The YouTube Social Network

Mirjam Wattenhofer
Google Zurich
mirjam@google.com

Roger Wattenhofer
ETH Zurich
wattenhofer@ethz.ch

Zack Zhu[*]
ETH Zurich
zazhu@ethz.ch

Abstract

Today, YouTube is the largest user-driven video content provider in the world; it has become a major platform for disseminating multimedia information. A major contribution to its success comes from the user-to-user social experience that differentiates it from traditional content broadcasters. This work examines the social network aspect of YouTube by measuring the full-scale YouTube subscription graph, comment graph, and video content corpus. We find YouTube to deviate significantly from network characteristics that mark traditional online social networks, such as homophily, reciprocative linking, and assortativity. However, comparing to reported characteristics of another content-driven online social network, Twitter, YouTube is remarkably similar. Examining the social and content facets of user popularity, we find a stronger correlation between a user's social popularity and his/her most popular content as opposed to typical content popularity. Finally, we demonstrate an application of our measurements for classifying YouTube Partners, who are selected users that share YouTube's advertisement revenue. Results are motivating despite the highly imbalanced nature of the classification problem.

Introduction

YouTube is a key international platform for socially-enabled media diffusion. According to public statistics, more than 48 hours of video content is uploaded every minute and 3 billion views are generated every day. To complement the content broadcast/consume experience, YouTube connects seamlessly with major online social networks (OSNs) such as Facebook, Twitter, and Google+ to facilitate off-site diffusion. In fact, 12 million users have linked their YouTube account with at least one such OSN for auto-sharing, and more than 150 years of YouTube content is watched on Facebook every day.

More importantly, YouTube serves as a popular social network on its own, connecting registered users through subscriptions that notify subscribers of social and content updates of the subscribed-to users. In this paper, we leverage full-scale data of the YouTube social network to answer practical questions from a graph theoretic point of view. We shed light on the following:

1. What can be observed from the complete YouTube social network topology? How does it compare with other social networks in terms of intrinsic properties and emergent observations?
2. On the YouTube platform, how do users connect and interact with each other? What is the relationship between the explicit and implicit social graphs describing the subscription and commenting relationships?
3. What constitutes popularity on YouTube? How does a user's topological (social) popularity correlate with his/her content popularity?

Our analyses will illustrate that YouTube deviates significantly from traditional OSN characteristics. However, it concurs with the observations of Kwak, Lee, Park, and Moon for the Twittersphere (2010). We will see a surprising dichotomy of content and social activities on the YouTube platform, indicating that YouTube is, distinctly, as much a social network as it is a content-diffusion platform. Finally, we note social popularity on YouTube correlates more with the maximum content popularity achieved as opposed to the summary measures of content popularity.

These observations lead to the conjecture that a new class of social network is emerging, a type that facilitates indirect socialization via a gluing content layer in between directed users-to-user interaction. Potentially, a paradigm shift is taking place for OSNs such that what constitutes "social" now incorporates user-content-user interaction in addition to the traditional user-user interaction. Through intrinsically different linking and interacting characteristics, these content-driven social networks create new social dynamics and necessitate further research to better understand their role in the processes of socialization and information diffusion.

Background and Related Work

Recent surge of OSN popularity has attracted the attention of researchers from a variety of fields. Here, the surveyed works are divided into two categories: OSN measurements and machine learning-based OSN applications.

[*] Zack Zhu conducted this research while visiting Google. He is currently in the Wearable Computing Laboratory at ETH Zurich.

OSN Measurements

Researchers have taken advantage of the YouTube Data API to measure a variety of metrics in relation to video popularity on YouTube. Studies by Cha et al. (2007), Benevenuto et al. (2009), and Cheng et al. (2008) all analyze this video corpus for the purposes of understanding video popularity. However, by measuring at the video-level, user-based social characteristics are largely omitted. We build on these works by aggregating YouTube's video corpus metrics at the user level to complement content metrics with social topology. In this way, we are able to make the connection between video content popularity and corresponding social popularity. Even though these works make efforts to avoid sampling biases while accessing the data through a rate-limiting YouTube API and/or online crawlers, the datasets collected are only fractions of the entire corpus. In this work, we obtain data from within YouTube to allow complete measurements.

Researchers Mislove, Marcon, Gummadi, Druschel, and Bhattacharjee (2007), Krishnamurthy et al. (2008), and Kwak et al. (2010), have reported measurements of various major online social networks. In the work of Mislove et al. (2007), an array of graph-based measurements are presented for multiple popular online social networks, including YouTube. They present a framework of measurements that we adopt here for ease of comparison. In their measurement methodology, a mix of API and HTML scraping techniques are used to obtain a sampled version of the social graph. However, as the authors themselves pointed out, their methodology is limited when trying to extrapolate observations to the entire YouTube population. This work addresses this concern and compares our results where appropriate. In Krishnamurthy et al. (2008), the authors examine topological features of a sampled Twitter network as well as content uploaded from users. This work is similar to what we present for YouTube as we analyze measurements from two social graphs and the video corpus. Recently, Kwak et al. (2010) conducts one of the first full crawls of a major online social network by measuring the entire Twittersphere. The size and coverage of their dataset is comparable to what we present here.

OSN Applications

In terms of user classification, De Choudhury et al. (2010) proposes threshold networks with non-arbitrary thresholds for increased accuracy in both link prediction and user classification. In this work, the idea of thresholding to prune real-world datasets is used to illustrate an interesting relationship between explicit and implicit social relationships. Hong et al. (2011) leverage network characteristics to successfully predict popular messages and Bakshy et al. (2011) classify "influential" users according to re-tweet quantities. A key similarity between these works is their use of various topological metrics calculated from the social graph. In our work, such features are utilized as well in our classification application. On top of the explicit social graph, topology measures of an implicit social graph and aggregated user-level metrics from the video corpus are used as well.

Table 1: Nodal feature descriptions

Name	Description
user	Encrypted user id
sub.out	Out degree on subscription graph
sub.in	In degree on subscription graph
avg.pub.out	Average out degree of users subscribed to
avg.pub.in	Average in degree of users subscribed to
avg.sub.out	Average out degree of subscribers
avg.sub.in	Average in degree of subscribers
reciprocal	# of reciprocal links on subscription graph
sub.pagerank	PageRank of the subscription graph
com.in	In degree on the comment graph
com.out	Out degree on the comment graph
com.pagerank	PageRank on the comment graph
max.fav	Max # of times any video is favourited
med.fav	Median # of times any video is favourited
min.fav	Min # of times any video is favourited
max.views	Max # of times any video is viewed
med.views	Median # of times any video is viewed
min.views	Min # of times any video is viewed
max.coms	Max # of comments any video received
med.coms	Median # of comments any video received
min.coms	Min # of comments any video received
max.raters	Max # of raters for any video
med.raters	Median # of raters for any video
min.raters	Min # of raters for any video
max.avg.rating	Max of average ratings for any video
med.avg.rating	Med of average ratings for any video
min.avg.ratings	Min of average ratings for any video
main.cat	Category that most videos are uploaded in
uploads	Number of videos posted

Measuring All of YouTube

As mentioned in the previous section, a multitude of work has sampled various major OSNs through online crawls and/or API usage. However, few measurement projects have captured whole social graphs without compromise. This work leverages the data and computing power available within Google to shed light on a major social platform.

The data collection process mainly utilizes MapReduce (Dean and Ghemawat 2008) and Pregel (Malewicz et al. 2010), a large-scale proprietary graph computing framework, to leverage Google's computing resources. Therefore, the runtime to capture entire datasets can be completed in tens of minutes, capturing and processing complete social graphs on the YouTube social network. We base our analyses of the YouTube social network on three main corpora of data: the explicit social graph depicting subscriptions, the implicit social graph depicting commenting activities, and aggregated metrics of user-uploaded content. These datasets were captured in August 2011. We removed axis labels on our plots to preserve data confidentiality.

We compose a directed graph to represent the subscription relationships of registered YouTube users. Each node represents one such user while a link points from a subscriber to the user subscribed-to. Therefore, this graph is composed of registered users who have subscribed to at least one user or received at least one subscription. Similarly, the comment graph is composed of users who have posted or received at least one comment. Again, links point from the commenter to the comment-receiving user. Both graphs contain nodes on the order of hundreds of millions and links on the order of billions, comparable to the measurement size of Kwak et al. (2010) for Twitter.

The third corpus of data is an aggregation of YouTube video metrics to the uploader level. For each uploader $\mu \in N$ who has uploaded at least one video, we can construct a video vector $\vec{v}_\mu$. Then, for each of our video-level metrics (e.g. view count, number of video comments, etc.) we aggregate by taking the minimum, median, and maximum of $\vec{v_\mu}$. For example, a specific user's median average rating refers to the scalar $m_\mu = median(v_\mu^1, v_\mu^2, ..., v_\mu^n)$, where each v^n denotes the average rating for the n^{th} video that user μ has uploaded. Table 1 presents the user features accumulated from the three datasets. The naming convention in Table 1 will be referred to consistently from here on.

Degree Distribution of Social Graphs

Starting off with basic degree distribution analysis of both social graphs, Figure 1 plots the log-log complementary cumulative distribution function (CCDF). As done in most OSN studies, the degree distributions are typically generated as log-log CCDF to better illustrate the tail behaviour on both ends. To interpret the plots, it can be understood, for an (x, y) pair, a fraction y of the population has a degree more than x. Noticeably, there is a sharp kink in the subscription out-degree curve at $x = x^*$. This is an artifact of the YouTube subscription rule that limits the number of subscriptions for users who do not have a significant number of subscribers themselves. In both graphs, there exists extremely popular users who have millions of subscribers and/or commenters. However, about half of the sampled population has one or zero subscriber and/or commenter. In the comment graph, the out-degree distribution also does not fit the power-law signature. Fitting the power-law distribution (via maximum likelihood estimation) to the in-degree of the subscription graph and the comment graph, scaling exponents of 1.55 and 1.44 are found, respectively. These exponents differ from the majority of real-world social networks, which have been measured between 2 and 3 (Kwak et al. 2010) as well as 1.99 (Mislove et al. 2007) for a sampled YouTube subscription graph.

A Content-Driven OSN

Traditionally, researchers have modelled social networks, online or offline, as undirected links between users. Intrinsically different for YouTube, the *de facto* mode of linking is through *directed* subscription links. Furthermore, user interaction is largely through uploaded video content. For example, as opposed to interacting directly (e.g. wall posts, direct messages), much of the interaction on YouTube takes place in a video-centered manner, such as rating another's video or leaving a video comment. Therefore, user interaction becomes very much a user-content-user relationship where users interact with each other through a gluing layer of uploaded content. These two inherent characteristics of a content-driven OSN, such as YouTube, may explain why significant differences exist from traditional OSNs.

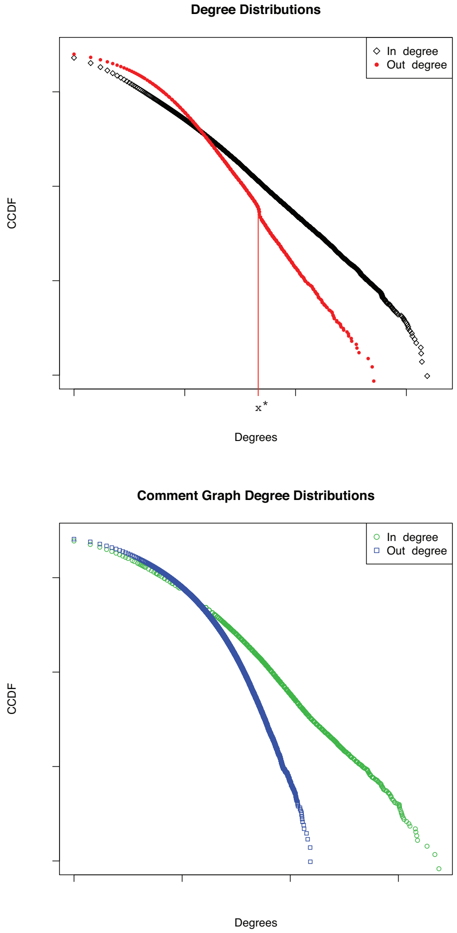

Figure 1: Degree distribution of subscription graph (top) and comment graph (bottom) plotted on log-log scales.

In our measurement, we find the YouTube subscription graph deviates in three aspects that mark traditional OSNs: assortative linking, user homophily, and reciprocity. In addition, contrary to traditional OSN studies that have illustrated social interactions as a strong signal for social links (Xiang, Neville, and Rogati 2010; Kahanda and Neville 2009), we note a surprising dichotomy of commenting activities and subscription links on the YouTube platform.

Assortativity, Reciprocity, and Homophily In Figure 2, the assortative linking property can be examined for the subscription graph. We log-bin user in-degrees on the x-axis and plot the distribution of average publisher (users receiving the subscription) in-degrees on the y-axis in the form of a box-plot. It is clear that the median of bins across most in-degrees hold steady as users consistently subscribe to those with much more popularity (via in-degree) than that of themselves. Although there are significant variances associated with the bins, the median level consistently show that

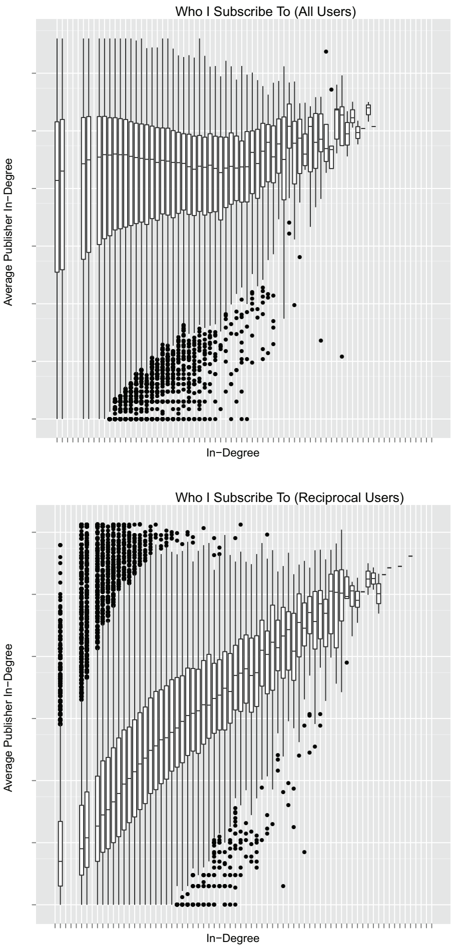

Figure 2: Who subscribes to whom (top) and Who subscribes to whom for reciprocal users (bottom). A significant difference in linking assortativity can be seen between the top and bottom plots.

users on YouTube tend to subscribe to power-users with indegrees orders of magnitude larger. Thus, assortative linking largely does not take place on the YouTube subscription network, as most of the population is contained in the lower indegree bins per the power-law distribution shown previously. This measurement agrees with an earlier measurement study on a sampled YouTube dataset (Mislove et al. 2007), where YouTube was found to be the only network with a negative assortativity coefficient (disassortative linking) among the measured social networks. Reciprocal linking on traditional OSNs like Facebook and Orkut exist by definition as the user-user links are undirected. However, with directed links such as subscription or following (on Twitter), two links between a pair of users are required to form a reciprocal relationship. Hence, a new dynamic emerges as users can now perceive the reception of subscription as a symbol of authority or interest from others. In effect, highly subscribed-to users of YouTube tend to have a high in-degree/out-degree ratio as they rarely subscribe to others. This phenomenon explains the low-levels of reciprocity observed in our dataset.

Measuring reciprocity on the YouTube subscription network, we found only 25.42% of the users to have one or more reciprocal links. This level is quite small when compared to measurements of other directed social networks such as Flickr at 68% (Cha, Mislove, and Gummadi 2009) and Yahoo! 360 at 84% (Kumar, Novak, and Tomkins 2006). However it is remarkably close to the 22.1% of the population on Twitter reported by Kwak et al. (Kwak et al. 2010). This shows that, like Twitter, YouTube users subscribe to the notion of influence via subscription links as opposed to real-life social relationships as links typically depict in traditional OSNs.

Defining reciprocative users as those with more reciprocative out-going links than non-reciprocative out-going links, approximately 15% of the user population are reciprocative users. The bottom plot of Figure 2 shows the same "Who I Subscribe To" plot for these users. Interestingly, we now observe assortative linking by the median of the in-degree bins, which is strikingly different compared to what is above. Although defying outliers exist, largely, reciprocative users are significantly more assortative in their subscription behaviour than non-reciprocative users. This suggests that user linking behaviour may be guided by the linking mechanism (directed or undirected) and ultimately affect the dynamics of the social network. On YouTube, even though users link directly and the majority link non-reciprocatively and disassortatively, there exists a subset of the measured population that demonstrate traditional social behaviour, despite the linking mechanism of YouTube.

To examine user homophily, we capture the video category that each user uploads the most content in. Then, this upload category is compared between linked users on the subscription graph and comment graph. Examining who each user is linked with (inward links *and* outward links), it is observed that on average, only 26.58% (*s.d.* of 3.33%) of a user's subscription neighbor set have the same mode upload category. Correspondingly, the average is 27.46% (*s.d.* of 3.50%) on the comment graph. Further, on the subscription graph, only 12.49% of users have more neighbours in the same main upload category than neighbours in another category. Similarly on the comment graph, 10% of the users have more than 50% of their neighbours in the same main upload category. Therefore, at the user-level, we observe a lack of homophily between linked users when comparing the mode upload category, which may be loosely used to represent user interest.

Again, this deviates from what is reported for traditional OSNs (McPherson, Smith-Lovin, and Cook 2001). In 2010, Weng et al. (2010) statistically tested for the presence of homophily in a sampled dataset of 6748 Singapore-based users on Twitter. They concluded, with high probability, that users linked with ”following” relationships are interested in similar topics and used it to as basis to explain observed reciprocity. However, both Cha et al. (2010) and Kwak et al. (2010), with a near-complete dataset of Twitter, found low rates of reciprocal linking and a lack of homophily on Twitter. Here, our results for YouTube show a lack of homophily and reciprocity to confirm this phenomenon of content-driven OSNs.

A Dichotomy of Interaction and Subscription Overlapping the subscription graph and the comment graph, we study the nodes who exist in both graphs and compare their incoming links. Formally, for each node μ, we can analyze the overlap of its commenter set $C_\mu \in \{c_1, c_2, ..., c_n\}$ and subscriber set $S_\mu \in \{s_1, s_2, ..., s_n\}$. Denoting a set $O_\mu := S_\mu \bigcap C_\mu$, we calculate the overlap percentage as $\rho_\mu = \frac{O_\mu}{S_\mu \bigcup C_\mu}$.

Averaging over all n nodes that exist in both graphs, we find $\bar{\rho} = 9.6\%$ (*s.d.* of 1.9%). Comparing the overlapped set with each of the two neighbourhood sets, it is found that, averaging over all users, $\bar{\alpha} = \frac{1}{N}\sum_{\mu=0}^{N} \frac{O_\mu}{S_\mu} = 18.1\%$ and $\bar{\beta} = \frac{1}{N}\sum_{\mu=0}^{N} \frac{O_\mu}{C_\mu} = 16.5\%$. These results show that, on average, only a small portion of a user's subscribers are also commenters and vice versa. This differs from traditional OSNs such as Facebook since links are formed based on social relationships and interaction is basis for relationship. Here, we observe the user-content-user relationship that differentiates the interaction dynamic for YouTube as a content-driven OSN.

Sensibly, it is expected that users leave comments on videos to address the video content, without necessarily implicating a social or influence relationship with the uploader. On the other hand, it is surprising that subscribers are also not likely to comment on videos uploaded by users they subscribe to, as indicated by a low $\bar{\alpha}$. At the same time, it is found that both the subscription graph and the comment graphs are as active as each other with similar number of nodes and number of unique (for comment graph) pairwise links. This suggests an interesting nature of YouTube—there is a dichotomy between the dynamics of "content" (user$\rightarrow$content) and the dynamics of "social" (user$\rightarrow$user), where content consumption and social influence are similarly active, but largely, separate components of the same system.

Naturally, commenting relationships between users can be captured as a directed graph with weights, where each weight represents the frequency of comments in the direction of the link. Therefore, a modified comment graph can be generated by pruning edges with a threshold, τ, below some frequency. Numerically, for $\tau = 1, 2, 5, 10$, the resulting $\bar{\rho}, \bar{\alpha}$, and $\bar{\beta}$ are plotted in Figure 3. Expectedly, the thresholded commenter set steadily decreases in its ability to recall subscription links ($\bar{\rho}$). However, the thresholding effect significantly increases the overlap proportion between the intersection and the commenter set, $\bar{\beta}$. From Figure 3, we can see that on average, >50% of commenters who have commented equal to or more than 10 times are subscribers, compared to less than 20% without thresholding. In effect, simply thresholding on comment frequency produces a smaller set of commenters who are more likely to be subscribers as well. This shows that although the majority of the commenter and subscriber sets follow the dichotomy mentioned above, a significant portion of *repeat* commenters break this dichotomy between commenters and subscribers.

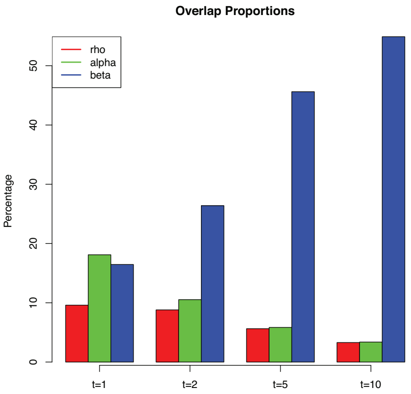

Figure 3: Implict-explicit social graph overlap for various comment threshold levels.

What is YouTube Popularity?

Although various methods exist, the proxy measure of popularity on YouTube comes from the number of subscribers. In this work, we measure and compute additional features that represent social and content popularity. In this section, we provide some analysis of how these features relate to each other.

The top plot of Figure 4 depicts in-degree of users (with at least one upload) against the number of videos uploaded (undeleted) on log-log scales. There is a linearly increasing trend in the median of each logscale-bin, suggesting that, typically, users increase the number of uploads as they become more popular. However, this increase in median flatlines and picks up again for extremely popular users. Despite the trend observed in the median of each bin, a large number of outliers exist on the lower-end of the in-degree scale. Amongst these, some users with 0 or 1 subscriber are uploading thousands of videos. This points to the fact that although many users take advantage of the subscription service to link to others, a significant number of users simply use YouTube as a content diffusion network without any need to connect "socially".

In the same figure, we plot the relationship of PageRank and In-Degree for the subscription graph. The main idea behind PageRank is to allow propagation of influence over the network (Easley and Kleinberg 2010). Instead of just counting the number of edges, it takes into account *who* subscribes. From the bottom plot of Figure 4, we see a linearly increasing relationship between a user's in-degree (also log-binned) and PageRank. This is not surprising as a user with more subscribers obtain "influentialness" from a larger number of subscribers. However, a large number of outliers with relatively high PageRank exist in bins containing low in-degree users. In these cases, the relatively few subscription links come from subscribers who are very influential. Even though these outlier users are still one or two

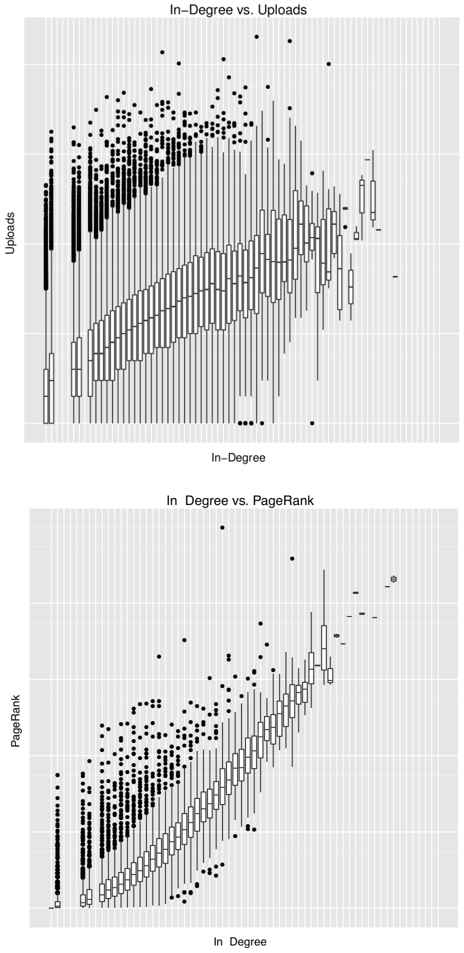

Figure 4: In-degree vs. quantity of uploaded content (top) and PageRank (bottom)

orders of magnitude less in PageRank than the top in-degree users, they may be emerging power-users in the future. We leave the study of evolutionary dynamics of these users to future work. As mentioned previously, measures in addition to subscriber quantity can capture other facets of user popularity. Figure 5 plots a heatmap of Spearman's rank correlation (Spearman 2011) for the various nodal features measured. This measure ranges between $[-1, 1]$ where -1 denotes perfect anti-correlation while 1 denotes perfect rank correlation. The metrics are arranged such that they are hierarchically arranged with dendograms to show the clusters by correlation strength. Note, as the minimum feature correlation is slightly greater than -0.5, the lower end of the color spectrum (red) depicts -0.5 and *not* perfect anti-correlation (-1).

In Figure 5, the dendogram can be pruned to show three main clusters based on correlation strength. Looking row-wise, the top group consists of features such as *sub.pagerank*, *com.pagerank*, *sub.in*, *com.in*, which are social features that gauge popularity. In the same group by correlation strength, there are content-based popularity measures that depict *maximum* content popularity attained: *max.coms*, *max.fav*, *max. views*, *max.avg.rating*, *max.raters*. In the second group by correlation, typical (median) content popularity and minimum content popularity measures are clustered together. Finally, the third group of correlated measures are measures of user actions such as the out degree on social graphs or upload quantity. From the correlation clusters, social popularity on YouTube is tied, not to *typical* performance in content popularity, but to *maximum* achieved content popularity. This suggests that a single "hit" could dramatically influence a user's social popularity despite a larger proportion of lacklustre uploads.

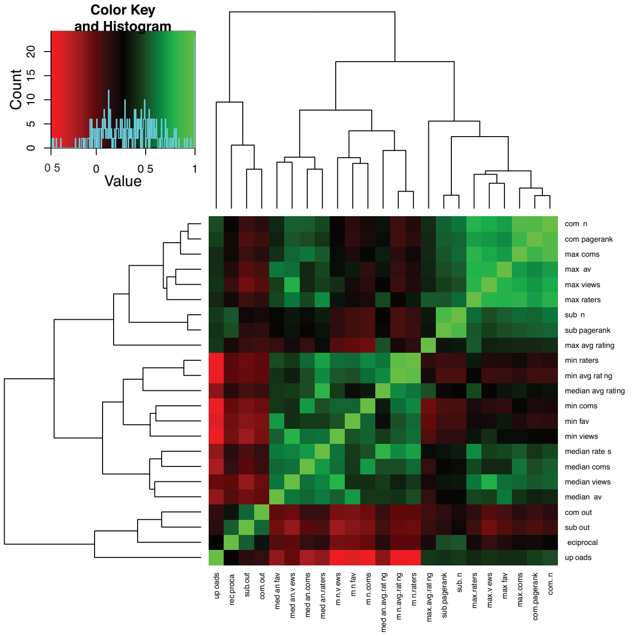

Figure 5: Heatmap of feature correlations

YouTube Partner Classification

Leveraging the explicit, implicit social graphs and content metrics, a simple supervised learning task is described in this section. Here, we will illustrate the effectiveness of the nodal user features in a highly-imbalanced binary classification task.

The YouTube Partner Program

Created in 2007, the YouTube Partner Program (YPP) has over 20,000 partners from 22 countries around the world. It is a program that promotes outstanding content on YouTube by sharing YouTube's advertisement revenues with content creators. To qualify for the program, users go through a manual selection process that considers criteria such as size of audience, quality of content, compliance with YouTube's Terms of Use, among others. To cope with increasing popularity of the YPP, we illustrate a classification system that pre-filters potential candidates to aid the manual selection process. There exists three categories of user partners in the YPP, which we denote as P_1, P_2, P_3. Typically, partners in

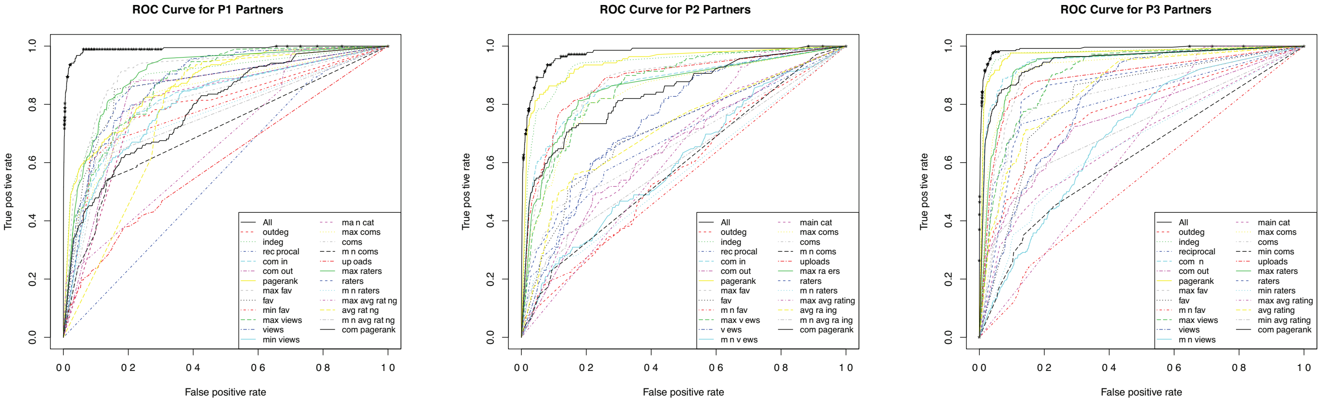

Figure 6: Receiver operator characteristic curve for YouTube partner classification

the P_1 group are private users who have gained significant popularity and consistently influences the YouTube community. P_2 and P_3 partners are mainly companies that establish contracts with YouTube for diffusing content as part of their business. In our setup, a binary-classifier is used to classify each category independently.

The YPP Classification Framework

A critical challenge to the YPP classification problem is the highly-skewed nature of YPP instances, which is a tiny fraction of the YouTube user population. Due to this imbalance, the classifier needs to carefully avoid over-classifying negatively such that promising users are not passed as false negatives. On the other hand, a higher number of false positives is not as critical since this tool serves as part of a discovery process, where manual selection follows.

Classifier Setup Formally, a feature vector $\bar{f}_\mu$ can be constructed for each user μ, where $\mu \in N$ of the set of users who have existing measurements for the three aforementioned data sources. Then, a feature matrix $F \in \Re^{d \times N}$ can be constructed from the N users with $d = 24$ features each. For each of the 3 types of partners, a vector $l \in \Re^{l \times N}$ contains the binary label the the partner type to be classified.

In the training and testing phases of the classifier, a 10-fold cross-validation setup is adopted where $\{F, l\}$ is column-split into $\{F_{TR}, l_{TR}\} \sim 90\%$ of N and $\{F_{TE}, l_{TE}\} \sim 10\%$ of N. Due to the computation limits of running the classifier process in the statistical program R (R Development Core Team 2010), F is uniformly sampled to be approximately 35% of the users who have complete feature information. We use the randomForest package (Liaw and Wiener 2002) in R to train and test the classifier. To ensure there is no shortage of positive instances exposed, the training data contains all positive instances in F_{TR} while negative instances are sampled to maintain a 1:1 ratio with the number of positive instances used. However, it should be noted that all testing results reflect the highly-imbalanced real-world ratio found in the original data.

Table 2: AUCs of YPP classification

	Mean AUC	**Std. Dev.**
P_1 **Partners**	0.9682098	0.02123823
P_2 **Partners**	0.9580012	0.03234237
P_3 **Partners**	0.9428346	0.04128529

Table 3: Top Gini entropy reducing features

Rank	**P1**	**P2**	**P3**
1	indeg	pagerank	pagerank
2	max.fav	indeg	indeg
3	pagerank	com.in	com.in

Classifier Performance In Figure 6, the ROC curve is shown for each of the three types of partners. High AUC is obtained when all features are used in the classification process. Also plotted are the ROC curves of the individual features when used as the only predictor. From the ROC curves, it is apparent that the trained model is able to successfully utilize a mixture of high and low performing features. Table 2 shows a tabulation of the averaged AUC scores in the testing phase for a 10-fold cross-validation setup. Table 3 lists the top 3 features for each partner category in terms of Gini entropy reduction. In the classification results observed, our classifier results in some false positives that drive down precision, however, false negatives remain close to 0 in all three classes. The classifier does a formidable job of correctly filtering out a large number of non-partner users. Therefore, as intended, this classifier may serve as a tool to pre-select real-world users for manual partner selection in the YPP.

Conclusions

In this work, we tie together three full-scale datasets to better understand the nature of the YouTube social network. Compared to recent work, this is one of the most comprehensive measurement studies of a major OSN to date. This work was possible due to the availability of data and computing resources from within Google.

We found that the content-driven nature of the YouTube social network differentiates itself from traditional social networks in terms of user linking and interaction behaviours. Comparing the subscription and comment graphs, we find very little overlap between commenters and subscribers, indicating a dichotomy of “social” and “content” activities within the same system. Examining popularity, we notice video “hits” are more . Finally, we successfully leverage our measurements to classify for pre-filtering potential YouTube partners for manual selection.

This work is one of the first steps towards full-scale OSN measurement and analysis. We propose three areas for future. First, many computationally intensive graph metrics, such as ones involving 2-hop ego networks, were not attempted due to the size of the dataset. These metrics provide additional insight to understand the topology of the social graphs and are invaluable in applications such as link prediction. Second, we do not discuss temporal dynamics of the social graphs in terms of their evolution. For example, tracking the evolution of low in-degree/high PageRank users may reveal interesting insights to how YouTube celebrities emerge. Finally, understanding the interrelation between social graphs and content broadcast/consumption patterns, including temporal analysis, could lead to insights on which/how “viral” videos spread on the YouTube network.

Acknowledgments

We would like to thank Salvatore Scellato for his insights.

References

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone’s an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM ’11, 65–74. New York, NY, USA: ACM.

Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J.; and Ross, K. 2009. Video interactions in online video social networks. *ACM Transactions on Multimedia Computing Communications and Applications* 5(4):1–25.

Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.-Y.; and Moon, S. 2007. I tube, you tube, everybody tubes: analyzing the world’s largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC ’07, 1–14. New York, NY, USA: ACM.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM).*

Cha, M.; Mislove, A.; and Gummadi, K. P. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, WWW ’09, 721–730. New York, NY, USA: ACM.

Cheng, X.; Dale, C.; and Liu, J. 2008. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, 229 –238.

De Choudhury, M.; Mason, W. A.; Hofman, J. M.; and Watts, D. J. 2010. Inferring relevant social networks from interpersonal communication. In *Proceedings of the 19th international conference on World wide web*, WWW ’10, 301–310. New York, NY, USA: ACM.

Dean, J., and Ghemawat, S. 2008. Mapreduce: simplified data processing on large clusters. *Commun. ACM* 51:107–113.

Easley, D., and Kleinberg, J. M. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Hong, L.; Dan, O.; and Davison, B. D. 2011. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW ’11, 57–58. New York, NY, USA: ACM.

Kahanda, I., and Neville, J. 2009. Using transactional information to predict link strength in online social networks.

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, WOSN ’08, 19–24. New York, NY, USA: ACM.

Kumar, R.; Novak, J.; and Tomkins, A. 2006. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’06, 611–617. New York, NY, USA: ACM.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web WWW 10*, 591. ACM Press.

Liaw, A., and Wiener, M. 2002. Classification and Regression by randomForest. *R News* 2(3):18–22.

Malewicz, G.; Austern, M. H.; Bik, A. J. C.; Dehnert, J. C.; Horn, I.; Leiser, N.; and Czajkowski, G. 2010. Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 international conference on Management of data*, SIGMOD ’10, 135–145. ACM.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27(1):415–444.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement IMC 07* 29.

R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Spearman, C. 2011. Spearman ’ s rank correlation coefficient. *Amer J Psychol* 15(September):1–5.

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. TwitterRank : Finding Topic-sensitive Influential Twitterers. *New York* Paper 504:261–270.

Xiang, R.; Neville, J.; and Rogati, M. 2010. Modeling relationship strength in online social networks. *Proceedings of the 19th international conference on World wide web WWW 10* 981.