Detecting Spammers and Content Promoters in Online Video Social Networks

Fabrício Benevenuto*, Tiago Rodrigues, Virgílio Almeida,
Jussara Almeida, and Marcos Gonçalves
Computer Science Department, Federal University of Minas Gerais
Belo Horizonte, Brazil
{fabricio, tiagorm, virgilio, jussara, mgoncalv}@dcc.ufmg.br

ABSTRACT

A number of online video social networks, out of which YouTube is the most popular, provides features that allow users to post a video as a response to a discussion topic. These features open opportunities for users to introduce polluted content, or simply pollution, into the system. For instance, *spammers* may post an unrelated video as response to a popular one aiming at increasing the likelihood of the *response* being viewed by a larger number of users. Moreover, opportunistic users - *promoters* - may try to gain visibility to a specific video by posting a large number of (potentially unrelated) responses to boost the rank of the *responded video*, making it appear in the top lists maintained by the system. Content pollution may jeopardize the trust of users on the system, thus compromising its success in promoting social interactions. In spite of that, the available literature is very limited in providing a deep understanding of this problem.

In this paper, we go a step further by addressing the issue of detecting video spammers and promoters. Towards that end, we manually build a test collection of real YouTube users, classifying them as spammers, promoters, and legitimates. Using our test collection, we provide a characterization of social and content attributes that may help distinguish each user class. We also investigate the feasibility of using a state-of-the-art supervised classification algorithm to detect spammers and promoters, and assess its effectiveness in our test collection. We found that our approach is able to correctly identify the majority of the promoters, misclassifying only a small percentage of legitimate users. In contrast, although we are able to detect a significant fraction of spammers, they showed to be much harder to distinguish from legitimate users.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services ; J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms: Human factors, Measurement

Keywords: social networks, social media, video response, video spam, video promotion, spammer, promoter.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

1. INTRODUCTION

With Internet video sharing sites gaining popularity at a dazzling speed, the Web is being transformed into a major channel for the delivery of multimedia. Online video social networks, out of which YouTube is the most popular, are distributing videos at a massive scale. As an example, according to comScore, in May 2008, 74 percent of the total U.S. Internet audience viewed online videos, being responsible for 12 billion videos viewed on that month (YouTube alone provided 34% of these videos) [1]. Additionally, with ten hours of videos uploaded every minute [3], YouTube is also considered the second most searched site in the Web [2].

By allowing users to publicize and share their independently generated content, online video social networks may become susceptible to different types of malicious and opportunistic user actions. Particularly, these systems usually offer three basic mechanisms for video retrieval: (1) a search system, (2) ranked lists of top videos, and (3) social links between users and/or videos. Although appealing as mechanisms to ease content location and enrich online interaction, these mechanisms open opportunities for users to introduce polluted content, or simply pollution, into the system. As an example, video search systems can be fooled by malicious attacks in which users post their videos with several popular tags [23]. Opportunistic behavior on the other two mechanisms for video retrieval can be exemplified by observing a YouTube feature which allows users to post a video as a response to a video topic. Some users, which we call *spammers*, may post an unrelated video as response to a popular video topic aiming at increasing the likelihood of the *response* being viewed by a larger number of users. Additionaly, users we refer to as *promoters* may try to gain visibility to a specific video by posting a large number of (potentially unrelated) responses to boost the rank of the *video topic*, making it appear in the top lists maintained by YouTube. Promoters and spammers are driven by several goals, such as to spread advertise to generate sales, disseminate pornography (often as an advertisement to a Web site), or just to compromise system reputation.

Polluted content may compromise user patience and satisfaction with the system since users cannot easily identify the pollution before watching at least a segment of it, which also consumes system resources, especially bandwidth. Additionally, promoters can further negatively impact system aspects, since promoted videos that quickly reach high rankings are strong candidates to be kept in caches or in content distribution networks [10].

In this paper, we address the issue of detecting video spammers and promoters. To do it, we crawled a large user data set from YouTube site, containing more than 260 thousands users. Then,

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

we created a labeled collection with users "manually" classified as legitimate, spammers and promoters. After that, we conducted a study about the collected user behavior attributes aiming at understanding their relative discriminative power in distinguishing between legitimate users and the two different types of polluters envisioned. Using attributes based on the user's profile, the user's social behavior in the system, and the videos posted by the user as well as her target (responded) videos, we investigated the feasibility of applying a supervised learning method to identify polluters. We found that our approach is able to correctly identify the majority of the promoters, misclassifying only a small percentage of legitimate users. In contrast, although we are able to detect a significant fraction of spammers, they showed to be much harder to distinguish from legitimate users. These results motivated us to investigate a hierarchical classification approach, which explores different classification tradeoffs and provides more flexibility for the application of different actions to the detected polluters.

The rest of the paper is organized as follows. The next section discusses related work. Section 3 describes our crawling strategy and the test collection built from the crawled dataset. Section 4 investigates a set of user attributes and their ability to distinguish promoters, spammers and legitimate users. Section 5 describes and evaluates our strategies to detect promoters and spammers. Finally, Section 6 offers conclusions and directions for future work.

2. RELATED WORK

Content pollution has been observed in various applications, including e-mail [15], Web search engines [13], blogs [29]. Thus, a number of detection and combating strategies have been proposed [9, 16, 25, 32]. Most of them rely on extracting evidences from *textual descriptions* of the content, treating the text corpus as a set of objects with associated attributes, and applying some classification method to detect spam [17]. A framework to detect spamming in tagging systems, a malicious behavior that aims at increasing the visibility of an object by fooling the search mechanism, was proposed in [23]. A few other strategies rely on image processing algorithms to detect spam in image-based e-mails [31].

Our proposal is complementary to these efforts for two reasons. First, it aims at detecting *users* who disseminate *video* pollution, instead of classifying the content itself. Content-based classification would require combining multiple forms evidences extracted from textual descriptions of the video (e.g., tags, title) and from the video content itself, which, in turn, would require more sophisticated multimedia information retrieval methods that are robust to the typically low quality of user-generated videos [7]. Instead, we explore attributes that capture the feedback of users with respect to each other or to their contributions to the system (e.g., number of views received), exploiting their interactions through video responses.

In a previous study, we analyzed the properties of the social network created by video response interactions in YouTube, finding evidence of pollution [5]. Additionally, we preliminarily approached this problem by creating a small test collection composed of spammers and legitimate users, and applying a binary classification strategy to detect spammers [6]. The present work builds on this preliminary effort by providing a much more thorough, richer and solid investigation of the feasibility and tradeoffs in detecting video polluters in online video sharing systems, considering a much larger test collection, a richer set of user attributes, as well as different types of malicious and opportunistic behaviors.

Our study is also complementary to other studies of the properties of social networks [4, 26] and of the traffic to online social networking systems, in particular YouTube. An in-depth analysis of popularity distribution and evolution, and content characteristics of YouTube and of a popular Korean service is presented in [10]. Gill *et al* [14] characterize YouTube traffic collected from an university campus network, comparing its properties with those previously reported for other workloads.

3. USER TEST COLLECTION

In order to evaluate our proposed approach to detect video spammers and promoters in online video social networking systems, we need a test collection of users, pre-classified into the target categories, namely, spammers, promoters and, in lack of a better term, legitimate users. However, to the best of our knowledge, no such collection is publicly available for any video sharing system, thus requiring us to build one.

Before presenting the steps taken to build our user test collection, we introduce some notations and definitions. We say a YouTube video is a *responded video* or a *video topic* if it has at least one video response. Similarly, we say a YouTube user is a *responsive user* if she has posted at least one video response, whereas a *responded user* is someone who posted at least one responded video. Moreover, we define as *spammer* a user who posts at least *one* video response that is considered unrelated to the responded video (i.e., a spam). Examples of video spams are: (i) an advertisement of a product or website completely unrelated to the subject of the responded video, and (ii) pornographic content posted as response to a cartoon video. A *promoter* is defined as a user who posts a large number of video responses to a *responded video*, aiming at promoting this *video topic*. As an example, we found promoters in our dataset who posted a long sequence (e.g., 100) of (unrelated) video responses, often without content (0 second) to a single video. A user that is neither a spammer nor a promoter is considered legitimate. The term *polluter* is used to refer to either a spammer or a promoter.

We build our user test collection by first crawling YouTube, one of the most popular social video sharing systems [1] (Section 3.1). Next, we carefully select and manually classify a subset of these users (Section 3.2).

3.1 Crawling YouTube

Our strategy consists of collecting a sample of users who participate in interactions through video responses, i.e, who post or receive video responses. These interactions can be represented by a *video response user graph* $G = (X, Y)$, where X is the union of all users who posted or received video responses until a certain instant of time, and (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 representative sample of the YouTube video response user graph, we build a crawler that implements Algorithm 1. The sampling starts from a set of 88 seeds, consisting of the owners of the top-100 most responded videos of all time, provided by YouTube. The crawler follows links of responded videos and video responses, gathering information on a number of different attributes of their contributors (users), including attributes of all responded videos and video responses posted by her.

The crawler ran for one week (01/11-18, 2008), gathering a total of **264,460** users, **381,616** responded videos and **701,950** video responses. This dataset produces a large weakly connected component of graph (X, Y), and is used as source for building our test collection, as described next.

3.2 Building a Test Collection

The main goal of creating a user test collection is to study the patterns and characteristics of each class of users. Thus, the desired

Algorithm 1 Video Response Crawler

```
Input: A list L of users (seeds)
for each User U in L do
    Collect U's info and list of videos (responded and responses);
    for each Video V in the video list do
        Collect info of V;
        if V is a responded video then
            Collect info of V's video responses;
            Insert the responsive users in L;
        end if
        if V is a video response then
            Insert the responded user in L;
        end if
    end for
end for
```

properties for our test collection include the following: (1) having a significant number of users of all three categories; (2) including, but not restricting to, spammers and promoters which are aggressive in their strategies and generate large amounts of pollution in the system; and (3) including a large number of legitimate users with different behavioral profiles. We argue that these properties may *not* be achieved by simply randomly sampling the collection. The reasons for this are twofold. First, randomly selecting a number of users from the crawled data could lead us to a small number of spammers and promoters, compromising the creation of effective training and test data sets for our analysis. Moreover, research has shown that the sample does not need to follow the class distribution in the collection in order to achieve effective classification [30]. Second, it is natural to expect that legitimate users present a large number of different behaviors in a social network. Thus, selecting legitimate users randomly may lead to a large number of users with similar behavior (i.e. post one video response to a discussed topic), not including examples with different profiles.

Aiming at capturing all these properties, we define three strategies for user selection (described below). Each selected user was then manually classified. However, this classification relies on human judgment on, for instance, whether a video is related to another. In order to minimize the impact of human error, three volunteers analyzed all video responses of each selected user in order to independently classify her into one of the three categories. In case of tie (i.e., each volunteer chooses a different class), a fourth independent volunteer was heard. Each user was classified based on majority voting. Volunteers were instructed to favor legitimate users. For instance, if one was not confident that a video response was unrelated to the responded video, she should consider it to be legitimate. Moreover, video responses containing people chatting or expressing their opinions were classified as legitimate, as we choose not to evaluate the expressed opinions. The volunteers agreed in about 97% of the analyzed videos, which reflects a high level of confidence to this human classification process. The three user selection strategies used are:

(1) In order to select users with different levels of interaction through video responses, we first defined four groups of users based on their in and out-degrees in the video response user graph (Section 3.1). Group 1 consists of users with low (≤ 10) in and out-degrees, and thus who respond to and are responded by only a few other users. Group 2 consists of users with high (> 10) in-degree and low out-degree, and thus receive video responses from many others but post responses to only a few users. Group 3 consists of users with low in-degree and high out-degree, whereas very interactive users, with high in and out-degrees, fall into group 4. One hundred users were randomly selected from each group[1], and manually classified, yielding a total of 382 legitimate, 10 spammers, and no promoter. The remaining 8 users were discarded as they had their accounts suspended due to violation of terms of use.

(2) Aiming at populating the test collection with polluters, we searched for them where they are more likely to be found. We first note that, in YouTube, a video v can be posted as response to at most *one* video at a time (unless one creates a copy of v and uploads it with a different ID). Thus, it is more costly for spammers to spread their video spam in YouTube than it is, for instance, to disseminate spam by e-mail. We conjecture, then, that spammers would post their video responses more often to popular videos so as to make each spam visible to a larger community of users. Moreover, some video promoters might eventually be successful and have their target listed among the most popular videos. Thus, we browsed the video responses posted to the top 100 most responded videos of all time, selecting a number of *suspect* users[2]. The classification of these suspect users led to 7 legitimate users, 118 spammers, and 28 promoters in the test collection.

(3) To minimize a possible bias introduced by strategy (2), we *randomly* selected 300 users who posted video responses to the top 100 most responded videos of all time, finding 252 new legitimate users, 29 new spammers and 3 new promoters (16 users with closed accounts were discarded).

In total, our test collection contains 829 users, including 641 classified as legitimate, 157 as spammers and 31 as promoters. Those users posted 20,644 video responses to 9,796 unique responded videos. Our user test collection aims at supporting research on detecting spammers and promoters. Since the user classification labeling process relies on human judgment, which implies in watching a significantly high amount of videos, the number of users in our test collection is somewhat limited. In future work, we plan to make our test collection available and also study collaborative ways to increase its size.

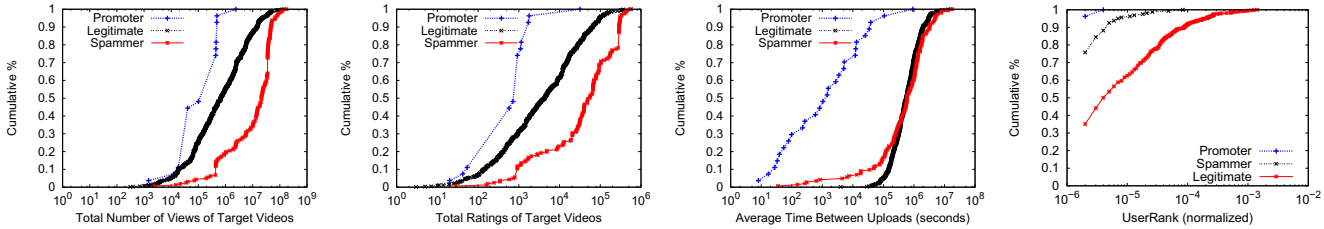
4. ANALYZING USER BEHAVIOR ATTRIBUTES

Legitimate users, spammers and promoters have different goals in the system, and, thus, we expect they also differ on how they behave (e.g., who they interact with, which videos they post) to achieve their purposes. Thus, our next step is to analyze a large set of attributes that reflect user behavior in the system aiming at investigating their relative discriminatory power to distinguish one user class from the others. We considered three attribute sets, namely, video attributes, user attributes, and social network (SN) attributes.

Video attributes capture specific properties of the videos uploaded by the user, i.e., each user has a set of videos in the system, each one with attributes that may serve as indicators of its "quality", *as perceived by others*. In particular, we characterize each video by its duration, numbers of views and of commentaries received, ratings, number of times the video was selected as favorite, as well as numbers of honors and of external links. Moreover, we consider three separate groups of videos owned by the user. The first group contains aggregate information of *all videos* uploaded by the user, being useful to capture how others see the (video) contributions of this user. The second group considers only *video responses*, which may be pollution. The last group considers only the *responded*

[1]Groups 1, 2, 3 and 4 have 162,546, 2,333, 3,189 and 1,154 users. Thus, homogeneous random selection from each one yields a bias towards group 4.

[2]As an example, the owner of a video with a pornographic picture as thumbnail but posted to a political debate video discussion topic.



(a) Total Views of Target Videos (b) Total Ratings of Target Videos (c) Avg. Time between Uploads (d) UserRank

Figure 1: Cumulative Distribution of User Behavior Attributes

videos to which this user posted video responses (referred to as *target* videos). For each video group, we considered the average and the sum of the aforementioned attributes, summing up 42 video attributes for each user, all of which can be easily derived from data maintained by YouTube. We explicitly choose not to add any attribute that would require processing the multimedia content itself.

The second set of attributes consists of individual characteristics of user behavior. We expect that legitimate users spend more time doing actions such as selecting friends, adding videos as favorites, and subscribing to content updates from others. Thus, we select the following 10 user attributes: number of friends, number of videos uploaded, number of videos watched, number of videos added as favorite, numbers of video responses posted and received, numbers of subscriptions and subscribers, average time between video uploads, and maximum number of videos uploaded in 24 hours.

The third set of attributes captures the social relationships established between users via video response interactions, which is one of the several possible social networks in YouTube. The idea is that these attributes might capture specific interaction patterns that could help differentiate legitimate users, promoters, and spammers. We selected the following node attributes extracted from the video response user graph, which capture the level of (social) interaction of the corresponding user: clustering coefficient, betweenness, reciprocity, assortativity, and UserRank.

The clustering coefficient of node i, $cc(i)$, is the ratio of the number of existing edges between i's neighbors to the maximum possible number, and captures the communication density between the user's neighbors. The betweenness is a measure of the node's centrality in the graph, that is, nodes appearing in a larger number of the shortest paths between any two nodes have higher betweenness than others [28]. The reciprocity $R(i)$ of node i measures the probability of the corresponding user u_i receiving a video response from each other user to whom she posted a video response, that is, $R(i) = \frac{|OS(i) \cap IS(i)|}{|OS(i)|}$, where $OS(i)$ is the set of users to who u_i posted a video response, and $IS(i)$ is the set of users who posted video responses to u_i. Node assortativity is defined, as in [9], as the ratio between the node (in/out) degree and the average (in/out) degree of its neighbors. We compute node assortativity for the four types of degree-degree correlations (i.e., in-in, in-out, out-in, out-out). Finally, we also applied the PageRank [8] algorithm, commonly used to assess the popularity of a Web page [24], to our video response user graph. The computed metric, which we refer to as UserRank, indicates the degree of participation of a user in the system through interactions via video responses. In total, we selected 8 social network attributes.

We assessed the relative power of the 60 selected attributes in discriminating one user class from the others by independently applying two well known feature selection methods, namely, information gain and χ^2 (Chi Squared) [34]. Table 1 summarizes the results, showing the number of attributes from each set (video, user, and social network) in the top 10, 20, 30, 40, and 50 most discriminative attributes according to the ranking produced by χ^2. Results for information gain are very similar and, thus, are omitted.

Attribute Set	Top 10	Top 20	Top 30	Top 40	Top 50
Video	9	18	25	30	36
User	1	2	4	7	9
SN	0	0	1	3	5

Table 1: Number of Attributes at Top Positions in χ^2 Ranking

Note that the 9 of the 10 most discriminative attributes are video-related. In fact, the most discriminative attribute (according to *both* methods), is the total number of views (i.e., the popularity) of the *target* videos. Figure 1(a) presents the cumulative distributions of this metric for each user class, showing a clear distinction among them. The curve for spammers is much more skewed towards a larger number of views, since these users tend to target popular videos in order to attract more visibility to their content. In contrast, the curve for promoters is more skewed towards the other end as they tend to target videos that are *still* not very popular, aiming at raising their visibility. Legitimate users, being driven mostly by social relationships and interests, exhibit an intermediary behavior, targeting videos with a wide range of popularity. The same distinction can be noticed for the distributions of the total ratings of target videos, shown in Figure 1(b), another metric that captures user feedback with respect to these videos, and is among the top 10 most discriminative attributes.

The most discriminative user and social network attributes are the average time between video uploads and the UserRank, respectively. In fact, Figure 1(c) and (d) show that, in spite of appearing in lower positions in the ranking, particularly for the UserRank attribute (see Table 1), these two attributes have potential to be able to separate user classes apart. In particular, the distribution of the average time between video uploads clearly distinguishes promoters, who tend to upload at a much higher frequency since their success depends on them posting as many video responses to the target as possible. Figure 1(c) also shows that, at least with respect to this user attribute, spammers can not be clearly distinguished from legitimate users. Finally, Figure 1(d) shows that legitimate users tend to have much higher UserRank values than spammers, who, in turn, have higher UserRank values than promoters. This indicates that, as expected, legitimate users tend to have a much more participative role (system-wide) in the video response interactions than users from the other two classes, which are much more selective when choosing their targets.

5. DETECTING SPAMMERS AND PROMOTERS

In this section, we investigate the feasibility of applying a supervised learning algorithm along with the attributes discussed in the

previous section for the task of detecting spammers and promoters. In this approach, each user is represented by a vector of values, one for each attribute. The algorithm learns a classification model from a set of previously labeled (i.e., pre-classified) data, and then applies the acquired knowledge to classify new (unseen) users into three classes: legitimate, spammers and promoters. Note that, in this paper, we do not address the labeling process. Labeled data may be obtained through various initiatives (e.g., volunteers who help marking video spam, professionals hired to periodically manually classify a sample of users, etc). Our goal here is to assess the *potential effectiveness* of the proposed approach as a first effort towards helping system administrators to detect polluters in online video social networks.

We start by presenting, in Section 5.1, the metrics used to evaluate our experimental results. Section 5.2 describes the classification algorithm, i.e., the classifier, and the experimental setup used. The classifier was applied according to two different strategies, referred to as flat and hierarchical classifications. In the flat classification, illustrated in Figure 2 (left), the users from the test collection are directly classified into promoters (P), spammers (S), and legitimate users (L). In the hierarchical strategy, the classifier is first used to separate promoters (P) from non-promoters (NP). Next, it classifies promoters into heavy (HP) and light promoters (LP), as well as non-promoters into legitimate users (L) and spammers (S), in a hierarchical fashion shown in Figure 2 (right). Results from our flat and hierarchical classifications are presented in Sections 5.3 and 5.4. Section 5.5 discusses the impact of reducing the attribute set on the classification effectiveness.

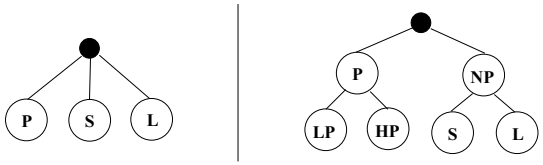


Figure 2: Classification Strategies: Flat (left) and Hierarchical (right)

5.1 Evaluation Metrics

To assess the effectiveness of our classification strategies we use the standard information retrieval metrics of recall, precision, Micro-F1, and Macro-F1 [33]. The recall (r) of a class X is the ratio of the number of users correctly classified to the number of users in class X. Precision (p) of a class X is the ratio of the number of users classified correctly to the total predicted as users of class X. In order to explain these metrics, we will make use of a confusion matrix [22], illustrated in Table 2. Each position in this matrix represents the number of elements in each original class, and how they were predicted by the classification. In Table 2, the precision (p_{prom}) and the recall (r_{prom}) of the class promoter are computed as $p_{prom} = a/(a+d+g)$ and $r_{prom} = a/(a+b+c)$.

		Predicted		
		Promoter	**Spammer**	**Legitimate**
	Promoter	a	b	c
True	**Spammer**	d	e	f
	Legitimate	g	h	i

Table 2: Example Confusion Matrix

The F1 metric is the harmonic mean between both precision and recall, and is defined as $F1 = 2pr/(p+r)$. Two variations of F1, namely, micro and macro, are normally reported to evaluate classification effectiveness. Micro-F1 is calculated by first computing global precision and recall values for all classes, and then calculating F1. Micro-F1 considers equally important the classification of *each user*, independently of its class, and basically measures the capability of the classifier to predict the correct class on a per-user basis. In contrast, Macro-F1 values are computed by first calculating F1 values for each class in isolation, as exemplified above for promoters, and then averaging over all classes. Macro-F1 considers equally important the effectiveness in *each class*, independently of the relative size of the class. Thus, the two metrics provide complementary assessments of the classification effectiveness. Macro-F1 is especially important when the class distribution is very skewed, as in our case, to verify the capability of the classifier to perform well in the smaller classes.

5.2 The Classifier and the Experimental Setup

We use a Support Vector Machine (SVM) classifier [20], which is a state-of-the-art method in classification and obtainned the best results among a set of classifiers tested. The goal of a SVM is to find the hyperplane that optimally separates with a maximum margin the training data into two portions of an N-dimensional space. A SVM performs classification by mapping input vectors into an N-dimensional space, and checking in which side of the defined hyperplane the point lies. SVMs are originally designed for binary classification but can be extended to multiple classes using several strategies (e.g. one against all [18]). We use a non-linear SVM with the Radial Basis Function (RBF) kernel to allow SVM models to perform separations with very complex boundaries. The implementation of SVM used in our experiments is provided with libSVM[3] [12], an open source SVM package that allows searching for the best classifier parameters using the *training* data, a mandatory step in the classifier setup. In particular, we use the *easy* tool from libSVM, which provides a series of optimizations, including normalization of all numerical attributes.

The classification experiments are performed using a 5-fold cross-validation. In each test, the original sample is partitioned into 5 sub-samples, out of which four are used as training data, and the remaining one is used for testing the classifier. The process is then repeated 5 times, with each of the 5 sub-samples used exactly once as the test data, thus producing 5 results. The entire 5-fold cross validation was repeated 5 times with different seeds used to shuffle the original data set, thus producing 25 different results for each test. The results reported are averages of the 25 runs. With 95% of confidence, results do not differ from the average in more than 5%.

In the following two sections, we discuss the results obtained with the two classification strategies (flat and hierarchical) using all 60 selected attributes, since, as discussed in Section 4, even attributes with low ranks according to the employed feature selection methods (e.g., UserRank) may have some discriminatory power, and may be useful to classify users. Moreover, SVMs are known for dealing well with high dimensional spaces, properly choosing the weights for each attribute, i.e., attributes that are not helpful for classification are given low weights by the optimization method used by the SVM [20]. The impact of using different subsets of the attributes on the classification effectiveness is analyzed in Section 5.5.

5.3 Flat Classification

Table 3 shows the confusion matrix obtained as the result of our

[3]For experiments involving the SVM J parameter (discussed in Section 5.4), we used a different implementation, called SVM light, since libSVM does not provide this parameter. Classification results are equal for both implementations when we use the same classifier parameters.

experiments with the flat classification strategy. The numbers presented are percentages relative to the total number of users in each class. The diagonal in boldface indicates the recall in each class. Approximately 96% of promoters, 57% of spammers, and 95% of legitimate users were correctly classified. Moreover, no promoter was classified as legitimate user, whereas only a small fraction of promoters were erroneously classified as spammers (3.87%). By manually inspecting these promoters, we found that the videos that they targeted (i.e., the promoted videos) actually acquired a certain popularity. In that case, it is harder to distinguish them from spammers, who target more often very popular videos, as well as from some legitimate users who, following their interests or social relationships, post responses to popular videos. Referring to Figure 1(a), these (somewhat successful) promoters are those located in the higher end of the curve, where the three user classes can not be easily distinguished.

		Predicted		
		Promoter	**Spammer**	**Legitimate**
	Promoter	**96.13%**	3.87%	0.00%
True	**Spammer**	1.40%	**56.69%**	41.91%
	Legitimate	0.31%	5.02%	**94.66%**

Table 3: Flat Classification

A significant fraction (almost 42%) of spammers was misclassified as legitimate users. In general, these spammers exhibit a dual behavior, sharing a reasonable number of legitimate videos (non-spam) and posting legitimate video responses, thus presenting themselves as legitimate users most of the time, but occasionally posting video spams. This dual behavior masks some important aspects used by the classifier to differentiate spammers from legitimate users. This is further aggravated by the fact that a significant number of legitimate users post their video responses to popular responded videos, a typical behavior of spammers. Therefore, as opposed to promoters, which can be effectively separated from the other classes, distinguishing spammers from legitimate users is much harder. In Section 5.4.1, we discuss an approach that allows one to trade a higher recall of spammers at a cost of misclassifying a larger number of legitimate users.

As a summary of the classification results, Micro-F1 value is 87.5, whereas per-class F1 values are 63.7, 90.8, and 92.3, for spammers, promoters, and legitimate users, respectively, resulting in an average Macro-F1 equal to 82.2. The Micro-F1 result indicates that we are predicting the correct class in almost 88% of the cases. Complementarily, the Macro-F1 result shows that there is a certain degree of imbalance for F1 across classes, with more difficulty for classifying spammers. Comparing with a trivial baseline classifier that chooses to classify every single user as legitimate, we obtain gains of about 13% in terms of Micro-F1, and of 183% in terms of Macro-F1. As a first approach, our proposed classification provides significant benefits, being effective in identifying polluters in the system.

5.4 Hierarchical Classification

Our flat classification results show that we can effectively identify promoters, but separating spammers from legitimate users is a harder task. This motivates us to experiment with a hierarchical classification strategy, illustrated in Figure 2 (right), which allow us to take advantage of a cost mechanism in the SVM classifier, specific for binary classification. In this mechanism, one can give priority to one class (e.g., spammers) over the other (e.g., legitimate users) by varying its J parameter[4] [27]. By varying J, we can study several tradeoffs and scenarios. In particular, we evaluate the tradeoffs between identifying more spammers at the cost of misclassifying more legitimate users (Section 5.4.1), and we further categorize promoters into heavy and light, based on their aggressiveness (Section 5.4.2). Splitting the set of promoters is also motivated by the potential for disparate behaviors with different impact on the system, thus requiring different treatments. On one hand, heavy promoters may reach top lists very quickly, requiring a fast detection. On the other hand, light promoters may conceal a collusion attack to promote the same responded video, thus requiring further investigation.

		Predicted	
		Promoter	**Non-Promoter**
	Promoter	**92.26%**	7.74%
True	**Non-Promoter**	0.55%	**99.45%**

Table 4: Hierarchical Classification of Promoters vs. Non-Promoters

The results for the first phase of the hierarchical classification (promoters versus non-promoters) are summarized in Table 4. Macro-F1 and Micro-F1 are 93.44 and 99.17, respectively. Similarly to the results with the flat characterization, the vast majority of promoters were correctly classified (both results are statistically indistinguishable). In fact, the absolute number of erroneously classified users in each run of a test is very small (mostly 1 or 0).

5.4.1 Non-promoters

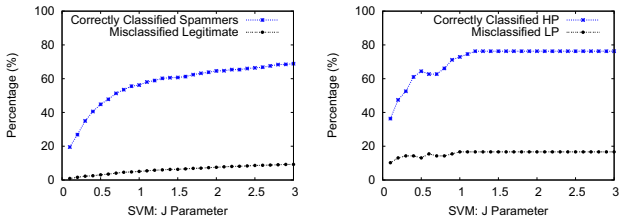
As previously discussed, there are cases of spammers and legitimate users acting similarly, making the task of differentiating them very difficult. In this section, we perform a binary classification of all (test) users identified as non-promoters in the first phase of the hierarchical classification, separating them into spammers and legitimate users. For this experiment, we trained the classifier with the original training data without promoters.

		Predicted	
		Legitimate	**Spammer**
	Legitimate	**95.09%**	4.91%
True	**Spammer**	41.27%	**58.73%**

Table 5: Hierarchical Classification of Non-Promoters

Table 5 shows results of this binary classification. In comparison with the flat classification (Table 3), there was no significant improvement on separating legitimate users and spammers. These results were obtained with J=1. Figure 3(a) shows that increasing J leads to a higher percentage of correctly classified spammers (with diminishing returns for $J > 1.5$), but at the cost of a larger fraction of misclassified legitimate users. For instance, one can choose to correctly classify around 24% of spammers, misclassifying only 1% legitimate users ($J = 0.1$). On the other hand, one can correctly classify as much as 71% of spammers ($J = 3$), paying the cost of misclassifying 9% of legitimate users. The best solution to this tradeoff depends on the system administrator's objectives. For example, the system administrator might be interested in sending an automatic warning message to all users classified as spammers, in which case they might prefer to act conservatively, avoiding sending the message to legitimate users, at the cost of reducing the number of correctly predicted spammers. In another situation, the system administrator may prefer to detect a higher fraction of spammers for manual inspection. In that case, misclassifying a few

[4]The J parameter is the cost factor by which training errors in one class outweigh errors in the other. It is useful, when there is a large imbalance between the two classes, to counterbalance the bias towards the larger one.



(a) Spammers vs. Legitimate (b) Heavy vs. Light Promoters

Figure 3: Impact of Varying the J Parameters

more legitimate users has no great consequence, and may be preferred, since they will be cleared out during inspection. It should be stressed that we are evaluating the potential benefits of varying J. In a practical situation, the optimal value should be discovered in the training data with cross-validation, and selected according to the system administrator goal.

5.4.2 *Heavy and Light Promoters*

In order to be able to further classify promoters into heavy and light, we need first a metric to capture the promoter "aggressiveness", and then we must label each promoter as either heavy or light, according to this metric. The metric chosen to capture the aggressiveness of a promoter is the *maximum number of video responses posted in a 24-hour period*. We expect that heavy promoters would post a large number of videos in sequence in a short period of time, whereas light promoters, perhaps acting jointly in a collusion attack, may try to make the promotion process imperceptible to the system by posting videos at a much slower rate. The k-means clustering algorithm [19] was used to separate promoters into two clusters, labeled heavy and light, according to this metric.

Out of the 31 promoters, 18 were labeled as *light*, and 13 as *heavy*. As expected, these two groups of users exhibit different behaviors, with different consequences from the system perspective. Light promoters are characterized by an average "aggressiveness" of at most 15.78 video responses posted in 24 hours, with coefficient of variation (CV) equal to 0.63. Heavy promoters, on the other hand, exhibit an average behavior of posting as much as 107.54 video responses in 24 hours (CV=0.61). In particular, after manual inspection, we found that all heavy promoters posted a number of video responses sufficient to boost the ranking of their targets to the top 100 most responded videos of the day (during collection period). Some of them even reached the top 100 most responded videos of the week, of the month and of all time. On the other hand, no light promoter posted enough video responses to promote the target to the top lists (during the collection). However, all of them participated in some collusion attack, with different subsets of them targeting different videos.

We performed a binary classification of all (test) users identified as promoters in the first phase of the hierarchical classification, separating them into light and heavy promoters. To that end, we retrained the classifier with the original training data containing only promoters, each one labeled according to the cluster it belongs to. The results are summarized in Table 6. Approximately 83% of light promoters and 73% of heavy promoters are correctly classified. Figure 3 (right) shows the impact of varying the J parameter, and how a system administrator can trade detecting more heavy promoters (HP) for misclassifying a larger fraction of light promoters (LP). A conservative system administrator may choose to correctly classify 36% of heavy promoters at the cost of misclassifying only 10% of light promoters ($J = 0.1$). A more aggressive one may choose to classify as much as 76% of heavy promoters, if she can afford misclassifying 17% of the light ones ($J \geq 1.2$).

		Predicted	
		Light Promoter	**Heavy Promoter**
	Light Promoter	**83.33%**	16.67%
True	**Heavy Promoter**	27.12%	**72.88%**

Table 6: Hierarhical Classification of Promoters

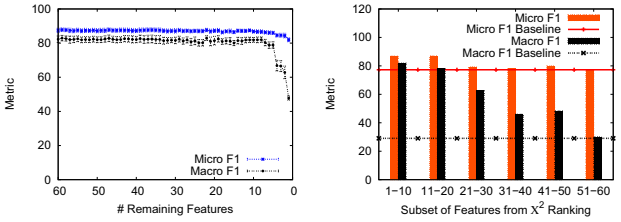
An interesting finding of our work is with respect to collusion of promoters (especially light promoters). Intuitively, if we identify one element of a collusion, the rest of the collusion can be also detected by analyzing other users who post responses to the promoted video. By inspecting the video responses posted to some of the target videos of the detected promoters, we found hundreds of new promoters among the investigated users, indicating that our approach can also effectively unveil collusion attacks, guiding system administrator towards promoters that are more difficult to detect.

5.5 Impact of Reducing the Attribute Set

Once we have understood the main tradeoffs and challenges in classifying users into spammers, promoters and legitimate, we now turn to investigate whether competitive effectiveness can be reached with fewer attributes. We report results for the flat classification strategy, considering two scenarios.

Scenario 1 consists of evaluating the impact on the classification effectiveness of gradually removing attributes in a decreasing order of position in the χ^2 ranking. Figure 4(a) shows Micro-F1 and Macro-F1 values, with corresponding 95% confidence intervals. There is no noticeable (statistical) impact on the classification effectiveness (both metrics) when we remove as many as the 40 lowest ranked attributes. It is worth noting that some of the most expensive attributes such as UserRank and betweenness, which require processing the entire video response user graph, are among these attributes. In fact, all social network attributes are among them, since UserRank, the best positioned of these attributes, is in the 30^{th} position. Thus, our classification approach is still effective even with a smaller, less expensive set of attributes. The Figure also shows that the effectiveness drops sharply when we start removing some of the top 10 attributes from the process.

Scenario 2 consists of evaluating our classification when subsets of 10 attributes occupying contiguous positions in the ranking (i.e., the first top 10 attributes, the next 10 attributes, etc) are used. Figure 4(b) shows Micro-F1 and Macro-F1 values for the flat classification and for the baseline classifier that considers all users as legitimate, for each such range. In terms of Micro-F1, our classification provides gains over the baseline for the first two subsets of attributes, whereas significant gains in Macro-F1 are obtained for all attribute ranges, but the last one (the 10 worst attributes). This confirms the results of our attribute analysis that shows that even low-ranked attributes have some discriminatory power. In practical terms, significant improvements over the baseline are possible even if not all attributes considered in our experiments can be obtained.



(a) Scenario 1 (b) Scenario 2

Figure 4: Impact of Reducing the Set of Attributes

6. CONCLUSIONS AND FUTURE WORK

Promoters and Spammers can pollute video retrieval features of online video social networks, compromising not only user satisfaction with the system, but also system resources and aspects such as caching. We propose an effective solution to the problem of detecting these polluters that can guide system administrators to spammers and promoters in online video social networks. Relying on a sample of pre-classified users and on a set of user behavior attributes, our flat classification approach was able to detect correctly 96% of the promoters, 57% of spammers, wrongly classifying only 5% of the legitimate users. Thus, our proposed approach poses a promising alternative to simply considering all users as legitimate or to randomly selecting users for manual inspection. We also investigated a hierarchical version of the proposed approach, which explores different classification tradeoffs and provides more flexibility for the application of different actions to the detected polluters. As example, the system administrators may send warning messages for the suspects or put the suspects in quarantine for further investigation. In the first case, the system administrators could be more tolerant to misclassifications than in the second case, using the different classification tradeoffs we proposed. Finally, we found that our classification can produce significant benefits even if only a small subset of less expensive attributes is available.

It is expect that spammers and promoters will evolve and adapt to anti-pollution strategies (i.e. using fake accounts to forge some attributes) [11]. Consequently, some attributes may become less important whereas others may acquire importance with time. Thus, labeled data needs also to be constantly updated and the classification models need to be re-learned. Periodical assessment of the classification process may be necessary in the future so that retraining mechanisms could be applied.

It is also natural to expect that our approach could benefit from other anti-pollution strategies. We choose three to discuss here. *(1) User Filtering*: If most owners of responded videos check their video responses to remove those which are polluted videos, video spamming would be significantly reduced. The challenge here is to provide users incentives that encourage them to filter out polluted video responses. *(2) IP Blocking*: Once a polluter is detected, it is natural to suspend her account. Additionally, blocking IP addresses to respond or to upload new videos (but not to watch content) could be useful to prevent polluters from continuing acting maliciously on the system with new accounts. *(3) User Reputation*: Reputation systems allow users to rank each other and, ideally, users engaging in malicious behavior eventually would develop low reputations [21]. However, current designs of reputation systems may suffer from problems of low robustness against collusion, and high implementation complexity.

Lastly, we envision two directions towards which our work can evolve. First, we aim at reducing the cost of the labeling process by studying the viability of semi-supervised learning methods to detect polluters. Second, we intend to explore other refinements to the proposed approach such as to use different classification methods (maybe combined). We believe that better classification effectiveness may require exploring other features which include temporal aspects of user behavior and also features obtained from other social networks formed by YouTube links.

7. REFERENCES

[1] comscore: Americans viewed 12 billion videos online in may 2008. `http://www.comscore.com/press/release.asp?press=2324`.

[2] The new york times: Search ads come to youtube. `http://bits.blogs.nytimes.com/2008/10/13/search-ads-come-to-youtube`.

[3] Youtube fact sheet. `http://www.youtube.com/t/fact_sheet`.

[4] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Int'l World Wide Web Conference (WWW)*, 2007.

[5] F. Benevenuto, F. Duarte, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Understanding video interactions in youtube. In *ACM Multimedia (MM)*, 2008.

[6] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross. Identifying video spammers in online social networks. In *Int'l Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[7] S. Boll. Multitube–where web 2.0 and multimedia could meet. *IEEE MultiMedia*, 14, 2007.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Int'l World Wide Web Conference (WWW)*, 1998.

[9] 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*, 2007.

[10] 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 *Internet Measurement Conference (IMC)*, 2007.

[11] F. Douglis. On social networking and communication paradigms. *IEEE Internet Computing*, 12, 2008.

[12] 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, 2005.

[13] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Int'l Workshop on the Web and Databases (WebDB)*, 2004.

[14] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: A view from the edge. In *Internet Measurement Conference (IMC)*, 2007.

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

[16] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Int'l. Conference on Very Large Data Bases (VLDB)*, 2004.

[17] 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, 2007.

[18] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. In *IEEE Transactions on Neural Networks*, volume 13, 2002.

[19] A. Jain, M. Murty, and P. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31, 1999.

[20] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, 1998.

[21] S. Kamvar, M. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Int'l World Wide Web Conference (WWW)*, 2003.

[22] R. Kohavi and F. Provost. Glossary of terms. *Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, Machine Learning*, 30, 1998.

[23] G. Koutrika, F. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating spam in tagging systems. In *Int'l Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2007.

[24] A. Langville and C. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[25] Y. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng. Detecting splogs via temporal dynamics using self-similarity analysis. *ACM Transactions on the Web (TWeb)*, 2, 2008.

[26] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Internet Measurement Conference (IMC)*, 2007.

[27] K. Morik, P. Brockhausen, and T. Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Int'l Conference on Machine Learning (ICML)*, 1999.

[28] M. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68, 2003.

[29] A. Thomason. Blog spam: A review. In *Conference on Email and Anti-Spam (CEAS)*, 2007.

[30] G. Weiss and F. Provost. The effect of class distribution on classifier learning: An empirical study. Technical report, 2001.

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

[32] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: Signatures and characteristics. In *ACM SIGCOMM*, 2008.

[33] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrival*, 1, 1999.

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