Understanding Video Interactions in YouTube

Fabricio Benevenuto[‡] fabricio@dcc.ufmg.br
Fernando Duarte[‡] fernando@dcc.ufmg.br
Tiago Rodrigues[‡] tiagorm@dcc.ufmg.br
Virgilio Almeida[‡] virgilio@dcc.ufmg.br
Jussara Almeida[‡] jussara@dcc.ufmg.br
Keith Ross[†] ross@poly.edu

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

ABSTRACT

This paper seeks understanding the user behavior in a social network created essentially by video interactions. We present a characterization of a social network created by the video interactions among users on YouTube, a popular social networking video sharing system. Our results uncover typical user behavioral patterns as well as show evidences of anti-social behavior such as self-promotion and other types of content pollution.

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, Videos

Keywords

video response, social networks, promotion

1. INTRODUCTION

Video content is becoming a predominant part of users daily lives on the Web. By allowing users to generate and distribute their own content to large audiences, the Web has been transformed into a major channel for the delivery of multimedia. In fact, a number of services in current Web 2.0 are offering video-based functions as alternative to text-based ones, such as video reviews for products, video ads and video responses [11]. Most part of this huge success of multimedia content is due to the change on the user perspective from consumer to creator. As a consequence, several multimedia issues should be revisited.

In fact, a recent discussion on the needs and challenges of multimedia research in the context of Web 2.0 pointed out that understanding how users typically behave (e.g., which interactions they establish) is of great relevance as users play an important role in the social network system [3]. As an example, the design of effective video content classification mechanisms seems crucial for automatic identification of videos with malicious content such as copyright protected, pornography or spams. However, content classification based solely on the *bare* content can be a challenging research problem due to the typically low quality of user generated videos [3] and the multitude of strategies one can make use of to publicize (malicious) content in a video (URL to website, static image or streaming). In contrast, understanding how users interact with each other in a social video sharing system may highlight aspects inherent to the way malicious users act, which, in turn, may be used in a much more effective way in the detection (and possibly removal) of malicious or unwanted content.

In this paper, we give a first step in this direction. Our goal is to understand user behavior in a social network created essentially by *video* interactions. Thus, we present a characterization of a social network created by the video response interactions among users in YouTube, the most popular social video-based media network today, generating high-volumes of Internet traffic (over 3.4 billion videos streamed in December 2007 [1]). The YouTube video response feature allows users to converse through video, by creating a video sequence that begins with an opening video and an array of responses from fans and detractors who respond with videos of their own. Our characterization highlights the social networking issues that influence the behavior of users interacting primarily with stream objects, instead of textual content traditionally available on the Web. Furthermore, our analysis reveals evidences of anti-social behavior in video interactions. To the best of our knowledge, this is the first effort towards understanding video interactions issues in social network.

The rest of the paper is organized as follows. Next section describes how we crawled YouTube. In Section 2 we present a characterization of video interactions in the social network formed by relationships between users. In Section 4 we discuss user behavior characteristics and anti-social behavior existent in video interactions. Finally, we conclude in Section 5.

2. CRAWLING A SOCIAL NETWORK

To collect data, we visit pages on the YouTube site and gather information about video responses and their contributors. We say a YouTube video is a *responded video* if it has at least one video response. A responded video has a sequence of video responses listed chronologically in terms of when they were created. We say a YouTube user is a *responded user* if at least one of its contributed videos is a responded video. Finally, we say that a YouTube user is a *responsive user* if it has posted at least one video response.

A natural user graph emerges from video responses. 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

[1]http://www.comscore.com/press/release.asp?press=1264

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

Algorithm 1: Crawler

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 $x_2 \in X$. Since YouTube does not provide a means to systematically visit all the responded videos, we design a sampling procedure that allows us to obtain a subgraph (A, B) of (X, Y) with the following properties: 1) Each connected component in (A, B) is a connected component in (X, Y); that is, the sampled subgraph (A, B) consists of (entire) connected components from (X, Y), which is important to analyze social networking aspects. 2) The subset A covers a large fraction of X. 3) The most responded users are included in A, which ensures that we are including the most important users, and only neglecting users who have few responded videos. To this end, we designed the sampling procedure described in Algorithm 1.

Using any starting seed set, the algorithm 1 ensures that the resulting graph has property 1. We ran this sampling procedure with two different seed sets. Our first seed set uses the contributors of the all-time top-100 responded videos. The sampled graph (A, B) obtained from this seed set is the graph analyzed on the next sections. Our second seed set consists on users obtained from the random sampling technique described in Algorithm 2. This second data set is used only to verify properties 2 and 3 of graph (A, B).

Of the 100 random seed users, we find that 67 of those users belong to A. Thus, our sampling scheme satisfies Property 2. To verify Property 3, we rank order the 10 most, 100 most and 1000 most responded users from our second data set. We find that (A, B) contains all 10 of the 10 most responded users, 98 of the 100 most responded users, and 951 of 1000 most responded users. Thus, Property 3 is verified as well. The basic statistics of our all-time top-100 crawl is provided in Table 1. The following sections present results only for this dataset, as the main conclusions hold for the random dataset as well.

```
input : A list of words from a dictionary
Select a random word from the dictionary;
Search tag using YouTube API, using the word as tag;
foreach Contributor C of the videos found do
    if C is a responded OR a responsive user then
        Add user to list;
    end
end
Randomly select 100 users from the final list;
```

Algorithm 2: Find random seeds

3. NETWORK CHARACTERISTICS

This section presents characteristics of social networks that emerge from the user graph (A, B). Table 2 presents the main statistics of graph (A, B) and its largest strongly connected component (SCC).

3.1 Degree Distribution

The key characteristics of the structure of a directed network are the in-degree (k_{in}) and the out-degree (k_{out}) distributions. As shown in Figure 1, the distributions of the degrees for the entire graph follow power laws $P(k_{in,out}) \propto 1/k_{in,out}^{\alpha^{in,out}}$, with exponent $\alpha^{in} = 2.096$ and $\alpha^{out} = 2.759$ with the following coefficient of determination: $R^2 = 0.98$ and $R^2 = 0.97$. The scaling exponents of the whole network lie in a range of 2.0 and 3.4, which is a very common range for social and communication networks [10]. Our results agree with previous measurements of many real-world networks that exhibit power law distributions.

The in-degree exponent is smaller than the exponent of the out-degree distribution, indicating that there are more users with larger in-degree than out-degree. This fact suggests a link asymmetry in the directed interaction network. Unlike other social networks that exhibit a significant degree of symmetry [9], the user interaction network shows a structure similar to the Web graph, where pages with high in-degree tend to be authorities and pages with high out-degree act as hubs directing users to recommended pages [8]. In order to investigate this point further, Figure 2 (left) shows the cumulative distribution of ratios between in-degree and out-degree for the user interaction network. The network has 60% of the users with out-degree higher than in-degree and 5% of the users with significantly higher in-degree than out-degree. This is evidence that a few users act as “authorities” and “hubs”. We have observed in our dataset that authority-like users (that is, highly responded users), with high in-degree, are typically media companies that upload professional content, including sports, entertainment video and TV series. Nodes with very high out-degree may indicate either very active users or spammers, i.e., users that distribute content that legitimate users have not solicited.

We now investigate assortative mixing, a graph theoretical quantity typical of social networks. A network is said to exhibit assortative mixing if the nodes with many connections tend to be connected to other nodes with many connections. Social networks usually show assortative mixing. The assortative (or disassortative) mixing is evaluated by the Pearson coefficient r, which is calculated as follows [10]:

$$r = \frac{\sum_i j_i k_i - M^{-1} \sum_i j_i \sum_{i'} k_{i'}}{\sqrt{[\sum_i j_i^2 - M^{-1}(\sum_i j_i)^2][\sum_i k_i^2 - M^{-1}(\sum_i k_i)^2]}}, \quad (1)$$

where j_i and k_i are the excess in-degree and out-degree of the vertices that the ith edge leads into and out of, respectively, and M is the total number of edges in the graph.

Table 2 shows values of r for the directed graph of the interaction network. The video response user graph has a disassortative mixing $r = -0.017$, where high degree nodes preferentially connect with low degree ones and vice versa. Analyzing only the largest SCC, we can observe an assortative mixing $r = 0.017$. The existence of a significant assortative mixing is associate with the notion of social communities [10]. So, the entire user interaction graph does not show evidence of formation of a social community, differently from its largest SCC.

3.2 Clustering Coefficient

It has been suggested in the literature that social networks possess a topological structure where nodes are organized into com-

characteristic	top-100
Period of sampling	09/21-09/26/07
# videos collected	3,436,139
# video responses	417,759
# views	20,645,583,524
# views of responses	2,826,822,374

Table 1: Summary of all-time top-100 Data Set

Characteristic	Dataset	Largest SCC
# nodes	160,074	7,776
# edges	244,040	33,682
Avg Clustering Coefficient	0.047	0.137
# nodes of largest SCC	7,776	7,776
# components	149,779	1
r	-0.017	0.017
Avg distance	8.40	8.40
Avg k_{in} (CV)	1.53 (9.38)	4.33 (3.14)
Avg k_{out} (CV)	1.53 (1.717)	4.33 (1.28)

Table 2: Summary of the Network Metrics

munities [10], a feature that can account for the values for the clustering coefficient and degree correlations. The clustering coefficient 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. The clustering coefficient of a network, CC, is the mean clustering coefficient of all nodes. The average CC over the whole network is $CC = 0.047$, whereas the mean clustering coefficient for a random graph with identical degree distribution but random links is $CC = 0.007$, which shows the presence of small communities in the video-response network. The leftmost part of Figure 3 shows the cumulative distribution of the clustering coefficient. The network contains a significant fraction of their nodes with zero clustering coefficient. Specifically, 80% of all nodes in the entire user interaction network have $CC = 0$. This feature indicates that there is a clear difference on average between clustering in the entire network and the components of the network. The right part of the figure shows how the clustering coefficient varies with the node out-degree. Higher values of the clustering coefficient occur among low degree-nodes, suggesting the lack of large communities around high-degree nodes. Our conjecture is that highly responsive users do not necessarily have social links with the contributors of the videos that they are responding to. Therefore, there may not exist a sense of community among the users that receive video responses from a single responsive user. Low degree nodes might explain the formation of very small communities, composed of a few people like a family or a group of friends.

Table 2 shows social characteristics of the largest SCC of (A, B). Figure 2 (right) shows the distribution of the size of the strongly connected components sorted from the largest component to the smallest one. The distribution suggests a general structure that includes the largest SCC, the middle components (i.e., 1,974), and a large number of components with just one node (i.e., 147,805). As we are working with a directed graph, these components with size one are nodes with links in only one direction. The middle components are groups of users which represent small size communities (e.g., families and groups of friends) that express their interests and establish communication via video responses. The largest SCC represents about 5% of the nodes, but it is considerably larger than the others. It concentrates 10% of the views and 22% of the video responses and deserves further analysis. Although it includes about 5% of the nodes, its size is comparable to the size of SCC in other networks, derived from blogosphere samples [1]. The differences in size of connected components may be due to time factors, that account for the adoption by users of specific features (i.e., video response) in social networking environments. In order to understand

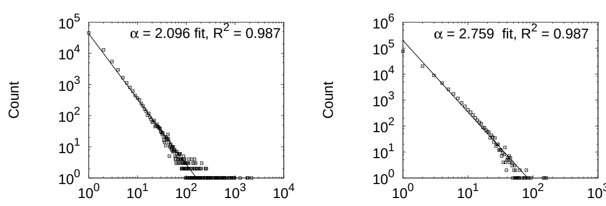

Figure 1: In-Degree and Out-Degree Distributions

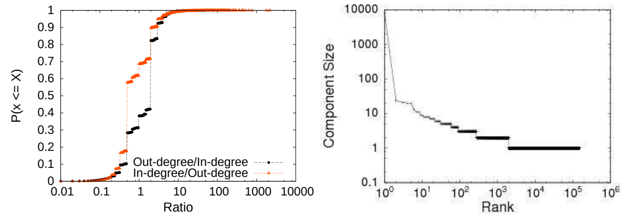

Figure 2: CDF of in-degree to out-degree ratio (left). Component Size vs Rank (right).

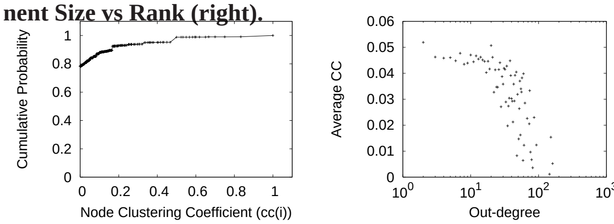

Figure 3: Distribution of CC (left) and average CC as a function of node out-degree (right)

its characteristics, we investigate network properties of the largest SCC in the video response user graph. The average clustering coefficient of the largest SCC is $CC = 0.137$, three times greater than the clustering coefficient of the entire network. Thus, user interactions captured by the largest SCC might form a more tightly connected community.

4. ANTI-SOCIAL BEHAVIOR

Different forms of unsolicited communication are taking a toll on users of social networking services [12]. Unsolicited communication opens a large gray area, where videos could be considered spam or promotion. The simplest form of spam occurs when users submit a video with a long list of misleading tags to describe its content in order to video searching mechanisms [7]. Another form of video spam occurs when a video is posted as a response to an opening video, but whose content is completely unrelated to the opening video [2]. On the other hand, promotion consists on users trying to boost the ranking of their videos to make them highly visible in the social network ranks. Due to its intrinsic nature, video response appears to be an attractive feature to users motivated to spam in order to promote specific content, advertise to generate sales, disseminate pornography (often as an advertisement) or simply compromise the system reputation. Unlike textual responses or comments, one has to start the streaming and view it to realize the specific video is some form of spam or promotion, consuming system resources, in particular bandwidth, and compromising user patience and satisfaction with the system. In this section we focus on the use of metrics that help to understand different types of users that participate in video-based interactions in social networks.

4.1 Characterizing User Behavior

Simple features can be used as a first cut in the identification of anti-social behavior. We define the inter-reference distance (IRD) on the sequence of users that upload video responses to video i as the total number of responses that appear between two video responses from the same user. In order to calculate a user's IRD we compute his IRD for each video responded by the user. Then, we compute the user's average IRD. Previous studies [6] on spam characterization refer to the importance of analyzing temporal issues. For example, whereas traditional e-mail traffic is concentrated on diurnal periods, the arrival rate of spam e-mails is roughly stable over time [6]. With IRD, we want to assess temporal patterns of the users' participation in a sequence of video responses. Figure 4 plots the average IRD for each responsive user as a function of the

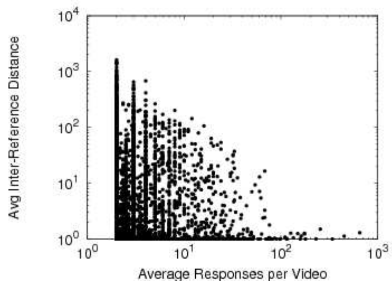

Figure 4: Temporal patterns interactions through videos

average number of video responses per video responded by the user. A user that uploads many video responses per video, one after the other, like a mechanical process, might be a candidate for further investigation. Thus, the combination of a large number of video responses per video and small IRD suggests the user has some type of anti-social behavior. We conducted an investigation to verify if the combination of two metrics (i.e., IRD and average number of responses per video responded) could accurately be used to identify users with anti-social behavior. Our experiment focus on users that are located on the rightmost part of Figure 4. In the dataset, a total of 298 users have average IRD less than 3 and average number of responses per video greater than 10. A group volunteers in our laboratory randomly selected 95 users, which represent the dataset with a confidence interval of 90% and an error of 7%. Our volunteers then viewed the users' video responses and classified responsive users into two categories, according to the content of their video: social or anti-social user. If at least one video response is considered spam or promotion, the responsive user is labeled as anti-social. A total of the 80% users that meet the specified requirements were classified as anti-social user, suggesting that the proposed metrics could be a starting point to develop heuristics to combat anti-social behavior in video interactions.

4.2 User Rank

The next step is to use the structure of the social network for detecting anti-social patterns. We use the PageRank [4] algorithm, on the video response user graph, to determine the importance of a user in the network. In the PageRank algorithm, a Web page has a high rank if the page has many incoming links or a page has links coming from highly ranked pages. We call the scores computed by the algorithm as UserRank, which could be used as an indicator of the importance of users in terms of their participation in video interactions. The correlation between UserRank and in-degree is 0.13, suggesting that the relationship between the two metrics is weak. Figure 5 (left) shows a strong correlation between UserRank and total number of views of the user's videos, suggesting the algorithm captures the importance of users in terms of number of views. We verified the profiles and the video responses of a few users with high and low rank. Users with high rank are among the most responded and viewed. Most of them are directors (i.e, a director account has special privilegies in YouTube). Low rank users have small number of views and post video responses to some videos but receive no or few video responses from the video community.

We now propose the use of UserRank to detect users boosting a video ranking (namely promoted videos). Intuitively, boosting a video ranking can bring some extra visibility to a video. However, the video can also be viewed as spam or undesired content for other users. Moreover, videos which quickly reach a high ranking are strong candidates to be kept in caches or in content distribution networks (CDNs) [5]. Thus, promoted videos can be confused with a popular video, impacting performance and scalability of online social video sharing services.

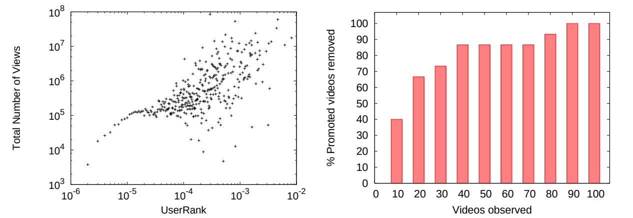

Figure 5: UserRank correlated with number of views (left) and detection of promoted videos (right)

Our strategy to remove promoted videos consists on identifying suspected videos based on the UserRank of their owners. Suppose a user have a promoted video on the top-100 most responded videos. The video had its rank boosted by a series of self responses or video responses posted by fake accounts with low rank. Clearly, the UserRank of the owner of this promoted video is low compared to the owners of the non promoted videos on the top-100 list, since they received posts from several different users with different ranks. In order to verify if the UserRank can be useful to point owners of promoted videos among the top-100 most responded videos, we conduct the following experiment. We progressively observed 10 videos from the top-list of YouTube, selecting the videos ordered by their owners UserRank. Figure 5 (right) reports the percentage of promoted videos identified from the total existent on the top-100 list. By observing 40 videos of the users with lowest UserRank we are able to identify 87% of the promoted videos existent among the top-100 most responded videos, which is much higher than if we had selected the videos to observe randomly.

5. CONCLUSION

Based on the analysis of video interactions, our work raises a number of questions about user behavior in a social network and shows evidence of anti-social behavior, such as self-promotion and other forms of content pollution. Our current and future work is focused on evaluating the use of network characteristics to identify spammers in online social networks.

6. REFERENCES

[1] N. Ali-Hasan and L. Adamic. Expressing social relationships on the blog through links and comments. In *ICWSM*, 2007.

[2] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross. Identifying video spammers in online social networks. In *Proc. of AIRWeb*, Beijim, China, 2008.

[3] S. Boll. Multitube–where web 2.0 and multimedia could meet. *IEEE MultiMedia*, 14(1):9–13, 2007.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of WWW Conference*, 1998.

[5] 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.

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

[7] 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.

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

[9] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of IMC*, 2007.

[10] M. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, 2003.

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

[12] A. Zinman and J. Donath. Is britney spears spam? In *Proc. of CEAS*, 2007.