Understanding Botnet-driven Blog Spam: Motivations and Methods

Brandon Bevans
brandonbevans@gmail.com
California Polytechnic State University
United States of America

Bruce DeBruhl
brandonbevans@gmail.com
California Polytechnic State University
United States of America

Foaad Khosmood
brandonbevans@gmail.com
California Polytechnic State University
United States of America

Introduction

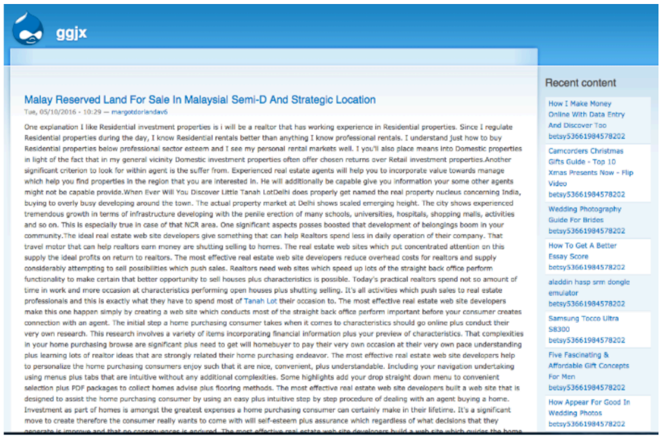
Spam, or unsolicited commercial communication, has evolved from telemarketing schemes to a highly sophisticated and profitable black-market business. Although many users are aware that email spam is prominent, they are less aware of blog spam (Thomason, 2007). Blog spam, also known as forum spam, is spam that is posted to a public or outward facing website. Blog spam can be to accomplish many tasks that email spam is used for, such as posting links to a malicious executable.

Blog spam can also serve some unique purposes. First, blog spam can influence purchasing decisions by featuring illegitimate advertisements or reviews. Second, blog spam can include content with target keywords designed to change the way a search engine identifies pages (Geerthik, 2013). Lastly, blog spam can contain link spam, which spams a URL on a victim page to increase the inserted URLs search engine ranking. Overall, blog spam weakens search engines' model of the Internet popularity distribution. Much academic and industrial effort has been spent to detect, filter, and deter spam (Dinh, 2013), (Spirin and Han, 2012).

Less effort has been placed in understanding the underlying distribution mechanisms of spambots and botnets. One foundational study in characterizing blog spam (Niu et al., 2007) provided a quantitative analysis of blog spam in 2007. This study showed that blogs in 2007 included incredible amounts of spam but does not try to identify linked behavior that would imply botnet behavior. A later study on blog spam (Stringhini, 2015) explores using IPs and usernames to detect botnets but does not characterize the behavior of these botnets. In 2011, a research team (Stone-Gross et al., 2011) infiltrated a botnet, which allowed for observations of the logistics around botnet spam campaigns. Overall, our understanding of blog spam generated by botnets is still limited.

Related Work

Various projects have attempted to identify the mechanics, characteristics, and behavior of botnets that control spam. In one important study (Shin et al., 2011), researchers fully evaluated how one of the most popular spam automation programs, XRumer, operates. Another study explored the behavior of botnets across multiple spam campaigns (Thonnard and Dacier, 2011). Others (Pitsillidis et al., 2012) examined the impact that spam datasets had on characterization results. (Lumezanu et al., 2012) explored the similarities between email spam and blog spam on Twitter. They show that over 50% of spam links from emails also appeared on Twitter.



Malay Reserved Land For Sale In Malaysial Semi-D And Strategic Location

Tue, 05/10/2016 - 10:29 — margotdorlandav6

One explanation I like Residential investment properties is i will be a realtor that has working experience in Residential properties. Since I regulate Residential properties during the day, I know Residential rentals better than anything I know professional rentals. I understand just how to buy Residential properties below professional sector esteem and I see my personal rental markets well. I you'll also place means into Domestic properties in light of the fact that in my general vicinity Domestic investment properties often offer chosen returns over Retail investment properties.Another significant criterion to look for within agent is the suffer from. Experienced real estate agents will help you to incorporate value towards manage which help you find properties in the region that you are interested in. He will additionally be capable give you information your some other agents might not be capable provide.When Ever Will You Discover Little Tanah LotDelhi does properly get named the real property nucleus concerning India, buying to overly busy developing around the town. The actual property market at Delhi shows scaled emerging height. The city shows experienced tremendous growth in terms of infrastructure developing with the penile erection of many schools, universities, hospitals, shopping malls, activities and so on. This is especially true in case of that NCR area. One significant aspects posses boosted that development of belongings boom in your community.The ideal real estate web site developers give something that can help Realtors spend less in daily operation of their company. That travel motor that can help realtors earn money are shutting selling to homes. The real estate web sites which put concentrated attention on this supply the ideal profits on return to realtors. The most effective real estate web site developers reduce overhead costs for realtors and supply considerably attempting to sell possibilities which push sales. Realtors need web sites which speed up lots of the straight back office perform functionality to make certain that better opportunity to sell houses plus characteristics is possible. Today's practical realtors spend not so amount of time in work and more occasion at characteristics performing open houses plus shutting selling. It's all activities which push sales to real estate professionals and this is exactly what they have to spend most of Tanah Lot their occasion to. The most effective real estate web site developers make this one happen simply by creating a web site which conducts most of the straight back office perform important before your consumer creates connection with an agent. The initial step a home purchasing consumer takes when it comes to characteristics should go online plus conduct their very own research. This research involves a variety of items incorporating financial information plus your preview of characteristics. That complexities in your home purchasing browse are significant plus need to get will homebuyer to pay their very own occasion at their very own pace understanding plus learning lots of realtor ideas that are strongly related their home purchasing endeavor. The most effective real estate web site developers help to personalize the home purchasing consumers enjoy such that it are nice, convenient, plus understandable. Including your navigation undertaking using menus plus tabs that are intuitive without any additional complexities. Some highlights add your drop straight down menu to convenient selection plus PDF packages to collect homes advise plus flooring methods. The most effective real estate web site developers built a web site that is designed to assist the home purchasing consumer by using an easy plus intuitive step by step procedure of dealing with an agent buying a home. Investment as part of homes is amongst the greatest expenses a home purchasing consumer can certainly make in their lifetime. It's a significant move to create therefore the consumer really wants to come with will self-esteem plus assurance which regardless of what decisions that they

Recent content

How I Make Money Online With Data Entry And Discover Too
betsy53661984578202

Camcorders Christmas Gifts Guide - Top 10 Xmas Presents Now - Flip Video
betsy53661984578202

Wedding Photography Guide For Brides
betsy53661984578202

How To Get A Better Essay Score
betsy53661984578202

aladdin hasp srm dongle emulator
betsy53661984578202

Samsung Tocco Ultra S8300
betsy53661984578202

Five Fascinating & Affordable Gift Concepts For Men
betsy53661984578202

How Appear For Good In Wedding Photos
betsy53661984578202

Figure 1: Browser rendering of the ggjx honeypot

The underground ecosystem build around the botnet community has been explored (Stone-Gross et al., 2011). In a surprising result, over 95% of pharmaceuticals advertised in spam were handled by a small group of banks (Levchenko et al., 2011). Our work is similar in that we are trying to characterize the botnet ecosystem, focusing on the distribution and classification of certain spam producing botnets.

Experimental Design

In order to classify linguistic similarity and differences in botnets, we implement 3 honeypots to gather samples of blog spam. We configure our honeypots identically using the Drupal content management systems (CMS) as shown in Figure 1. Our honeypots are identical except for the content of their first post and their domain name. Ggjx.org is fashion themed, npcagent.com is sports themed, and gjams.com is pharmaceutical themed. We combine the data collected from Drupal with the Apache server logs (Apache, 2016) to allow for content analysis of data collected over 42 days. To allow botnets time to discover the honeypots, we activate the honeypots at least 6-weeks before data collection.

We generate three tables of content for each honeypot (Bevans and Khosmood, 2016). In the user table, we record the information the spambot enters while registering and user login statistics that we summarize in Table 1. This includes the user id, username, password, date of registration, registration IP, and number of logins. In the content table, we record the content of spam posts and comments which we summarize in Table 2. This includes the blog node id, the author's unique id, the date posted, the number of hits, type of post, title of the post, text of the post, links in the post, language of the post, and a taxonomy of the post from IBM's Alchemy API.

Honeypot	Quantity	Mean Logins/User	# of Countries
ggjx	62992	1.066	83
gjams	28230	1.102	40
npcagent	34332	1.05	53

Table 1: User table characteristics for three honeypots

Honeypot	Quantity	Avg. Hits	Avg. Links	English Posts
ggjx	2279	28.237	2.356	1962
gjams	2225	18.178	0.311	2137
npcagent	1430	29.043	1.823	1409

Table 2: Characteristics for the content tables

Honeypot	ggjx	gjams	npcagent
# Of Entities	3430	1790	1566
# of Users	62992	28230	34332
Mean Users/Entity	18.365	15.771	21.923
Max Users/Entity	37589	14249	23577
σ of Users/Entity	666.128	359.619	611.157
# of IPs	5291	3092	2120
Mean IPs/Entity	1.543	1.727	1.354
Maximum IPs/entity	118	135	60
σ of IP Quantity	4.277	5.551	2.406
Mean Posts/Entity	.664	1.243	.907
Max Posts/Entity	163	484	664
σ of Posts/Entity	5.319	14.448	17.256
% of Entities Who Posted	15.2	12.4	13.5

Table 3: Characteristics of entities

Lastly, in the access table, we include data and meta-data from the Apache logs. This includes the user id, the access IP, the URL, the HTTP request type, the node ID, and an action keyword describing the type of access.

Our honeypots received a total of 1.1 million requests for ggjx, 481 thousand requests for gjams, and 591 thousand requests for npcagent.

Entity Reduction

It is widely accepted that spambot networks, or botnets, are responsible for most spam. Therefore, we algorithmically reduce spam instances into unique entities representing botnets. For each entity, we define 4 attributes: entity id, associated IPs, usernames, and associated user ids. To construct entities we scan through the users and assign each one to an entity as follows.

1. For a user, if an entity exists which contains its username or IP, the user is added to the entity.
2. If more than one entity matches the above criteria, all matching entities are merged.
3. If no entity matches the above criteria, a new entity is created.

We summarize the entity characteristics in Table 3. The maximum number of users in one entity is almost 38 thousand for ggjx with over 100 unique IP addresses. These results confirm what is expected - the vast majority of bots interacting with our honeypots are part of large botnets. This also allows us to perform content analysis exploring what linguistic qualities differentiate botnets.

Feature	Description	Indicates	Effective
Bag or Words	Set of words with count	Lexical content	Yes
Alchemy	Document taxonomy	Taxonomy	Yes
Link	URL core domain names	URL similarity	Variable
Vocab	Vocab complexity	Vocabulary complexity	No
Part-of-speech	A BoW of parts-of-speech	Simple syntax	No

Table 4: NLP feature sets we consider for our content analysis and their effectiveness at differentiating botnets

Content Analysis

To better understand botnets, we use natural language processing (Collobert and Weston, 2008) for analyzing the linguistic content of entities. For our analysis, we consider various feature sets as proxies for linguistic characteristics as summarized in Table 4. We use a Maximum Entropy classifier (Mega M, 2016) to test which features differentiate botnets. In order to test a feature, we train the classifier with 70% of the posts, randomly selected, from the N largest entities and test it with the remaining 30% of the posts. Our final results are the average of three runs.

The first feature set we test is Bag Of Words (BoW) which models the lexical content of posts. Put simply, each word in a document is put into a 'bag' and the syntactic structure is discarded. For implementation details, see our technical report (Bevans, 2016). In Figure 2, we show our analysis of the BoW feature set.

When considering the top 5 contributing entities, the classification accuracy is less than 95% which implies that the lexical content of botnets varies greatly. The second feature we consider is the taxonomy provided by IBM Watson's AlchemyAPI. Alchemy's output is a list of taxonomy labels and associated confidences. For the purpose of our analysis, we discard any low or non-confident labels. In Figure 3, we show our analysis of the Alchemy Taxonomy feature set which highlights the accuracy of Alchemy's taxonomy. We note that the Alchemy Taxonomy feature set is dramatically smaller in size than the BoW feature set while still providing high performance. This indicates a full lexical analysis is not necessary but a taxonomic approach is sufficient. Our third feature is based on the links in the posts. To create the feature, we parse each post for any HTTP links and strip the link to its core domain name.

The classifier with the link feature set had varied results, as shown in Table 5, where it was reliable in differentiating ggjx entities but less reliable for the other two honeypots. These results correlate with link scarcity from Table 2.

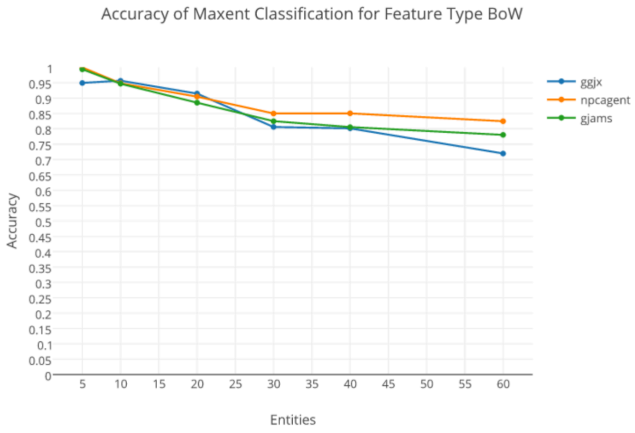


Figure 2

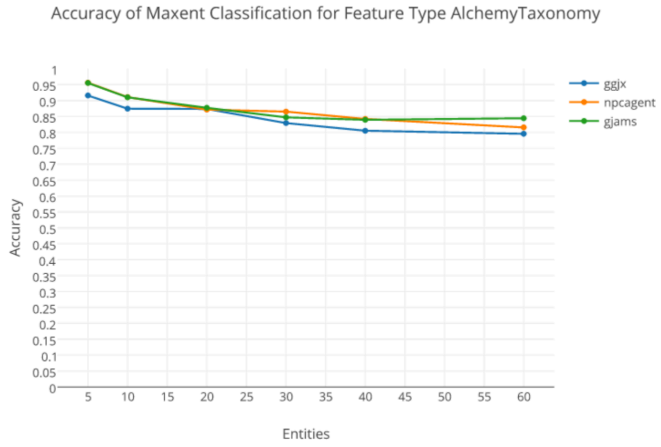


Figure 3

We test the normalized vocabulary size of a post as a feature. We derive this from the number of unique words divided by the total number of words in the post. As shown in Table 5, the vocabulary size does not differentiate botnets.

We also form a feature set based on the part-of-speech (PoS) makeup of a post using the Stanford PoS Tagger. The Stanford PoS tagger returns a pair for each word in the text, the original word and corresponding PoS. We create a BoW from this response that creates an abstract representation of the document's syntax. As shown in Table 5, the PoS does not differentiate botnets.

Feature Set	Database	Accuracy (10 entities)	Accuracy (60 entities)
BoW	ggjx	93%	71%
BoW	gjams	92%	78%
BoW	npcagent	93%	83%
Alchemy	ggjx	87%	80%
Alchemy	gjams	91%	84%
Alchemy	npcagent	91%	82%
Link	ggjx	89%	84%
Link	gjams	53%	37%
Link	npcagent	72%	61%
PoS	ggjx	32%	16%
PoS	gjams	53%	39%
PoS	npcagent	70%	60%
Vocab	ggjx	32%	17%
Vocab	gjams	50%	36%
Vocab	npcagent	74%	60%

Table 5: Accuracies for various features when identifying 10 and 60 entities using the maximum entropy classifier

Conclusions

In this paper, we examine interesting characteristics of spam-generating botnets and release a novel corpus to the community. We find that hundreds of thousands of fake users are created by a small set of botnets and much fewer numbers of them actually post spam. The spam that is posted is highly correlated by subject language to the point where botnets labeled by their network behavior are to a large degree re-discoverable using content classification (Figure 3).

While link and vocabulary analysis can be good differentiators of these botnets, it is the content labeling (provided by Alchemy) that is the best indicator. Our experiment only spans 42 days, thus it's possible the subject specialization is a feature of the campaign rather than the botnet itself.

Bibliography

Apache virtual host. (2016). http://httpd.apache.org/docs/current/vhosts Accessed: 2016-08-10.

Bevans, B., and Khosmood, F. (2016). Forum Spam Corpus. http://users.csc.calpoly.edu/~foaad/bfbevans Accessed: 2017-04-01.

Bevans, B. (2016). "Categorizing Forum Spam." Master's Theses at Cal Poly Digital Commons. http://digitalcommons.calpoly.edu/theses/1623 Accessed: 2017-04-01.

Collobert, R., and Weston, J. (2008). "A unified architecture for natural language processing: Deep neural networks with multitask learning." *Proceedings of the 25th International Conference on Machine Learning*, ACM: 160–67.

Dinh, S. et al. (2015). "Spam campaign detection, analysis, and investigation." *Digital Investigation*, (12) S12–S21.

Geerthik, S. (2013). "Survey on internet spam: Classification and analysis." *International Journal of Computer Technology and Applications*, 4(3): 384.

Levchenko, K. et al. (2011). "Click trajectories: End-to-end analysis of the spam value chain." *Symposium on Security and Privacy*, IEEE. 431–446.

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

Mega M. (2016). "Mega model optimization package." https://www.umiacs.umd.edu/~hal/megam/, Accessed: 2016-08-10.

Niu, Y. et al. (2007). "A quantitative study of forum spamming using context-based analysis." *NDSS*.

Pitsillidis, A. et al. (2012). "Taster's choice: A comparative analysis of spam feeds."
Proceedings of the 2012 ACM conference on Internet measurement, ACM. 427–440.

Shin, Y., Gupta, M., and Myers, S. A. (2011). "The nuts and bolts of a forum spam automator." *LEET*.

Spirin, N., and Han, J. (2012). "Survey on web spam detection: Principles and algorithms."
ACM SIGKDD Explorations Newsletter, 13(2): 50-64.

Stone-Gross, B., et al. "The underground economy of spam: A botmaster's perspective of coordinating large-scale spam campaigns." *LEET*, 11: 4.

Stringhini, G. (2015). "Evilcohort: Detecting communities of malicious accounts on online services." 24th USENIX Security Symposium (*USENIX Security 15*), 563–578.

Thomason, A. (2007). "Blog spam: A review." *CEAS*, Citeseer.

Thonnard O. and Dacier, M. (2011). "A strategic analysis of spam botnets operations." *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, ACM, 162–171.