Identifying Video Spammers in Online Social Networks

Fabricio Benevenuto[‡*] fabricio@dcc.ufmg.br
Tiago Rodrigues[‡] tiagorm@dcc.ufmg.br
Virgilio Almeida[‡] virgilio@dcc.ufmg.br
Jussara Almeida[‡] jussara@dcc.ufmg.br
Chao Zhang[†] chao@cis.poly.edu
Keith Ross[†] ross@poly.edu

[‡]Computer Science Department, Federal University of Minas Gerais, Brazil
[†]Polytechnic University, Brooklyn, NY, USA

ABSTRACT

In many video social networks, including YouTube, users are permitted to post video responses to other users' videos. Such a response can be legitimate or can be a *video response spam*, which is a video response whose content is not related to the topic being discussed. Malicious users may post video response spam for several reasons, including increase the popularity of a video, marketing advertisements, distribute pornography, or simply pollute the system.

In this paper we consider the problem of detecting video spammers. We first construct a large test collection of YouTube users, and manually classify them as either legitimate users or spammers. We then devise a number of attributes of video users and their social behavior which could potentially be used to detect spammers. Employing these attributes, we apply machine learning to provide a heuristic for classifying an arbitrary video as either legitimate or spam. The machine learning algorithm is trained with our test collection. We then show that our approach succeeds at detecting much of the spam while only falsely classifying a small percentage of the legitimate videos as spam. Our results highlight the most important attributes for video response spam detection.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; H.4.m [**Information Systems Applications**]: Miscellaneous

General Terms

Human factors, Measurement, Videos

Keywords

social network, video response, video spam

AIRWeb '08, April 22, 2008 Beijing, China.

1. INTRODUCTION

Recently, online social networking services such as Facebook, Wikipedia and YouTube are experiencing a dramatic growth in terms of popularity. In particular, video content is becoming a predominant part of users' daily lives on the Web. By allowing users to generate and distribute their own multimedia content to large audiences, the Web is being transformed into a major channel for the delivery of multimedia. Video pervades the Internet and supports new types of interaction among users, including political debates, video chats, video mail, and video blogs. A number of Web services are offering video-based functions as alternative to text-based ones, such as video reviews for products, video ads and video responses [18]. In particular, the video response feature allows users to converse through video, by creating a video sequence that begins with an opening video and then followed with video responses from fans and detractors.

By allowing users to publicize and share their independently generated content, social video sharing systems may become susceptible to different types of malicious and opportunistic user actions, such as self-promotion, video aliasing and video spamming [6]. We define a video response spam as a video posted as a response to an opening video, but whose content is completely unrelated to the opening video. Video spammers are motivated to spam in order to promote specific content, advertise to generate sales, disseminate pornography (often as an advertisement) or compromise the system reputation.

Spamming has been observed in several different contexts, including email [11], Web search engines [9] and blogs [19]. A number of spam detection techniques exploit characteristics present in the text (e.g., email body, commentaries in a blog) [14]. Moreover, users of such systems can quickly learn to identify some text spams (e.g., URLs to suspect Web sites), skipping or ignoring them. On the other hand, video spamming, particularly in social video sharing systems, can be much more challenging to detect and combat. Content-based detection techniques are not easily applied to non-textual video objects. On the other hand, exploiting characteristics of the traffic to specific videos, such as number of views and number of comments received, may not, by

* Fabricio is supported by UOL (www.uol.com.br), through *UOL Bolsa Pesquisa* program, process number 20080125143100a.

itself, be enough to distinguish spam from unpopular user-generated content. Ultimately, users can not easily identify a video spam before watching at least a segment of it, thus consuming system resources, in particular bandwidth, and compromising user patience and satisfaction with the system. Thus, identifying video spam is a challenging problem in social video sharing systems. However, we are not aware of any video spam detection technique.

This paper gives a first step in this direction. However, instead of identifying video spam content, our goal is to detect *video spammers*, i.e., users who post video spam as responses to other videos. We propose and evaluate a video spammer detection mechanism that classifies a user as a spammer based on the user's profile, the user's social behavior in the system, and the videos the user has posted. These attributes capture characteristics that are inherent to the user behavior and thus may better distinguish legitimate users from malicious video spammers.

In order to design and evaluate our proposed mechanism, we start by crawling a large user data set from YouTube, a pioneer social media sharing system which generates high volumes of Internet traffic and includes many social networking characteristics. A test collection is then built by carefully selecting users from the crawled data and manually classifying each user as either legitimate or spammer. Our test collection consists of 592 users, 119 of which are classified as spammers. We then characterize several user and video attributes from our test collection, selecting those that may better distinguish spammers from legitimate users. The selected attributes are grouped into three subsets: user attributes, social network attributes, and video attributes. The user attributes, extracted from the user profile, expresses how the user typically uses the system (e.g., number of videos uploaded, number of friends, and so on). The social network attributes express how the user interacts with other users through video responses, whereas the video attributes capture the interests of other users in the content posted by user (e.g., number of views or comments to the videos posted by the user). Finally, using our test collection, we evaluate the effectiveness of our detection mechanism using the selected attributes. We also evaluate the relevance to the classification of each subset of attributes.

We found that, using the complete set of attributes, our mechanism can correctly classify a significant fraction of the video spammers (44%) while incurring in 2% of misclassification of legitimate users. We also found that the social network attributes of a user as well as the attributes of the videos he/she posts are the most relevant to spammer detection. In fact, classifying users based only on one of the two subsets has an effectiveness close to when all three subsets of attributes are jointly used. Since video spamming is still an unexplored problem, and, to the best of our knowledge, no previous detection mechanism is available in the literature, we believe these results are significant and point towards a promising direction for future research.

In summary, the main contributions of this paper are:

- Quantitative evidence of video spamming activity (as defined above) in social online video sharing systems, particularly YouTube. To the best of our knowledge, this is the first work to show such evidence.
- The identification and characterization of a set of user and video attributes that can be used to distinguish video spammers from legitimate users.
- A test collection of users from YouTube, classified as spammers or legitimate users.
- A video spammer detection mechanism based on a classification algorithm, which showed to produce reasonably good results, detecting a significant fraction of video spammers with 2% of misclassification of legitimate users.

The rest of the paper is organized as follows. The next section briefly discusses related work. Section 3 describes our Youtube crawling strategy and the test collection built from the crawled dataset. Section 4 presents a characterization of several attributes of the users included in our test collection, particularly those that can be used to distinguish spammers from legitimate users. Section 5 describes and evaluates our video spammer detection mechanism. Finally, Section 6 offers conclusions and directions for future work.

2. RELATED WORK

Mechanisms to detect and identify spam and spammers have been largely studied in the context of Web [5, 13] and email spamming [12]. In particular, Castillo *et al* [5] proposed a framework to detect Web spamming which uses social network metrics. A framework to detect spamming in tagging systems, which is a type of attack that aims at raising the visibility of specific objects, was proposed in [17]. Although applicable to social media sharing systems that allow object tagging by users, such as YouTube, the proposed technique exploits a specific object attribute, i.e., its tags. Our approach is complementary to these efforts as it aims at detecting *video* spammers, using a combination of different categories of attributes of both objects and users.

A survey of approaches to combat spamming in Social Web sites is presented in [14]. Many existing approaches are based on extracting evidence from the content of a text, treating the text corpus as a set of objects with associated attributes and using these attributes to detect spam. These techniques, based on content classification, can be directly applied to textual information, and thus can be used to detect spam in email, text commentaries in blogs, forums, and online social networking sites. Additionally, detection of email spam based on image content was also studied previously [2, 22]. However, content classification is much harder to do for video objects. Our approach to detect video spammers consists on classifying users, instead of the their videos, and relies on a set of attributes associated to the user actions and social behavior in the system as well as attributes of their videos. However, our detection scheme does not involve video content which is difficult to automatically process and analyze. Rather, we focus on attributes that capture the interest of other users on the video (e.g., number of comments posted by other users, number of times the video was viewed, and so on).

Complementary to our effort, the characterization of the traffic to online video sharing systems, in particular YouTube, has also been the focus of some studies. An in-depth analysis of popularity distribution, popularity evolution and content characteristics of YouTube and of a popular Korean video sharing service is presented in [6]. The authors also analyze mechanisms to improve video distribution, such as caching and peer-to-peer distribution schemes. Gill *et al* [10]

present a characterization of the YouTube traffic collected from the University of Calgary campus network and compare its properties with those previously reported for Web and media streaming workloads. Both studies focus on traffic and video characterization. We are not aware of any effort to characterize video spamming. Towards this end, this paper presents a characterization of user and video attributes that can be used to distinguish spammers from legitimate users in YouTube.

3. YOUTUBE MEASUREMENTS

Our ultimate goal is to design a mechanism to classify users of social video sharing systems into legitimate and video spammers, using a set of their attributes and of their contributed videos. Towards this goal, we crawled data from YouTube, one of the most popular social media networking sites today [1]. A test collection, including a sample of the crawled data, was then built and used to evaluate the effectiveness of our classification approach. Section 3.1 describes our crawling strategy, whereas Section 3.2 presents the criteria used to select users for the test collection.

3.1 YouTube Data Collection

YouTube includes several social networking features. Since our focus is on *video response spamming*, we are interested in sampling information about users who participate in video-based interactions. In other words, our crawling strategy is driven by users who have responded to other users by uploading videos. This type of interaction is enabled by the *video response* feature, which allows a registered YouTube user to post a video as response to a pre-existing YouTube video. We say a YouTube video is a *responded video* if it has at least one video response. Similarly, we say a YouTube user is a *responded user* if at least one of its contributed videos is a *responded video*. Finally, we say a YouTube user is a *responsive user* if it has posted at least one video response.

A very natural user graph emerges from video response interactions. At a given instant of time t, let X be the union of all responded users and responsive users. The set X is, of course, a subset of all YouTube users. We denote the *video response user graph* as the directed graph (X, Y), where (x_1, x_2) is a directed arc in Y if user $x_1 \in X$ has responded to a video contributed by user $x_2 \in X$.

In order to obtain a large set of responded and responsive users, and ultimately build a video response user graph, we use the sampling procedure described in Algorithm 1. Using as seeds the list of top-100 most responded videos of all time, provided by YouTube, we followed links of responded videos and video responses, collecting information about each user and its posted videos and video responses. The dataset gathered with this crawler, summarized in Table 1 produces a large (most likely the largest) weakly connected component of (X, Y), and is used as source for building our test collection (next section).

3.2 Test Collection Definition

A test collection, containing a set of YouTube users each pre-classified as legitimate or video spammer, is required to evaluate the effectiveness of our classification approach.

```
      input : A list of users (seeds)
1.1   foreach User U in the crawler list do
1.2       Collect U's info using the YouTube API;
1.3       Collect U's video list using the API;
1.4       foreach Video V in the video list do
1.5           Copy the HTML of V;
1.6           if V is a responded video then
1.7               Copy the HTML of V's video
                  responses;
1.8               Insert the responsive users in the
                  crawler list;
1.9           end
1.10          if V is a video response then
1.11              Insert the responded user in the
                  crawler list;
1.12          end
1.13      end
1.14  end
```

Algorithm 1: Crawler Algorithm for Video Responses

However, as far as we know, no such collection is publicly available (neither for YouTube nor for any other video sharing system). But how do we create a large and representative test collection? Relying on random sampling to select a reasonable number of users from the crawled data would not be advisable as it could yield a very small fraction of spammers, preventing a sound analysis of the results.

Thus, we define three strategies, described below, that aim at not only selecting different types of legitimate users but also include users who are more likely to be spammers. Each selected user is then classified as either spammer or legitimate user. We define as a spammer a user who posts at least *one* video response that is considered unrelated to the responded video. Examples of video responses that are considered unrelated to the responded video, and thus, are classified as video spams are: (1) an advertisement of a product or website completely unrelated with the topic of the responded video, (2) pornographic content posted as response to a cartoon video, and (3) videos with no content (with duration equal to 0 seconds), probably posted by automatic tools.

The definition of a video spammer is thus subjective, as it relies on human judgment as to whether a video is related to another. In order to minimize the impact of human error, three volunteers independently classified each user. All users (and their videos and video responses) were analyzed and independently classified by two volunteers. The third

Characteristic	Video Response Dataset
Sample Period	01/11/2008 - 01/18/2008
video-data	
# video responses	701,950
# responded videos	381,616
# views of video responses	5,397,904,689
# views of responded videos	16,721,814,009
User-data	
# users collected	264,460

Table 1: Summary of Video Response Data Set

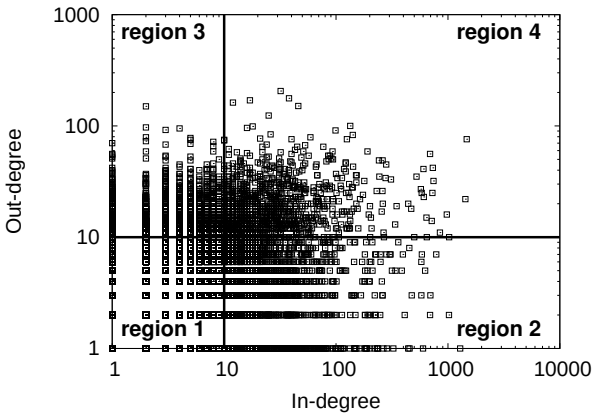

Figure 1: Classes of Users Based on Interactions via Video Responses

one was used whenever the two classification results differ. In case of doubt of whether a video response was or not related to the responded video, the volunteers were instructed to classify it as a legitimate response (i.e., non-spam). As an example, video responses with people chatting or expressing their opinions were classified as legitimate, as we choose the conservative approach of not evaluating the expressed opinions.

The three strategies used to build our test collection are:

1. In order to select users with different levels of interaction through video responses, we consider the video response user graph built from our entire dataset, as described in Section 3. Figure 1 shows the in-degree and out-degree of each user. We define four regions in this graph, representing four different classes of users with different levels of interaction. Region 1 consists of users with low in-degree and out-degree, and thus who respond and are responded by only a few other users in YouTube. They, thus, present a very low level of video-based interaction with other users. Region 2 consists of users with high in-degree and low out-degree. These users receive video responses from a large number of other users, but post responses to only a small number of peers, acting as authorities [16]. Region 3 consists of users with low in-degree and high out-degree, acting as hubs [16]. Intuitively, region three is the most likely region to have video spammers, who try to disseminate their content by posting video responses to many other users. Lastly, region 4 consists of very interactive users, with high in-degree and out-degree. The four regions were defined considering a threshold equal to 10 for both in-degrees and out-degrees. We then randomly selected 100 users from each region[1].

   Out of the 400 selected users, 381 were manually classified as legitimate, and only 11 were classified as spammers. The remaining 8 users were not included in our collection as they had had their accounts suspended by YouTube due to violation of terms of use. The spammers found by this strategy are spread across the four regions as follows: 3 are in region 1, 1 is in region 2, 5 are in region 3, and 2 were found in region 4.

2. Our second strategy is based on the assumption that a video spammer is more likely to post video responses to the most popular videos in order to make his spam visible to a larger community of users. YouTube provides the ranking of videos according to several criteria such as most viewed and most responded. We choose to randomly select 100 users from those who posted video responses to videos in the top 100 most responded videos of all time. Out of these, 8 were classified as spammers and 92 as legitimate users.

3. Our last strategy was devised to increase the number of spammers in our test collection. It is based on our observation that some spammers can be easily identified by analyzing the thumbnails of the video responses posted to videos occupying top positions in the rankings kept by YouTube. We browsed the video responses posted to the top 100 most responded videos of all time, selecting a large number of suspect users for manual inspection. This strategy led to the insertion of 100 more users classified as spammers in our test collection.

In total, our test collection contains 592 users, out of which 473 were classified as legitimate users and 119 as spammers. Throughout the rest of the paper, we will refer to these users simply by legitimate users and spammers, taking our manual classification as baseline of comparison for evaluating the effectiveness of our spammer detection mechanism. The users in our test collection posted a total of 16,611 video responses to 8,710 different videos.

4. SPAMMERS AND LEGIMATE USERS

Unlike legitimate users of social networking sites, people who spam aim at commercial intent (e.g., advertising), self-promotion, and belittlement of ideas and reputation [14]. Thus, the behavior of spammers differs from that of legitimate users. This section presents characteristics that underscore the differences between the two classes of users. Initially, we study characteristics related to the individual behavior of users, such as: number of videos watched, number of subscriptions, and number of videos added as favorites. Intuitively, we expect that legitimate users spend more time interacting with YouTube interfaces, doing actions like choosing friends, watching and uploading videos, and setting favorite videos. In order to verify this intuition, we looked at the characteristics of the users of the test collection. Figure 2 shows the cumulative distribution function for three individual characteristics: number of friends, number of favorites and number of uploads. We notice from the figure that legitimate users do exhibit a higher level of interaction with the system. They have more friends, they contribute with more videos and they have larger lists of favorites. For example, 19% of the legitimate users have less than 10 friends while 56% of the spammers have less than 10 friends.

We also looked at the quality of the contributions made by the two classes of users. Each video uploaded by a user has a set of attributes such as number of views, number of video responses, number of comments received, number of times

[1]Note that the numbers of users falling into regions 1, 2, 3 and 4 are 162,546, 2,333, 3,189 and 1,154, respectively. Thus, randomly selecting users from each region yields a sample biased towards region 4

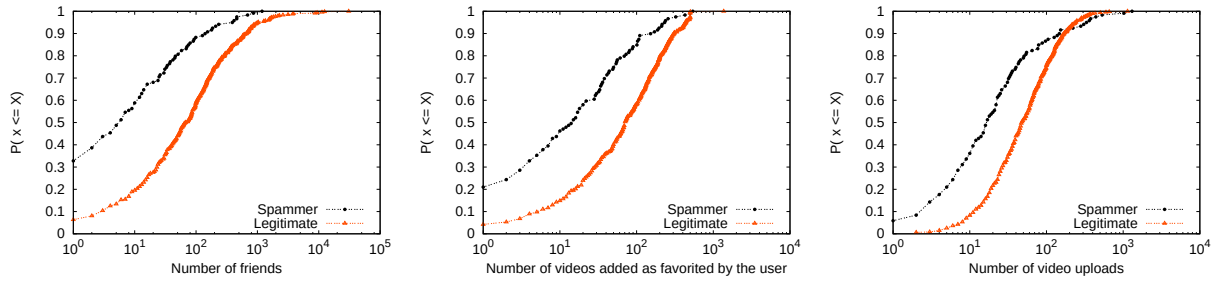

Figure 2: Number of Friends (left). Number of favorite videos (center). Number of uploads (left).

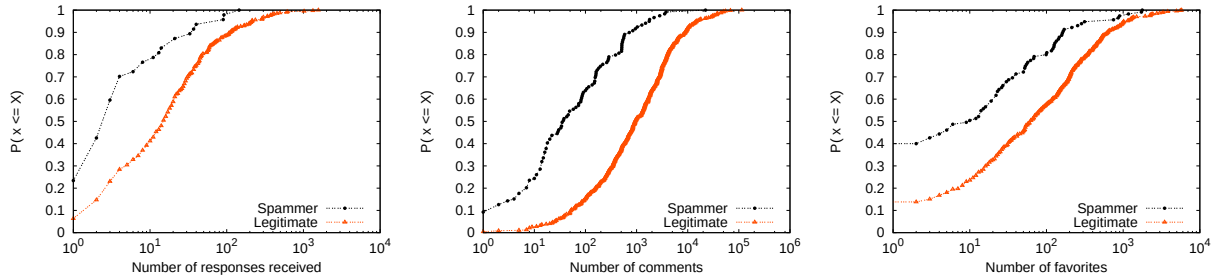

Figure 3: Number of video responses received (left). Number of videos watched (center). Distribution of the number of favorites of the video responses (right)

it was selected as favorite, among others. These attributes can be viewed as proxies for the quality of the user generated content. In order to have a better understanding of the quality of the contributions, we calculate the aggregated characteristics for two sets of videos: all videos uploaded by the user and video responses only. The reason for the aggregation into two classes stem from the fact that the video response is a mechanism used by spammers to distribute their video spam. Figure 3 shows the cumulative function for number of video responses received by all videos, number of comments received by all videos uploaded by the user and number of times the video responses were selected as favorites by other users. The three plots of the figure show how other users "view" the quality of videos contributed by the two classes of users. Videos uploaded by spammers receive fewer comments and fewer video responses than those contributed by legitimate users. Furthermore, we see that the number of times a video response is selected as a favorite is much smaller for video responses posted by spammers. In the test collection, 87% of the video responses uploaded by legitimate users are selected at least once as a favorite. The proportion for spammers is 60%. The perceived quality of the videos contributed by a user is also an important discriminator for classifying spammers and legitimate users. Other video characteristics such as number of views, number of honors, and number of links can also provide useful information to discriminate between the two classes of users.

Now we turn to the social characteristics of the users. These characteristics are derived from the structure of the video response graph, which is one of the many possible social networks in YouTube. There are several social metrics associated with the network that could indicate the level of interaction of a user in the social network, including clustering coefficient, reciprocity, in-degree and out-degree distributions.

The clustering coefficient [7] of a node i $cc(i)$ is the ratio of the number of existing edges over the number of all possible edges between i's neighbors. It measures the density of communication, not only between two users but among neighbors of neighbors. Figure 4 (center) shows the cumulative distribution of the clustering coefficient. As we can see legitimate users have higher clustering coefficient than video spammers. The average clustering coefficient over the whole network is 0.050 for legitimate users, whereas the average clustering coefficient for spammers is 0.014. Another interesting metric to observe is the reciprocity of each user. The reciprocity (R) for the video response user graph is given by:

$$R(x) = \frac{|OS(x) \cap IS(x)|}{|OS(x)|} \quad (1)$$

where OS(x) is the set of users that receive a video response from user x and IS(x) is the set of users that send video responses to x. Reciprocity measures the probability of a user receiving a video response from each user he/she sent a video response. Figure 4 (right) shows the fraction of spammers that have reciprocity greater than 0 is very low (i.e., 0.05), while the percentage of legitimate users that have reciprocity greater than 0 is more than 50%. Therefore, spammers are naturally associated with small (but potentially non-zero) reciprocity, whereas legitimate users, whose behavior is characterized by social relationships, are associated with the highest reciprocity.

We also use the Pagerank [4] algorithm, on the video response user graph, to determine the "user rank". Basically, a user has a high rank if the he/she has many incoming links or the user has links coming from highly ranked users. We call the scores computed by the pagerank algorithm as UserRank, which could be used as an indicator of the importance of users in terms of their participation in video interactions. We randomly selected a few users that have ei-

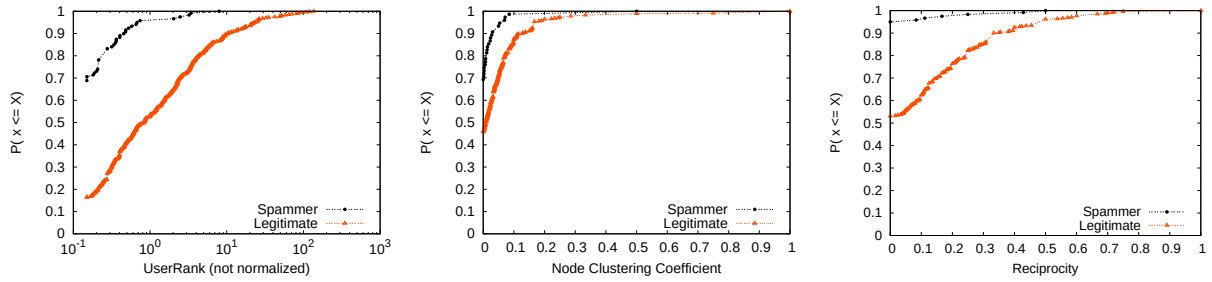

Figure 4: UserRank (left). Node clustering coefficient (center). Reciprocity (right).

ther high rank or low rank. Users with high rank are among the most viewed and most subscribed users. Most of them are directors (a director accounts have special advanced options in YouTube). Low rank users have small number of views and some of them exhibit some of the characteristics already discussed in this section, namely, they post video responses to many videos but receive no video responses from the video community. The left part of Figure 4 confirms the legitimate users have higher pagerank than spammers.

These differences suggest mechanisms to differentiate legitimate users from spammers on the basis of network structure and individual characteristics. In the next section we propose using these differences as the basis of the video spammer classification algorithm.

5. SPAMMER DETECTION MECHANISM

Our spammer detection method relies on a machine learning approach for classifying our dataset. In this approach, the classification algorithm "learns" with part of the data and then applies its knowledge to classify users into two types: legitimate or spammers. This section discusses details of the classification scheme for identifying video spammers in an online social network. First, we present the features used by the classifier. We also define a set of metrics that are used to evaluate the efficiency of the classification scheme. We then show the results obtained by the classification scheme when applied to the test collection.

5.1 Features

User-based Features: For each user in our test collection, we associate a number of features that correspond to characteristics of the user profile. The features are: number of videos uploaded, the number of friends, number of videos watched, number of videos added as favorites, number of video responses posted, number of video responses received, number of subscriptions, number of subscribers. We selected 8 attributes based on user characteristics.

Video-Based Features: The video-based features consist of aggregated characteristics of the set of videos uploaded by a user. For each attribute we calculate the total and the average value of the characteristic for the set of videos. We consider two sets of videos: 1) All videos uploaded by the user. 2) Only the video responses uploaded by the user. For both sets, we consider the following attributes: number of views, duration, number of ratings, number of comments, number of favorites, number of honors, number of external links. Considering the two sets, we have 14 attributes. Since we also consider the total and the average value for each attribute, we end up with 28 attributes based on video characteristics.

Social Network Features: We use a number of social network features based on the video response graph: node in-degree, out-degree, clustering coefficient, userrank, betweenness, reciprocity and assortativity. The node assortativity is defined as the ratio between the degree of the node and average degree of its neighbors [5]. We calculate node assortativity for the four types of degree-degree correlations (i.e., in-in, in-out, out-in, and out-out). In total, we have 10 attributes based on social characteristics.

5.2 Spam Metrics

In order to define the metrics used to evaluate the proposed heuristics, we consider the following measures:

		Prediction	
		Legitimate	**Spammer**
True Label	**Legitimate**	a	b
	Spammer	c	d

Let a represent the number of legitimate users correctly classified as legitimate, b the number of legitimate users falsely classified as spammer, c the spammers falsely classified as legitimate, and d the number of spammers correctly classified as spammers. In order to evaluate the classification algorithms, we consider the following metrics, commonly used on Machine Learning and Information Retrieval [3]:

- True positive rate TP, or recall: $R = \frac{d}{c+d}$.
- True negative rate: $TN = \frac{a}{a+b}$.
- False positive rate: $FP = \frac{b}{a+b}$.
- False negative rate: $FN = \frac{c}{c+d}$.
- Accuracy $= \frac{a+d}{a+b+c+d}$
- F-measure: $\mathrm{F} = 2 \cdot \frac{P \cdot R}{P+R}$, where P is the precision $P = \frac{d}{b+d}$.

We report all the metrics listed above since they have direct interpretation in practice. The true positive (or recall) can be understood as the rate at which spammers are predicted to be spammers whereas the true negative is the rate at which legitimate users are predicted as non-spammers. On the other hand, the false positive is the rate at which legitimate users are predicted to be spammers, and false negative the rate at which spammers are predicted as legitimate. The accuracy provides the rate at which the classifier predicts results correctly. We also use the F-measure to

compare results, since it is a standard way of summarizing precision and recall. The maximum value of F-measure is 1, which means that the prediction is perfect.

5.3 Classification

The SVM [20] (Support vector machine) methods are a well-known class of algorithms for data classification. We choose to use the SVM methods as the classifier for our dataset. Basically, SVM performs classification by mapping input vectors to an N-dimensional space. The goal is to find the optimal hyperplane that separates the data into two categories, each one constructed on each side of the hyperplane. We use a binary non-linear SVM with RBF kernel to allow SVM models to perform separations with very complex boundaries. We choose to use the implementation of SVM provided with libSVM [8], an open source SVM package that allows searching for the best classifier parameters (i.e. cost and gamma) in order to define the best SVM configuration for the dataset. Particularly, we use a tool from libSVM called *easy*, which provides a series of optimizations, including normalization of all numerical attributes. Since our dataset contains 592 users, we use a 5-fold cross-validation in order to avoid test folders with a very small number of users. In a 5-fold cross validation, the original sample is partitioned into 5 sub-samples. Of the 5 sub-samples, one sub-sample is used for testing the classifier, and the remaining 4 sub-samples are used as training data. The process is then repeated 5 times, with each of the 5 sub-samples used exactly once as the test data. The final result reported is the average of 5 runs.

In our analysis, we build the classifier using each set of correlated features (i.e. user-based, video-based, social network, and using all features together). The results are shown in Table 2. Analyzing the results for the classifier using all features, we observe that SVM obtained 0.439 for true positive rate, meaning that 43.9% of the spammers are correctly classified as spammers and could be correctly removed from the system. For the legitimate users, 98.1% are classified correctly. The accuracy obtained is 0.87, meaning that our approach classified erroneously 13% of all users. Clearly, the majority of the users classified erroneously are spammers since our test collection contains about 4 times more legitimate users than spammers.

Metric	User	Video	SN	ALL
TP	0.054	0.426	0.375	0.439
TN	0.998	0.922	1	0.981
FP	0.002	0.078	0	0.019
FN	0.946	0.574	0.625	0.561
Accuracy	0.821	0.821	0.874	0.870
F-measure	0.094	0.484	0.540	0.558

Table 2: Results SVM classification

Observing the false positive rate, we note that SVM classified only 1.9% of legitimate users as spammers. For example, the 1.9% of legitimate users classified as spammers could have their accounts suspended or their videos excluded. Using only social network attributes, the SVM classified all legitimate users as legitimate users, at a cost of having a smaller true positive rate compared with the result which uses all attributes. Depending on the system objectives, it may be better to have a smaller FP than a higher TP.

We investigated the user attributes and YouTube profile of each user classified erroneously. For the spammers classified as legitimate users (false negative of 56.1%), we observed that most of these users use their account as a legitimate user and act as a spammer only for some of its video responses. Most of these users have video responses that are not spam, tricking the classifier in some attributes. Also, some of the videos marked as spam are very popular (i.e. high number of subscribers, views, comments, favorites), tricking the classifier and reaching the spammer objectives. For the legitimate users classified as spammers (false positive of 1.9%) we observed that most have small in-degree and pagerank, high out-degree, and their videos have low popularity. However, these users need further investigation, since these characteristics are common to several other legitimate users who were correctly classified.

Comparing all subsets of attributes, the average values of F-measure show that the social-network attribute set is the most important of three sets for the SVM classification, followed by video-based and user-based attributes. However, statistically these average numbers are not different, except for user-based attributes. We execute the t-test [15] with 90% confidence interval for each two sets of attributes, concluding that all results for social network attributes, video-based attributes, and the set of all attributes are not statistically significant. The user-based attributes turn out to be the least important of the three sets of attributes.

In order to verify the ranking of importance of these attributes we use three feature selection methods available on Weka [21]. We choose three methods, described on the literature [23, 24], namely: χ^2 (Chi Squared), Information Gain, and Symmetrical Uncert. Table 3 presents the 10 most important features for each feature selection method used. All three feature selection methods have 9 attributes in common, 6 of video-based attributes and 3 social network attributes. Moreover, note that all the attributes in the first places are social network attributes, and that the information gain and χ^2 methods have 10 attributes in common.

In conclusion, social network metrics and video-based attributes are the most important set of attributes considered by SVM classifier, which are also important for other feature selection algorithms.

6. CONCLUSIONS AND FUTURE WORK

In this paper we studied video spam in a popular online social video network, namely YouTube. Our study relies upon a dataset collected from YouTube. We crawled the YouTube site to obtain an entire component of the video response user graph. By *manual inspection*, we created a test collection with users classified as spammers or legitimate. We provided a characterization of the users on this test collection which raises several attributes useful to characterize the social or anti-social behavior of users. Using a classification technique, we proposed a video spam detection mechanism which is able to correctly identify significant fraction of the video spammers (44%) in our test collection while incurring in 2% of misclassification of legitimate users. Furthermore, our results were able to bring to light the most important attributes.

As future work, we plan to improve our mechanism to detect spammers and also develop heuristics to identify users that exhibit other kinds of anti-social behavior including causing a video to enter into the list of top videos. Furthermore, we intend to evaluate other social networks formed by YouTube links in order to identify new attributes that can

Position	χ^2	Information Gain	Symetrical Uncert
1	Out-degree	Out-degree	Out-degree
2	# comments total (all videos)	PageRank	# responses created
3	Duration total (all videos)	# comments total (all videos)	Duration mean (all videos)
4	# comments mean (video responses)	In-degree	In-degree
5	PageRank	# comments mean (video responses)	# ratings total (all videos)
6	In-degree	Duration total (all videos)	# comments total (all video)
7	Duration mean (all videos)	# responses received	PageRank
8	# comments mean (all videos)	# comments mean (all videos)	Duration total (all videos)
9	# ratings total (all videos)	# ratings total (all videos)	# Comments mean (video responses)
10	# responses received	duration mean (all videos)	Comments mean (all videos)

Table 3: Ranking of attributes for three different methods

be used to identify video spammers.

7. ACKNOWLEDGEMENTS

The authors would like to thank Prof. Marcos Gonçalves (Federal University of Minas Gerais - Brazil) for his valuable suggestions in Section 5.

8. REFERENCES

[1] Alexa. `http://www.alexa.com`.

[2] H. Aradhye, G. Myers, and J. Herson. Image analysis for efficient categorization of image-based spam e-mail. In *Proc. of the Int'l Conf. on Document Analysis and Recognition (ICDAR)*, 2005.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of World Wide Web Conf. (WWW)*, 1998.

[5] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. In *Int'l ACM SIGIR*, pages 423–430, 2007.

[6] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proc. of IMC*, 2007.

[7] S. Dorogovtsev and J. Mendes. *Evolution of Networks: from Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[8] R. Fan, P. Chen, and C. Lin. Working set selection using the second order information for training svm. *Journal of Machine Learning Research (JMLR)*, 6:1889–1918, 2005.

[9] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proc. of WebDB*, 2004.

[10] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: A view from the edge. In *Proc. of IMC*, 2007.

[11] L. Gomes, J. Almeida, V. Almeida, and W. Meira. Workload models of spam and legitimate e-mails. *Performance Evaluation*, 64:690–714, 2007.

[12] L. Gomes, F. Castro, V. Almeida, J. Almeida, R. Almeida, and L. Bettencourt. Improving spam detection based on structural similarity. In *Proc. of SRUTI*, 2005.

[13] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Int'l. Conf. on Very Large Data Bases*, pages 576–587, 2004.

[14] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11(6):36–45, 2007.

[15] R. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley and Sons, INC, 1991.

[16] J. Kleinberg. Hubs, authorities, and communities. *ACM Computing Surveys*, 31, 1999.

[17] G. Koutrika, F. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating spam in tagging systems. In *Proc. of AIRWeb*, 2007.

[18] M. Shannon. Shaking hands, kissing babies, and...blogging? *Communications of the ACM*, 50, 2007.

[19] A. Thomason. Blog spam: A review. In *Proc. of Conf. on Email and Anti-Spam (CEAS)*, 2007.

[20] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research (JMLR)*, 6:1453–1484, 2005.

[21] I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[22] C. Wu, K. Cheng, Q. Zhu, and Y. Wu. Using visual features for anti-spam filtering. In *Proc. of IEEE Int'l Conf. on Image Processing (ICIP)*, 2005.

[23] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of the Int'l Conf. on Machine Learning (ICML)*, 1997.

[24] L. Yu and H. Liu. Redundancy based feature selection for microarray data. In *Proc. of ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, 2004.