User Groups in Social Networks: An Experimental Study on YouTube

Michael Steve Stanley Laine, Gunes Ercal, and Bo Luo
Department of EECS / ITTC, University of Kansas
msteve@ku.edu; gunes@ittc.ku.edu; bluo@ittc.ku.edu

Abstract

While some social networking sites such as Orkut or Friendster are purely social, others such as YouTube, Flickr, and LiveJournal are highly content-oriented yet maintaining a social component. The nature of the interaction between content and connections is fundamentally important not just from a social science perspective but also to answer how the relevant content and connections can be found more easily. YouTube recently added the ability for users to form groups, in which explicit category affiliation is noted too. YouTube is ripe for consideration of how content and contacts are related. This is amongst the seminal works dealing with YouTube groups in general, not only in the context of categories, and we are amongst the first to study what motivates group membership in YouTube in the context of other observable group activities. We also investigate the role of users in groups, how groups evolve and the structure of these groups organized under a category change over time. Finally we find what form of linkage motivates new members to join these groups.

1. Introduction

Social networking sites (SNS) in which explicit connections are made between parties interested in exchanging information and content increasingly gain popularity. While some SNS (e.g. Orkut) are "purely social", others such as YouTube, Flickr, and LiveJournal are highly content oriented while maintaining a social component too. The nature of the interaction between content and connections is fundamentally important not just from a social science perspective but also to answer how the relevant content and connections can be found more easily. YouTube in particular allows various content types: videos, images, music, and text. In addition to the variety of content, it further allows two types of social relationships: subscriptions and friendships. Thus, it is undoubtedly a site ripe for consideration in the question of how content and contacts are related, and indeed has been studied in this context. Moreover, YouTube enables users to form explicit groups, in which explicit category affiliation is noted. This adds a very different dimension to the social activities and phenomena that have hitherto been scarcely considered.

In other SNS, user behavior was noted to vary extremely depending upon the stated category of user interest. For example, the semantic differentiation inherently provided by categories proved to be a fundamentally structural differentiator in question and answer networks such as Yahoo! Answers. A higher degree of reciprocity and short cycles were noted in categories such as Wrestling than categories such as Programming, plausibly relating to the common sense intuition that the users frequenting the Programming category would be more motivated by the expertise involved in the exchange whereas the Wrestling category users may be more socially motivated.

Naturally, this begets a similar question in all online social networks: to what extent does category of interest influence the actual network structure? We particularly delve into this question with regards to grouping behavior in YouTube, statistically analyzing various group properties conditional upon category. This is also amongst the seminal works dealing with YouTube groups in general, not only in the context of categories, and we are amongst the first to study what motivates group membership in YouTube in the context of other observable group activities.

In this paper, our primary goal is to explore the characteristics of social activities and communities in a content-oriented social network, and to discover the role of content in the context. We investigate two types of ties: *subscription*, which is a content-oriented relationship; and *friendship*, which is socialization-oriented. We also study the characteristics of social groups in different categories, which are determined by the content. Moreover, we compare explicit groups with random groups and artificial groups. We study the role of owners and key members in the community and how their roles influence the content. We also study the dynamics of groups in YouTube as to what fraction of groups flourish, remain stable, and show decline in the number of members, how linkage with existing group members influence users to join a particular group. We also discuss how the structural properties of the group network and category network change with

time. Through carefully designed experiments, we have discovered some interesting phenomena: (1) in a content-oriented SNS, content-oriented social activities and relationships are more intense compared with socialization-oriented activities and relationships, which indicates the primary motivation and goal of the majority of users is the content instead of socialization. (2) Users from explicit groups demonstrate stronger connections and activities. (3) Social connections, activities, and grouping behaviors are significantly shaped by their social context. (4) At least 50% of the users in every category are singletons (do not share any social/content relationship with other members). (5) Other than the owners, users at the center of the group's network play a significant role in contributing videos. (6) 7-10% of the users (owners and users at the center of the group network) contribute at least 60% of the videos to the group. (7) Category networks become denser over time: showing increasing average degree and shrinking diameter. (8) New members of the group are within 3 hops from the existing members and mostly from the subscription fringe.

2. Related work

A wide spectrum of research efforts have been devoted to SNS. Social network analysis (SNA) uses mathematical and/or computational methods to study network structures and topology. Topics in this category include: network identification and mining [27, 20, 12, 14, 25], community evolution and growth [15, 21, 13, 24], topological measurement [26, 1, 32, 28], etc. Studies of social network users have also been introduced, e.g. user behavior[29], user activities[16]. Meanwhile, discoveries from social science community have been tested on large-scale real world data. For instance, the well-known six-degrees of separation has been tested over MSN instant messaging network [22] and DBLP co-authorship network [11]. Existing works on communities in social network mostly focus on "implicit groups", i.e., a small set of users (nodes) demonstrate close relationships, but never explicitly declare themselves as a "group". Graph mining techniques [12, 24] have been employed to "discover" such a closely related group of nodes through network topology or social activities (e.g. blog comments, trackbacks). On the other hand, there exist "explicit" communities as well: many social networks allow users to create and join groups, and socialize in this small community (e.g. send group messages, make content only available to group members). [26] measures and analyzes important group features: distribution of group sizes (power law), clustering coefficients inside groups (higher than average), etc. [5] studies the growth and evolution of explicit groups in LiveJournal and DBLP. Particularly, it uses a decision tree to predict the propensity of users joining groups and group expansion. [33] work on 20 manually-selected groups from YouTube, and study features related to individuals in the groups: number of videos (per member), number of subscribers, etc. While they have made interesting findings on group subgraphs, their analysis at inter-group and category levels are not statistically significant due to small sample size. Our work on "explicitly-defined" groups is significantly different from others. First, we conduct our analysis at three levels: category, group, and group member (individual), and hence discover interesting phenomenon and social significance at different levels. Second, we have done comprehensive measurements and conducted novel experiments, e.g., compare with synthetic groups. We work on YouTube network, where content (video) sharing is the primary goal, and socialization is built on top of contents. In this way, social features (e.g. friendship) are studied in association with content features (e.g. videos, scores). Finally, there has been research on other types of social networks and/or social groups, for instance, organizational network [2], physical (not on-line) social networks [18, 17], physical social groups [8]. These topics are not within the scope of this paper, however, their methodologies and theories could be adapted to for our future research.

3. Experiments

YouTube allows users to create their own groups with three different levels of privacy and category association which are specified during the creation of the group. Each group has a set of members who have explicitly chosen to join the group, collection of videos submitted by the group members, topics which the members wish to discuss, discussions or notes is the individual post or comment each member wishes to say to a particular topic. YouTube initially had twelve categories, and recently added three: Education, Science & Technology, and Nonprofits & Activism. We focus more on newer categories as they are unlikely to have "dead" groups. The distribution of the groups in various categories is shown in Figure 1.

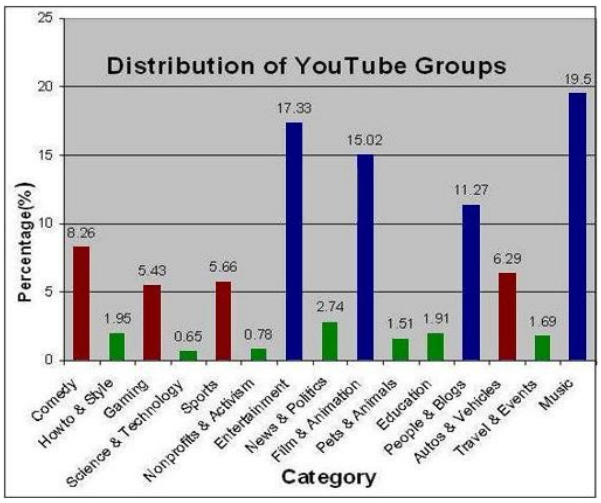

Figure 1: Distribution of YouTube Groups

The data was crawled from YouTube using automated scripts written in Python and Java in five phases.

Phase one: We first crawled random YouTube users from the social network of YouTube using snowball sampling, starting with one high degree user as seed and adding the user's friends to the seed database and crawled the other users from this database. These users were crawled for experiments to create artificial and random groups. The crawler was entirely written in Python. YouTube API was used to obtain user information such as number of friends, subscribers, subscriptions, location, etc. YouTube limits the number of friends, subscribers that can be collected using the API to 20 we resorted to screen-scraping methods by accessing the channel page of each user to collect a particular user's friends, subscribers and subscriptions.

Phase two: we then collected all the group information such as group name, number of members, topics, videos, date of creation. Data is collected from all the fifteen categories using screen-scraping methods.

Phase three: We crawled all groups with three or more members from the category Education and Science & Technology. Each group was crawled to obtain all the members and videos from group's homepage. The data was collected using crawlers written in Python by screen-scraping methods, as there is no provision in the YouTube API to obtain such data. All the group members were also crawled to collect information about user's friends, subscribers and subscriptions.

Phase four: To compare how categories differ from each other, we crawled five other categories: (1) News & Politics: the highest average number of members, videos and discussions; (2) Music: the largest of all categories (Since it had too many groups so we randomly crawled 3000 groups from the category); (3) Pets & Animals: high correlations among the group variables(the number of members, notes, topics, videos); (4) Nonprofits & Activism: low correlation among group variables; and (5) Sports: high average number of members, videos and discussions and also high correlations among the group variables. This crawling was conducted five months after the previous phase. YouTube had redesigned the group's homepage using HTML and AJAX. Therefore, a new crawler had to be written in JAVA using the crawljax package. In this phase we crawled the whole pages, and parse them with Python to extracted group members. The member data were crawled as in the previous phases.

Phase five: To track the evolution of groups over time, the groups and the group members in the category Education and Science & Technology were crawled again after two months from the previous crawl.

In the following tables and figures, C denotes average clustering coefficient, δ denotes average degree, SPL denotes average shortest path length in the largest connected component (denoted by LCC). LCC-F/LCC-S denotes the LCC of friendship/subscription network, N the number of nodes, E the number of edges, D the diameter and R the radius.

4. Small Worlds and even Smaller Worlds

The high clustering coefficient and short average path lengths observed in the group-friendship in all categories (Tables 4 and 5) confirm the small world phenomenon [30, 3, 17] often observed in social networks. The friends of a user within a group are also likely to be friends with each other, and a significant fraction of group members are connected by very short paths. As clustering coefficient is indicative of a sense of locality, we further speculate that C values when restricted to the subnetwork corresponding to a well-defined community be notably larger than C values throughout a general network. As group membership indicates locality of user interests (in the choice to join) and exchange (in the group platforms for videos and discussions), we expect that C restricted to friends-network of a YouTube group be notably larger than C of the general YouTube friends network as well as that of a subnetwork extracted from sets of random users of the general YouTube friends network, namely an artificial network comparable size of random YouTube users. We confirm the latter comparison in Table 2, where the random users' network has negligible C. Regarding the former, though there are somewhat differing results from the literature, the clustering coefficients for YouTube groups we have studied are two to three times larger than that found in the

literature on any general large online social networks including YouTube [7, 19, 6, 26], and are an order of magnitude greater than some results on transitivity in YouTube [19]. These confirm the stronger community-nature of groups, also noted in [26], and is thus expected. We further observe from Table 4 that the average shortest path lengths are extremely small in groups in YouTube network results in the literature, indeed establishing groups as networks of much smaller worlds. The group-subscription network in all the categories is also characterized by short average path lengths and high clustering coefficient confirming the small world phenomenon in a network based purely on content. We also observe that the subscription network on average connects more users than the friendship network which again emphasis the greater influence of content over friendship among the group members. The somewhat surprising aspect of our group experiment involves comparison to the artificial groups extracted from the friend's network of a particular user (i.e. a star-shaped subgraph centered around the user). The centers of the artificial groups are randomly chosen from YouTube users with the same number of friends as the number of members in the YouTube group that we compare to. For the same reason that general SNS and YouTube in particular, exhibit small-world characteristics, one might further

Table 1: Pearson coefficients between group variables

Category	Mbrs-Vidoes	Mbrs-Topics	Mbrs-Notes	Videos-Topics	Videos-Notes
Pets & Animals	0.885	0,850	0.886	0.684	0.989
Autos & Vehicles	0.811	0.491	0.774	0.312	0.962
Travel & Events	0.694	0.730	0.780	0.385	0.949
Sports	0.842	0.698	0.870	0.639	0.863
People & Blogs	0.810	0.284	0.534	0.117	0.508
Comedy	0.752	0.790	0.604	0.506	0.684
Film & Animation	0.787	0.059	0.201	0.031	0.635
Science & Technology	0.571	0.769	0.263	0.243	0.451
Entertainment	0.702	0.329	0.529	0.218	0.635
Nonprofits & Activism	0.654	0.044	0.174	0.079	0.259
Howto & Style	0.410	0.254	0.553	0.176	0.559
Music	0.546	0.149	0.353	0.143	0.468
Gaming	0.537	0.261	0.259	0.117	0.268
News & Politics	0.501	0.043	0.128	0.031	0.186
Education	0.195	0.211	0.229	0.205	0.287
p-value	0.000	0.000	0.000	0.000	0.000

Table 2: C & δ for random & artificial groups

Group type	C	δ
Artificial YouTube group	0.0623	0.7624
Random YouTube users	0.0	0.0078

Table 3: C & δ for different categories

Link types		Friends	Subscribers	Subscriptions
Education	C	0.3720	0.2475	0.3109
	δ	1.7737	1.1348	1.1348
Music	C	0.3147	0.1896	0.2563
	δ	1.0097	0.5219	0.5219
Pets & Animals	C	0.3488	0.2520	0.2886
	δ	1.1653	0.7558	0.7558
Sports	C	0.3185	0.2164	0.2567
	δ	0.9377	0.5247	0.5247
Science & Technology	C	0.3839	0.2549	0.3230
	δ	1.4471	1.0029	1.0029
News & Politics	C	0.3133	0.2534	0.3429
	δ	1.1263	0.6703	0.6703
Nonprofits & Activism	C	0.3746	0.2300	0.2870
	δ	1.3929	0.8503	0.8503

Table 4: LCC of Group network

	Science & Technology	Education	Nonprofits & Activism	News & Politics	Pets & Animals	Music	Sports
N	17.22	25.27	22.36	30.50	16.79	25.05	20.23
E	60.20	122.43	107.624	118.12	49.69	72.09	66.01
SPL	1.619	1.62	1.628	1.780	1.690	1.718	1.641
C	0.3287	0.3535	0.3228	0.2696	0.3069	0.2685	0.2748
δ	1.78	2.37	1.71	1.85	1.62	1.51	1.48
D	3.28	3.29	3.28	3.72	3.45	3.56	3.31
R	1.87	1.92	1.89	2.13	1.97	2.02	1.89

Table 5: LCC of Category network

	Science & Technology		Education		Nonprofits & Activism		Pets & Animals		Music		News & Politics		Sports	
	LCC-F	LCC-S	LCC-F	LCC-S	LCC-F	LCC-S	LCC-F	LCC-S	LCC-F	LCC-S	LCC-F	LCC-S	LCC-F	LCC-S
N	2857	1206	8425	3669	6665	6803	6932	7816	19395	18121	21574	21883	25531	26752
E	10696	9858	39262	36769	41959	96420	20549	46234	65411	103250	104682	162992	102695	190332
SPL	7.0	7.21	6.19	6.92	6.42	5.95	6.62	6.24	6.52	6.50	5.53	5.55	6.40	6.79
C	0.3349	0.2752	0.2913	0.2262	0.3192	0.3005	0.2569	0.2658	0.2296	0.1895	0.2686	0.2167	0.2593	0.2304
δ	3.74	8.17	4.66	10.02	6.29	14.17	2.96	5.91	3.37	5.70	4.85	7.45	4.02	7.11

D	22	23	26	24	24	24	25	20	25	24	18	23	21	21
R	11	10	13	13	12	12	13	10	13	12	9	12	11	11

expect higher clustering from a subset of users in that social network when all of those users are friends of a particular user (a form of locality). While this may be true based on some literature results [7], nonetheless, the relative values in Table 2 comparing Artificial YouTube group with the actual groups in the Education and Science & Technology categories indicates that the group association is a much stronger tie of local socialization and community formation.

Table 5 on category networks yield C values of categories to be highly comparable to C for the Artificial YouTube group as constructed above, yielding similar comparisons to YouTube group networks which are also highly clustered. And, again, belonging to the same group strictly dominates both in its small-world properties, despite that the category network too is a small world network as further exhibited by very short average path lengths and high C values. Surely, we cannot proceed without a closer examination of our results in the perspective of the seminal work of [26] which gave initial analysis of a sample of YouTube groups not restricted to a particular category. [26] noted that the clustering coefficient of 0.34 in YouTube groups was indeed around thrice as large as that of the general YouTube network. In fact, as much as we have confirmed the smaller world phenomenon in groups even when restricted to groups of certain categories, the C for the categories we have exhaustively considered ranges from 0.26 to 0.33 compared to 0.34 of the random sample of groups considered in [26]. In light of the great variance in potential motivations for user group membership and activity differentiated by group Category as exhibited by Table 1, this discrepancy in C values indicate varying degrees of socialization as differentiated by category affiliation, while maintaining the smaller-worlds property conditional upon group affiliation. The C values of the category subscription network ranges from 0.19 to 0.30 exhibiting lower C values than their corresponding category friendship network. Our C values for the subscription network is much higher for all the categories and differs remarkably from the results on transitivity in [7].

5. Category networks

5.1. Structure of category network

The structure of the category network consists of 3 regions: the giant component, middle region and the singleton nodes as stated in the work by [21].

Giant component is the single largest component of the category. It contains 70-95% of the links and 9-35% of the nodes in the category networks. The LCC of the categories in both the subscription and friendship network exhibits high clustering coefficient, shorter average path lengths and diameters relative to the size of the network, thus clearly displaying small world properties. The clustering coefficient of the friendship network is higher than the subscription network, due to the forced symmetry of friendship links and the fact that users subscribe to users who generate videos. This indicates that there are few content producers in YouTube groups which are discussed in section 7. The size and the structural properties of the giant component differ with each category which is an indicator of varying degrees of socialization [31].

The middle region is marked by the presence of smaller, tightly clustered and very dense components which indicate a stronger social cohesion among the members. The middle region consists of 5-30% of the links and 10-18% of the nodes. Though the middle region is characterized by the presence of very highly clustered and dense components, the average degree of the middle region is smaller than the giant component.

The singleton region refers to users who are not part of the group's social network (no friends or subscriptions within the group). At least more than 50% of the nodes are singletons in all categories indicating majority of YouTube users are more interested in videos rather than social relationships. This confirms our assumption that YouTube is a content-oriented network where content is most important and the social activities revolve around it. Similar results were observed in Yahoo360 and Flickr [21].

5.2. Effects of semantic differentiation

Though different categories share lots of similarity in structure, the minute difference in structure, the way a particular category attracts more users or become popular over other categories conveys that semantic differentiation plays a key role in shaping the network and has an effect on the user behavior. Most users of YouTube use it to watch videos. Hence categories like Music and Entertainment has the largest number of groups. News & Politics is the most popular category among users because of the newer and sensitive content it generates. Meanwhile, two categories with similar content, Education and Science & Technology, exhibit similar structure standing out from rest of the categories. The giant component in the

subscription network is much smaller comparing to the giant component in the friendship network. Nearly 80% of the nodes are in the singleton region which indicates members in a group do not subscribe to other members as much as members in other categories do. Also, in all other categories, the clustering coefficient of the middle region of the subscription network is lower than that of the giant component, unlike the clustering coefficient of the middle region of Education and Science & Technology. This indicates these two categories have isolated cohesive group of users closely tied with each other because of content and we do not observe similar behavior in case of the friendship network where the clustering coefficient of the middle region is always lower than that of the giant component.

Another interesting behavior is that 95% of the links are in the giant component in both friendship and subscription n/w of the News & Politics category. This is attributed to the presence of very active high degree nodes who are key contributors of the group's resources (ref. Section 7). This category is the most active and popular as it displays the highest average number of members, videos, and discussions. It is the hot favorite of all the categories because of the content and the greater reach of the high degree users which also plays a role in roping new members to the group. Thus semantic differentiation plays key role in shaping the structure and user behavior in online social networks.

6. Different forms of socialization: Content Versus Personal

Friendship links are inherently more personal compared to the often asymmetric, content-oriented subscriber relation. The closer relationship between friends relative to subscribers is also shown in the differing C and SPL values of Table 3. Both forms of linkage may result in information exchange between the two users, and moreover a tendency for one form may create a tendency in another, as observed in Figure 3.

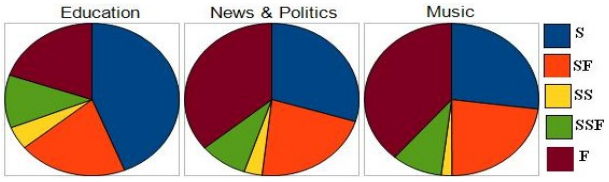

Figure 3: Distribution of different types of connections; S: subscription, SS: mutual subscription, F: friendship

Nonetheless, the intent of the two types of links is different and is indicator of different forms of social exchange. From Figure 3 which demonstrates distribution of different types of social connections and results from the other four categories we can clearly conclude that: (1) there still exists purely socialization-oriented connections, with no preference of content sharing in all the categories; (2) purely social relationships are dominated by purely content-oriented connections in smaller categories such as Education, Science & Technology, Nonprofits & Activism; (3) content oriented connections are highly asymmetric, which indicates that content producers/distributers are different from consumers; (4) Smaller categories are more active in content sharing compared to larger categories; and (5) Purely content-oriented connections are dominated by purely social relationships in larger categories like Music, Sports, News & Politics; With an increase in the size of the categories there is a shift in domination from purely content-oriented to purely social-oriented links. Also, from the temporal data in Education and Science & Technology, we observe a similar shift and the change in distribution of links, we learn that being in the same group favors tie formation. Users initially establishing one way content-oriented relationship tend to establish social relationship or a two way content-oriented relationship with other group members over time.

7. Contributors to the groups

Some group members contribute more videos to the groups than others. One may expect the owners and moderators to be the key contributors. However, there is one set of users who contribute significantly more than the moderators. These are the users who are at the center of the groups (nodes whose eccentricity is equal to radius). This interesting phenomenon is observed over all the categories. Owners of the groups are the maximum contributors (most video submissions to the group) in at least 55% of all the groups in all categories, followed by users at the center of the group's friendship network and group's subscription network who are also the maximum contributors in at least 19% of the groups, where as moderators are the maximum contributors in less than 1% of the groups. It's interesting to observe that users at the center contribute significantly larger than the moderators. These users at the center have many friends, high indegree (subscribers) and are very popular. These users position in the network is epiphenomenon to their contribution to the group. This is in accordance with "methodological individualism" [4] because the only reason they are in the center of the network is because of the content they generate and is not the case otherwise. Groups serve as a platform for these popular users to showcase their videos and get more popular. In the News & Politics category the users at the center of

the friendship and subscription networks are the maximum contributors in 25% of the groups which is higher when compared to the other categories where it varies between 19-21%. This could be a significant factor as to why News & Politics is the most popular and active among all the categories as these groups have the highest average number of members, videos, and discussions. The fraction of videos contributed by group owners and users at the center of the groups vary depending on the category where as the fraction of videos contributed by the moderators to the category is always less than 0.1. Except for the categories News & Politics and Pets & Animals the fraction of videos contributed by the corresponding group owners in all other categories is around 0.4 and by users who are at the center of the group varies between 0.10 to 0.23. In the categories News & Politics and Pets & Animals the contribution from the owners is 0.23 which is remarkably low compared to other categories. This deficit by the owners is compensated by the users at the center who contribute about 0.41 and 0.42 in the News & Politics and Pets & Animals categories respectively. The group owners and the users at the center of the friendship and subscription networks who are just 7-10% of the category's population contribute nearly 60% of the videos in all the categories thus confirming that there are few content producers in YouTube groups.

8. Richness of sets of Categories of Groups

It has been observed that the indegree and outdegree in many SNS exhibit power law [21, 26, 6]. We investigated the power law behavior in the number of members, videos, topics & notes over individual categories and also over all the categories. Using Kolmogorov-Smirnov goodness of fit metrics and the maximum likelihood method [10], we confirm that the distribution of members and topics over all categories follows power law. Additionally, the plot of rank vs. frequency on a doubly logarithmic axis gives a straight line, which is a necessary condition for a distribution that exhibits power law behavior. The power-law coefficients for both members and topics are equal to 1.85 and 2.11 respectively.

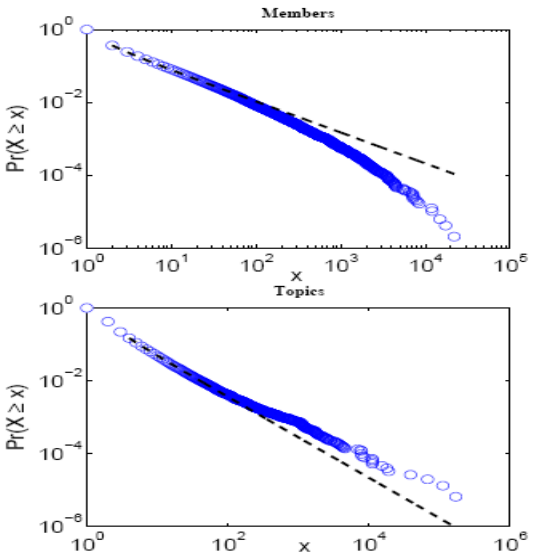

Figure 2: Log-log plot of Members and Topics of complementary cumulative distribution functions.

These values are similar to that observed for the community size for the Amazon co-purchasing network [9]. Distributions of videos and notes exhibit very high cut-off under the K-S test to qualify these distributions as power-law. Members within different categories were found to follow power law with similar coefficient and lower cutoff except for the categories of Entertainment, Film & Animation, Nonprofits & Activism and Education for which the power law coefficient deviated significantly from the values computed over all the categories and the lower cutoff was much higher which again results in dropping of many sample points. Topics also exhibit power law behavior over all the categories with the power law coefficients and the lower cutoff of each category deviating slightly from that obtained over all the categories. Even though notes do not obey power law over all the categories, it does follow power law in the categories like Pets & Animals and Sports. This may be a likely reason for the high correlation between all of the group variables in these categories. The existence of power law in members may due to the fact that groups with more members have a better chance to be larger, similar to the findings in [5]. Groups with larger member base tend to get larger because they have more outreach ties, so the existence of the group is well known to a larger member base through friendship and subscriber linkages when compared to smaller groups. Also the topics in a group depend on the number of members in the group and hence they tend to follow a similar distribution.

9. Group Dynamics

In this section we first discuss as to what fraction of groups flourish (show increase in the number of members), remain stable (no increase in the number of members), and show a decline in the number of members. We then answer how users join groups in YouTube as it does not provide any page where users can explicitly browse or search to find communities that match their interests unlike other SNS. Finally we discuss how the structural properties of the group network and category network change with time.

9.1. Group Activity

We observed that nearly 25% and 15% of the groups flourished, 65% and 80% of the groups remained stable at the end of the fifth and seventh month from the initial crawl respectively and the remaining fraction of groups showed decrease in number of members.

9.1.1. Flourishing Groups. These groups showed significant activity in terms of video submissions and notes. 87% and 91% of the groups were active (meaning there was at least more than one video submission or notes) and only 45% and 52% of the groups in the Science & Technology and Education categories respectively had video submissions where as 86% and 91% of these groups had discussions going on. Notes was the only form of activity in nearly 50% and 40% of the active groups in the Science & Technology and Education category respectively and it was not the same case otherwise with the videos which clearly indicates discussions is an important group activity than video submissions in groups based on videos.

9.1.2. Static Groups. These groups showed a marked reduction in the activity. Only 61% and 71% were active and nearly 86% and 83% of the groups in Science & Technology and Education categories did not have a single video submission to the group. Also in these groups, notes seemed to be the major cause of the activity. Notes were the sole form of activity in nearly 78% and 76% of these active groups in Science & Technology and Education categories respectively.

9.1.3. Dwindling Groups. Nearly 80% and 84% of the groups were active in Science & Tech and Education categories respectively which is higher than the static groups. It has been observed that nearly 66% and 59% of the groups in Science & Technology and Education categories did not have a single video submission to the group, which is much lower than the static groups. Even among these groups, notes were the sole form of activity in nearly 57% and 51% of the active groups in Science & Technology and Education categories.

Greater number of groups flourish and are active in the Education when compared to Science & Technology which is primarily due to the larger member base; and discussions is the main source of activity in groups as most users in groups are content consumers as opposed to few content producers.

9.2. Growth of groups and its effect on the network structure

The growth of the groups in YouTube primarily depends on the size of the fringe (users who have friends, subscribers or subscriptions to members in the group but they themselves are not in the group). These users in the fringe are the potential group members [5] and particularly in case of YouTube groups. We observe that links within the groups grow much faster than the nodes because links from existing group members already exist outside the group's network and are brought into the group's network once these new members join these groups. It indicates that new members are the members in the fringe of the groups who have existing subscriptions to users in the group and/or friends who are already part of the group. This contributes to the humungous growth of links over the number of nodes. We also observe that the growth exponent (Growth of edges/Growth of nodes) is increasing during successive crawls in both the categories for both types of links which indicates a very rapid densification of the category network.

Table 6: Growth of groups

	Education		Science & Tech	
	5 months	7 months	5 months	7 months
Nodes(initial)	21747		10737	
Nodes(final)	29438	31622	13262	13934
Growth - nodes	35.37%	45.41%	23.52%	29.78%
Links(initial)	21071		7371	
Links(final)	45587	56229	14138	16291
Growth – friendship	116.35%	166.86%	91.81%	121.05%
Links(initial)	38885		11164	
Links(final)	70775	90722	20175	25360
Growth - subscription	82.01%	133.31%	80.71%	127.16%

The nodes in the Education category grow much faster when compared to Science & Technology and this can be attributed to the larger fringe size and larger member

base. This is in accordance with the rich get richer phenomena.

With time, the nodes and edges shift to the giant component from both the middle and singleton region. This indicates users in the groups establish social contact over time or subscribe to other users in the group and form a part of this giant component. This is supported by the increasing clustering coefficient and average degree and decreasing average path length and diameter even though the size of the nodes double in this region. The percentage of nodes in the giant component increases from 18.7 to 32.8 and edges from 72.4 to 90.1 in the friendship network. Similar shift is also observed in the subscription network of Education category and also in the friendship and subscription network in the Science & Technology category. The average path lengths is seen to drop from 8.62 to 5.75 in the friendship network and from 7.68 to 5.97 in the subscription network of Education category over a period of seven months despite the huge growth of the network. Similar results were observed with Science & Technology too. These interesting results were observed in citation graph for U.S. patents, the graph of the Internet etc.[23]. This increasing average degree and decreasing diameter indicate densification of the network which is contrary to the conventional belief that the diameters increase slowly as O(logn) or O(log(logn)) and average degree remains constant [23].

The structural properties of the LCC of the group network changes differently when compared to the LCC of the category network. Unlike the LCC of the category n/w which shows a decrease in the average shortest path length and diameter of the LCC of the group n/w shows a slight increase in these parameters where as the average degree always displays a steady increase like that of the LCC of the category n/w.

9.3. New members in the Groups

With YouTube not providing explicit pages for users to explore groups and join them. It is of good interest to find how new members join groups. Links to groups can be found on the homepage of existing group members. There is a high possibility for users to join groups by browsing the homepage of their friends, subscribers or subscriptions, who are already part of the group. In this section we examine how new users are distributed and find which form of linkage contributes significantly towards this regard and also find the distance of new members to the existing members.

Table 7 shows the percentage of new members reachable through various numbers of hops from the existing group members. The members reachable by first hop were not considered for the subsequent hops and so on. We observe that the majority of the new members are reached over the first three hops.

Table 7: Percentage of new members reachable from the group members over various hops

Hop	Education		Science & Technology	
	5 months	7 months	5 months	7 months
1	22.58%	39.58%	25.87%	34.01%
2	14.69%	17.11%	19.02%	19.09%
3	12.81%	6.38%	14.93%	8.06%
4	0.29%	1.47%	4.87%	3.52%
5	0.73%	0.39%	1.20%	0.83%
6	0.13%	0.06%	0.54%	0.37%

We clearly see that subscription dominates over all other links – majority of the new members subscribes to at least one other group member. It is an indicator that users are driven by content than any other form of links. Subscription link contributing to the maximum number of new members also conveys one other important information about the users in the groups. These groups usually have very few popular users to whom many users subscribe to but not vice versa. The graph below shows the number of hops over which the new members can be reached from the current group members.

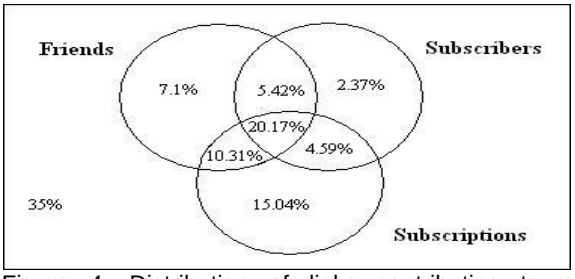

Figure 4: Distribution of links contributing to new members of the group (between the 5th and 7th month) in the Education category

After the second crawl, 66% and 51% of the new members can be reached via either one of friendship, subscriber or subscription links in the Science & Technology and Education category respectively. After the third crawl, 66% and 65% of the new members can be reached via either one of friendship, subscriber or subscription links in the two categories respectively.

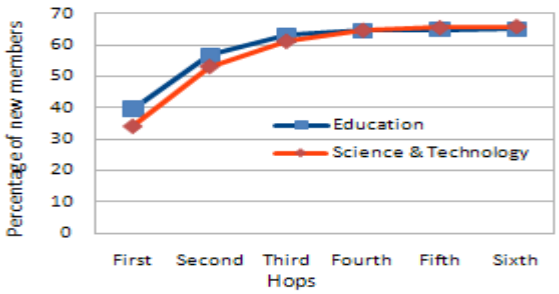

Figure 5: New members in the groups at the end of 7th month

So nearly 65% of the new group members who join the group are just three hops away from the existing group members and majority of them are from the subscription fringe (shows that popular content producers are the ones who rope in most members). The remaining fraction of new members may be one or two hops away from the users (users who are in the fringe of the group) who are not part of the group.

10. Conclusion and Future Directions

In this paper we have explored various aspects of the relationships between content and structure in the context of grouping behaviors in a content-oriented network: YouTube. Our discoveries that the category of group content has a strong impact on motivation for group membership and activities, while topics and membership are strongly correlated across all categories with both following power-law, have implications on predictors of group growth. Our results concerning the small-world characteristics groups, even in comparison to artificial groups extracted from YouTube as well as the corresponding category networks (which exhibited similar characteristics to the group network), yields that the group and category association is indeed very strong. Social connections and activities, as well as grouping behaviors, are significantly shaped by their social context: the content. Majority of the users in all categories are singleton nodes who do not share any social or content based relationship with other group members. Owners and users at the center of the group's network (only 7-10% of the total population) contribute at least 60% of the videos to the group. Most of the new group members are within 3 hops from the existing members and are from the subscription fringe which illustrates that popular content producers rope in most members. Aside from the obvious immediate direction of extending the analysis to other categories, this study may be extended in several interesting directions: (1) examining the groups' influences in content (video) propagation and discussions over the general YouTube. (2) Understanding the privacy issues caused by the knowledge of group membership (and even category alone) as it may reveal user's interest, geographical location, and other sensitive information. (3) Improving recommendation systems based on comparisons of more elaborately discovered artificial groups to actual groups and build recommendation systems for potential group members based on the knowledge of the user's membership to similar groups, membership of user's friends, subscribers/subscriptions. (4) Associating our findings with sociology research on user behaviors and physical (off-line) social groups (e.g. [8]) to further understand the underlying forces and mechanisms behind the phenomena. Also, further compare online grouping behaviors with off-line grouping behaviors. (5) Better understanding the interplay between the dynamics of local networks (e.g. groups and categories) and the global network; test whether group growth is consistent with existing models (e.g. [8]) and if it is not consistent with existing models, develop better models to capture community growth.

Acknowledgements

Bo Luo is partially supported by a University of Kansas General Research Fund (GRF:).

11. References

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon and H. Jeong, *Analysis of topological characteristics of huge online social networking services*, *ACM WWW,* 2007.

[2] M. K. Ahuja and K. M. Carley, "Network Structure in Virtual Organizations", Organization Science, 10 (1999), pp. 741-757.

[3] R. Albert and A. L. Barabási, "Statistical mechanics of complex networks", Reviews of Modern Physics, 74 (2002), pp. 47-97.

[4] K. J. Arrow, "Methodological Individualism and Social Knowledge", The American Economic Review, 84 (1994).

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg and X. Lan, *Group formation in large social networks: membership, growth, and evolution*, *ACM KDD,* 2006.

[6] F. Benevenuto, F. Duarte, T. Rodrigues, V. A. F. Almeida, J. M. Almeida and K. W. Ross, *Understanding video interactions in youtube*, *16th ACM international conference on Multimedia*, Vancouver, British Columbia, Canada, 2008, pp. 761-764.

[7] J.-I. Biel, "Please, subscribe to me! Analysing the structure and dynamics of the YouTube network", (2009).

[8] K. Carley, "A Theory of Group Stability", American Sociological Review, 56 (1991), pp. 331-354.
[9] A. Clauset, M. E. J. Newman and C. Moore, "Finding community structure in very large networks", (2004).
[10] A. Clauset, C. R. Shalizi and M. E. J. Newman, "Power-law distributions in empirical data", (2009).
[11] E. Elmacioglu and D. Lee, "On six degrees of separation in DBLP-DB and more", SIGMOD Rec., 34 (2005), pp. 33--40.
[12] G. W. Flake, S. Lawrence and C. L. Giles, *Efficient identification of Web communities*, 2000.
[13] D. Gibson, J. Kleinberg and P. Raghavan, *Inferring Web communities from link topology*, Pittsburgh, Pennsylvania, United States, 1998.
[14] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", Proceedings of the National Academy of Sciences (PNAS), 99 (2002), pp. 7821-7826.
[15] K. I. Goh, Y. H. Eom, H. Jeong, B. Kahng and D. Kim, "Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions", Physical Review E (Statistical, Nonlinear, and Soft Matter Physics, 73 (2006), pp. 8.
[16] V. Gomez, A. Kaltenbrunner and V. Lopez, *Statistical analysis of the social network and discussion threads in slashdot*, Beijing, China, 2008.
[17] C. W. Harrison, "Search Parameters for the Small World Problem", Social Forces, 49 (1970), pp. 259-264.
[18] C. W. Harrison, S. A. Boorman and R. L. Breiger, "Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions", The American Journal of Sociology, 81 (1976), pp. 730-780.
[19] A. Java, X. Song, T. Finin and B. Tseng, Why we twitter: understanding microblogging usage and communities, *WebKDD/SNA-KDD*, San Jose, CA, 2007.
[20] Z. Kou and C. Zhang, "Reply networks on a bulletin board system", Phys. Rev. E, 67 (2003), pp. 6.
[21] R. Kumar, J. Novak and A. Tomkins, *Structure and evolution of online social networks*, Philadelphia, PA, USA, 2006.
[22] J. Leskovec and E. Horvitz, *Planetary-scale views on a large instant-messaging network*, Beijing, China, 2008.
[23] J. Leskovec, J. Kleinberg and C. Faloutsos, *Graphs over time: densification laws, shrinking diameters and possible explanations*, ACM, Chicago, Illinois, USA, 2005.
[24] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram and B. L. Tseng, *Facetnet: a framework for analyzing communities and their evolutions in dynamic networks*, Beijing, China, 2008.
[25] N. Matsumura, D. E. Goldberg and X. Llor, "Mining directed social network from message board", WWW '05: Special interest tracks and posters, 2005.
[26] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel and B. Bhattacharjee, "Measurement and analysis of online social networks," *ACM SIGCOMM conference on Internet measurement*, San Diego, CaA, USA, 2007, pp. 29-42.
[27] M. K. Reiter and A. D. Rubin, "Crowds: anonymity for Web transactions", *ACM TISSEC*., 1 (1998), pp. 66-92.
[28] G. Robins, P. Pattison, Y. Kalish and D. Lusher, "An introduction to exponential random graph (p*) models for social networks", *Social Networks*, 29 (2007), pp. 173 - 191.
[29] P. Singla and M. Richardson, *Yes, there is a correlation: - from social networks to personal behavior on the web*, Beijing, China, 2008.
[30] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks", Nature, 393 (1998), pp. 440-442.
[31] B. Wellman, *For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community*, *Proceedings of the 1996 ACM SIGCPR/SIGMIS conference on Computer personnel research*, ACM, Denver, CO, 1996, pp. 1-11.
[32] M. Wilson and C. Nicholas, *Topological analysis of an online social network for older adults*, *SSM '08: Proceeding of the 2008 ACM workshop on Search in social media*, ACM, Napa Valley, CA, USA, 2008, pp. 51-58.
[33] P. Yu, M. Hu and N. Kim, *Social network analysis: YouTube*, 2007.