between malicious and benign Facebook *pages* based on spatial characteristics, temporal behavior, and other information associated with these *pages* is a hard and challenging problem. Using all the aforementioned insights and observations obtained from our dataset, we construct a

[10] Exact description for each of these attributes can be found at https://developers.facebook.com/docs/graph-api/reference/page/

diverse and robust feature set, and attempt to automate the task of identifying malicious *pages* from benign *pages* using supervised learning, as described in the next section.

5 Automatic detection of malicious pages

Past research has shown that URL blacklists and reputation services are ineffective initially, and take time to update [36]. 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 multiple supervised learning algorithms on our dataset of malicious *pages* in an attempt to create an effective model for automatic detection of malicious Facebook *pages*, independent of third party reputation services.

Feature set: We extracted a total of 96 features, 55 features from *page* information, and 41 from the posts published by the *pages* in our dataset, to train and evaluate the aforementioned algorithms. Table 10 shows a list of all these features, along with their category and feature type.

Classification algorithms: We experimented with a variety of classification algorithms – Naive Bayesian, Logistic Regression, Decision Trees, Random Forests, and Artificial Neural Networks. We used balanced training and test sets containing equal numbers of positive and negative examples (627 malicious *pages*, and a random sample of 627 benign *pages* picked from the dataset of 1,278 benign *pages*), so random guessing results in an accuracy, as well as an area under the receiver operating characteristic (ROC) curve (AUC) of 50%. All models were validated using 10-fold cross validation. Although our actual dataset is highly unbalanced, we use a balanced dataset for our experiments in order the obtain a model for better classification of new data, as opposed to a model that would represent our dataset better.

We also performed experiments with unbalanced classes. Note that collecting posts along with their likes, comments, and shares (which we use as features for our analysis) is a time consuming and computationally expensive task. We therefore performed undersampling on our dataset of malicious *pages* as opposed to collecting more data for benign *pages*. We obtained unbalanced classes in the ratio 1:10 for malicious:benign classes by picking up 10 random samples, each of size 128 ($1/10^{th}$ of the total 1,278 benign *pages*) from the malicious class. For each such random sample, we performed 10-fold cross validation and averaged out the results across all samples and folds.

In addition, we trained and evaluated bag-of-words models obtained using the textual content present in the posts published by these *pages*. We used the most recent 100 posts published by Facebook *pages* in our dataset for calculating post features and building our bag-of-words. We did not find any explicit distinctive features in our dataset to separate the malicious class from benign, thus making effective automation a hard goal to achieve. We thus tried to build an extensive feature set to capture as much characteristics as possible.

5.1 Supervised learning with *page* and post features

Table 9 shows the accuracy and ROC AUC values for various classification algorithms that we applied on the *page* and post level features. We considered post features extracted from the most recent 100 posts generated by the *pages*. A combination of post and *page* level features performed the best, signifying that both the characteristics, and posting behavior of *pages* need to be recorded for efficient automatic detection of malicious *pages*. The Logistic Regression classifier achieved highest accuracy of 76.71% with an area under the ROC curve of 0.846.

The algorithms performed poorly under unbalanced setting. The most probable reason for this poor performance is the lack of data in the malicious class due to undersampling. We therefore preferred models trained on balanced classes.

Table 9. Classification accuracy and ROC AUC values for automatically detecting malicious Facebook *pages*. Logistic Regression classifier performed the best. We used a 1:10 split for malicious:benign classes for unbalanced classes.

		Balanced		Unbalanced	
Classifier	Feature set	Acc. (%)	ROC AUC	Acc. (%)	ROC AUC
Naive Bayesian	*Page*	63.95	0.685	54.37	0.670
	Post	69.61	0.753	77.29	0.741
	Page + post	70.81	0.776	70.76	0.764
Logistic Regression	*Page*	67.38	0.745	90.81	0.613
	Post	76.55	0.825	90.97	0.740
	Page + post	**76.71**	**0.846**	91.07	0.800
Decision Trees	*Page*	65.55	0.668	90.85	0.493
	Post	71.37	0.720	88.63	0.599
	Page + post	70.81	0.758	90.11	0.660
Random Forest	*Page*	67.86	0.750	90.92	0.745
	Post	74.95	0.829	91.41	0.832
	Page + post	75.27	0.837	91.17	0.856

We performed further experiments by varying the number of most recent posts we considered for generating post features. Figure 10 shows the ROC AUC values achieved by the Logistic Regression classifier with varying post history. We started the experiment by considering 20 most recent posts for post features, and observed an overall increasing trend in performance as we increased the number of most recent posts to 100. We did not go beyond 100 for the post history because our ground truth dataset for malicious and benign *pages* was derived based on this limit.

The classifier achieved a maximum ROC AUC value of 0.85 (and an accuracy of 77.67%) with a post history size of 80 posts using a combination of *page* and post features. Performance remained unchanged with respect to *page* features,

Table 10. Page and post level features used for training supervised learning models.

Category	Feature type	Feature
Page (55)	Boolean (19)	Affiliation, birthday, can post, cover picture, current location, working hours, description present, location, city, street, state, zip, country, latitude, longitude, personal interests, phone number, public transit, website field
	Numeric (34)	Average sentence length for description, average word length for description, parking capacity, category list length, check-ins, no. of email IDs in description, fraction of HTTP URLs in description, description length, fraction of URLs shortened, fraction of URLs active, likes, page name length, no. of subdomains in URLs, path length of URLs, no. of redirects in URLs, no. of parameters in URLs, [no. of !, no. of ?, no. of alphabets, no. of emoticons, no. of English stop words, no. of English words, no. of lower case characters, no. of upper case characters, no. of newline characters, no. of words, no. of unique words, no. of sentences, no. of total characters, no. of digits, no. of URLs] in description, description repetition factor, talking-about count, were-here count
	Nominal (2)	Category, description language
Posts (41)	Numeric (41)	Daily activity ratio, audience engaged, [average no. of upper case characters, average length, average word length, no. of English words, no. of English stop words] for description, message, and name fields, no. of posts containing the field [description, message, name], no. of comments, no. of likes, no. of shares, no. of posts with status_type [added_photos, added_video, created_event, created_note, mobile_status_update, published_story, shared_story, wall post], no. of posts with type [event, link, music, note, offer, photo, video, status], total no. of URLs, total no. of unique domains

since change in the size of post history does not affect *page* features (and is thus not reported in the figure).

5.2 Supervised learning with bag-of-words

In addition to *page* and post level features, we used a term frequency based bag-of-words model to automatically identify malicious Facebook *pages*. 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)[11]. We performed experiments by calculating term frequencies of unigrams, bigrams, and trigrams, and limited our vocabulary size to the top 10,000 features.

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

[11] https://developers.facebook.com/docs/graph-api/reference/v2.6/post

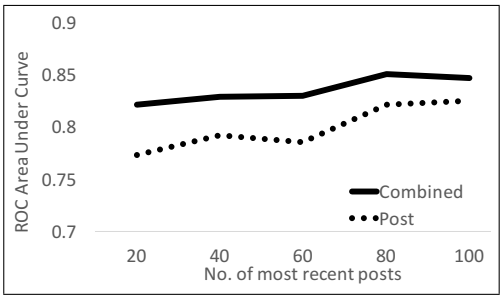


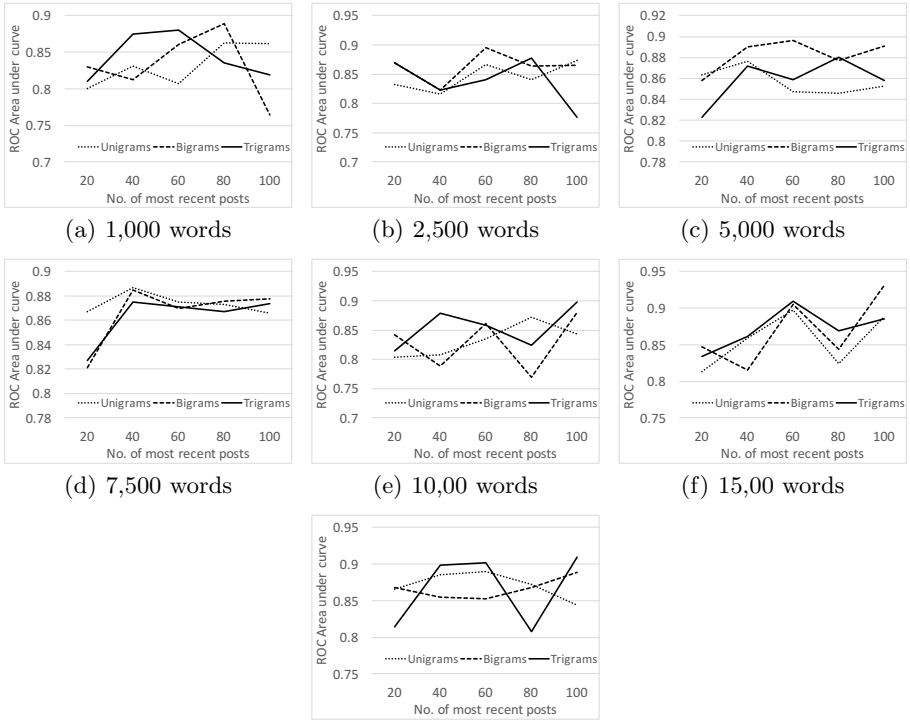
Fig. 10. ROC area under curve values for Logistic Regression classifier corresponding to different sizes of post history. We observe an overall increase in performance as we increase the number of most recent posts used for computing post features.

fast and effective classification. We chose Sparsenn for this task[12]. 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 11 presents the results of our experiments.

Table 11. Classification accuracy and ROC AUC values for automatically detecting malicious Facebook pages using bag-of-words. Artificial neural networks performed the best. We used a 1:10 split for malicious:benign classes for unbalanced classes.

		Balanced		Unbalanced	
Classifier	Feature set	Acc. (%)	ROC AUC	Acc. (%)	ROC AUC
Naive Bayesian	Unigrams	68.27	0.682	87.10	0.571
	Bigrams	69.06	0.690	88.13	0.611
	Trigrams	69.77	0.697	88.48	0.609
Logistic Regression	Unigrams	74.18	0.795	90.47	0.763
	Bigrams	74.34	0.791	90.47	0.767
	Trigrams	73.93	0.789	90.48	0.782
Decision Trees	Unigrams	68.12	0.678	87.71	0.641
	Bigrams	67.05	0.678	88.41	0.655
	Trigrams	66.63	0.672	88.64	0.658
Random Forest	Unigrams	72.26	0.794	91.64	0.773
	Bigrams	71.80	0.802	91.63	0.778
	Trigrams	72.18	0.794	91.63	0.766
Neural Networks	Unigrams	81.74	0.862	92.41	0.812
	Bigrams	84.12	0.872	92.26	0.734
	Trigrams	**84.13**	**0.900**	92.97	0.826

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



(a) 1,000 words (b) 2,500 words (c) 5,000 words

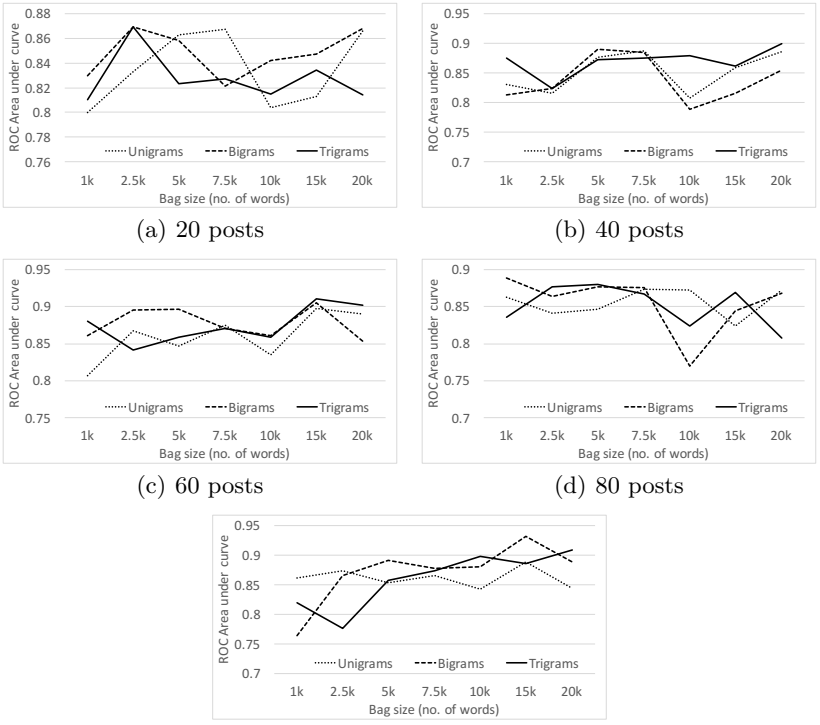
(d) 7,500 words (e) 10,00 words (f) 15,00 words

(g) 20,00 words

Fig. 11. ROC AUC values obtained by neural networks trained on a bag of words for different sizes of bag of words.

Under balanced settings, Neural networks (hidden units = 64, and learning rate = 0.07, determined experimentally) on trigrams performed the best, achieving an accuracy of 84.13% with an area under the ROC curve of 0.9. This signified that artificial neural networks trained on the top 10,000 trigrams outperformed all the other learning models including our previous models trained on *page* and post level features (discussed in Section 5.1).

We extended our experiments by performing a grid search over post history (number of most recent posts) and bag of words size. Using default values for hidden units (16) and learning rate (0.05), we varied the size of the bag of words from 1,000 through 20,000, and post history from 20 through 100 most recent posts published by the *page*. All these experiments were performed using unigrams, bigrams and trigrams. Figure 11 and Figure 12 show the varying values of area under the ROC curve for different sizes of post history and bag of words respectively. We achieved a maximum ROC AUC value of 0.931 using bigrams with a bag size of 15,000 words and post history size of 100.



(a) 20 posts (b) 40 posts

(c) 60 posts (d) 80 posts

(e) 100 posts

Fig. 12. ROC AUC values obtained by neural networks trained on a bag of words for different sizes of post history.

6 Discussion and limitations

We now discuss the implications of our findings.

Politically polarized entities: Our analysis revealed the presence of some politically polarized entities in our dataset of malicious *pages*. We verified the actual presence of all such entities in our dataset through manual verification. Interestingly, the presence of such politically polarized entities was obvious despite the fact that none of the events we analyzed, had any direct connection with politics, for example, elections. One potential reason for this phenomenon could be a large section of users with polar ideologies marking websites promoting / supporting opposing political views as untrustworthy on Web of Trust.

Entities involved in politics tend to be followed by masses with similar orientation, and is a global phenomenon in the real world. It is likely that such activity exists on online platforms other than Facebook too. We do not propose to debar such activity. However, we believe that extremely polar content should be moderated both online and offline, in order to maintain stability among the masses. An easy way to moderate such entities can be to display nudges or

warning messages to users before they subscribe to such *pages* on any online platform [42].

Beyond *pages*: *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. Moreover, Facebook has a common definition of "*Page* Spam" for *pages*, groups and events, and explicitly states that *Pages, groups or events that confuse, mislead, surprise or defraud people on Facebook are considered abusive.* Our analysis and results can thus be easily extended to study malicious groups and events as well.

Automatic identification of malicious *pages*: Our findings shed some light on subtle differences (like temporal behavior, content type, etc.) between malicious and benign *pages*, which we used to train various supervised learning algorithms to automatically differentiate between malicious and benign *pages*. These findings, however, are 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.

Comparison with past work: Most past work in the space of detecting malicious content on Facebook focuses on individual posts rather than the entities generating these posts (users, *pages*, etc.). This made it non trivial to compare the performance of our techniques with techniques proposed previously. Further due to the scarcity of publicly available datasets for Facebook data, we were not able to cross evaluate our models. To this end, we intend to anonymize our dataset and make it publicly available for research purposes in the near future.

7 Conclusion and Future work

In this chapter, we identified and characterized Facebook *pages* posting malicious URLs. We looked beyond traditional types of malicious content like unsolicited bulk messages, spam, phishing, malware, etc., and studied a broader section of content that is deemed as malicious by community standards and *Page* Spam definitions established by Facebook. We focused on Facebook *pages* because of their public nature, vast audience, and inflated malicious activity [9]. Our observations revealed presence of politically polarized entities among malicious *pages*. We also found a substantial number of malicious *pages* dedicated to promote content from a single malicious domain. Further, we observed that malicious *pages* were more active than benign *pages* in terms of hourly, daily, and weekly activity. Network analysis revealed presence of collusive behavior among malicious *pages* that engaged heavily in promoting each others' content. We applied multiple machine learning algorithms on our dataset to automate the detection of

malicious *pages*, using *page* and post level features, and bag-of-words. Our experiments showed that artificial neural networks trained on bag-of-words work best in detecting malicious *pages* automatically. We believe that our findings will enable researchers to better understand the landscape of malicious Facebook *pages* that have been hiding in plain sight and promoting malicious content seemingly unperturbed.

In future, we would like to expand our analysis to identify malicious "groups" and "events" on Facebook which largely remain unexplored.

8 Acknowledgements

We would like to thank all the members of Precog Research Group and Cybersecurity Education and Research Centre (CERC) at IIIT Delhi for their constant support and feedback for this work.

References

1. Aggarwal, A., Rajadesingan, A., Kumaraguru, P.: Phishari: Automatic realtime phishing detection on twitter. In: eCrime Researchers Summit (eCrime), 2012. pp. 1–12. IEEE (2012)
2. Ahmed, F., Abulaish, M.: An mcl-based approach for spam profile detection in online social networks. In: IEEE TrustCom. pp. 602–608. IEEE (2012)
3. Akoglu, L., Chandy, R., Faloutsos, C.: Opinion fraud detection in online reviews by network effects. ICWSM 13, 2–11 (2013)
4. Carlisle, J.E., Patton, R.C.: Is social media changing how we understand political engagement? an analysis of facebook and the 2008 presidential election. Political Research Quarterly 66(4), 883–895 (2013)
5. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: WWW. pp. 675–684. ACM (2011)
6. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: The million follower fallacy. ICWSM 10, 10–17 (2010)
7. De Choudhury, M., Monroy-Hernandez, A., Mark, G.: Narco emotions: affect and desensitization in social media during the mexican drug war. In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems. pp. 3563–3572. ACM (2014)
8. Dewan, P., Bagroy, S., Kumaraguru, P.: Hiding in plain sight: Characterizing and detecting malicious facebook pages. In: Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on. pp. 193–196. IEEE (2016)
9. Dewan, P., Kumaraguru, P.: Towards automatic real time identification of malicious posts on facebook. In: Privacy, Security and Trust (PST), 2015 13th Annual Conference on. pp. 85–92. IEEE (2015)
10. Douceur, J.R.: The sybil attack. In: Peer-to-peer Systems, pp. 251–260. Springer (2002)
11. Douglas, S., Maruyama, M., Semaan, B., Robertson, S.P.: Politics and young adults: The effects of facebook on candidate evaluation. In: Proceedings of the 15th Annual International Conference on Digital Government Research. pp. 196–204. dg.o '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2612733.2612754

12. Facebook: What is page spam? https://www.facebook.com/help/116053525145846 (2015)
13. Facebook.com: Facebook community standards. https://www.facebook.com/communitystandards (2015)
14. Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting burstiness in reviews for review spammer detection. ICWSM 13, 175–184 (2013)
15. Friggeri, A., Adamic, L.A., Eckles, D., Cheng, J.: Rumor cascades. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (2014)
16. Gao, H., Chen, Y., Lee, K., Palsetia, D., Choudhary, A.N.: Towards online spam filtering in social networks. In: NDSS (2012)
17. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B.Y.: Detecting and characterizing social spam campaigns. In: Internet Measurement Conference. pp. 35–47. ACM (2010)
18. Guardian, T.: Facebook's failure: did fake news and polarized politics get trump elected? https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories (2016)
19. Gupta, A., Kumaraguru, P.: Credibility ranking of tweets during high impact events. In: PSOSM. ACM (2012)
20. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: Tweetcred: Real-time credibility assessment of content on twitter. In: Social Informatics, pp. 228–243. Springer (2014)
21. Gupta, M., Zhao, P., Han, J.: Evaluating event credibility on twitter. In: SDM. pp. 153–164. SIAM (2012)
22. Jiang, M., Cui, P., Beutel, A., Faloutsos, C., Yang, S.: Catching synchronized behaviors in large networks: A graph mining approach. ACM Transactions on Knowledge Discovery from Data (TKDD) (2016)
23. Jiang, M., Cui, P., Faloutsos, C.: Suspicious behavior detection: Current trends and future directions. IEEE Intelligent Systems 31, 31–39 (2016)
24. Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. pp. 219–230. ACM (2008)
25. Karma, F.: Study: Average growth of facebook fan pages. http://blog.fanpagekarma.com/2013/03/20/infographic-average-growths-facebook-fan-pages/ (2013)
26. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 435–442. ACM (2010)
27. Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., Lauw, H.W.: Detecting product review spammers using rating behaviors. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 939–948. ACM (2010)
28. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics pp. 50–60 (1947)
29. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: Can we trust what we rt? In: Proceedings of the first workshop on social media analytics. pp. 71–79. ACM (2010)
30. Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., Ghosh, R.: Spotting opinion spammers using behavioral footprints. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 632–640. ACM (2013)

31. Mukherjee, A., Liu, B., Glance, N.: Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st international conference on World Wide Web. pp. 191–200. ACM (2012)
32. Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., Booth, R.J.: The development and psychometric properties of liwc2007 (2007)
33. Rahman, M.S., Huang, T.K., Madhyastha, H.V., Faloutsos, M.: Efficient and scalable socware detection in online social networks. In: USENIX Security Symposium. pp. 663–678 (2012)
34. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., Menczer, F.: Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th international conference companion on World wide web. pp. 249–252. ACM (2011)
35. Ratkiewicz, J., Conover, M., Meiss, M.R., Gonçalves, B., Flammini, A., Menczer, F.: Detecting and tracking political abuse in social media. ICWSM 11, 297–304 (2011)
36. Sheng, S., Wardman, B., Warner, G., Cranor, L., Hong, J., Zhang, C.: An empirical analysis of phishing blacklists. In: Sixth Conference on Email and Anti-Spam (CEAS) (2009)
37. Stieglitz, S., Dang-Xuan, L.: Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior. In: System Science (HICSS), 2012 45th Hawaii International Conference on. pp. 3500–3509. IEEE (2012)
38. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: ACSAC. pp. 1–9. ACM (2010)
39. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Election forecasts with twitter: How 140 characters reflect the political landscape. Social Science Computer Review p. 0894439310386557 (2010)
40. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of the fourth international AAAI conference on weblogs and social media. pp. 178–185 (2010)
41. Vitak, J., Zube, P., Smock, A., Carr, C.T., Ellison, N., Lampe, C.: It's complicated: Facebook users' political participation in the 2008 election. CyberPsychology, behavior, and social networking 14(3), 107–114 (2011)
42. Wang, Y., Leon, P.G., Scott, K., Chen, X., Acquisti, A., Cranor, L.F.: Privacy nudges for social media: an exploratory facebook study. In: Proceedings of the 22nd international conference on World Wide Web companion. pp. 763–770. International World Wide Web Conferences Steering Committee (2013)
43. WOT: Web of trust api. https://www.mywot.com/en/api (2014)
44. Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B.Y., Dai, Y.: Uncovering social network sybils in the wild. ACM Transactions on Knowledge Discovery from Data (TKDD) 8(1), 2 (2014)
45. Zhang, X., Zhu, S., Liang, W.: Detecting spam and promoting campaigns in the twitter social network. In: Data Mining (ICDM), 2012 IEEE 12th International Conference on. pp. 1194–1199. IEEE (2012)