Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media

Network Analysis of Recurring YouTube Spam Campaigns

Derek O'Callaghan, Martin Harrigan, Joe Carthy, Pádraig Cunningham
School of Computer Science & Informatics, University College Dublin
{derek.ocallaghan,martin.harrigan,joe.carthy,padraig.cunningham}@ucd.ie

Abstract

As the popularity of content sharing websites has increased, they have become targets for spam, phishing and the distribution of malware. On YouTube, the facility for users to post comments can be used by spam campaigns to direct unsuspecting users to malicious third-party websites. In this paper, we demonstrate how such campaigns can be tracked over time using network motif profiling, i.e. by tracking counts of indicative network motifs. By considering all motifs of up to five nodes, we identify discriminating motifs that reveal two distinctly different spam campaign strategies, and present an evaluation that tracks two corresponding active campaigns.

Introduction

The usage and popularity of content sharing websites continues to rise each year. For example, YouTube now receives more than four billion views per day, with sixty hours of video being uploaded every minute; increases of 30% and 25% respectively over the prior eight months[1]. Unfortunately, such increases have also resulted in these sites becoming more lucrative targets for spammers hoping to attract unsuspecting users to malicious websites. Opportunities exist for the abuse of the YouTube facility to host discussions in the form of video comments (Sureka 2011), given the availability of bots that can be used to post spam comments in large volumes.

Our investigation has found that bot-posted spam comments are often associated with orchestrated campaigns that can remain active for long periods of time, where the primary targets are popular videos. An initial manual analysis of data gathered from YouTube revealed activity from a number of campaigns, two of which can be seen in Figure 1. As an alternative to traditional approaches that attempt to detect spam on an individual (comment) level (e.g. domain blacklists), this paper presents an evaluation of the detection of these recurring campaigns using network analysis, based on networks derived from the comments posted by users to videos. This approach uses the concept of *network motif profiling* (Milo et al. 2004), where motif counts from the derived networks are tracked over time. The objective is to discover certain discriminating motifs that can be used to identify particular strategies and the associated users as they periodically recur.

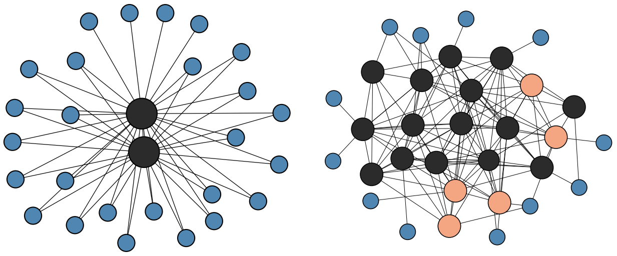

Figure 1: Strategies of two spam campaigns targeting YouTube since 2011. Smaller (blue) nodes are videos, larger nodes are spam user accounts, with dark-grey nodes being those with comments marked as spam, and lighter (orange) nodes being those whose comments were not marked accordingly.

This paper begins with a description of related work in the domain. Next, the methodology used by the detection approach is described, from derivation of the comment-based networks to the subsequent network motif profile generation. The results of an evaluation from a seventy-two hour time period are then discussed. These results demonstrate the use of certain discriminating motifs to identify strategies employed by two separate campaigns we have discovered within this time period. Finally, the overall conclusions are discussed, and some suggestions for future work are made.

Related Work

In relation to analysis of spam within YouTube, Benevenuto et al. (2009) created a directed network based on videos and associated responses, and used derived features as part of a larger set for spam user detection. Separate spam investigations include that of Sureka (2011). Regarding other online social networking websites, Gao et al. (2010) investigated spam within Facebook "wall" messages, using networks based on message similarity. The shortcomings of URL blacklists for Twitter spam prevention were highlighted by Grier et al. (2010), with related analysis of short-

[1] http://youtube-global.blogspot.com/2012/01/holy-nyans-60-hours-per-minute-and-4.html

ened URLs by Chhabra et al. (2011).

Network motifs (Milo et al. 2004) are structural patterns in the form of interconnected n-node subgraphs that are considered to be inherent in many varieties of network, indicating certain network characteristics while also permitting network comparison using *significance* profiles. They have also been used in the detection of spam. Boykin and Roychowdhury (2005) found that a low clustering coefficient (number of triangle structures) may indicate the presence of spam within email address networks. Becchetti et al. (2008) used the number of triangles and clustering coefficient in the detection of web spam. Motifs of size three (triads) have also been used to detect spam comments in blog-interaction networks (Kamaliha et al. 2008).

Methodology

Comment processing and network generation

A data set[2] was collected by periodically retrieving the details and comments of popular YouTube videos on a continual basis, as this subset of videos generally has a higher probability of attracting attention from spammers. From this data, we can generate a network to represent the comment posting activity of users during a specified time period. Having selected a set of comments, a number of pre-processing steps must be executed beforehand, as spammers often try to obfuscate comment text in order to bypass detection by any filters. To counteract these efforts, each comment is converted to a set of tokens, from which a modified comment text is generated (O'Callaghan et al. 2012).

Using these modified comment texts, we generate an undirected network with two categories of node, *users* and *videos*. An unweighted edge is created between a user and a video if at least one comment has been posted by the user to the video. To capture the relationship between the users involved in a particular spam campaign, unweighted edges are created between user nodes based on the similarity of their comments. Each modified comment text is converted to a set of hashes using the Rabin-Karp rolling hash method, with a sliding window length of 3. A pairwise distance matrix, based on Jaccard distance, can then be generated from these comment hash sets. For each pairwise comment distance below a threshold, an edge is created between the corresponding users if one does not already exist. Following experiments with values ranging from 0.2 to 0.8, a value of 0.6 was chosen for this threshold as it ensured increased connectivity between spam users without an over-abundance of edges between regular users.

Any users whose set of adjacent nodes consists solely of a single video node are removed. Since these users have commented on only one video, and are in all likelihood not related to any other users, they are not considered to be part of any spam campaign. The resulting network tends to consist of one or more large connected components, with a number of considerably smaller connected components based on videos with a relatively minor amount of comment activity. Finally, an approximate labelling of the user nodes is performed, where users are labelled as spam users if they posted at least one comment whose *spam hint* property is set to true. All other users are labelled as regular users.

Network motif profiles

Once the network has been generated, a set of *egocentric* networks can be extracted. In this context, an *ego* is a user node, where its egocentric network is the induced *k-neighbourhood* network consisting of those user and video nodes whose distance from the ego is at most *k* (currently 2). Motifs from size three to five within the egocentric networks are then enumerated using FANMOD (Wernicke and Rasche 2006). A profile of motif counts is maintained for each ego, where a count is incremented for each motif found that contains the ego. As the number of possible motifs can be relatively large, the length of this profile will vary for each network generated from a selection of comment data, rather than relying upon a profile with a (large) fixed length. For a particular network, the profiles will contain an entry for each of the unique motifs found in the entirety of its constituent egocentric networks. Any motifs not found for a particular ego will have a corresponding value of zero.

As an alternative to the *significance* profiles proposed by Milo et al. (2004), a *ratio* profile rp (Wu, Harrigan, and Cunningham 2011) is created for each ego, thus permitting the comparison of egocentric networks with each other. In this profile, the ratio value for a particular motif is based on the counts from all of the egocentric networks, i.e.:

$$rp_i = \frac{nmp_i - \overline{nmp_i}}{nmp_i + \overline{nmp_i} + \epsilon} \tag{1}$$

Here, nmp_i is the count of the i^{th} motif in the ego's motif profile, $\overline{nmp_i}$ is the average count of this motif for all motif profiles, and ϵ is a small integer that ensures that the ratio is not misleadingly large when the motif occurs in only a few egocentric networks. To adjust for scaling, a normalized ratio profile nrp is then created for each ratio profile rp with:

$$nrp_i = \frac{rp_i}{\sqrt{\sum rp_j^2}} \tag{2}$$

The generated set of normalized ratio profiles usually contain correlations between the motifs. Principal components analysis (PCA) is used to adjust for these, acting as a dimensionality reduction technique in the process. We can visualize the first two principal components as a starting point for our analysis. This is discussed in the next section.

Evaluation

For the purpose of this evaluation, the experiments were focused upon tracking two particular spam campaigns that we discovered within the data set. The campaign strategies can be seen in Figure 1, i.e. a small number of accounts each commenting on many videos (Campaign 1), and a larger number of accounts each commenting on few videos (Campaign 2). A period of seventy-two hours was chosen where these campaigns were active, starting on November 14th, 2011 and ending on November 17th, 2011.

[2] The data set is available at `http://mlg.ucd.ie/yt`

In order to track the campaign activity over time, this period was split into twelve windows of six hours each. For each of these windows, a network of user and video nodes was derived using the process described in the previous section. A normalized ratio profile was generated for each ego (user), based on the motif counts of the corresponding egocentric network. PCA was then performed on these profiles to produce 2-dimensional spatializations of the user nodes, using the first two components. These spatializations act as the starting point for the analysis of activity within a set of time windows.

Visualization and initial analysis

Having inspected all twelve six-hour windows, a single window containing activity from both campaigns has been selected for detailed analysis here, Window 10 from November 17th, 2011 (04:19:32 to 10:19:32). The derived network contained 263 video nodes, 295 user nodes (107 spam) and 907 edges. A spatialization of the first two principal components of the normalized ratio profiles for this window can be found in Figure 2. Users posting at least one comment marked as spam are in dark-grey, all other users are in orange (lighter). Spam campaign users have been highlighted accordingly. From the spatialization, it can be seen that:

1. The vast majority of users appear as overlapping points in the large cluster on the right.
2. There is a clear distinction between the two different campaign strategies, as these points are plotted separately (both from regular users and each other).
3. The inaccuracy of the *spam hint* comment property is demonstrated as the Campaign 2 cluster contains users coloured in orange (lighter), i.e. none of their comments were marked as spam. Conversely, there are users coloured in dark-grey in the large cluster of regular users.

Apart from the highlighted campaign clusters, other spam nodes have been correctly marked as such. For example, the five users ("Other spam users") that are separated from the normal cluster appear to be isolated spam accounts having similar behaviour to the Campaign 1 strategy, but on a smaller scale. They are not considered further during this evaluation as they are not part of a larger campaign. Further analysis of Campaign 2 revealed activity from a large number of users, each posting similar comments that are clearly from the same campaign. For example:

> Don't miss this guys, the CEO of apple is releasing
> ipads on Thursday: osapple.co.nr
>
> dont miss out. November 17 - new apple ceo is shipping
> out old ipad and iphones
> Not a lie. Go to this webpage to see what I mean:
> bit.ly\vatABm

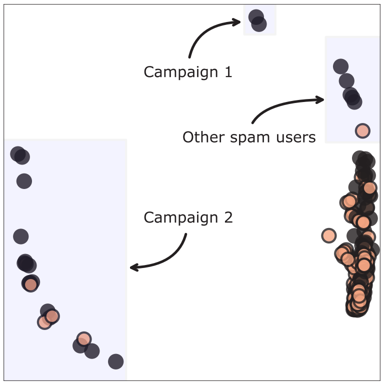

Figure 2: Window 10 spatialization (dark-grey nodes are users with spam comments, lighter (orange) nodes are all other users).

Both of these comments were made by the same user. However, while the first comment was accurately marked as spam, the second was **not**. An assumption here could be that the URL in the first comment is on a spam blacklist, while the shortened URL in the second enables such a list to be bypassed. Similar shortcomings are discussed in earlier work (Chhabra et al. 2011).

Discriminating motifs

An inspection of the individual motif counts found that certain motifs have relatively higher counts for users involved in the spam campaigns, than those found for regular users. These motifs may be considered indicative of different campaign strategies, and a subset can be found in Table 1.

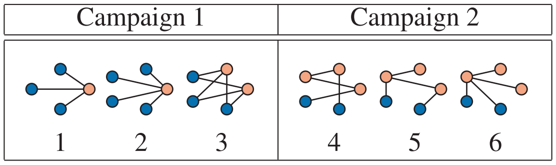

Table 1: Selected discriminating motifs – darker (blue) nodes are videos, lighter (orange) nodes are users.

These discriminating motifs would appear to correlate with the existing knowledge of the campaign strategies. Campaign 1 consists of a small number of users commenting on a large number of videos, and so it would be expected that motifs containing user nodes with a larger number of video node neighbours have higher counts for the users involved, as is the case here. The motifs considered indicative of Campaign 2 are more subtle, but all three highlight the fact that users appear to be more likely to be connected to other users rather than videos. This makes sense given that with this campaign, a larger number of users tend to comment on a small number of videos each, and the potential for connectivity between users is higher given the similarity of their comments. These motifs would also appear to indicate that users in the campaign don't comment on the same videos, as no two users share a video node neighbour.

Figure 3 contains a plot of the counts for a single discriminating motif in each of the six-hour windows. The counts were normalized using the edge count of the corresponding window networks followed by min-max normalization. The fluctuation in counts across the windows appears to track the

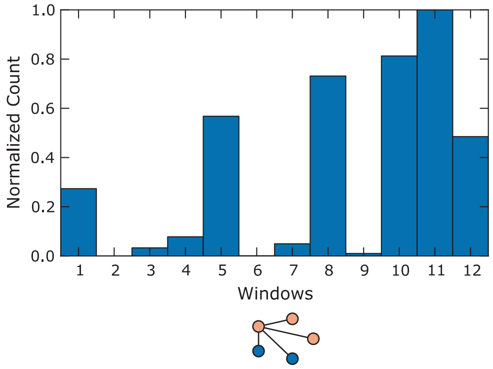

Figure 3: Tracking the recurring activity of Campaign 2 using a single discriminating motif.

recurring periodic activity of this campaign, as confirmed by separate analysis of the data set. This would appear to corroborate the *bursty* nature of spam campaigns (Gao et al. 2010).

Conclusions and Future Work

YouTube spam campaigns typically involve a number of spam bot user accounts controlled by a single spammer targeting popular videos with similar comments over time. We have shown that dynamic network analysis methods are effective for identifying the recurring nature of different spam campaign strategies, along with the associated user accounts. While the YouTube comment scenario could be characterized as a network in a number of ways, we use a representation comprising user and video nodes, user-video edges representing comments and user-user edges representing comment similarity. The discriminating power of these motif-based characterizations can be seen in the PCA-based spatialization in Figure 2. It is also clear from Figure 3 that histograms of certain discriminating motifs show the level of activity in campaign strategies over time.

For future experiments, it will be necessary to annotate the entire data set with labels corresponding to any associated spam campaign strategies, thus permitting evaluation of the motif-based characterization with strategies other than the two discussed in this paper. Feature selection of a subset of discriminating motifs could also be performed which would remove the current requirement to count all motif instances found in the user egocentric networks, as this can be a lengthy process. As the comment similarity distance threshold used to generate edges between users has an influence on the subsequent discovery of discriminating motifs, we will evaluate other means of generating these edges. We will also work to develop a mechanism to detect any evolution in campaign strategies over time.

Acknowledgements

This work is supported by 2CENTRE, the EU funded Cybercrime Centres of Excellence Network and Science Foundation Ireland under grant 08/SRC/I140: Clique: Graph and Network Analysis Cluster.

References

Becchetti, L.; Boldi, P.; Castillo, C.; and Gionis, A. 2008. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, 16–24. New York, NY, USA: ACM.

Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J.; and Gonçalves, M. 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*, SIGIR '09, 620–627. New York, NY, USA: ACM.

Boykin, P., and Roychowdhury, V. 2005. Leveraging social networks to fight spam. *Computer* 38(4):61 – 68.

Chhabra, S.; Aggarwal, A.; Benevenuto, F.; and Kumaraguru, P. 2011. Phi.sh/$ocial: The phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, CEAS '11, 92–101. New York, NY, USA: ACM.

Gao, H.; Hu, J.; Wilson, C.; Li, Z.; Chen, Y.; and Zhao, B. Y. 2010. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th Annual Conference on Internet Measurement*, IMC '10, 35–47. New York, NY, USA: ACM.

Grier, C.; Thomas, K.; Paxson, V.; and Zhang, M. 2010. @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, 27–37. New York, NY, USA: ACM.

Kamaliha, E.; Riahi, F.; Qazvinian, V.; and Adibi, J. 2008. Characterizing Network Motifs to Identify Spam Comments. In *Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, 919–928. Washington, DC, USA: IEEE Computer Society.

Milo, R.; Itzkovitz, S.; Kashtan, N.; Levitt, R.; Shen-Orr, S.; Ayzenshtat, I.; Sheffer, M.; and Alon, U. 2004. Superfamilies of Evolved and Designed Networks. *Science* 303(5663):1538–1542.

O'Callaghan, D.; Harrigan, M.; Carthy, J.; and Cunningham, P. 2012. Network Analysis of Recurring YouTube Spam Campaigns (long version). *CoRR* abs/1201.3783.

Sureka, A. 2011. Mining User Comment Activity for Detecting Forum Spammers in YouTube. *CoRR* abs/1103.5044.

Wernicke, S., and Rasche, F. 2006. FANMOD: A Tool for Fast Network Motif Detection. *Bioinformatics* 22(9):1152–1153.

Wu, G.; Harrigan, M.; and Cunningham, P. 2011. Characterizing Wikipedia Pages using Edit Network Motif Profiles. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*, SMUC '11, 45–52. New York, NY, USA: ACM.