Angel or Demon? Characterizing Variations Across Twitter Timeline of Technical Support Campaigners

Srishti Gupta[1], Gurpreet Bhatia[1], Saksham Suri[1], Dhruv Kuchhal[1], Payas Gupta[2], Mustaque Ahamad[3], Manish Gupta[4] and Ponnurangam Kumaraguru[1]

[1] *Indraprastha Institute of Information Technology, Delhi*

[2] *Pindrop Technologies, Atlanta*

[3] *Georgia Institute of Technology, Atlanta*

[4] *Microsoft, India*

ABSTRACT

Technical Support spam, which abuse Web 2.0 and carry out social engineering attacks have been in existence for a very long time, despite several measures taken to thwart such attacks. Although recent research has looked into unveiling tactics employed by spammers to lure victims, damage done on Online Social Networks is largely unexplored. In this paper, we perform the first large-scale study to understand the behavior of technical support spammers, and compare them with the legitimate technical support offered to OSN users by several brands such as Microsoft, Facebook, Amazon.

We analyze the spam and legitimate accounts over a period of 20 months, and provide a taxonomy of the different types of spammers that are active in Tech Support spam landscape. We develop an automated mechanism to classify spammers from legitimate accounts, achieving a precision, recall of 99.8%. Our results shed light on the threats associated with billions of users using OSNs from Tech Support spam, and can help researchers and OSN service providers in developing effective countermeasures to fight them.

ISSN 2332-4031; DOI 10.34962/jws-67

1 Introduction

The increasing popularity of Online Social Networks (OSNs) has attracted a cadre of criminals who craft large-scale phishing and spam campaigns targeted against OSN users. Traditionally, spammers have been driving traffic to their websites by luring users to click on URLs in their posts on OSNs Grier *et al.*, 2010; Gao *et al.*, 2010; Thomas *et al.*, 2011b. A significant fraction of OSN spam research has looked at solutions driven by URL blacklists Gao *et al.*, 2010; Thomas *et al.*, 2011a, manual classification Benevenuto *et al.*, 2009, and honeypots Lee *et al.*, 2010; Stringhini *et al.*, 2010. Since defence mechanisms against malicious / spam URLs have already matured, cybercriminals are looking for other ways to engage with users. Telephony has become a cost-effective medium for such engagement, and phone numbers are now being used to drive call traffic to spammer operated resources (e.g., call centers and Over-The-Top applications like WhatsApp).

In this paper, we explore a data-driven approach to understand OSN abuse that makes use of phone numbers as action tokens in the realization / monetization phase of spam campaigns. Internet crime reports suggest that people fell victim to phone scams leading to a loss of $7.4B in 2015 for Americans alone [1], which further suggests that telephony has turned out to be an effective tool for spammers. Specifically, in the phone-based abuse of OSNs, spammers advertise phone numbers under their control via OSN posts and lure OSN users into calling these numbers. Since spammers use phone calls to trap victims, it is safe to assume that spammers would provide real phone numbers under their control. In addition, advertising phone numbers reduce spammers' overhead of finding the set of potential victims which can be targeted via the phone. Over phone conversations, they try convincing the victims that their services are genuine, and deceive them into making payments Miramirkhani *et al.*, 2017. To maximize their reach and impact, we observe that spammers disseminate similar content across multiple OSNs.

While URLs help spammers attract victims to websites that host malicious content, phone numbers provide more leverage to spammers. Due to the inherent trust associated with the telephony medium and the impact of human touch over phone calls, spammers using phone numbers stand a better chance of convincing and hence are likely to make more impact. Besides, they can use fewer phone numbers as compared to URLs; a large number of URLs are required to evade filtering mechanisms incorporated by OSNs. [2] Moreover, the monetization and advertising channel in phone-based campaigns i.e., (Phone) and (Web) respectively is different as compared to a single channel (Web) used in URL-based campaigns. Hence, phone-based spam requires correlation of abuse information across channels which makes it harder for OSN service providers to build effective solutions. Finally, since the modus operandi in URL-based and phone-based spam campaigns is different,

[1] https://blog.truecaller.com/2017/04/19/truecaller-us-spam-report-2017/

[2] https://support.twitter.com/articles/90491

leaving phone-based spams unexplored can limit OSN service providers' ability to defend their users from spam. While extensive solutions have been built to educate users about URL-based spam Kumaraguru *et al.*, 2009, limited education is available for phone-based attacks. This is evident from several well publicized and long running Tech Support spam campaigns (since 2008) that use phone numbers to lure victims leading to huge financial losses in the past, as reported by the Federal Bureau of Investigation. Although detecting and avoiding OSN abuse using phone numbers is more critical now than ever, to the best of our knowledge, this space is largely unexplored.

In this paper, we address this gap by *identifying* and *characterizing* spam campaigns that abuse phone numbers across multiple OSNs. Studying phone-based spam across multiple OSNs provides a new perspective and helps in understanding how spammers work in coordination to increase their impact. From 22M posts collected from Twitter, Facebook, GooglePlus, YouTube, and Flickr, we identify 202 campaigns running across different countries, leveraging 806 unique abusive phone numbers. Finally, we study how legitimate and spam tech support campaigns and their associated accounts differ by analyzing the accounts used by spammers that have been identified over a period of 20 months on Twitter. We discuss the difference in the way legitimate and malicious accounts operate on Twitter, in addition to the way phone numbers are being used. Studying these campaigns, we make the following key observations:

1. We find that the cross-platform phone based spam campaigns originate from more than 16 countries, but most of them come from Indonesia, United States of America (USA), India, and United Arab Emirates (UAE). These campaigns are supported by less number of phone numbers as compared to URLs, perhaps due to (a) the high cost of acquiring a phone number, and (b) weak defense mechanisms against phone - based spam. Victims that fall prey to these campaigns are offered banned filmography, personal products and a variety of other services; but the services are not delivered even after successful payment.

2. As reported in earlier research Ghosh *et al.*, 2012, we also find evidence that suggests spammers collude to maximize their reach either by creating multiple accounts or promoting other spammers' content. To evade suspension strategies of each OSN, spammers keep the volume per account low. Our results show that accounts are suspended after being active for 33 days (on average); while literature suggests that spammers involved in URL-based spam campaigns, on the other hand, could survive only for three days after their first post Thomas *et al.*, 2011b. In addition, 68.7% of spammer accounts are never suspended; it suggests a crucial need to build effective solutions to combat phone-based spam.

3. Our analysis also suggests that OSN service providers should work together in the fight against phone-based spam campaigns. By examining phone numbers involved in campaigns across OSNs, we find that although all OSNs are consistently being abused, Twitter is the most preferred OSN for propagating a phone campaign. By analyzing spammers' multiple identities across OSNs, we find that Twitter is able to suspend 93.3% more accounts than Facebook. *cross-platform intelligence* can be useful in preventing the onset and reducing the lifetime of a campaign on a particular network with good accuracy. We estimate that cross-platform intelligence can help protect 35,407 victims across OSNs, resulting in potential savings of $8.8M.

4. We discuss the difference in the way legitimate and malicious accounts operate on Twitter, in addition to the way phone numbers are being used. We design an automated classification system based on machine learning, and apply multiple features to classify tech support scam campaigns. Our experimental evaluation demonstrates the efficacy of the proposed classification system achieving 99.8% precision and recall.

Altogether, our results shed light on phone-based spam campaigns where spammers are using one channel (OSN) to spread their content, and the other channel (voice / SMS / message via phone) to convince their victims to fall prey to their campaigns. Given that no timely and effective filters exist on either channel to combat such spam, there is an imperative need to build one.

2 Related Work

Spam is a growing problem for OSNs, and several researchers have looked at different ways to combat it. In this section, we present prior research in detecting spam campaigns.

Handling non-phone based spam: There has been a large body of work that reports the existence of spam on multiple OSNs like YouTube Benevenuto *et al.*, 2009, Twitter Grier *et al.*, 2010, and Facebook Gao *et al.*, 2010. Thomas et al. studied the characteristics of suspended accounts on Twitter Thomas *et al.*, 2011b. With an in-depth analysis of several spam campaigns, they reported that 77% spam accounts suspended by Twitter were taken down on the day of their first tweet. Apart from this, there has been work done to differentiate a spammer from a non-spammer Yardi *et al.*, 2009; Benevenuto *et al.*, 2010; Wang, 2010; Lee *et al.*, 2011; Amleshwaram *et al.*, 2013. Lumezanu et al. studied the spread of URL campaigns on email and Twitter and found that spam domains receive better coverage when they appear both on Twitter and email Lumezanu and Feamster, 2012. In addition to characterizing URL-based spam, methods have been proposed for detecting Lee *et al.*, 2010; Webb *et al.*, 2008; Chu *et al.*, 2012 and preventing Rahman *et al.*, 2012; Faloutsos, 2013 such campaigns. While a lot of work has been done on characterizing and detecting URL-based spam campaigns, campaigns abusing phone numbers have been largely ignored.

Handling phone based spam: A large fraction of phone spam includes robocalling and spoofing, wherein spammers call the victims and trick them into giving personal or financial information. [3] Studies have shown that, in spam activities,

[3] https://www.consumer.ftc.gov/articles/0076-phone-scams

phone numbers are more stable over time than email, and hence can be more helpful in identifying spammers Costin *et al.*, 2013; Isacenkova *et al.*, 2014. Christin et al. analyzed a type of scam targeting Japanese users, threatening to reveal the users' browsing history, in case they do not give them money Christin *et al.*, 2010. In studies mentioned above, the authors relied on publicly available datasets to perform their analyses. In contrast, we develop an infrastructure to collect millions of posts from OSNs, cluster them into campaigns, and conduct our analyses. Researchers have investigated phone number abuse by analyzing cross-application features in Over-The-Top applications Gupta *et al.*, 2016, cross-channel SMS abuse Srinivasan *et al.*, 2016, characterizing spam campaigns on Twitter Gupta *et al.*, 2018, and by characterizing honeypot numbers Gupta *et al.*, 2015; Gupta *et al.*, 2014; Balduzzi *et al.*, 2016; Marzuoli *et al.*, 2016. Recently, Miramirkhani et al. studied the Tech Support campaign that abuse phone numbers, from the perspective of domains that were used to host malicious content Miramirkhani *et al.*, 2017. The authors also interacted with spammers to understand their social engineering tactics. While they focused on URLs and domains abused by spammers, we study the cross-platform spread of phone-based spam campaigns across OSNs, along with strategies adopted by spammers for sustainability and visibility. Besides, we highlight how cross-platform intelligence about spam accounts can be shared across OSNs to aid in spam detection.

3 Dataset

In this section, we discuss our methodology for collecting phone numbers, posts and other metadata; which we use later to find campaigns on OSNs. These campaigns are then tagged as benign or spam. Figure 1 shows the architecture of our data collection subsystem that is used to collect phone numbers across multiple OSNs. We picked Twitter as the starting point to find phone numbers, as it provides easier access to large amounts of data as compared to other online social networks Osborne and Dredze, 2014. We set up a framework to collect a stream of tweets containing phone numbers. For each unique phone number received every day, a query was made to other OSNs viz. Facebook, [4] GooglePlus, Flickr, and YouTube, and for every search, we stored the following details: user details (user ID, screen name, number of followers and friends), post details (time of publication, text, URL, number of retweets, likes, shares, and reactions), and whether the ID were suspended. The data collection ran over a period of six months, between April 25, 2016 and October 26, 2016. Our system collected 22,690,601 posts containing 1,845,150 unique phone numbers, posted by 3,365,017 unique user accounts on five different OSNs. After removing noise (i.e., the posts which do not contain a phone number), the filtered set was used for finding campaigns.

We acknowledge that our dataset may contain two kinds of bias: (1) Only 1% sample of all public tweets is available from the Twitter Streaming API; it can underestimate the spam campaigns observed on Twitter. (2) Since we treat Twitter as the starting point, we may miss some campaigns which are popular on other social networks, but not on Twitter. However, Twitter provides best access to user posts, justifying our choice.

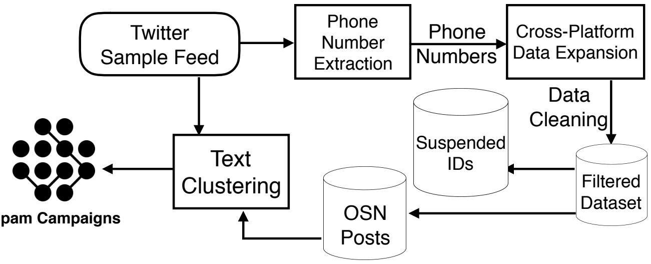


Figure 1: System Architecture for Data Collection across Multiple OSNs.

Campaigns: A *campaign* is defined as a collection of posts made by a set of users sharing similar text and phone numbers. To make sure that we do not tag any benign campaign as spam, we filtered out the phone numbers used by even one Twitter verified account. Every phone number, say *ph1*, is represented by a set of frequent unigram tokens which occur around the phone number. All posts that contain at-least 33% tokens from the representative token set are put together in a cluster; indicating posts related to the phone number. Different phone numbers, say *ph1* and *ph2*, are put together in the same cluster if the average Jaccard coefficient between the corresponding set of posts is greater than 0.7. We calculated different values of Jaccard coefficient and average silhouette scores to measure quality of clusters Almeida *et al.*, 2011, and found 0.7 as knee point for corresponding value of silhouette score as 0.8. All users that post about any phone number in the clustered set are put together. A cluster thus formed is marked as a campaign. Using this method, we found 22,390 campaigns in the dataset, collectively amounting to ∼10.9M posts.

Spam Campaigns: We flag a campaign as *spam* if it meets the following criteria: (a) phone number involved in the campaign is present in the United States Federal Trade Commission's Do Not Call (DNC) dataset [6], or (b) even if one OSN account involved in the campaign is suspended. Further, to be able to characterize the spam campaigns in detail, we focused only on campaigns with at least 5000 posts. With this, we identified 6,171 out of 22,390 campaigns as spam. From this set of campaigns, we did a manual inspection to verify if the campaign is indeed spam. This results in a working dataset of *202 campaigns* comprising of *∼4.9M posts.* During manual inspection, we also assigned topics to the 202 campaigns, where multiple campaigns could be assigned the same topic. For instance, a campaign selling shoes and other selling jackets would be assigned the topic – "Product Marketing".

[4] Collecting data from Facebook was challenging. In April 2015, Facebook deprecated their post-search API end-point [5], so we used an Android mobile OAuth token to search content using the Graph API Gupta *et al.*, 2016.

[6] https://www.ftc.gov/site-information/open-government/data-sets/do-not-call-data

4 Characterizing Spam Campaigns

In this section, we focus on the following research questions. Where do spam campaigns originate from? Do spammers use automation when posting phone numbers or answering "phone calls"? What does a spammer OSN account suspension depend on? What is the typical modus operandi of the spammers?

4.1 Where does Phone Spam Originate?

It is important to know from which countries does the spam originate; it can be used in developing anti-spam filtering solution. We assume that the country associated with a phone number is the source country. For the analysis, we need to extract the country of the spam phone number. This is done either by identifying (a) the language of the post containing the spam phone number via the 'lang' field in the tweet object, or (b) by the country code using Google's phone number library. [7] These two methods helped in identifying countries for 127 campaigns. For rest of the campaigns, we called up the top two frequently occurring phone numbers in the campaign using Tropo [8], a VoIP software that can be used to make spoofed calls. We recorded all the calls and used Google's Speech API [9] to detect language and country of the campaign. We could identify origin country for 26 more campaigns; for the remaining 49, the country is unknown. Table 1 presents topic distribution across various campaigns originating from different countries along with the average number of posts being made in each campaign. While majority of the spam was similar to advanced-fee scam, where spammers trick victims to make payments in advance, there were certain different type of campaigns observed in the dataset as well: Hacking (Tech Support) and Alternating Beliefs (Love Guru). In the *Love-Guru* campaign, astrologers promise victims to fix their love and marriage related problems. In the *Tech Support* campaign, spammers pose as technical support representatives or claim to be associated with big technological companies (like Amazon, Google, Microsoft, Quebec, Norton, Yahoo, Mcafee, Dell, HP, Apple, Adobe, TrendMicro, and Comcast) and offer technical support fixes.

Top four source countries selected by the volume of campaigns viz. Indonesia, United States of America (USA), India, and United Arab Emirates (UAE) show interesting characteristics (see Figure 2).

Table 1: Distribution of Campaigns across Topics and Source Countries. (#C denotes number of campaigns).

Country	Campaign Topics	#C	#Posts
Argentina	Party Reservations	1	39,476
	Pornography	1	30,751
Chile	Delivering Goods	1	6,691
Columbia	Hotel Booking	1	18,228
	Pornography	1	5,324
India	Hotel Booking	1	10,986
	Alternating Beliefs (Marriage)	1	15,128
	Hacking(Tech Support)	1	43,552
Indonesia	Hotel Booking	1	8,291
	Product Marketing	75	2,689,616
	Pornography	4	164,382
	Alternating Beliefs (Marriage))	7	101,799
	Purchasing Followers	15	406,713
	Finance, Real Estate	3	23,700
	Selling Adult Products	5	48,109
	Uncategorized	3	29,043
Nigeria	Alternating Beliefs (Marriage)	1	29,226
Pakistan	Finance, Real Estate	1	16,058
UAE	Escorts	5	69,263
USA	Party Reservations	8	172,090
	Product Marketing	1	22,804
	Pornography	1	19,653
	Alternating Beliefs (Marriage)	1	12,936
	Escorts	1	9,652

4.2 Do Spammers use Automation?

While investigating further, we found that 99.3% pairs of consecutive posts related to the same campaign appeared on Twitter in less than 10 minutes. Given that a major fraction of content appeared within a few minutes, it is likely that content generation is automated. To ascertain this, we looked at the information of the client (provided by the Twitter API) used by spammers to interact with the Twitter API or their web portal. We found that most of the content was generated using 'twittbot.net', a popular bot service, known to be used by spammers Thomas *et al.*, 2011b. Apart from the bot service, several other clients like RoundTeam (0.25%), IFFTT (0.03%), Buffer (0.017%), and Botize (0.016%), were used for Twitter. Besides, we found that volume per phone number was also high in Indonesian campaigns; 80% phone numbers had more than 1000 posts. One would assume that volume per phone number would be low since there are humans at the other end to service the requests. However, by processing the text in the posts created in this campaign, we found that spammers requested users to communicate via SMS or WhatsApp ($\sim$ 71% posts). This explains why spammers would be able to handle the load of interacting with victims. There are many other advantages of using these messaging services – spammers can further send phishing messages to victims, communicate with them unmonitored, and potentially use automated bots to reply to SMSs or Whatsapp messages.

4.3 What Governs Spammers' Suspension?

As expected, we find that the visibility (number of likes, shares, and retweets) of a post is positively correlated with the number of posts (Pearson correlation coefficient = 0.97). While this may sound intuitive, the number of accounts that were suspended within a campaign were not positively correlated with the number of posts. We noticed that even though the volume generated by Indonesian campaigns was 98.2% higher than Indian campaigns, the fraction of users suspended in Indian campaigns was 85.6% higher. Further, we observed that the account suspension is dependent on the nature of cam-

[7] https://github.com/googlei18n/libphonenumber

[8] https://www.tropo.com/

[9] https://cloud.google.com/speech/

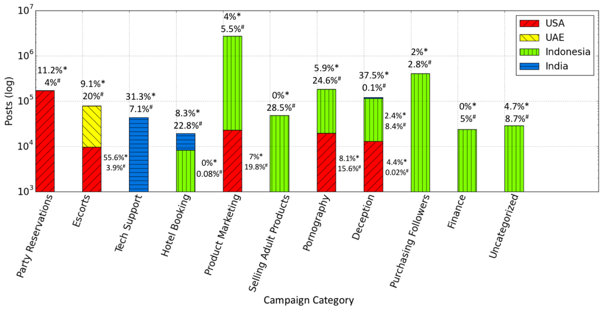
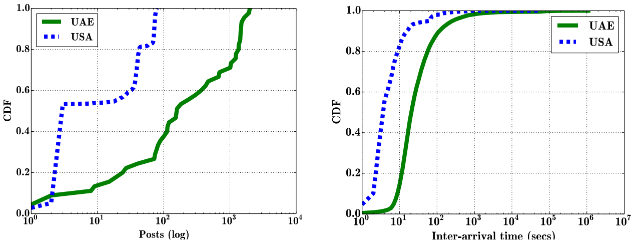


Figure 2: Comparison of campaigns running in the top 4 countries – Indonesia, USA, India, and UAE across different campaign categories. While visibility that a post receives is positively correlated with volume, account suspension in a campaign is not. Escort service and Tech Support campaigns had largest percentage of suspended accounts. The number of users suspended is represented by * and # denotes the fraction of posts getting visibility.



(a)

(b)

Figure 3: Comparing Escort service campaign in USA vs. UAE. Even though volume generated per USA account is lower than UAE accounts (a), inter-arrival time between two consecutive posts in the USA is smaller which could be a potential reason for suspension of accounts (b).

paigns; campaigns providing escort services or technical support services had more accounts suspended.

Surprisingly, for similar escort service campaign running in two different countries, USA and UAE, there was a significant difference in the number of accounts suspended. Before concluding that the country plays a major role in account suspension, we performed detailed analysis as follows.

The number of posts generated by escort campaign running in the USA (9,652) was lower than that running in UAE (69,263), but 55.6% user accounts were suspended in the USA in comparison to only 9.1% accounts suspended in UAE. We looked at several reasons which could potentially lead to account suspension – volume generated per user or URLs used in the posts. We noticed that volume per user was higher for UAE users (Figure 3(a)), number of URLs shared in UAE campaign was higher, and words used in both the campaigns had a good overlap. Also, from Figure 3(b), we observed that inter-arrival time between two consecutive posts made by all the users in the USA (41s on an average) is smaller than that of posts made in the UAE campaign (392s on an average).

4.4 What is the Modus Operandi?

To ascertain the attack methodology the victims faced, we performed an experiment after receiving our institute's Institutional Review Board (IRB) approval. Pretending to be a potential victim, we called up phone numbers mentioned in campaigns selling adult (Viagra) pills in USA and UAE. In Indonesia, we interacted with spammers selling herbal products, and in India with those promoting tech support and astrology services (providing solutions to marriage and love problems). To avoid time zone conflict, we called the spammers in their local time of the day. Overall, we made 41 calls to different phone numbers from Indonesia, India, USA and UAE. Apart from Indonesia, campaigns from other countries had an IVR deployed, before reaching a spammer. We posit this can help in load balancing between limited human resources on the spammers' end. Due to language limitation in Indonesia, spammers preferred chatting over platforms like WhatsApp, where they were extremely responsive.

The campaigns in USA and UAE were not limited by any delivery location; they had a usual delivery time of 2–4 weeks. These campaigns were operating solely over the phone and had no option of visiting an online portal to make the transaction. The attackers confidently asked for the credit card details over the phone even though banks advise otherwise. Spammers from Indonesia told that they would start delivery only after receiving the payment, which was to be done via bank transfer. During the interactions, spammers were persuasive in selling products by claiming their products to be the best as compared to similar products in the market. Tech support campaigns in India were providing service to users remotely over the Internet and charged over call once the issue was 'fixed'. The catch was that the spammers pretended that there was a problem with the victims' computer and then tried to convince the victim to pay them to fix it, as reported in several complaints [10]. Another astrology based spam campaign running in India tricked by promising to fix users' marriage and love related problems within 48 hours [11]. We called 4 numbers in different Indian states. Interestingly, all the spammers had a similar way of dealing with the problem, where they asked to send personal details over WhatsApp. It is evident that spammers running campaigns in different countries deploy similar mechanisms to let the victim reach them (posts on social media), to set up the product / service delivery operation (product delivery post payment and service delivery prior to payment), and model of payment (details transfer via phone, WhatsApp, verbal). It is the product delivery operation that creates deliberate confusion for a victim; intuitively, the delivery mechanism is similar for benign campaigns. Spammers leverage the advantage of similar delivery mechanisms, offer fake promises and later do not deliver.

Sample of Transcribed Calls with Spammers

IVR: Press 1 to know about our products, 2 to check the status of previous order and 3 for other inquiries

[10] https://800notes.com/Phone.aspx/1-800-549-5301/2

[11] https://www.complaintboard.in/complaints-reviews/vashikaran-fake-vashikaran-fraud-cheater-money-taker-l149781.html

Victim: *pressed 1*

IVR: Press 1 to know more about <company-name> viagra pills and 2 for other products.

Victim: *pressed 1*

IVR: *call forwarding to human*

Scammer: Hello, I'm <name>, speaking from <company-name>, what would you like to know about the <brand> viagra pills.

Victim: What are the various packs I can buy and how much does it cost?

Scammer: We have only one variant which costs $99 - $119 for the pills and $20 for delivery.

Victim: Okay. How can I pay for the order if I decide to order? Do you have a web portal where I can make an online transaction?

Scammer: No sir, currently, we're operating only over phone, so you can provide your VISA card details to me, and I'll be happy to place the order for you.

Victim: Is phone the only option? I would like to make the payment through the web portal.

Scammer: Sorry sir, but we operate only over phone.

Victim: Okay, what are the product guarantees you offer?

Scammer: Yes, sir please be assured that we provide 100% return guarantee.

Victim: Can I get some samples before placing the order?

Scammer: I am sorry sir, we don't provide samples. Should I place an order for you?

Victim - No, thank you for the information.

5 Characterizing Cross-Platform Spam Campaigns

In this section, we aim to answer the following research questions. Are spam campaigns run in a cross-OSN manner? How does the content cross-pollinate across OSNs? How do spammers maximize visibility? To what extent OSNs are able to detect phone based spam? Can existing intelligence on URL based spam be trivially adapted to handle the growing phone based spam problem?

5.1 Do Phone-based Spam Campaigns run in a Cross-OSN Manner?

We observed that spam campaigns do not limit themselves to one OSN and are rather present on multiple networks. The distribution of posts across platforms in top 3 spam campaigns: Loveguru (from Alternating Beliefs category), Tech Support, and Indonesian Herbal Product (from Product Marketing category) is shown in Table 2. Even though Twitter has the largest fraction (possibly thanks to the first data source bias in our data collection method), all OSNs are abused to carry out spam campaigns.

Table 2: Top Cross-Platform Spam Campaigns

Campaign	TW	FB	G+	YT	FL
Tech Support	28,984	2,151	7,830	2,850	1,737
LoveGuru	6,934	1,418	4,257	101	63
Indonesia Herbal Product	1,443,619	9,238	21	46	336

Due to lack of space, in this section, we focus on studying in detail the Tech Support campaign. The details for other campaigns are available at *http://bit.ly/phcamp-dash*. Tech support scams have been around for a long period [12],incurring financial losses of $2.2M to victims in 2016 alone, as reported by the US Federal Bureau of Investigation (FBI). Earlier, attackers used to call victims offering to fix their computer or PC. Now, attackers have changed their strategy; instead of calling victims, attackers float their phone numbers on OSNs and ask users to call them in case they need any technical assistance related to their computers. Once the victim calls the phone number, the attacker asks for remote access to their machine to diagnose the problem. The attacker fudges the expected problems with victim's machine and convinces her to get it fixed. The reason this campaign is identified as spam, is because attackers deceive in believing that there exists some problem with their PC and charge money in return. Previous work has focused on the methods used by attackers to convince the victim and to make money Miramirkhani *et al.*, 2017. In this paper, we are interested in looking at the cross-platform behavior of such tech support scam campaigns. Over the course of six months of data collection, we got a total of 43,552 posts spread across all the five OSNs propagating to the extent of 41 phone numbers. Table 3 shows the complete dataset description for tech support campaigns.

Table 3: Statistics for Tech Support Campaign

Features	TW	FB	G+	YT	FL
Total Posts	28,984	2,151	7,830	2,850	1,737
Posts with URLs	25,245	1,391	5,714	227	1,503
Distinct Phone Numbers	41	33	37	39	20
Distinct User IDs	748	289	360	433	79
Distinct Posts	16,142	1,797	6,570	2,050	1,449
Distinct URLs	68	951	3,189	80	293

As phone numbers are one of the primary tokens used by spammers, we examined carrier information tied to each number to identify what kind of phone numbers spammers use viz. landline, mobile, VoIP, or toll-free). We derived this information from several online services like Twilio (mobile carrier information) [13], Truecaller (spam score assigned to the phone number) [14], and HLR lookups (current active location of the phone number). [15] We found that all the phone numbers used in the Tech Support campaign were toll-free numbers. Using a toll-free number offers several advantages to a spammer: (1) increased credibility: it does not incur a cost to the person calling, hence people perceive it to be legitimate, (2) it provides international presence: spammers can be reached from any part of the world. Further, we found that spammers used services like ATL, Bandwidth, and, Wiltel Communications to obtain these toll-free numbers and that a majority of them were registered between 2014 and 2016.

5.2 How does Content Cross-pollinate?

Now, we answer the following question: *Is a particular OSN preferred to start the spread of a campaign? Is there a specific pattern in the way spam propagates on different OSNs?*

Figure 4(a) shows the temporal pattern of content across OSNs. Note that our data collection was done over a period of six months while a campaign may have existed before and / or after this period. Hence, while the longest detected active time for a campaign in our dataset is 186 days, the actual time may be greater.

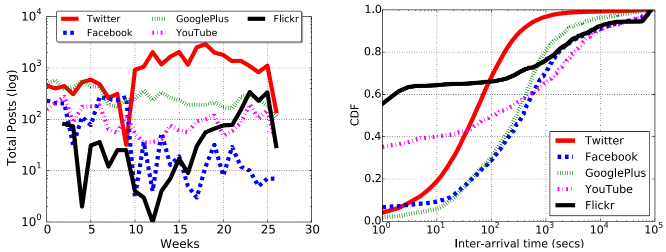
A majority of these posts are densely packed into a small number of short time bursts, while the entire campaign spans a much

[12] https://blog.malwarebytes.com/tech-support-scams/

[13] https://www.twilio.com/

[14] http://truecaller.com/

[15] https://www.hlr-lookups.com/



(a) Posts across OSNs

(b) Inter-arrival Time of Posts appearing on OSNs

Figure 4: Temporal properties of Tech Support Campaign across OSNs – all OSNs are abused to spread the campaign but volume is maximal on Twitter. Inter-arrival time between two consecutive posts is minimal for Twitter. Spammers began to heavily abuse Flickr towards the end of our data collection.

longer period. Though the volume of content is significantly higher on Twitter, all OSNs are consistently being abused for propagation. Inter-arrival time, i.e., the average time between two successive posts is observed to be least on Twitter (308s), as shown in Figure 4(b). It is interesting to note that a few campaigns on Flickr have an inter-arrival time between two posts close to 1s, even though the average inter-arrival time is highest on Flickr. As Figure 4(a) shows, the volume on Flickr increased during the last few weeks of our data collection period. We divided the inter-arrival time into two time windows; first 15 weeks, and last 11 weeks. We observed that the average inter-arrival time in latter time window dropped from 9786s to 2543s which means spammers had started heavily abusing Flickr to spread the Tech Support campaign. It is hard to ascertain the motivation of the spammers in sending high volume content on Twitter, but, we speculate one of the reasons could be the public nature of the Twitter platform, as compared to closed OSNs like Facebook. For all the phone numbers, we analyzed the appearance of phone numbers on different OSNs, and the order in which they appear, as reported in Table 4.

Table 4: Distribution of phone numbers according to their first appearance amongst OSNs. Flickr is never chosen as a starting point and there is no particular sequence in which spam propagates across OSNs.

Starting OSN	#Cases	Most common sequence
Twitter (TW)	12	TW → G+ → YT
GooglePlus (G+)	10	G+ → TW → YT → FB → FL
Facebook (FB)	6	FB → G+ → TW → YT
YouTube (YT)	13	YT → G+ → TW → FB

For each network that is picked as the starting point, we identified the most common sequence in which phone numbers appeared subsequently on other OSNs. We found that Flickr was *never* chosen as the starting OSN to initiate the spread of a phone number. Further, we noticed that the posts originating from YouTube took the maximum time to reach a different OSN with an average inter-OSN time of 5 hours.

To summarize, we observed that all OSNs were abused to spread the Tech Support campaign, and no particular OSN was preferred to drive the campaign. In addition, there was no particular sequence in which spam propagated across OSNs.

5.3 How to Maximize Visibility?

We observed various strategies adopted by spammers to increase the dissemination of their posts. In this section, we discuss those strategies and their effectiveness.

The *Visibility* of a post is defined as the action performed by the user (consumer of the post) in terms of liking or sharing the post, which accounts for traction a particular post received. For each network, we define the value of visibility as follows: number of likes and reshares on Facebook, +1s and reshares on GooglePlus, number of likes and retweets on Twitter, and video like count on YouTube. We did not consider Flickr in our analysis since Flickr API gives only the view count of the image posted on the platform. A user only viewing an image cannot be assumed to be a victim of the campaign. To calculate visibility in all scenarios, we collected the *likes / retweets, plus-oners / reshares,* and *likes* from Twitter, GooglePlus, and Facebook respectively using their APIs. Apart from calculating values for each visibility attribute, we also collected properties of the user accounts involved, i.e., the IDs of user accounts involved in retweeting / liking / resharing the content. Due to rate limiting constraints on each of the APIs, we could not fetch visibility information daily. We collected this data six months after our data collection period, as posts take time to reach their audience. Due to this, (1) we might have missed information of tweets posted by suspended accounts, and (2) our total visibility values represent a lower bound.

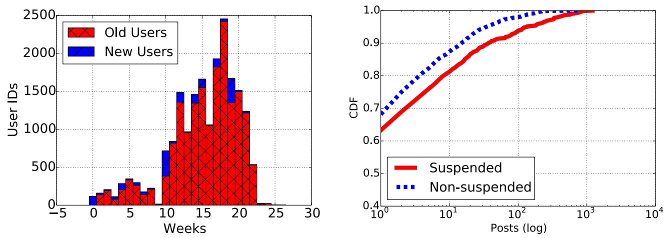
To increase the visibility of content, we observed that the spammers use the following tricks: 67% of posts contained hashtags (for marketing Carrascosa *et al.*, 2013, gaining followers), 82.7% of posts contained URLs (for increased engagement with potential victims), 12.1% of posts contained short URLs (for obfuscating the destination of a URL and getting user engagement analytics), and 72% of posts contained photos (as visual content gathers more attention). We also noticed collusion between accounts and cross-referenced posts to increase the visibility of the campaign.

Cross-referenced posts: We call a post cross-referenced if it was posted to OSN X, but contains a URL redirecting to OSN Y. For instance, a Twitter post containing a link 'fb.me/xxxx' which would redirect to a different OSN, Facebook. Spammers either direct victims to existing posts or to another profile which is propagating the same campaign on a different OSN. In the Tech Support campaign, we observed that 3.2% of Facebook posts redirected to YouTube, and 1.78% of posts redirected from GooglePlus to YouTube.

Collusion between accounts: In the Tech Support campaign, we observed traces of collusion, i.e., spammers involved in a particular campaign, *like / share* each other's posts on OSNs or like their content to increase reachability. Collusion helps in cascading information to other followers in the network.

We calculated the visibility received by all the posts after removing likes / reshares / retweets by the colluders (i.e., accounts spreading the campaign already present in the dataset). We noticed that the posts containing the above-mentioned attributes (hashtags, URLs, short URLs, photos, cross-referencing, and collusion) garnered around ten times more visibility than posts not containing them. Around 10% of the posts saw traces of collusion, contributing to 20% of the total visibility. Maximum visibility (22.1% of total visibility) was observed for posts containing hashtags. In addition, we observed that a major chunk of visibility came from GooglePlus, followed by Facebook. This shows that the audience targeted influences the visibility garnered by a particular campaign, as GooglePlus is known to be consumed mostly by IT professionals [16].

[16] https://insight.globalwebindex.net/



(a) New users created from time to time for campaign sustainability.

(b) Volume per user kept low to evade suspension.

Figure 5: New user accounts created from time to time and volume per ID kept low, to avoid suspension in the Tech Support Campaign.

5.4 What Fraction Suspended?

To aid in the propagation of a campaign, spammers manage multiple accounts to, garner a wider audience, withstand account suspension, and in general increase the volume. Individual spammer accounts can either use automated techniques to aggressively post about a campaign or use hand-crafted messages. In this section, we examine the behavior of user accounts behind the Tech Support campaign. Spammers want to operate accounts in a stealth mode, which requires individual accounts to post few posts. It costs effort to get followers to a spam account, and the number of 'influential' accounts owned by a spammer is limited. Thus, the spammer tends to repeatedly use accounts to post content keeping volume low per account (Figure 5(b)), while creating new accounts once in a while (Figure 5(a)).

Long-lived user accounts: During our data collection, we found that 68.7% (1,305) of the accounts were *never* suspended or taken down on any of the five OSNs. This is in stark contrast to the URL based campaigns Thomas *et al.*, 2011b, where the authors observed that 92% of the user accounts were suspended within three days of their first tweet. To take into account delays in the OSNs' account suspension algorithm, we queried all the accounts six months after the data collection to determine which accounts were deleted / suspended. This process consists of a bulk query to each OSN's API with the profile ID of the account. [17] For each of these accounts, we looked at the time stamp of the first and last post within our dataset, after which we assumed that the account was suspended immediately. Out of the accounts which were suspended, around 35% of the accounts were suspended within a day of their first post; the longest lasting account was active for 158 days, before finally getting suspended. On an average, accounts got suspended after being active for 33 days. This is in clear contrast to users getting suspended within three days for URL based spam campaigns, and thus, focused efforts are needed to strengthen defense from evolving phone-based spam campaigns.

5.5 Is Existing Intelligence based on URLs Useful to Handle Phone-based Spam?

Apart from creating accounts to propagate content, and using phone numbers to interact with victims, spammers also need a distinct set of URLs to advertise. In this section, we look at the domains, subdomains and URL shorteners used by spammers. Of all the posts, we had 4,581 unique URLs and 594 distinct domains. Of all the URLs, 12.1% were shortened using bit.ly; 3% of them received over 69,917 clicks (data collected from bit.ly API), showing that the campaign was fairly successful.

Given the prevalence of spam on OSNs, we examined the effectiveness of existing blacklists to detect malicious domains. Specifically, we used Google safe browsing [18] and Web of Trust (WOT) [19] to see if they were effective in flagging domains as malicious. Web of Trust categorizes the domains into several reputation buckets along with the confidence to assign a category. Please note that one domain may be listed in multiple categories. We marked a domain as malicious if the domain appeared in any of the following categories – negative (malware, phishing, scam, potentially illegal), questionable (adult content). We checked the URLs and domains even after six months of data collection since blacklists may be slow in updating response to new spam sites. We marked a URL malicious if it was listed as malicious either by Google safe browsing or WOT. We checked these domains against the blacklists, finding that 10% of the domains were blacklisted by WOT, none by Google safe browsing. Overall, we found that existing URL infrastructure was ineffective to blacklist URLs used in phone-based spam campaigns.

5.6 Is Cross-Platform Intelligence Useful?

Given that existing URL infrastructure is ineffective, we study if cross-platform intelligence across OSNs can be used. To this end, we look at the spam user profiles across OSNs to figure out which OSN is most effective in building the intelligence.

Homogeneous identity across OSNs: Simply analyzing users' previous posts might not be sufficient, as users can switch between multiple identities, making it hard for OSN service providers to detect and block them. Moreover, spammers may appear legitimate based on the small number of posts made by a single identity. The challenge remains in analyzing the aggregate behavior of multiple identities. To understand how user activity is correlated across OSNs, we pose the question: *do users have a unique identity on a particular OSN or do they share identities across OSNs? Within the same network, can we find the same users sharing multiple identities?*

To answer this, we looked at user identities across different OSNs in *aggregate* (multiple identities of the same user across different OSNs) and *individual* (multiple identities of the same user on a single OSN) forms. If the *same* user has multiple identities, sharing similar name or username, it is said to exhibit a homogeneous identity. To define user identity in a particular campaign, we used two textual features: *name* and *username* Ottoni *et al.*, 2014. Since networks like YouTube and Google Plus do not provide the username, we restrict matching to identities sharing the same name. We used Levenshtein distance to find similarity in usernames. LD(s_i, s_j) is the Levenshtein edit distance between usernames s_i and s_j. Here, LD(s_i, s_j) = 1 means the strings are identical, while LD(s_i, s_j) = 0 means they are completely different. After manual verification by comparing profile images across

chart-of-the-day-who-is-most-likely-to-use-google

[17] If the account is deleted / suspended, (a) Twitter redirects to http://twitter.com/suspended, and returns error 404, (b) Youtube returns 'user not found', (c) Facebook returns error 403 in case the account is suspended, (d) GooglePlus throws a 'not found' error, (e) Flickr responds with a 'user not found' error.

[18] https://developers.google.com/safe-browsing/v4/lookup-api

[19] https://www.myWOT.com/wiki/API

OSNs, we found users having LD >= 0.7 are homogeneous identities. We found four cases where multiple user identities were found for the same user within the same network, and in 65 instances, multiple user identities were present for the same user in more than two networks. Specifically, we found 51 users sharing multiple identities across two different OSNs, and 10 users sharing multiple identities across 3 OSNs. We noticed that these accounts shared same phone numbers across OSNs; some accounts post more phone numbers that are part of tech support campaign.

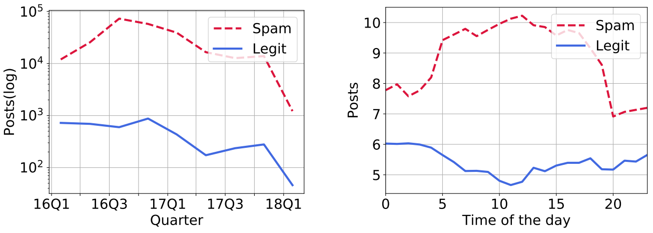
We found that the total number of posts made by these accounts was highest on GooglePlus (2696), followed by Twitter (1776), Facebook (577), Flickr (387), and YouTube (323). Out of all the homogeneous identities, the following are the percentages of accounts suspended on each OSN – Twitter (60%), YouTube (48%), GooglePlus (32%) Flickr (33%), and Facebook (4%). Our data is insufficient to determine whether account suspension is due to dissemination of content across OSNs or other unobserved spammers' properties. Notwithstanding, the association between user identities across OSNs, strengthens the fact that sharing information about spammer accounts across OSNs could help OSNs to detect spammers accurately.

Reducing financial loss and victimization: The actual number of users that are impacted depends on how many victims called spammers and bought the products advertised by campaigns. Since it is hard to get this data, we provide a rough estimate of the number of victims falling for campaigns identified in our dataset. We find reputation of spammers in terms of their followers count on Twitter, friends / page likes on Facebook, circle count on GooglePlus, and subscriber count on Youtube. As these users have subscribed to spammers to get more content, they are likely to fall for the spam. Some of the users would be the ones who aren't aware of the campaign being spam, while some followers / friends could be spammers themselves who have followed other spammers' accounts. We again collected this data after 6 months of our data collection and recorded 637,573 followers on Twitter, 21,053 friends on Facebook, 11,538 followers on GooglePlus, and 2,816 likes on YouTube amounting to a total of 670,164 users. Please note that this number is a lower bound, as we were not able to retrieve statistics for suspended / deleted accounts. Assume that we transfer knowledge from Twitter to other OSNs and prevent the onset of campaigns on other OSNs, we analyzed how much money and victims could be saved. Looking only at the friends, followers, and likers on Facebook, GooglePlus, and YouTube respectively, we could save 35,407 (21,053 + 11,538 + 2,816) unique victims and $8.8M (35,407 * $290.9) by transferring intelligence across OSNs. We used the average cost of the Tech Support Spam to be $290.9 per victim, as reported by Miramirkhani et al. Miramirkhani *et al.*, 2017.

6 Legitimate vs. Spam Tech Support

In this section, we compare the characteristic difference between the accounts involved in propagating spam and legitimate Tech support campaigns.

Within the Tech support campaign, we curated the list of brands / organizations that were being targeted by the spammers to coerce victims. We found 16 such brands viz Microsoft, Gmail, Facebook, Yahoo, McAfee, etc. To find the legitimate dataset, for each brand, we searched the official verified website on Google and took the official phone numbers that were being used for handling respective technical support. Further, we took all the phone numbers used by legitimate handles and searched for tweets containing those phone numbers using Twitter Streaming API. Table 5 presents basic statistics for the two campaigns. The relative sizes of these posts illustrate the scale of the problem: spam is approximately 47 times larger than the legitimate posts received.



(a) Volume Per Quarter (b) Volume Per Hour

Figure 6: (a) Volume generated in spam campaigns is higher than that generated in legitimate campaigns to maximize reach. (b) Hours of operation in both the campaigns is complementary.

Table 5: Characteristics of attributes for spam and legitimate Tech support campaigns.

Category	Spam	Legitimate
#Posts	269,652	5,712
#Unique Phone Numbers	1,164	279
#Unique IDs	6,077	794
#Suspended IDs	67,757	47

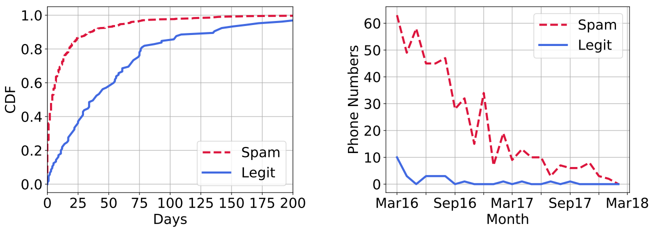
6.1 General Characteristics

As seen in Figure 6(a), proportion of the data collected for both legitimate and spam campaigns remain the same through out the period of collection; volume generated in spam campaigns is higher than the volume generated in legitimate campaign. This is because spammers need to maximise the reach and target as many victims as possible, hence the large number of tweets. Figure 6(b) shows the difference between the two classes in terms of time of day when tweets were posted from these accounts. Hours of operation are almost complementary in both the campaigns. In addition, we observed that spam campaigns pick up early and decay during the day in terms of volume, but legitimate campaigns consistently post during the day without long spikes.

6.2 Phone Number Reusability

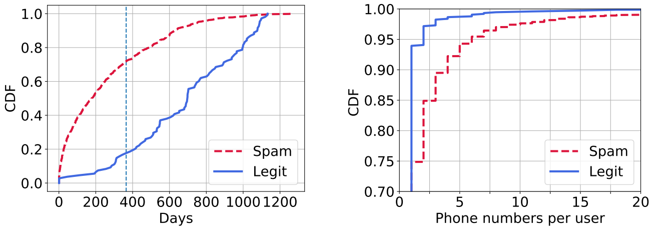
For each phone number, we calculated the number of days between consecutive occurrences for a phone number, defined as *reusability*. We observed that 50% phone numbers appeared again in less than 5 days. In addition, 70% phone numbers used in spam campaigns were being reused within 10 days of the first appearance. This shows that the pool of phone numbers is not kept for long and is replenished in sometime (see Figure 7(a)). Figure 7(b) shows that new phone numbers appear every month; spammers keeping switching between phone numbers and not use a particular phone number for a very long time.

Figure 8(a) shows that the same spam phone number is not used for a very long time whereas legitimate phone numbers have a comparatively higher lifetime (difference between first and last post made using that phone number). Previous studies have also indicated that most of the spam comes from IP addresses that are extremely short-lived to avoid detecting behavioral patterns from



(a) Phone number reusability (b) New numbers every month

Figure 7: (a) Spam phone numbers are reused more; one phone number is not used at a stretch. (b) Spam phone number pool is replenishd with new phone numbers every month to avoid pattern detection.



(a) Lifetime of a phone number (b) Phone numbers per user

Figure 8: (a) Lifetime of a spam phone number is smaller than legitimate phone number since new phone numbers are added in the pool. (b) Nearly 95% of the legit users have only one phone number whereas a lot of spam users employee multiple phone numbers to maximise reach and regulate volume per phone number to avoid detection.

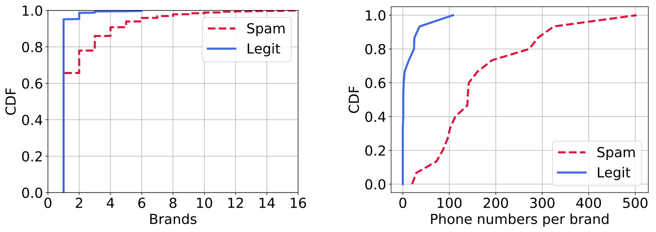
historical data Venkataraman *et al.*, 2007. In addition, the phone numbers used per spam account is more than a legitimate user, as shown in Figure 8(b). Spammers could imploy such tactic to regulate the volume per phone number to avoid detection by OSNs. Further, since there are physical entities handling the phone call requests, number of phone numbers per spam account is not very high.

In conclusion, spam phone numbers have shorter lifetime and are more reused, i.e., a single phone number is not used for a very long time at a stretch, but reappears in sometime.

6.3 Brand Propagation

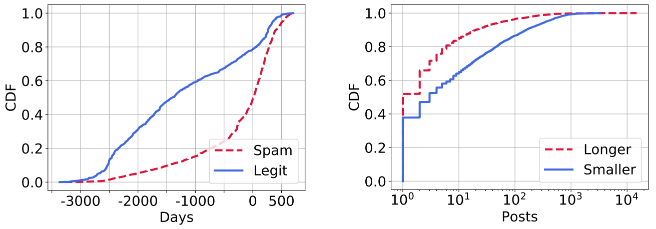
Brands are the top companies / organizations a user tweets about. We first selected the top 15 brands that represented majority of our data. Products belonging to the same companies were grouped together for e.g. Instagram and WhatsApp were clubbed along with Facebook as the parent brand (company) is the same. Any tweet that does not mention any of these brands was taken as *Others*. Figure 9(a) shows that majority of the legitimate users only talk about one brand but nearly 35% of the spam users mention more than one brand. This could be because it increases spammers' probability of receiving a tech support related victim phone call. On the other hand, the legitimate tech support only talks of the brand they are serving.

In addition, the number of phone numbers used to propagate



(a) Brands per user (b) Phone numbers per brand

Figure 9: (a) Spammers tweet about multiple brands, use multiple phone numbers for a single brand. (b) On the other hand, legitimate users tweet about a single brand and in more than 90% cases, use one phone number per brand.



(a) CDF of Lifetime of a User. (b) Volume of users who have stayed lesser vs. longer time.

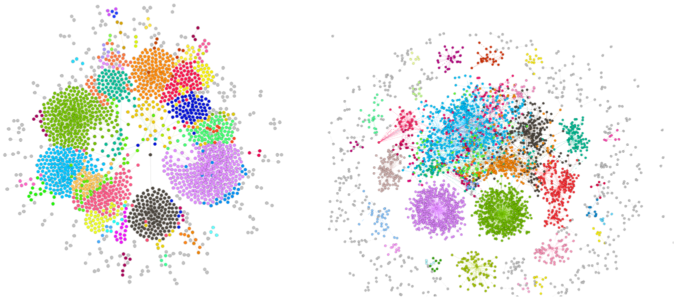
Figure 10: (a) Lifetime of a spam account tends be much smaller than legitimate account because Twitter suspends spam accounts due to high volume of tweets. (b) Accounts that have posted more tweets were suspended by Twitter sooner.

Tech Support about any one brand in spam data was much greater than legitimate. More than 90% legitimate brands had a few phone numbers while spammers used several hundred phone numbers in operation, using 3 or 4 phone numbers in the same tweet. Using more phone numbers per brand helps spammers to handle more requests for a particular brand, thereby maximizing the reach.

Even though phone numbers per spammer were less (see Figure 8(b)), overall, the number of unique phone numbers used in propagating spam campaigns were higher than the phone count used in legitimate campaigns.

6.4 Lifetime of Spammers

Lifetime of a spammer and legitimate account was calculated by taking March 1, 2016 as the starting date (beginning of data collection) and represented as the number of days user between the date user created an account on Twitter and starting date. Negative lifetime means the user account was created before the start date. As Figure 10(a) shows, legitimate accounts have been on Twitter much longer than spam accounts. Since the volume posted by legitimate users is smaller than spammers, they did not get suspended by Twitter. In addition, we observed that the spam accounts which were not suspended by Twitter posted fewer tweets, as shown in Figure 10(b).



(a) Legitimate users (b) Spammers

Figure 11: Network Graph using *user mentions* shows high modularity for spammers, compared to legitimate users. Each color represents different communities. (a) Legitimate users have loose community, modularity coefficient < 0.5, while (b) spammers have a dense structure where they mention each other in their tweets achieving a high modularity coefficient (0.85).

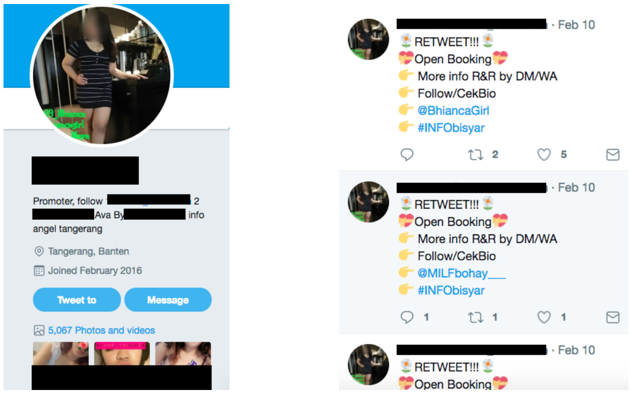
6.5 Network Characteristics

We analyzed the frequency at which spam and legitimate users interact within their own group. Spammers operate in cohorts. They post similar content, target the same brands (major tech companies like Apple, Google, Microsoft), retweet each other and share phone numbers. Group of spammers collude so that their malicious content spreads out effectively. The high density of connections among spam users suggest a strong modularity, as shown in Figure 11(b). In this network graph, each node represents one user and nodes are connected if one user mentions the other and a user can be tech support provider (spammer / legitimate user) or someone looking for tech support information online. The network graph generated using Louvain Algorithm [20] showed high modularity for spammers. In contrast, the legitimate clustering graph depicted well defined boundaries. Legitimate tech support providers have little incentive to collaborate since they only provide tech support for their own products.

7 Detecting Spammers

A critical stage in developing effective classification model is the identification of features that can separate one class from another. This section discusses the features used for both legitimate and spam account detection. Figure 12 shows a spammer account that clearly violates the Twitter policy by promoting and posting repeated, pornographic content.

We present the design philosophy of the classification system. In particular, we introduce the classification features used to distinguish between scam and legitimate Tech support campaign and also show the performance we obtained using various machine learning algorithms. The features extracted are common to both legitimate and spam technical support campaigns and were collected for each campaign / handle separately. User level features describe the basic characteristics used to define a user account while content level features study the behavioral patterns of social network accounts around the tweets posted by the users. The specific description of each feature for a account / user is shown in Table 6.



[redacted]

Promoter, follow [redacted] 2 [redacted] Ava By [redacted] info angel tangerang

Tangerang, Banten

Joined February 2016

5,067 Photos and videos

[redacted]

[redacted] · Feb 10
RETWEET!!!
Open Booking
More info R&R by DM/WA
Follow/CekBio
@BhiancaGirl
#INFObisyar

[redacted] · Feb 10
RETWEET!!!
Open Booking
More info R&R by DM/WA
Follow/CekBio
@MILFbohay___
#INFObisyar

[redacted] · Feb 10
RETWEET!!!
Open Booking

(a) Bio of the spammer account claiming itself as a promoter of pornographic content.

(b) Timeline of the spammer mentioned in (a) indicating similar content is being posted repeatedly.

Figure 12: An example spammer account that has not been suspended by Twitter yet, but our system could detect it as spammer.

7.1 Experimental Setup and Results

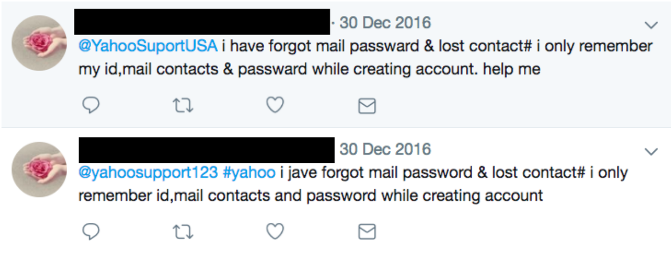
Our dataset consisted of 768 legitimate users and 6,238 spam users. To account data imbalance caused by spam class, we split the spam users into 9 disjoint sets with 693 users in 8 of these sets and 694 in the last set using random split. We added the legitimate users to each of these sets and trained separate classifiers for the data splits obtained. To perform the training and testing we did an 80:20 train / test split with 80% of each sets data being used to train the classifier while the held out 20% used to test the performance of the trained model. The performance of five (5) machine learning algorithms viz. Random Forest, SVM, Logistic Regression, KNN, and Extra Trees classifier was evaluated to identify the best classifier that is suitable for the proposed unified framework. We used 10-fold cross validation and report the mean accuracy, precision, and recall of each of the classifiers. The data was standardised to reduce the negative effect of outliers and brought the data values in ranges which helped the classifier learn and converge better.

We observed that **Random Forest** performed well for account detection, achieving 99.8% precision and recall. To avoid overfitting, we applied L2 regularisation on the dataset after standardisation process. After conducting the classification experiments we tried to analyse and see the most important features by analyzing feature importance for the Extra Trees Classifier. This was done by calculating the average feature importance across the 9 splits made above. The top 5 features were: Ratio of unique numbers (27.10%), user status (5.65%), presence of URLs in tweets (5.29%), age of account (4.71%), and mean phone number count (3.41%). During our introspection, we found that our system could not identify spammers who were posting smaller number of posts at a given point in time as compared to accounts that posted bursts of posts at the same time.

8 Discussion

Providing feedback via Twitter is seen as one of the powerful tool for prompt grievance redressal, where anybody with a grievance against a company can be heard by fellow customers. The company

[20] https://perso.uclouvain.be/vincent.blondel/research/louvain.html



[redacted] · 30 Dec 2016

@YahooSuportUSA i have forgot mail passward & lost contact# i only remember my id,mail contacts & passward while creating account. help me

[redacted] 30 Dec 2016

@yahoosupport123 #yahoo i jave forgot mail password & lost contact# i only remember id,mail contacts and password while creating account

Figure 13: People tagging wrong handles for complaint redressal.

can instantly find information on the consumer before deciding whether and how to respond. A study done by Simply Measured showed that 99% of brands are on Twitter, and 30% of them have a dedicated customer service handle. The average response time was 5.1 hours with 10% of companies answering within an hour, and 93% of companies answering within 48 hours. [21] While the social media interaction helps in strengthening the customer-brand relationship, our work sheds light on it's exploitation by spammers. Figure 13 depicts a Twitter user tagging a spam Tech Support handle to get their issue resolved.

9 Conclusion

With the convergence of telephony and the Internet, the phone channel has become an attractive target for spammers to exploit and monetize spam conducted over the Internet. This paper presents a large-scale study of cross-platform spam campaigns that abuse phone numbers. We collect ~23 million posts containing ~1.8 million unique phone numbers from several OSNs over a period of six months. We identified 202 campaigns running from all over the world with Indonesia, United States, India, and the United Arab Emirates being the highest contributors. By examining campaigns running across OSNs, we showed that Twitter could suspend ~93% more accounts spreading spam as compared to Facebook. Therefore, sharing intelligence about spam user accounts across OSNs can aid in spam detection; ~35K victims and $8.8M could be saved based on exploratory analysis of our data. We used account and content level features and built a machine learning model to identify spammers, achieving a precision and recall of 99.8%.

References

Almeida, H., D. Guedes, W. Meira, and M. J. Zaki. 2011. "Is there a best quality metric for graph clusters?" In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer.

Amleshwaram, A. A., N. Reddy, S. Yadav, G. Gu, and C. Yang. 2013. "Cats: Characterizing automation of twitter spammers". In: *Communication Systems and Networks (COMSNETS), 2013 Fifth International Conference on*. IEEE.

Balduzzi, M., P. Gupta, L. Gu, D. Gao, and M. Ahamad. 2016. "MobiPot: Understanding Mobile Telephony Threats with Honeycards". In: *Proceedings of the 11th ACM SIGSAC Symposium on Information, Computer and Communications Security*. *ASIA CCS '16*. Xian, China: ACM.

Benevenuto, F., G. Magno, T. Rodrigues, and V. Almeida. 2010. "Detecting spammers on twitter". In: *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*. Vol. 6. 12.

Benevenuto, F., T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. 2009. "Detecting spammers and content promoters in online video social networks". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM.

Carrascosa, J. M., R. González, R. Cuevas, and A. Azcorra. 2013. "Are trending topics useful for marketing". *Proc. COSN*.

Christin, N., S. S. Yanagihara, and K. Kamataki. 2010. "Dissecting one click frauds". In: *Proceedings of the 17th ACM conference on Computer and communications security*. ACM. 15–26.

Chu, Z., I. Widjaja, and H. Wang. 2012. "Detecting social spam campaigns on twitter". In: *International Conference on Applied Cryptography and Network Security*. Springer. 455–472.

Costin, A., J. Isacenkova, M. Balduzzi, A. Francillon, and D. Balzarotti. 2013. "The role of phone numbers in understanding cyber-crime schemes". In: *Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on*. IEEE. 213–220.

Faloutsos, M. 2013. "Detecting malware with graph-based methods: traffic classification, botnets, and facebook scams". In: *Proceedings of the 22nd International Conference on World Wide Web*. ACM. 495–496.

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

Ghosh, S., B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi. 2012. "Understanding and combating link farming in the twitter social network". In: *Proceedings of the 21st international conference on World Wide Web*. ACM. 61–70.

Grier, C., K. Thomas, V. Paxson, and M. Zhang. 2010. "@ spam: the underground on 140 characters or less". In: *Proceedings of the 17th ACM conference on Computer and communications security*. ACM. 27–37.

Gupta, P., M. Ahamad, J. Curtis, V. Balasubramaniyan, and A. Bobotek. 2014. "M3AAWG Telephony Honeypots: Benefits and Deployment Options". *Tech. rep.*

Gupta, P., R. Perdisci, and M. Ahamad. 2018. "Towards Measuring the Role of Phone Numbers in Twitter-Advertised Spam". In: *Proceedings of the 13th ACM on Asia Conference on Computer and Communications Security*. *ASIA CCS '18*. Incheon, Republic of Korea: ACM. ISBN: 978-1-4503-5576-6. DOI: 10.1145/3196494.3196516. URL: http://doi.acm.org/10.1145/3196494.3196516.

Gupta, P., B. Srinivasan, V. Balasubramaniyan, and M. Ahamad. 2015. "Phoneypot: Data-driven Understanding of Telephony Threats." In: *NDSS*.

Gupta, S., P. Gupta, M. Ahamad, and P. Kumaraguru. 2016. "Exploiting Phone Numbers and Cross-Application Features in Targeted Mobile Attacks". In: *Proceedings of the 6th Workshop on Security and Privacy in Smartphones and Mobile Devices*. ACM. 73–82.

[21] http://get.simplymeasured.com/rs/simplymeasured2/images/CompleteGuidetoTwitterAnalyticsSimplyMeasured.pdf

Isacenkova, J., O. Thonnard, A. Costin, A. Francillon, and D. Balzarotti. 2014. "Inside the scam jungle: A closer look at 419 scam email operations". *EURASIP Journal on Information Security*. 2014.

Kumaraguru, P., L. F. Cranor, and L. Mather. 2009. "Anti-phishing landing page: Turning a 404 into a teachable moment for end users". *Conference on Email and Anti-Spam*. URL: http://precog.iiitd.edu.in/Publications_files/APWGLandingPage-Turning404intoEducation.pdf.

Lee, K., J. Caverlee, and S. Webb. 2010. "Uncovering social spammers: social honeypots+ machine learning". In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 435–442.

Lee, K., B. D. Eoff, and J. Caverlee. 2011. "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter." In: *ICWSM*.

Lumezanu, C. and N. Feamster. 2012. "Observing common spam in Twitter and email". In: *Proceedings of the 2012 ACM conference on Internet measurement conference*. ACM. 461–466.

Marzuoli, A., H. A. Kingravi, D. Dewey, and R. Pienta. 2016. "Uncovering the Landscape of Fraud and Spam in the Telephony Channel". In: *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE. 853–858.

Miramirkhani, N., O. Starov, and N. Nikiforakis. 2017. "Dial One for Scam: A Large-Scale Analysis of Technical Support Scams". In: *Proceedings of the 24th Network and Distributed System Security Symposium (NDSS)*.

Osborne, M. and M. Dredze. 2014. "Facebook, Twitter and Google Plus for breaking news: Is there a winner?" In: *ICWSM*.

Ottoni, R., D. B. Las Casas, J. P. Pesce, W. Meira Jr, C. Wilson, A. Mislove, and V. A. Almeida. 2014. "Of Pins and Tweets: Investigating How Users Behave Across Image-and Text-Based Social Networks." In: *ICWSM*.

Rahman, M. S., T.-K. Huang, H. V. Madhyastha, and M. Faloutsos. 2012. "Frappe: detecting malicious facebook applications". In: *Proceedings of the 8th international conference on Emerging networking experiments and technologies*. ACM. 313–324.

Srinivasan, B., P. Gupta, M. Antonakakis, and M. Ahamad. 2016. "Understanding Cross-Channel Abuse with SMS-Spam Support Infrastructure Attribution". In: *European Symposium on Research in Computer Security*. Springer. 3–26.

Stringhini, G., C. Kruegel, and G. Vigna. 2010. "Detecting spammers on social networks". In: *Proceedings of the 26th Annual Computer Security Applications Conference*. ACM. 1–9.

Thomas, K., C. Grier, J. Ma, V. Paxson, and D. Song. 2011a. "Design and evaluation of a real-time url spam filtering service". In: *2011 IEEE Symposium on Security and Privacy*. IEEE. 447–462.

Thomas, K., C. Grier, D. Song, and V. Paxson. 2011b. "Suspended accounts in retrospect: an analysis of twitter spam". In: *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM. 243–258.

Venkataraman, S., S. Sen, O. Spatscheck, P. Haffner, and D. Song. 2007. "Exploiting network structure for proactive spam mitigation".

Wang, A. H. 2010. "Don't follow me: Spam detection in twitter". In: *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*. IEEE. 1–10.

Webb, S., J. Caverlee, and C. Pu. 2008. "Social Honeypots: Making Friends With A Spammer Near You." In: *CEAS*.

Yardi, S., D. Romero, G. Schoenebeck, *et al.* 2009. "Detecting spam in a twitter network". *First Monday*. 15(1).

Table 6: Content and account based features used in classifying normal users from spammers.

Feature Type	Feature Name	Description
Content Level	Total tweets	The total number of tweets
	Tweets per day	Ratio of total number of tweets to total number of days between the first and last tweets of that user
	Links in tweets	Total number of links / URLs in the tweets
	Number of hashtags	Total number of hashtags in the tweets
	Number of mentions	Total number of mentions of other users in the tweets
	Popularity score	Ratio of sum of the total number of retweets and total favorite count to the the total number of tweets of that user
	Std. phone number count	Population deviation of phone numbers
	Mean phone number count	Mean population deviation of phone numbers
	Hashtag Ratio	The ratio of total hashtags to total tweets
	Links ratio	The ratio of total number of links / URLs to total number of tweets
	Mention ratio	The ratio of total number of mentions to total number of tweets
	Retweeted ratio	The ratio of tweets which were retweeted to total number of tweets
	Maximum number of hashtag	The maximum number of hashtags the user has used in any of his tweets.
	Mean tweet length	The mean length of tweets calculated based on the number of words.
	Deviation in tweet length	Standard deviation in tweet lengths.
	Deviation in Retweeted tweets	Standard deviation in number of tweets which were retweeted.
	Deviation in mentions	Standard deviation in mentions per tweet.
	Deviation in URLs	Standard deviation in number of URLs which the user has used in each tweet.
	Deviation in hashtags	Standard deviation in hashtags used per tweet.
	Screen name length	The length of the screen name for that user.
	Presence of URLs in tweets	Indicates whether the users tweets contain URLs.
	Hashtags per word	The ratio of total hashtags used by the user to the total number of words in the tweets.
	URLs per word	The ratio of total URLs used by the user to the total number of words in the tweets.
	Time per tweet	The average time between two consecutive tweets.
	Number of tweets replied	The count of tweets which were in reply to some other tweets.
	Repeated tweets	Subtracting the total number of tweets by the user by the unique tweets of the user.
	Ratio of unique numbers	Ratio between unique phone numbers in the users' tweets to unique phone numbers.
Account Level	Number of retweets	Total number of tweets of that user which were retweeted
	Time zone	Indicates whether the the tweets of the user contain a time zone
	Number of retweets	Total number of tweets of that user which were retweeted.
	Mean favorite count	The mean value of favorite count for all of the users tweets.
	Followers to friends ratio	The ratio between total number of followers to total number of friends of that user.
	Friends to followers ratio	The ratio between total number of friends to total number of followers of that user.
	Geo enabled	Indicates whether the user has enabled the geo-tagging of the tweets.
	Profile url	Indicates whether the user has provided a url in association with their profile.
	User status	Indicates whether the user account still exists or was suspended.
	Age of account	Indicates the number of days since when the account was activated indicating the accounts age in days.
	Followers count	Count of the number of followers the user has.
	Friends count	Count of the number of friends the user has.
	Statuses count	Indicates the total number of tweets including retweets issued by the user.
	Favorites count	Total number of tweets the particular user has liked.
	User description	Indicates whether the user has given a description for the profile or not.
	Default profile	Indicates whether the user altered the theme or background of their profile.
	Verified profile	Indicates whether the user has a verified account or not.
	Default profile image	Indicates whether the user has uploaded a profile image or whether a default image is being used instead.
	Listed count	A count of the public lists the user is part of.