Prevalence and Mitigation of Forum Spamming

Youngsang Shin, Minaxi Gupta, Steven Myers
School of Informatics and Computing
Indiana University, Bloomington
Email: {shiny, minaxi}@cs.indiana.edu, samyers@indiana.edu

***Abstract*—Forums on the Web are increasingly spammed by miscreants in order to attract visitors to their (often malicious) websites. In this paper, we study the prevalence of forum spamming and find that Internet users are at a high risk of encountering forums with spam links posted on them. To mitigate the problem, we examine the characteristics of 286 days of forum spam posted at a research blog and develop light-weight features based on spammers' IP, commenting activity and the anatomy of their posts. We find that an SVM classifier trained on these features can achieve a 99.81% precision and 92.82% recall in identifying forum spam.**

I. Introduction

The Internet boasts over 234 million active websites [1], [2]. Of these, 47 million were added in 2009. The rapid growth of websites makes the problem of attracting visitors especially challenging. Thus, website operators are always on the lookout for new mechanisms to make their websites discoverable, particularly through popular search engines. This is particularly true for operators of malicious websites. A popular technique to achieve this goal is *web spamming*. It exploits algorithms used by popular search engines in order to gain undeserved high rankings with respect to other sites on the Web.

While many tricks are used to achieve web spamming, *forum spamming*, is an increasingly popular one. In forum spamming, miscreants put links to their websites on forums frequented by Internet users. The definition of a forum is quite permissive: A *forum* is a website where visitors can contribute content. Examples of forums include webboards, blogs, wikis, and guestbooks. Forums increasingly play a major role in the Web landscape, particularly as online social networks like Facebook and Twitter gain popularity.

Forum spamming benefits spammers in two ways. First, it helps drive forum visitors to spammers' websites directly. Second, it increases search-engine rankings for their websites. Forums are an attractive target because search engines cannot simply blacklist them. Site take-down is also not an option because many forums are legitimate and contain valuable information. While search engines can identify some forum spam, and a long line of research, including [3], [4], [5], [6] has been done on the topic, it still does not help forum operators keep their forums free of spam. The problem of identifying spam is worsened by the fact that in many forums the attempts to post spam significantly outnumber the legitimate postings. This makes it hard for forum administrators to prevent spam from being posted or to manually remove it.

There are two types of forum spamming. In the first kind, miscreants build a forum solely for the purpose of spamming. Such a forum is commonly referred to as *splog*. [1] The second type of forum spamming involves posting spam in legitimate forums frequented by Internet users. This category can be further split into two types based on the direction of communication. Communication in forums happens either through a *post* of a new topic or through a *response* to an existing post. Examples of post spamming include a blog post, a new topic in a webboard, a new article in Wiki, or a message left at a guestbook. Examples of the response spamming include comments on a blog post, a reply to a topic on a webboard or guestbook, or a modification to a wiki page. While forum spamming targets both posts and responses on forums, it is more efficient for spammers to target commenting on existing posts which may already have a high rank in a search engine [7]. It also helps avoid detection if they craft their spam message in a manner that is coherent with the original post. In fact, various forum spam automators, including XRumer [8] do exactly this. Subsequently in this paper, we use the term *comment spam* to refer to forum spam appearing as comments to existing posts.

The focus of this paper is on comment spam. We begin by studying the prevalence of comment spam. Using labeled comment spam from a research blog for a period of 286 days, we find spam URLs posted on thousands of forums on the Web, spread across hundreds of TLDs. Forum platforms running phpBB or WordPress are among the most spammed. Also, spammers seem to target active forums more often than inactive ones. Further, we find that spam URLs get posted on more forums as time progresses.

Next, we examine various characteristics of comment spam in a quest to find light-weight features that can be used by a forum server to train a classifier which can identify comment spam. We focus on features based on spammers' origin information, commenting activity, and the text and URL content of the comment spam itself. We find that features based on URLs are not very effective but those based on the other categories yield excellent performance. Specifically, an SVM classifier using features based on spammer origin and commenting activity yields a precision of 99.81% with a 92.82% recall.

The rest of this paper is organized as follows. We show

[1] Although *splog* is used for a spam blog, we use the term for all types of spam forums.

the overview of our collected data in Section II. Section III presents the prevalence analysis of forum spam. We analyze spammers' activity and identify features for spam classification in Section IV. Section V explains the spam classification result. Related work is discussed in Section VI. Finally, Section VII concludes the paper.

II. Data Overview

We were motivated to study aspects of comment spam for a live blog on the Internet. Hence, we chose not to set up our own blog. Instead, we collected comments and their logs to posts on an active blog maintained by the security research group at the Computer Laboratory at University of Cambridge [2] over 286 days. The blog is built on the WordPress [9] software platform. It uses two mechanisms for filtering spam. First, it requires that posters provide a properly parsable email address prior to commenting. Second, it runs the Akismet [10] plugin for filtering spam. We assume that a comment on this blog is spam if it is categorized as such by Akismet [3] and non-spam otherwise.

Each comment and its log have various helpful fields, including the source IP address of the poster, posting date, author's name, author's URL, and the body of the comment. The comment body may include URLs. The author-related fields are unreliable, as a poster could put in any values they choose without effecting the comment they post. As such, we do not make use of them.

Table I presents an overview of collected data. We started with 240 initial blog topics. Over our observation period there were 49 new topics being discussed by the bloggers, resulting in an average of one new topic per week. There were three orders of magnitude more comments posted in the same duration, both to old and new topics. Of these comment posts 98.43% were classified as spam by Akismet. Clearly, the problem of spam on forums is at least as bad as email spam, where estimates are that at least 90% of email is spam [11].

TABLE I: Overview of data

Collection Period	8/19/09 - 5/31/10 (286 days)	
# of new blog posts		49
# of old blog posts		240
# of comments	29,243	(100%)
# of spam	28,783	(98.43%)
# of non-spam	460	(1.57%)

III. Prevalence of Forum Spam

Work in [12] confirmed that users searching for popular keywords on prominent search engines were likely to encounter forum spam. In order to motivate our work, we first examine various aspects of forum spam's prevalence.

Toward this goal, we extracted malicious URLs present in our data, once a day for the first 112 days. Using these URLs as keywords, we conducted daily searches using the Google search engine [4]. Note that older URLs in our dataset get searched over a larger number of days than ones added more recently. This is because we searched the entire set of URLs each day. The results of these searches are pages containing the same spam URLs as in our data. Although these could be web pages set up by spammers, a manual investigation of a random sample revealed them to be all forums, so we assume them to be forums. Our analysis subsequently in this Section to infer the type of forum supports this conclusion.

TABLE II: Forum pages with known comment spam

Collection Period	9/11/2009 - 12/31/2009 (112 days)
# of unique malicious URLs searched	77,321
% producing search results	92%
# of unique web pages containing them	1,854,039

92% of malicious URLs we searched produced pages indexed by Google for an average of 26 search results per URL. Table II shows the total URLs and the corresponding search results. Figure 1 shows the percentage of URLs that yield a given number of search results The URL with the maximum number of search results had 10,086 different hits. Clearly, forum spam is prevalent and some spammers are quite aggressive about posting it. This is not surprising since tools like XRumer [8] are available to automate the process.

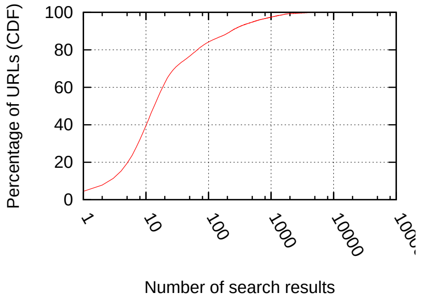


Fig. 1: Number of search results per malicious URL. 3% URLs produce more than 1000 results.

Spammed forum pages belonged to 16 generic top level domains (TLDs) and 198 country-code TLDs. To put it in perspective, there are 21 gTLDs and 260 ccTLDs as of now [13]. This highlights the diversity of spammed forum pages. The TLD `.com` accounted for over 50% of spammed pages, which is expected since it accounts for over 50% of the domains in the Internet [14].

Since we searched for all spam URLs each day, we can see how forum spam grows over time. The growth is shown in Figure 2. Even though this Figure underestimates the growth of URLs that were added later on, it shows that a large percentage

[2] http://www.lightbluetouchpaper.org

[3] Akismet is a web service to filter forum spam. The details of its classification algorithm are proprietary and its accuracy is not formally measured. However, false positives from our data set have already been excluded by the blog's administrator.

[4] Their University Research Program allows high-volume programmatic access to Google searches.

of URLs continue to appear at new forums as time passes. We note that a caveat in this statement is that since we do not know Google's crawl strategy, part of the growth may simply be a manifestation of crawling rather than spammer's posting strategy. However, since this is what Internet users see, forum spam seems to grow with time.

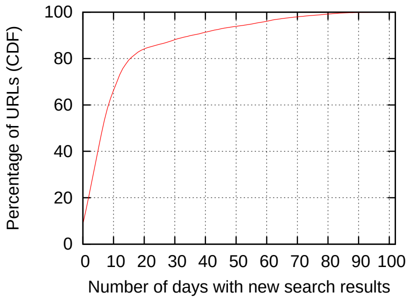


Fig. 2: Growth of forums at which malicious URLs are posted

Next, we try to understand the type of forums that get spammed. We use forum software as our identification metric. We take a two-step approach to identifying forums. In the first step, we try to find forums by looking for common keywords used in their URLs. For example, words, 'forum', 'board', or 'bbs' are often found in URLs belonging to forums. Table III shows the keywords we use. In the second step, we identify forum software by visiting each forum URL and parsing its HTML. We parse for 24 popular forum platforms and 13 blogging platforms, using the law of diminishing returns as our guide. Some platforms we parse for are used world wide and, in contrast, some are only used in specific countries (examples follow shortly). Parsing HTML for popular platforms involved creating heuristics based on unique features of each platform, including keywords, tag usage, comments, and others.

TABLE III: Keywords used to identify various types of forums

Type	Keywords
Webboard	forum, board, bbs, thread
Blog	blog, journal, diary, tag
Guestbook	guestbook, gbook, gb
Wiki	wiki

We pick 112,717 spammed forum pages collected on September 29, 2009 for the above investigation. Our heuristics categorized 70.55% of them. Of these, 69% were webboards, 16% blogs, and 14% guestbooks. There were a small number, 431, of wikis as well. Many of the 29.45% unclassified forums appeared to be built on proprietary software. This is especially true for commercial sites, which account for a large number of URLs in our data. Table IV shows each of the 24 webboard- and 13 blog-publishing softwares we identified, along with the number of pages for each and specific countries where the software is popular. We note that the pages listed against each type of forum platform represent many unique forums since our searches typically yielded only one search result per forum. The table proves that forum spam is posted on forums built on a wide variety of platforms, are with phpBB and WordPress being the most popular victims based on the number of pages they target. Also, the number of pages spammed suggests the use of forum spam automators, such as XRumer [8].

TABLE IV: Categorization of spammed pages

Webboard	# of pages
phpBB	11,853
vBulletin	1,201
Zeroboard (Korea)	1,189
YaBB	516
Discuz! (China)	397
Smf	267
Yuku/Ezboard	183
Simple Machines	172
Web Wiz Forums	135
bbPress	123
PunBB	108
IGN Boards	43
UBB.threads	35
IP.Board	29
Ikonboard	23
SINA.com.cn (China)	17
Sify.com (India)	16
Burning Board (Germany, Poland)	15
MyBB	11
Vanilla	8
163.com (China)	2
Jive Forums	1
MesDiscussions.net	1
Sohu.com (China)	1

Blog	# of pages
WordPress	6,195
WordPress.com	289
Movable Type	228
Tistory.com (Korea)	24
Blogger.com	13
Textcube (Korea)	8
Drupal	6
BIGADDA.com (India)	5
Bokee.net (China)	4
Sify.com (India)	3
Blogsmith	2
Sulekha.com (India)	1
BlogGang.com (Thailand)	1

Next, we check if spammed forums were active or not. Toward this goal, we randomly choose 5,000 spammed forums built on the phpBB software and parsed them to find last posting dates on pages with spam links. The main page of a phpBB webboard shows categories and each category leads users to a list of topics. Categories could themselves have multiple levels. We follow each category until we see a list of topics under each. Then, we extract the last posting date for each topic, restricting ourselves only to the first page in each topic since that is where the latest posts occurs. On each page, we check for posting dates in 12 different common date formats. Many non-English forums use different formats than the 12 we examined. We excluded them from our inspection. Further, we ruled out forums with mm/dd/yyyy or dd/mm/yyyy formats to avoid the confusion between date and month. This caused us to rule out 3.7% of the forums.

We regard a forum as active if more than 50% of topics have a posting date within one month of our visit. We find that only 45% of the forums were available to investigate. The rest were either removed or their hosts were unreachable. Of the successfully parsed forums among the available 2,065 forums, 71% were active while 29% were not. *This finding suggests that spammers target active forums more often than inactive ones perhaps to attract visitors and to command better search engine rankings*. Somewhat surprisingly, this contradicts the observation made by previous work where authors found that spammers posted comments after the blog topics became inactive [12], perhaps targeting unmanaged forums.

IV. Characterization of Comment Spam

In this Section we examine characteristics of comment spam with the goal of identifying heuristics that can be used by a forum server to identify it in real time. We examine characteristics in four categories: 1) spammer origin, 2) profile of commenting activity, 3) malicious URLs present in spam comments and 4) (non-URL) content of the comment spam itself.

A. Spammer Origins

An obvious way to locate the spammer is with the IP address contained in the post. However, DHCP and NAT effects can skew any estimations derived from IP addresses, as was noted in a recent paper which found that IP-based estimation overestimated the Torpig botnet population in by an order of magnitude [15]. Due to this, we explore the effectiveness of using autonomous system number (ASN), BGP prefix and geographical location to identify spammers, each of which presents aggregate views and helps eliminate the pitfalls of using individual IPs.

a) ASNs and BGP Prefixes of Spammers: We used a BGP routing table from the RouteViews Project [16] toward this goal. We chose the day of January 15, 2010 to map the source IP addresses to BGP prefixes and ASNs since it is in the middle of our data collection period and should give us a good estimate of the routing information from the duration of data collection. Table V provides an overview of IP addresses, BGP prefixes and ASNs for both spammers and non-spammers. Of the 4,492 posting IP addresses, 94% posted spam. There was an intersection of only 7 IP addresses that sent posted both spam and non-spam comments. This indicates that either spammers are not using compromised user machines or users of compromised machines are not visiting the same forums as those that spammers are posting to. On average, each spammer IP posted 6.8 comments, while in contrast non-spammer IPs posted only 1.7 comments. These numbers indicate that spammer IPs are separate from non-spammer IPs and that comment spammers are not trying to send low-volume spam per IP, in a fashion similar to traditional email spammers.

Table V shows that—just as posters of non-spam have an order of magnitude fewer source IP addresses than those of spam posters—their ASNs and BGP prefixes are also an order of magnitude smaller. Next, we examine if ASN and BGP prefixes are shared by spammers and non-spammers. Figure 3a shows ASNs corresponding to spam and non-spam IPs. It contains a couple of insights. First, very few ASNs contain both spammers and non-spammers and commonality is observed only in cases where a larger number of IPs originate from an ASN. Second, a large fraction of ASNs have only one IP. It is more the case for spammers than non-spammers. The story is similar in Figure 3b, except that the commonality in spammer and non-spammer prefixes is lesser than it was for ASNs. Overall, both these figures lead us to believe that ASNs and BGP prefixes would serve as reasonable features for spammer identification, though more will be required to address the issue of potential false positives.

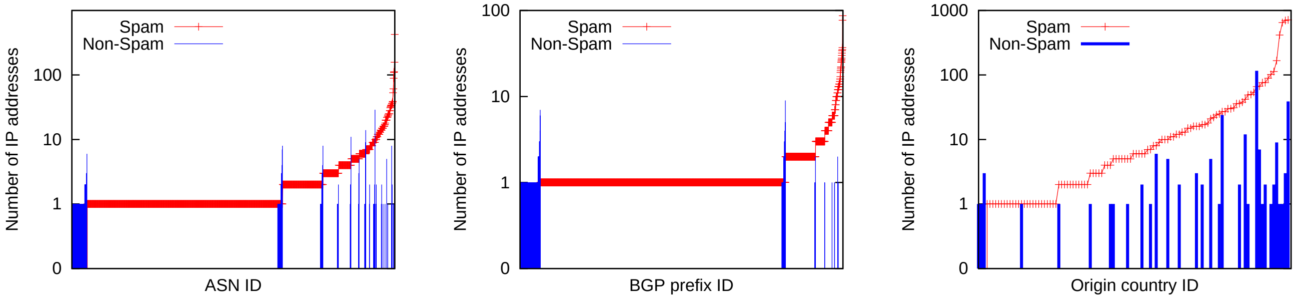
TABLE V: Comparison of spammer and non-spammer origins. Only 7 IP addresses were shared between spammers and non-spammers.

Unique IP addresses of all posters	4,492
Spam comments	
# of unique source IP addresses	4,229
# of ASNs	1,106
# of BGP prefixes	2,273
Non-Spam comments	
# of unique source IP addresses	270
# of ASNs	123
# of BGP prefixes	192

b) Geographic Location of Spammers: Table VI presents the geographic distribution of posters' source IP addresses. The data is derived using the *IP address to geolocation database* from [17]. The spammer IP addresses originate from 106 countries while non-spammer IPs arise from only 33. This trend is similar to what we observed for ASNs and BGP prefixes. Most non-spammer IPs come from the United Kingdom, which is where the blog is located. However, top spammers are located in the U.S., Russia, and Ukraine. Figure 3c shows the breakdown of spammer and non-spammer IPs by country. There is more commonality in spammer and non-spammer origin countries. This suggests geographic location is a worse metric to identify spammers by than is either ASNs or BGP prefixes. However, our results in Section V show that the geographic metric complements the other two, and so is useful for a classifier.

TABLE VI: Top-10 origin countries for spammer and non-spammer IP addresses

Spammers		Non-spammers	
Country	% of IPs	Country	% of IPs
United States	16.9%	United Kingdom	43.0%
Russia	16.6%	United States	14.4%
Ukraine	15.4%	European Union	8.9%
China	9.9%	Canada	4.4%
Germany	3.9%	Germany	3.3%
India	2.7%	France	2.6%
Israel	2.4%	Austria	2.2%
Latvia	2.1%	Australia	1.9%
Brazil	1.8%	Sweden	1.9%
Netherlands	1.8%	Russia	1.1%
other 96 countries	26.5%	other 23 countries	16.3%
Total	100%	Total	100%



(a) ASN (b) BGP (c) Origin country

Fig. 3: Comparison of ASNs, BGP prefixes, and origin countries for spammer and non-spammer IPs

B. Commenting Activity of Spammers

We now analyze the characteristics of commenting activity. Table VII shows an overview of commented blog posts during the data collection period. Spammers targeted all 289 posts, including both new ones posted during our data collection as well as old ones. However, non-spammers commented on only 62 posts during the period. Also, the average number of spam comments per post were close to 100 when legitimate comments were only 7.4. Finally, the number of spam comments per blog post had a much higher standard deviation.

TABLE VII: Overview of commented blog posts during the data collection period

Total number of total blog posts	289
Number of spammed blog posts	289
Average number of spam comments per blog post	99.69
Standard deviation of spam comments per blog post	514.04
Number of legitimately commented blog posts	62
Average number of legitimately commented blog posts	7.42
Standard deviation of legitimately commented blog posts	13.52

a) Commenting Time: Figure 4a shows a CDF of the posting times of spam and non-spam comments. Time in our data set is in GMT. To obtain accurate posting times, we convert the posting time of each comment to the time zone of each commenter. We note that non-spammers post comments during a few peak times, including 8:30am, 10:30am∼12:00pm, 4:00pm∼6:30pm, and 11:00pm. They also have distinct idle times, such as between 3:00am∼7:00am. This roughly corresponds to work schedules and normal sleep times. In contrast, commenting times for spammers do not have any obvious peak times although there is a slight decline in the early morning. These differences can be used to differentiate spam and non-spam comments to some extent.

b) Relationship between Posting and Commenting Times: We also examine the relationship between posting times of original blog post and the corresponding comments. Legitimate forum users tend to keep up with posts on their favorite forums. Therefore, we expect them to comment mainly on recent posts. In contrast, spammers find target forums through search engines and tend to comment on posts which are close to spammers' interest or popular. Therefore, we expect that spammers would not necessarily only comment on recent posts. These thoughts are corroborated by Figure 4b, which shows that 76.5% of non-spammers commented on posts within two weeks of the original post while over 95% of spammers put their comment spams on posts after at least 43 days after their original posting.

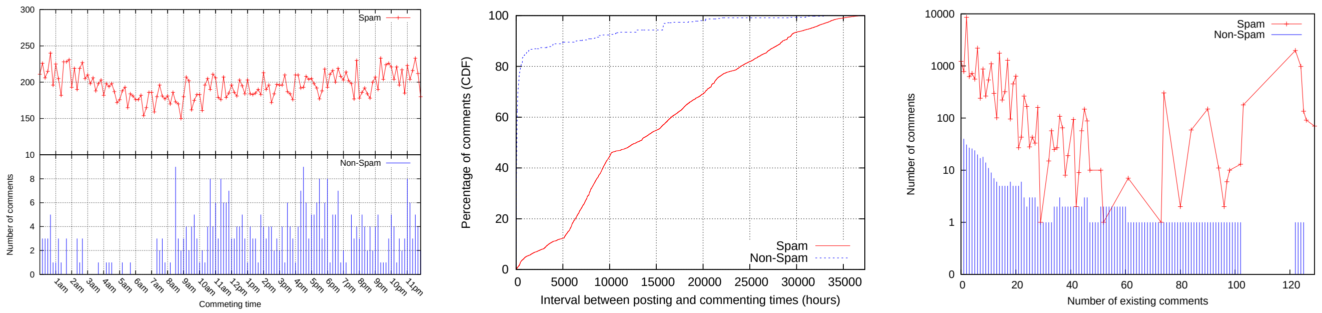
c) Post Popularity: Next, we investigate if spammers' comments are concentrated in more popular posts. Since our data is from a blog, each post does not have a hit count like webboards do. Thus, to measure the popularity of each post, we count the number of existing comments at the time a new comment request is submitted. Figure 4c shows the results. 66% of the non-spam comments were submitted to posts with less than 30 comments. For posts with more than 30 comments, non-spam comments did not have any notable trends. Spam comments followed a similar pattern up to posts with less than 30 comments. In contrast, for posts with more than 30 comments, more spam comments were submitted. This suggests that forum spammers may be posting to threads that have a high number of comments to reap the advantages of linking to a popular topic that may get indexed with a high rank in a web search engine.

C. URLs in Spam Comments

Spam comments usually contain both text and URLs. Here, we examine URLs, leaving the discussion of text to Section IV-D. We extract URLs contained in spam and non-spam comments from our entire data set and examine their characteristics. Table VIII shows that 92% of spam comments contain one or more URLs while only 20% of non-spam comments have URLs. This implies that the presence of URL itself may be taken as an indication that the comment is spam.

a) Number of URLs: Figure 5a shows the number of URLs in comments. We find that over 40% of spam comments have more than 10 URLs in them; furthermore, around 1% of even have more than 100 URLs. In comparison, 80% of non-spam comments have no URLs, another 15% have only one URL, 3% have 2 URLs, and the remaining 2% of non-spam comments have between three and eight URLs. Thus, presence of multiple URLs increases the probability that a comment is spam.

b) Length of URLs: Next, we investigate the length of URLs. Figure 5b shows that over 80% of URLs for spam

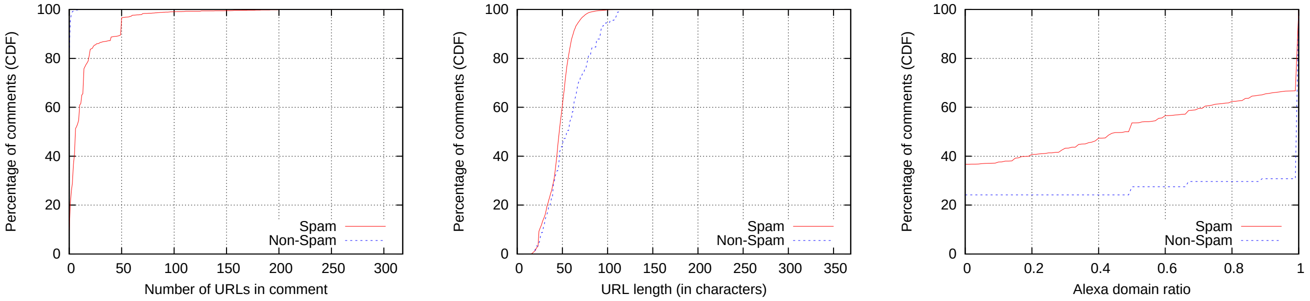


(a) Time of comments

(b) Interval between posting and commenting times

(c) Post popularity

Fig. 4: Characteristics of commenting activity



(a) Number of URLs per comment (Non-spam hugs the y-axis)

(b) URL lengths

(c) Ratio of Alexa domains in URLs per comment

Fig. 5: Characteristics of URLs in comments

TABLE VIII: Number of URLs in spam and non-spam comments

Total comments	29,243	
Spam comments	28,783	100%
Spam comments with URL(s)	26,726	92.85%
Spam comments without URL(s)	2,057	7.15%
Non-spam comments	460	100%
Non-spam comments with URL(s)	91	19.78%
Non-spam comments without URL(s)	369	80.22%

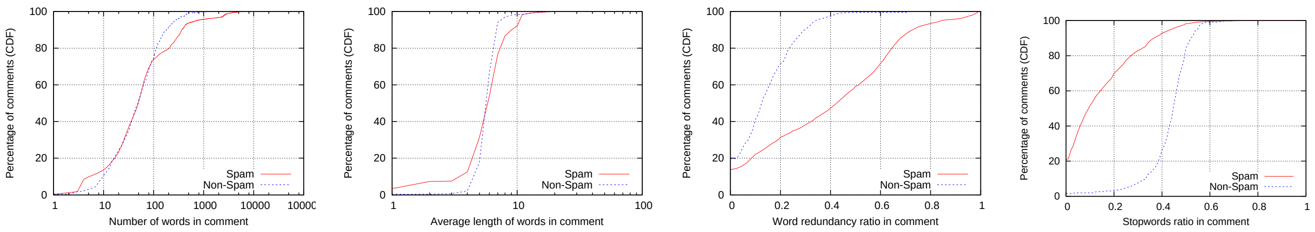
comments are less than 60 character long while over 50% of URLs for non-spam comments are longer than 60 characters. This is counter-intuitive at the beginning since previous work [18] in the context of phishing found malicious URLs to be three times or longer than regular URLs, which had a median length of 22 characters. However, we note that the difference may arise because legitimate forum users may be inserting complete URLs pointing to specific web pages to support their opinions. On the other hand, spammers' URLs may not have many sub-paths or parameters, leading to a shorter URL. Thus, the length of URL may be used as a feature in comment spam identification.

c) Presence of Popular Domains: Lastly, we examine the domain names contained in URLs. This is motivated by the observation that several randomly sampled spam comments in our data contained URLs from popular domains. There are multiple reasons for legitimate domains to be present in spam comments. First, spammers often set up spam forums, *splogs*, at popular forum services, such as `blogger.com` [12] in order to escape detection. They also often exploit any web service offering a profile page for its users whose content can be customized. As an example, consider a spam link found in our data. It was nothing but a link to a spammer's profile page at Amazon.com, which is a popular domain. The URL of the spam site was then contained in the profile page. Another way we find legitimate domains in spam comments is when spammers insert URLs belonging to popular domains along with spam URLs simply to make their posts look legitimate.

We check how many spam URLs in our data are found in the Alexa [19] list, which contains popular Internet domains. Excluding comments that do not contain any URLs, we find that 70% of non-spam comments contain URLs belonging only to Alexa domains while 24% do not have any Alexa domains. The rest have both. In contrast, 33% of spam comments contain URLs belonging only to Alexa domains and 37% do not contain any Alexa domains. The rest have both. This implies that the lack of Alexa domains can be used as a metric to flag spam comments but when spam comments contain them, as they easily can, the utility of this metric will diminish. Indeed, our results in Section V confirm this.

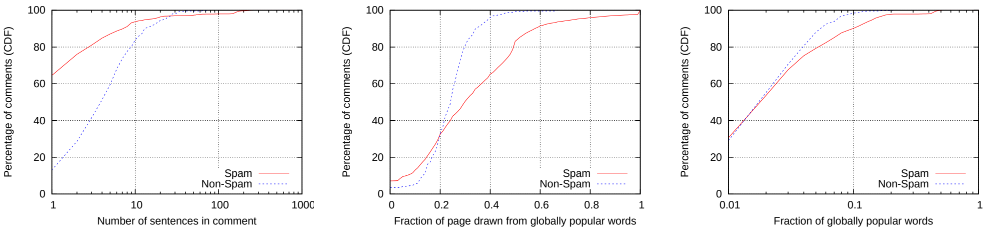
D. Textual Content of Spam Comments

Here, we look for characteristics of the text in spam comments. We examine if it contains features that can help in distinguishing spam comments from legitimate comments. Instead of re-inventing the wheel and developing our own



(a) Number of words (b) Average length of words (c) Word redundancy ratio (d) Stopwords ratio

Fig. 6: Characteristics of first feature set based on comment text



(a) Number of sentences (b) Fraction of page drawn from globally popular words (c) Fraction of globally popular words

Fig. 7: Characteristics of second feature set based on comment text. For *Fraction of pages drawn from globally popular words* and *Fraction of globally popular words*, we used 100, 200, 500, and 1000 globally popular words from comments in our data. We show only the results with 500 globally popular words here. Others had a similar trend.

features, we, instead, use a subset of features developed by [20], [7]. Specifically, we focus on light-weight features when choosing the subset. This is to maintain the ability for spam identification to be carried out in real time. We also rule out features such as number of words in the page title, amount of anchor text and fraction of visible content since they are specific to HTML encoding and our data is not HTML.

We group the chosen features into two sets. The first set of features require word segmentation; the second set requires sentence segmentation. The second set also maintains a globally popular word list across all comments, in addition to the previously mentioned word segmentation. The second set of features are comparatively more computationally expensive. While there are no trends for several of the features, we find that spam comments have higher word redundancy ratio, low stopword [5] count and a lower total number of sentences. We test the efficacy of these features in identifying spam comments in Section V.

V. Spam Classification

We model spam classification as a binary classification problem where our goal is to classify each comment as spam or non-spam using the features investigated in Section IV. The full feature set is listed in Table IX. We use a support vector machine (SVM) classifier toward this goal. Specifically, we run SVMlight [21] with a linear kernel function. We conduct a stratified 10-fold cross-validation [22] to reduce the variability in classification results. In each iteration, 10% of data is used for testing and the remaining for training. We judge the performance of the classifier using well-known metrics, precision, recall, and F-measure, which are defined in Figure 8.

TABLE IX: Features used for comment classification

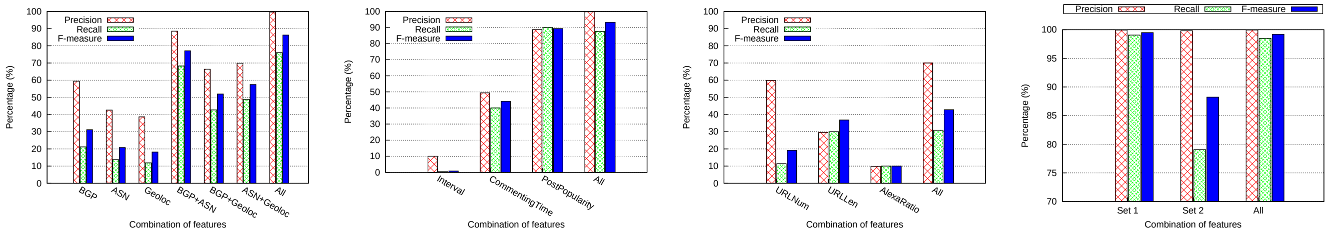
Feature Set	Description
Spammer Origins	ASN BGP prefix Geolocation (country)
Commenting Activity	Time of comment Interval between posting and commenting times Post popularity
URLs in Spam Comments	Number of URLs Length of URLs Presence of Alexa domains
Textual Content Feature Set 1	Number of words Average length of words Word redundancy ratio Stopwords ratio
Textual Content Feature Set 2	Number of sentences Fraction of page drawn from globally popular words Fraction of globally popular words

		Classified class	
		Non-spam	Spam
Correct class	Non-spam	*True Negative*	*False Positive*
	Spam	*False Negative*	*True Positive*

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$
$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$
$$F\text{-}measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Fig. 8: Performance metrics used in judging classifier performance in identifying forum spam

[5] Stop words are words which are filtered out prior to, or after, processing of natural language text. Examples include"the", "is", "at", and "which".



(a) Spammer origin (b) Commenting activity (c) URL characteristics (d) Comment text

Fig. 9: Spam classification results by feature sets

We explore the classifier with various feature sets listed in Table IX separately to understand which feature sets work best. We find that all but URL characteristics-based feature sets yield a close-to-perfect precision. (URL characteristics-based feature set produces only 69.99% precision.) However, the recall and F-measure for individual feature sets are less than desirable as the recall values are less than 90% in most cases. Also, the (computationally expensive) second content-based feature set is not worth the cost, for it fails to improve any of the performance metrics.

TABLE X: Feature-set combinations

Combination #	Features
Without content-based features	
Combination 1	Spammer origin, Commenting activity
Combination 2	Spammer origin, URL characteristics
Combination 3	Combination 1 & Combination 2
With content-based features	
Combination 4	Combination 1 & Content-based feature set 1
Combination 5	Combination 2 & Content-based feature set 1
Combination 6	Combination 3 & Content-based feature set 1

Next, we try different combinations of feature sets to see if recall and F-measure improve. The combinations are described in Table X. Even without the content-based features and URL characteristics-based features, Combination 1 has a 99.81% precision, 92.82% recall and an F-measure of 96.19%. That is, *features based only on spammer origin and commenting activity provide good classification performance*. Combinations including URL characteristics have a worse performance. Further, adding Content-Based Feature Set 1 to Combination 1 improves the precision, recall, and F-measure of the classifier to 99.95%, 98.0%, and 98.96% respectively. These results lead us to conclude that simple, light-weight features can help in classifying comment spam with good performance.

VI. Related Work

Unlike email and web spam, forum spam is a relatively less studied problem. Niu et al. in [12] regard forum spamming as a major new method of web spamming, and explore the impact of web spamming through forum spamming. They investigate how often Internet users encounter forum spam while searching using a major search engine. They also inspect the prevalence of spam blogs, *splogs*, at four popular forum services. Further, they set up three honey blogs to observe the activity of forum spammers. Their analysis focuses on observing trends. We complement their work by developing features that allow forum servers to identify forum spam in a light-weight manner. Bhattarai et al. in [7] propose to classify forums, as we do. However, they focus only on text present in comment body. The content-based features used in their work are variants of features identified by Ntoulas et al. in [20] in the general context of web spamming.

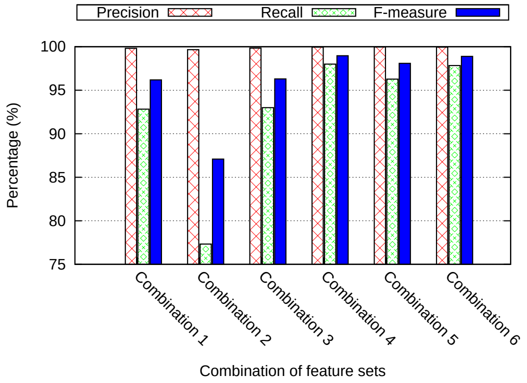


Fig. 10: Spam classification results for various feature-set combinations

For email or web spam, various types of information have been examined for mitigation purposes. The efforts can roughly be categorized into three kinds: network-level features of spammers, content-level features based on email or web page, and anatomy of malicious URLs. Network-level features were examined for spam detection in [23], [24]. Hao et al., in [23], propose a reputation system for email senders based on the features of a sender's source IP address, its ASN and other characteristics in email sending behavior. They combine this with white-listing, grey-listing and content-based spam detection system. Qian et al. in [24] also examine the network-level aggregation of spammers for such reputation systems. They compare clustering results based on BGP prefixes and ASNs. They show that BGP prefixes provide an adequate granularity for spammer aggregation. They can filter out 30% - 50% of spam emails only by using a reputation system which combines BGP prefix with reverse DNS information.

Content analysis is explored for either forum or web spam classification in [20], [5], [4], [25], [26]. It is based primarily on language model disagreement. The authors find that spam tends to be randomly composed from a dictionary or popular keywords; it consists of repeated popular keywords instead of complete sentences and others. They also investigated additional features related to HTML such as title, anchor text, links, and meta tag.

Many works, including [5], [4], [3], [6], [12], [27], [28], [29], investigate identifying malicious links contained in emails or on web pages. The works [28], [29] proposes an online learning system to detect malicious URLs by using lexical and host-based features of the URLs. In contrast, works in [5], [4], [3], [6], [12], [27] follow or crawl URLs for links. [5], [4] builds host-level link graphs to detect link farms and combines information with content-based features. Next, they decide if a host is a spammer. Niu et al. in [12] follow links and determine if they contain redirections or cloaks. If so, these features can be used to detect forum spam. Whittaker et al. in [27] crawl URLs and extracts content-based features. They then use the features along with lexical and host-based features of URLs for a phishing website classifier.

VII. Concluding Remarks

This paper is a proof-of-concept for techniques aimed at efficiently and in real-time classifying forum spam. Though there is little reason to believe that the blog we studied would have a profile different from a random forum on the Web, our current study is limited in that it uses comment spam from only one blog. This shortcoming is unlikely to impact spam comments, variance in legitimate forum users' commenting behavior may exist due to the user base of different types of forums. Therefore, we propose to extend this study to other types of forum platforms, and add a mix of popular and unpopular forums in order to study the problem in greater detail in the future.

Acknowledgments

We would like to thank Steven Murdoch from University of Cambridge for providing the data and useful comments on this work. We also thank Google for allowing high-volume programmatic access to Google searches through the Google University Research Program and IPinfoDB for the IP address geolocation database. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. CNS-0831988. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

[1] "December 2009 Web Server Survey," http://news.netcraft.com/archives/2009/12/24/december_2009_web_server_survey.html.

[2] "Internet 2009 in numbers," http://royal.pingdom.com/2010/01/22/internet-2009-in-numbers/.

[3] D. Zhou, C. J. Burges, and T. Tao, "Transductive link spam detection," in *WWW International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, May 2007.

[4] Q. Gan and T. Suel, "Improving web spam classifiers using link structure," in *WWW International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, May 2007.

[5] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: Web spam detection using the web topology," in *ACM Special Interest Group on Information Retrieval (SIGIR) Conference*, July 2007.

[6] J. Abernethy, O. Chapelle, and C. Castillo, "Web spam identification through content and hyperlinks," in *WWW International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, May 2008.

[7] A. Bhattarai, V. Rus, and D. Dasgupta, "Characterizing comment spam in the blogosphere through content analysis," in *IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*, 2009.

[8] "XRumer," http://www.botmasternet.com.

[9] "WordPress," http://wordpress.org.

[10] "Akismet," http://akismet.com.

[11] "MessageLabs intelligence: March 2010," http://www.messagelabs.com/mlireport/ MLI_2010_03_Mar_FINAL-EN.pdf.

[12] Y. Niu, Y.-M. Wang, H. Chen, M. Ma, and F. Hsu, "A quantitative study of forum spamming using context-based analysis," in *Internet Society Annual Network and Distributed System Security Symposium (NDSS)*, February 2007.

[13] "IANA root zone database," http://www.iana.org/domains/root/db/.

[14] "The domain name industry brief, volume 7, issue 2, june 2010," http://www.verisign.com/domain-name-services/domain-information-center/domain-name-resources/domain-name-report-june10.pdf.

[15] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, "Your botnet is my botnet: Analysis of a botnet takeover," in *ACM Conference on Computer and Communications Security (CCS)*, 2009.

[16] "Route views project," http://www.routeviews.org.

[17] "IPInfoDB," http://www.ipinfodb.com.

[18] D. K. McGrath and M. Gupta, "Behind Phishing: An Examination of the Phisher Modi Operandi," in *USENIX Workshop on Large-Scale Exploits and Emergent Threats*, 2008.

[19] Alexa Internet, Inc, "Alexa top sites," 2010. [Online]. Available: http://www.alexa.com/topsites

[20] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *WWW*, May 2006.

[21] "SVMlight support vector machine," http://svmlight.joachims.org.

[22] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Join Conference on Artificial Intelligence (IJCAI)*, 1995.

[23] S. Hao, N. A. Syed, N. Feamster, A. G. Gray, and S. Krasser, "Detecting spammers with SNARE: Spatio-temporal network-level automatic reputaion engine," in *USENIX Security Symposium*, 2009.

[24] Z. Qian, Z. M. Mao, Y. Xie, and F. Yu, "On network-level clusters for spam detection," in *Internet Society Annual Network and Distributed System Security Symposium (NDSS)*, 2010.

[25] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement," in *WWW International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, May 2005.

[26] P. Kolari, T. Finin, and A. Joshi, "SVMs for the blogosphere: Blog identification and splog," in *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[27] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in *Internet Society Annual Network and Distributed System Security Symposium (NDSS)*, 2010.

[28] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious URLs: An application of large-scale online learning," in *International Conference on Machine Learning*, 2009.

[29] ——, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.