Effects of Group Categories on the Structure of Online Social Networks

Michael Steve Stanley Laine

Submitted to the graduate degree program in the Department of Electrical Engineering & Computer Science and the Graduate Faculty of the University of Kansas School of Engineering in partial fulfillment of the requirements for the degree of Master of Science

Thesis Committee:

_____
Dr. Gunes Ercal Chairperson

_____
Dr. Bo Luo

_____
Dr. Prasad Kulkarni

_____
Date Defended: 08/24/2010

The Thesis Committee for Michael Steve Stanley Laine certifies

That this is the approved version of the following thesis:

Effects of Group Categories on the Structure of Online Social Networks

Committee:

_____
Dr. Gunes Ercal Chairperson

_____
Dr. Bo Luo

_____
Dr. Prasad Kulkarni

_____
Date Approved: 08/24/2010

Abstract

Over the past few years there has been increasing research interests spent on online social networks. While some social networking sites such as Orkut, Facebook and Friendster 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 more recently added the ability for users to form explicit groups, in which explicit category affiliation is noted too. YouTube is ripe for consideration of how content and contacts are related. We study YouTube groups in general not only in the context of categories but also 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 in YouTube.

Contents

Acceptance Page i

Abstract ii

List of Figures v

List of Tables vi

1. Introduction 1

2. Background 4

    2.1. Online Social Networks 4

    2.2. Why study online social networks? 5

    2.3. Analyzing online social networks 6

3. Related Work 9

    3.1. Related work 9

4. Crawling and Dataset 11

    4.1. Crawling 11

        4.1.1. First phase 11

        4.1.2. Second phase 11

        4.1.3. Third phase 12

        4.1.4. Fourth phase 13

        4.1.5. Fifth phase 14

    4.2. Dataset 15

    4.3. Richness of sets of Categories of Groups 15

    4.4. Notations & Tables 17

5. Characteristics of Group and Category networks 19

5.1. Structure of Category Network 19

5.1.1. Giant Component 19

5.1.2. Middle Region 20

5.1.3. Singleton Region 20

5.2. Structure of Group Network 22

5.3. Small and even Smaller Worlds 23

5.4. User and Group Behavior 25

5.4.1. Different forms of socialization: Content Versus Personal 26

5.4.2. Effects of semantic differentiation 27

5.4.3. Contributors to the groups 28

6. Dynamics of Group and Category Networks 31

6.1. Group Activity 31

6.1.1. Flourishing groups 32

6.1.2. Static groups 32

6.1.3. Dwindling groups 32

6.2. Growth of groups 33

6.3. New members in the groups 37

7. Conclusion & Future Directions 40

7.1. Conclusion 40

7.2. Future Directions 41

8. References 42

List of Figures

4.1 Distribution of YouTube Groups

4.2 Log-log plot of Members and Topics of complementary cumulative distribution functions (CCDF)

5.1 Structure of friendship network of the group “youtubehelp”

5.2 Distribution of different types of social connections in different categories

6.1 Distribution of links contributing to new members of the group (at the end of 7^{th} month) in the Education and Science & Technology category

6.2 New members in the groups at the end of 5th month

6.3 New members in the groups at the end of 7th month

List of Tables

4.1 Pearson coefficients between group variables

4.2 C & δ for random & artificial groups

4.3 C & δ for various categories

4.4 LCC of Group network

4.5 LCC of Category network

5.1 Distribution of nodes and edges in the friendship network

5.2 Distribution of nodes and edges in the subscription network

6.1 Growth of groups

6.2: Distribution of nodes and edges in the friendship network during successive crawls

6.3: Distribution of nodes and edges in the subscription network during successive crawls

6.4 Structural properties of the Giant Component of friendship and subscription network of Education Category

6.5 Structural properties of the Giant Component of friendship and subscription network of Science & Technology Category

6.6 Percentage of new members reachable through various numbers of hops from the group members.

Chapter 1

1. Introduction

Social networking sites in which explicit connections are made between parties interested in exchanging information and media content increasingly gain popularity. While some social networking sites such as Orkut, Facebook and Friendster 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 for a variation in content type: videos, images, music, and text. In addition to the variety of content types, it further allows two types of social relationships: subscriptions and friendships. Thus, YouTube 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 provides the ability for users to form explicit groups, in which explicit category affiliation is noted too. This adds a very different dimension to the social activities and phenomena that have hitherto been scarcely considered. Each group has a set of members who have explicitly chosen to join the group, collection of videos submitted by members in the group, topics which the group members wish to discuss, discussions or notes is the individual post or comment each member wishes to say to a particular topic in the group. While initially YouTube had only twelve categories, recently three new categories were added. The newer categories are Education, Science & Technology and Nonprofits & Activism. To study the dynamics of groups

as to how these groups change with time, we focus more on newer categories in this work as they are unlikely to have "dead" groups.

In other social networking sites, 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 as well 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 other online social networks that are not necessarily question and answer oriented: 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. We study YouTube groups in general not only in the context of categories but also study what motivates group membership in YouTube in the context of other observable group activities.

In this thesis, 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 such activities and communities. We have investigated two types of social ties: the subscription, which is a content-oriented relationship; and the friendship, which is socialization-oriented. We also study the characteristics of social communities in different categories, which are determined

by the content. Moreover, we compare explicit groups with random groups and artificial groups. Further we study the role of owners and key members in the community and how their roles influence the content and 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 users join groups and what forms of linkage with existing group members influence them to join a particular group in YouTube as it does not provide any page where users can explicitly browse and find groups unlike other online social networks where users can search or browse to find communities that match their interests. 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 network, 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 social connections and activities. (3) Social connections and activities, as well as grouping behaviors, are significantly shaped by their social context: the content. (4) At least 50% of the users in every category are singleton nodes they do not share any form social or content based relationship with other group members. (5) Other than the owners of the group, users at the center of the group's network play a significant role in contributing videos. (6) Category networks become denser over time displaying increasing average degree and shrinking diameter. (7) New members of the group are within 3 hops from the existing group members and most of them are from the subscription fringe.

Chapter 2

BACKGROUND

This chapter begins with an overview of online social networks followed by drafting the reasons for studying online social networks. Finally, concepts and terminologies that are relevant to the analysis carried out in this thesis are explained.

2.1 Online Social Networks

Online social network refers to websites which allow individuals to display information about oneself (profiles), connect with other individuals and to exchange/share messages or media among themselves. This is accomplished through websites known as Social Networking Sites (SNS). These sites generate significant traffic over the Internet and have evolved as powerful means to share media, exchange messages, form communities and establish social, content-oriented or professional contacts. According to Alexa, a company that ranks sites based on web traffic Facebook and YouTube are next to Google in the amount of traffic they generate. Twitter a micro-blogging site also finds its position in the top eleven. This soaring popularity and rapid growth has turned the head of researchers and big corporate giants in this direction. With this, there arise many interesting questions about these SNS as to what causes the rapid growth of the number of users, content and how users locate relevant content

Some of the most popular SNS are Facebook, Orkut, MySpace, Flickr, YouTube, LiveJournal, Twitter etc. Some of the SNS are “purely social” meaning these sites are build completely based on social connections between individuals like Facebook, Orkut and MySpace

while others are based on content as well as social connections such as Flickr a photo sharing site, YouTube a video sharing site, LiveJournal a blogging site etc. With this an immediate questions that arise are how the content and social connections are related, how social connection and content-oriented connections are related and do they influence each other.

Communities or groups in online social networks represent a closed group of individuals who share similar interests. Communities may be explicitly or implicitly defined. Explicitly defined communities refer to communities which have been explicitly created by individuals where as implicitly defined communities refer to users who share similar interests or geographic location or any other criteria. These explicitly defined communities have owners, moderators who play an important role in shaping these communities. These communities serve as a platform for users to connect with other users with similar interests and share content and establish contacts. SNS are evolving and constantly changing so as to generate more interest among individuals to be on the site. It will be quite interesting to observe the direction in which this evolution of SNS is heading to.

2.2 Why study Online Social Networks?

There are many reasons as to why we study online social networks. SNS have become one of the most indispensable means of establishing/maintaining social and professional contacts, exchanging messages and locating interesting content. Hence, it is important that we study them as they are going to be an important means of communication in the future. The popularity and rapid growth of users provides the research community with real world graphs so as to study the dynamics of networks at a very large scale. It provides a good understanding of

network structure, enables prediction of future growth and simulation of network systems of large size. These large real world graphs can also be used to develop search algorithms and plays a significant role in network analysis and planning [10, 28]. Understanding the formation and evolution of communities and the process that underline how new members join communities can be used to predict what type of communities grow and attract new users as opposed to what communities will dwindle over time. Identifying key influential members in the community who are connected to many other members can be of use in viral marketing [10]. SNS have become a business and there have been significant investment and work on linking these SNS to shopping sites as knowledge of people's social, professional and content-based connections, and information of people in the same communities or group in these SNS can be used to recommend products as they are more likely to share similar interests. So this connection of SNS with marketing and shopping further impacts the need to study the dynamics and growth patterns of these SNS. Also since these SNS are constantly evolving, it is important to understand how they change with time so as to predict the impact of these SNS on the people and the internet in the future.

2.3 Analyzing Online Social Networks

In this section, the various ways in which online social networks can be analyzed, relevant terminologies and phenomenon generally observed in social networks are presented. All analysis of social networks or complex networks involves visualizing the whole network as a graph with users as nodes and the relationship between them as edges or links. The links can be either directed or undirected. Friendship links are undirected meaning link in one direction implies link exists in the reverse direction whereas links like followers and following as in

twitter or subscribers and subscriptions as in YouTube are directed meaning link existing one direction does not imply link exist in the reverse direction.

Some of the common terminologies used in the analysis of online social networks are Degree, Clustering Coefficient, Shortest Path length, Diameter, Radius, Diameter, Eccentricity, Center, Connected Components, Small world, Power law.

Degree of a node refers to the number of outgoing or incoming links incident in that node in case of undirected graphs. In case of directed graphs, the number of outgoing links from a node is referred to as out-degree and the number of incoming links from a node is referred to as in-degree of that node.

In most of the work in this thesis, Average degree of a group or community is used which is nothing but the number of links existing in the groups normalized by the number of nodes.

Clustering coefficient refers to how closely the neighborhood of a particular node is connected. It is defined as the ratio of the number of links existing between the node's neighbors to the total number of possible links that can exist between the node's neighbors or the ratio of the number of existing triangles to the total number of possible triangles in a network.

Average clustering coefficient of group is the ratio of sum of all the clustering coefficient of the nodes to the total number of nodes in the group.

Average shortest path length refers to the sum of path lengths between all pairs of nodes normalized by n*(n-1) where n is the number of nodes in the graph with the assumption that the length is zero if one node is not reachable from the other node.

Radius refers to the minimum of all pairs shortest path.

Diameter refers to the maximum of all pairs shortest path.

Eccentricity refers to the maximum of shortest paths to all other nodes.

Center refers to the set of nodes with eccentricity equal to radius.

Small world networks are networks where users though not directly connected with each other but can be reached via other users by smaller number of hops. These networks are said to display small world phenomenon if they exhibit smaller average shortest path length and larger clustering coefficient when compared to random graphs of same size.

In undirected graphs, a connected component refers to the disjoint subgraphs of a larger graph. A connected component is a subset of a larger graph such that there exists a path between all pairs of nodes in the subgraph. In directed graphs, strongly connected component refers to a subgraph where a path exists between all pairs of nodes in the subgraph and weakly connected component refer to a subgraph where a path exists between all pairs of nodes in the subgraph when the graph is viewed as undirected graph.

Chapter 3

3. Related work

A wide spectrum of research efforts have been devoted to online social networks. First, 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 [1-5], community evolution and growth [6-9], topological measurement [10-13], etc. Studies of social network users have also been introduced, e.g. user behavior [14], user activities [15]. Meanwhile, discoveries from social science community have been tested on large-scale real world data. For instance, the well-known six-degrees of separation have been tested over MSN instant messaging network [16] and DBLP co-authorship network [17].

Existing work on communities in social network mostly focus on implicit groups. In such scenario, a small set of users (nodes) demonstrate close relationships, however, they never explicitly declare themselves as a group. Graph mining techniques [3, 9] 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). [10] measures and analyzes important group features: distribution of group sizes (power law), clustering coefficients inside groups (higher than average), etc. [18] 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. [19] 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 more comprehensive measurements and conducted novel experiments, e.g., compare with synthetic groups. Finally, 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).

Chapter 4

In this chapter, the details & issues involved in crawling and details regarding the dataset are presented.

4.1. Crawling

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

4.1.1. First phase:

The first phase involved crawling random YouTube users who were in the social network of YouTube. The crawler was written in Python and data was stored in MySQL database and in the form of text files organized into various folders named based on that particular YouTube user. YouTube API was only used to obtain user information like date joined, number of friends, subscribers, subscriptions, location and other such information about the users. User's friends, subscribers, subscriptions were crawled using screen scraping methods by accessing the channel page of each user because YouTube limits the number of friends, subscribers that can be collected using the API to 20. In order to obtain greater coverage of the graph we resorted to screen-scraping methods. Even by this method YouTube limits the number of friends, subscribers that can be obtained. This limit was 1020.

The users were crawled using snowball sampling, starting with one high degree user as a seed and adding the user's friends to the seed database and crawling the other users from this seed database. These users were crawled for experiments to create artificial and random groups.

4.1.2. Second phase:

This phase involved collecting all the groups from all the fifteen categories. First we extracted the group name from all the fifteen categories. The group information like number of members in the group, number of topics, numbers of videos, date of creation and other such information regarding the groups were collected for all the groups in all the fifteen categories.

The data was collected using crawlers written in Python by screen-scraping methods.

4.1.3. Third phase:

This phase involved our first collection of data on groups. We chose all groups with three or more members from the category Education and Science Technology. Groups were crawled one at a time obtaining all the members and videos by visiting the 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 group data.

After the completion of the group crawl, all the users in the group were crawled using the method described in the first phase.

4.1.4. Fourth phase:

This phase involved collecting all the groups with three or more members from the seven chosen categories. In order to compare how different categories compared with each other apart from crawling the groups in the category Education and Science & Technology we crawled five other categories namely News & Politics which had the highest average number of members, video submissions and discussions, Music the largest of all categories (Since it had too many groups so we randomly crawled 3000 groups from the category), Pets & Animals which had high correlations among the group variables(the number of members, number of notes, number of videos), Nonprofits & Activism which had low correlation among group variables and Sports which had high average number of members, videos and discussions and also high correlations among the group variables.

This collection was done after approximately five months from the previous phase. YouTube had redesigned the group's homepage using HTML and AJAX. So there was an issue obtaining data using the previous crawlers written in Python. So a new crawler had to be written in JAVA using the crawljax package which had made our work easier. We had to write plugins using the crawljax package in JAVA to collect our data this time. With the new design of the group's homepage we could obtain only the most recent 1200 members from the groups. There were only a few groups which had members greater 1200. So we omitted these few groups in the five categories. Interestingly the groups in the Education and Science & Technology category

which had less than this limit in the previous phase crossed this limit during this phase. Since we were able to obtain the newest 1200 members of the category, we could collect all the members even from those groups which had more than 1200 members in these two categories. In this phase we collected the whole page and saved it as HTML document. The members and other necessary relevant data were obtained from these pages by parsing these pages using scripts written in Python

After obtaining all the members in the groups, the users were crawled using the method described in the first phase.

4.1.5. Fifth phase:

To track the evolution of groups over time, the groups and the users belonging to these groups in the category Education and Science & Technology were crawled once again after two months from the previous crawl using methods described in the above phases.

4.2. Dataset

Our crawl encompasses seven categories of YouTube consisting of 193,602 members, 43 million subscription links, 53 million friendship links and 18,027 groups. We also have temporal group data from the categories Education and Science & Technology to study the dynamics of

groups over time. The second and third crawl of the groups in these two categories was after the fifth and seventh month from the initial crawl.

The distribution of the groups among various categories can be seen in figure.

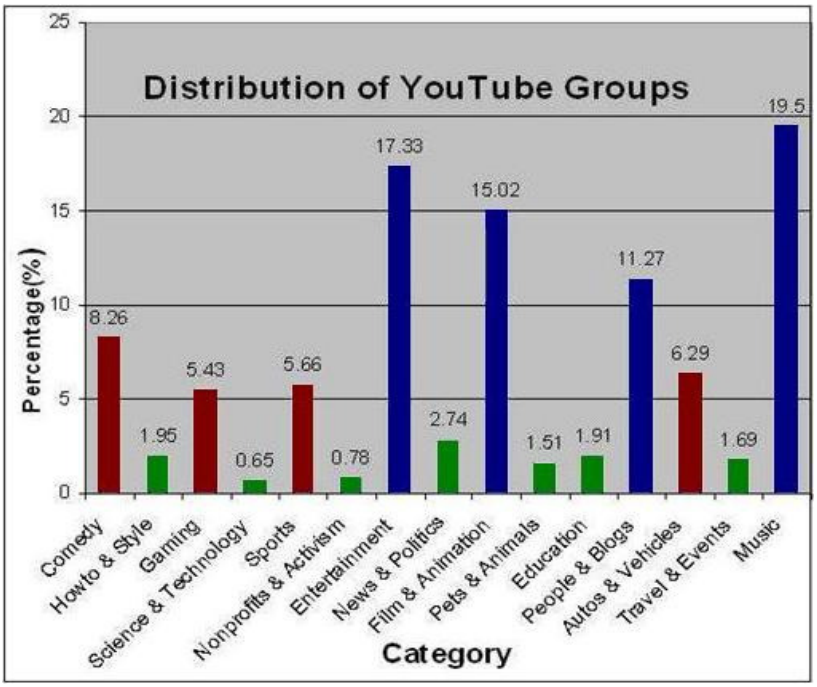

Figure 4.1: Distribution of YouTube Groups

4.3. Richness of sets of Categories of Groups

It has been observed that the indegree and outdegree in many online social networks exhibit power law behavior [7, 10, 24]. We investigated the power law behavior in the number of members, videos, topics & notes over individual categories and also over all the categories. From Kolmogorov-Smirnov goodness of fit metrics and using the maximum likelihood method based on [20], we confirm that the distribution of members and topics over all the 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. These values are similar to that observed for the community size for the Amazon co-purchasing network [21]. Distributions of videos and notes exhibit very high cut-off under the Kolmogorov-Smirnov test to qualify these distributions as power-law. We further investigated for power law within categories. Members within different categories were found to follow power law with similar power law 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 be due to the fact that groups with more members have a better chance to grow larger, similar to the findings in [18]. 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.

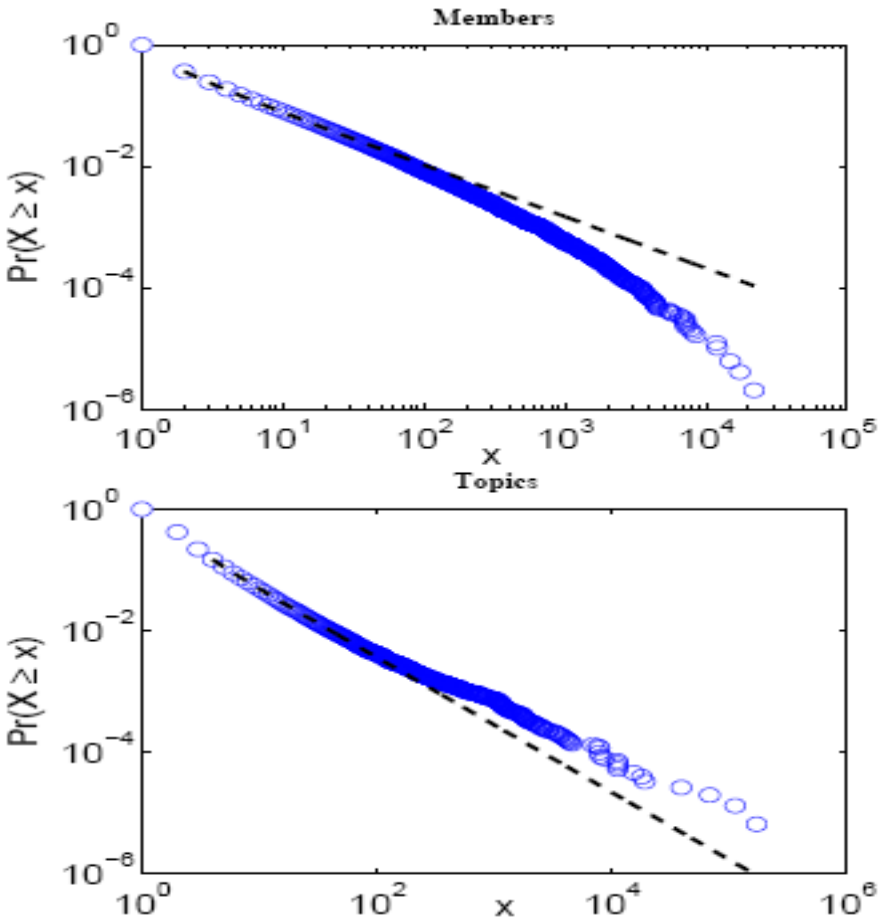

Figure 4.2: Log-log plot of Members (top) and Topics (bottom) of complementary cumulative distribution functions (CCDF)

4.4. Notations & Tables:

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

Table 4.1: 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 4.2: 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 4.3: C & δ for various 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 4.4: LCC of Group network

	Science & Technology	Education	Nonprofits & Activism	Pets & Animals	Music	News & Politics	Sports
	LCCF	LCCF	LCCF	LCCF	LCCF	LCCF	LCCF
N	17.22	25.27	18.60	20.23	16.79	25.05	20.23
E	60.2	122.4	108.2	66.0	49.7	72.1	66.0
SPL	1.619	1.62	1.737	1.641	1.690	1.718	1.641
C	0.329	0.354	0.288	0.275	0.310	0.269	0.275
δ	1.78	2.37	3.22	1.48	1.62	1.51	1.48
D	3.28	3.29	3.64	3.31	3.45	3.56	3.31
R	1.87	1.92	2.07	1.89	1.97	2.02	1.89
	Science & Technology	Education	Nonprofits &Activism	Pets & Animals	Music	News & Politics	Sports
	LCCS	LCCS	LCCS	LCCS	LCCS	LCCS	LCCS
N	17.88	16.79	30.50	20.23	18.60	28.02	22.18
E	149.1	49.7	118.1	66.0	108.2	137.3	127.9
SPL	1.700	1.690	1.780	1.641	1.737	1.846	1.738
C	0.356	0.310	0.270	0.275	0.288	0.217	0.254
δ	4.47	1.62	1.85	1.48	3.22	2.78	2.84
D	3.93	3.45	3.72	3.31	3.64	3.89	3.57
R	1.96	1.97	2.13	1.89	2.07	2.19	2.03

Table 4.5: LCC of Category network

	Science & Technology	Education	Nonprofits & Activism	Pets & Animals	Music	News & Politics	Sports
	LCCF	LCCF	LCCF	LCCF	LCCF	LCCF	LCCF
N	2857	8425	6665	6932	19395	21574	25531
E	10696	39262	41959	20549	65411	104682	102695
SPL	7.0	6.19	6.42	6.62	6.52	5.53	6.40
C	0.3349	0.2913	0.3192	0.2569	0.2296	0.2686	0.2593
δ	3.74	4.66	6.29	2.96	3.37	4.85	4.02
D	22	26	24	25	25	18	21
R	11	13	12	13	13	9	11
	Science & Technology	Education	Nonprofits & Activism	Pets & Animals	Music	News & Politics	Sports
	LCCS	LCCS	LCCS	LCCS	LCCS	LCCS	LCCS
N	1206	3669	6803	7816	18121	21883	26752
E	9858	36769	96420	46234	103250	162992	190332
SPL	7.21	6.92	5.95	6.24	6.50	5.55	6.79
C	0.2752	0.2262	0.3005	0.2658	0.1895	0.2167	0.2304
δ	8.17	10.02	14.17	5.91	5.70	7.45	7.11
D	23	24	24	20	24	23	21
R	10	13	12	10	12	12	11

Chapter 5

In this chapter, the structural properties of the group and category networks in various categories is presented and important findings of the user and group behavior conditional upon category is discussed.

5.1. Structure of category network

The structure of the category network can be divided into three regions: the giant component, middle region and the singleton nodes as stated in the work by [7]. The tables below show the distribution of nodes and edges in the friendship and subscription network.

Table 5.1: Distribution of nodes and edges in the friendship network

	Science & Technology		Education		Nonprofits & Activism		Pets & Animals		Music		News & Politics		Sports	
	N	E	N	E	N	E	N	E	N	E	N	E	N	E
Giant	21.5	75.7	28.6	86.1	33.5	90.0	30.8	86.6	29.9	89.0	35.5	95.0	26.0	91.4
Middle	17.8	24.3	14.4	13.9	15.6	10.0	13.5	13.4	9.9	11.0	7.8	5.0	9.9	8.6
Singleton	60.7		56.9		50.9		55.7		60.2		56.7		64.1	

Table 5.2: Distribution of nodes and edges in the subscription network

	Science & Technology		Education		Nonprofits & Activism		Pets & Animals		Music		News & Politics		Sports	
	N	E	N	E	N	E	N	E	N	E	N	E	N	E
Giant	9.08	71.1	12.5	84.6	34.1	92.1	34.6	89.4	18.0	88.8	35.5	95.5	27.3	90.2
Middle	10.3	28.9	8.2	15.4	17.2	7.9	12.4	10.6	10.2	11.2	8.1	4.5	10.5	9.8
Singleton	80.6		79.3		48.7		53.0		61.8		56.4		62.2	

5.1.1. Giant component

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 network which varies depending on the category. 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 which may be due to the forced symmetry of friendship links. The size and the structural properties of the giant component differ with each category which is an indicator of varying degrees of socialization among the members of the category.

5.1.2. Middle region

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 in the category network depending on the category. 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.

5.1.3. Singleton region

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 [7].

5.2. Structure of Group network

Both the friendship and subscription group network have high clustering coefficient and very small average shortest path lengths and this confirms the small world phenomenon in these groups. The value of clustering coefficient is higher than the value stated by Misolve et. al for YouTube groups[10].

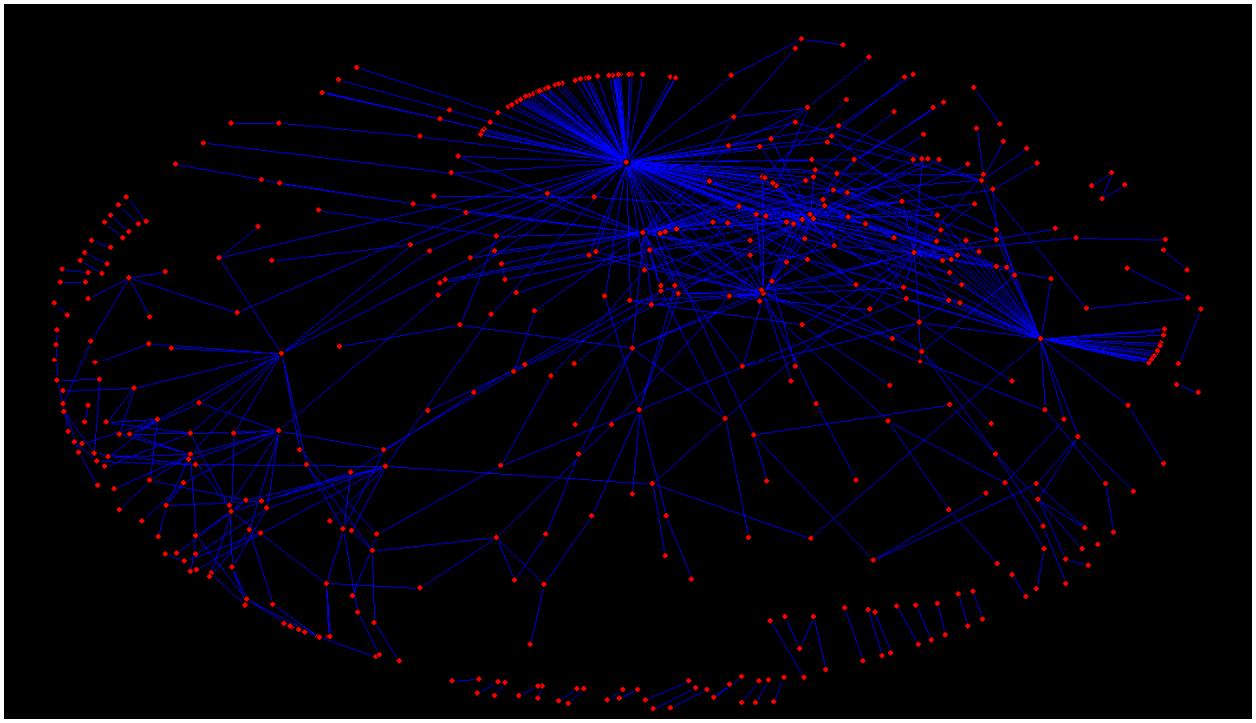

Figure 5.1: Structure of friendship network of the group “youtubehelp”

Again the two categories Education and Science & Technology stand out from rest of the categories. The clustering coefficient exhibited by these two categories in the subscriptions network is higher than the friendship network which is the case otherwise in the rest of the other categories. This is probably due to the reason there are fewer groups in these categories that have subscriptions to other members in the group which can be seen from the presence of many singleton nodes. So there are few groups in which members subscribe to other group members. However in groups where subscription links exist, users exhibit greater clustering coefficient.

5.3. Small Worlds and even Smaller Worlds

The high clustering coefficient and short average path lengths observed in the group-friendship and group-subscription network in all categories (Tables 4.4 and 4.5) confirm the small world phenomenon [20-21] 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 sub-network 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 4.2, 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 [10, 22-24], and are an order of magnitude greater than some results on transitivity in YouTube [23]. These confirm the stronger community-nature of groups, also noted in [10], and is thus expected. We further observe from Table 4.4 that the average shortest path lengths are extremely small in groups in the categories we exhaustively considered, compared to the general 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 and 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 online social networks and YouTube in particular, exhibit small-world characteristics, one might further 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 [22], nonetheless, the relative values in Table 4.2 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 4.5 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 [10] which gave initial analysis of a sample of YouTube groups not restricted to a particular category. [10] 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 [10]. In light of the great variance in potential motivations for user group membership and activity differentiated by group Category as exhibited by Table 4.1, 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 [22].

5.4. User and Group Behavior

The semantic differentiation plays a key role in shaping the structure of the category network and also affects the various forms of connection or linkage the members of the category share with each other and the way the members of the group contribute videos to their respective groups in the category.

5.4.1 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 5.4.3). 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.

5.4.2. Different forms of socialization: Content Versus Personal

All analysis above based on friendship links is an inherently more personal form of linkage in their forced symmetry and mere denotation than links between users based on the often asymmetric, content-oriented subscriber relation. The closer relationship between friends relative to subscribers is also expressed in the differing C and SPL values of Table 4.3. 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 5.2.

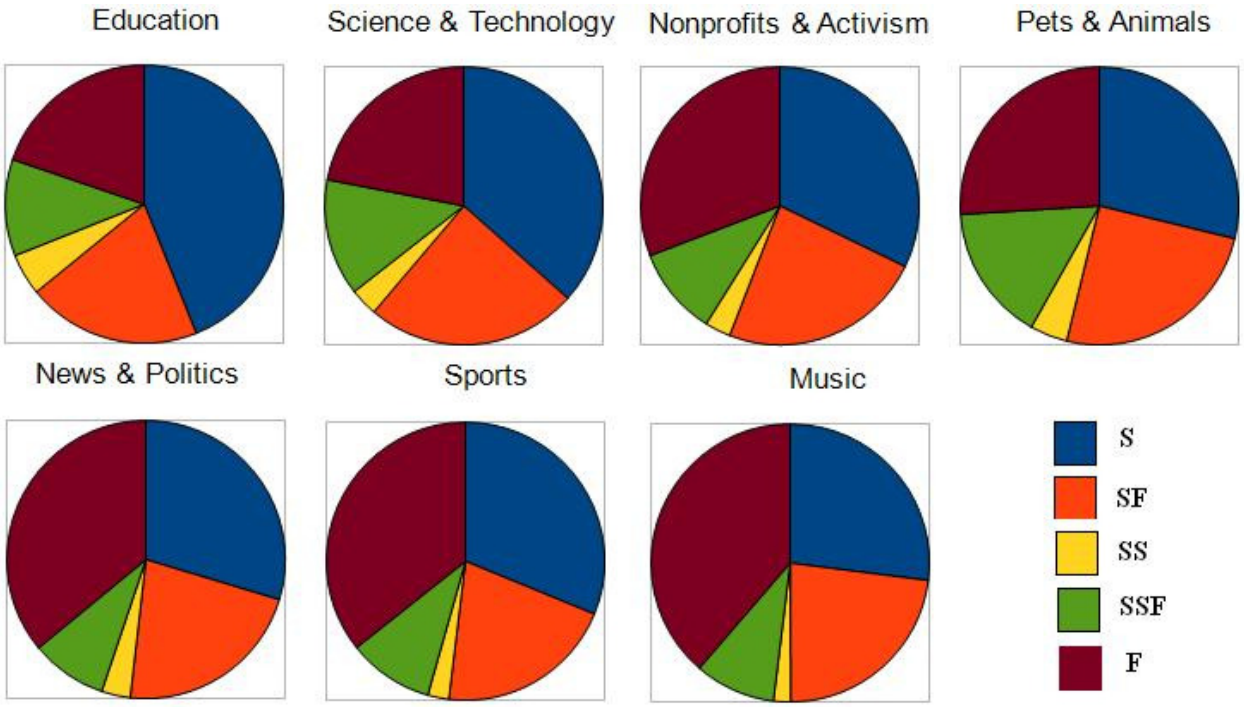

Figure 5.2: Distribution of different types of social connections in different categories: S: subscription, SS: mutual subscription, F: friendship

Nonetheless, the intent of the two types of links is different and is indicator of different forms of social exchange. From Figure 5.2 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.

5.4.3 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" [28] 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.

Chapter 6

Group Dynamics

In this section we study the dynamics of groups in YouTube. First we discuss as to what fraction of groups flourish (groups that show increase in the number of members), remain stable (groups with no increase in the number of members), and show a decline in the number of members. Then, we discuss how the structural properties of the group network and category network change with time. Finally, 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 online social networks.

6.1. Group Activity

In this subsection we discuss what fraction of groups flourish, remain stable or dwindle and discuss the significant source of activity in these groups. 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.

6.1.1. Flourishing Groups

Groups that flourished 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.

6.1.2. Static Groups

The groups that remained stable 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.

6.1.3. Dwindling Groups

Groups which displayed a decrease in the number of members showed significant activity even more than groups that were stable. Nearly 80%and 84% of the groups were active in

Science & Technology and Education categories respectively. 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.

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

Table 6.1: Growth of groups

	Education		Science & Technology	
	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%
Friendship links(initial)	21071		7371	
Friendship links(final)	45587	56229	14138	16291
Growth –friendship links	116.35%	166.86%	91.81%	121.05%
Subscription links (initial)	38885		11164	
Subscription links(final)	70775	90722	20175	25360
Growth - subscription	82.01%	133.31%	80.71%	127.16%

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 new members in online social networks [18] 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. This indicates that potential 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. 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.

Table 6.2: Distribution of nodes and edges in the friendship network during successive crawls (N – Nodes & E - Edges)

	Education						**Science & Technology**					
	First		**Second**		**Third**		**First**		**Second**		**Third**	
	N	E	N	E	N	E	N	E	N	E	N	E
Giant	18.7	72.4	28.6	86.1	32.8	90.1	9.8	35.2	21.5	75.7	24.1	79.4
Middle	18.7	27.6	14.5	13.9	12.6	9.9	22.4	64.8	17.8	24.3	16.8	20.6
Singleton	62.6		56.9		54.6		67.8		60.7		59.1	

Table 6.3: Distribution of nodes and edges in the subscription network during successive crawls (N – Nodes & E - Edges)

	Education						**Science & Technology**					
	First		**Second**		**Third**		**First**		**Second**		**Third**	
	N	E	N	E	N	E	N	E	N	E	N	E
Giant	5.6	58.1	12.	84.6	15.6	89.4	1.2	8.0	9.1	71.1	11.7	79.1
Middle	13.2	41.9	8.2	15.4	7.7	10.6	16.0	92.0	10.3	28.9	9.7	20.9
Singleton	81.2		79.3		76.7		82.8		80.6		78.6	

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 category too.

Table 6.4 Structural properties of the Giant Component of friendship and subscription network of Education Category

Education						
	Friendship			**Subscription**		
	First	**Second**	**Third**	**First**	**Second**	**Third**
Nodes	4060	8425	10358	1226	3669	4946
Edges	15258	39262	50670	14105	36769	53768
Avg. SPL	8.62	6.19	5.75	7.68	6.92	5.97
Clust. Coeff.	0.269	0.291	0.289	0.275	0.226	0.224
Avg. Degree	3.76	4.66	4.89	11.51	10.02	10.87
Diameter	29	26	23	22	24	21
Radius	15	13	12	10	13	12

Table 6.5 Structural properties of the Giant Component of friendship and subscription network of Science & Technology Category

Science & Technology						
	Friendship			**Subscription**		
	First	**Second**	**Third**	**First**	**Second**	**Third**
Nodes	1052	2857	3351	126	1206	1623
Edges	2596	10696	12993	614	9858	14582
Avg. SPL	7.19	7.0	6.53	5.41	7.21	6.21
Clust. Coeff.	0.239	0.335	0.317	0.200	0.275	0.296
Avg. Degree	2.47	3.74	3.88	4.87	8.17	8.99
Diameter	18	22	20	15	23	21
Radius	9	11	10	6	10	10

These interesting results were observed in citation graph for U.S. patents, the graph of the Internet etc [27]. 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 [27].

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

6.3. 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 these groups. Links to groups can be found on the homepage of users who have already joined these groups. There is a high possibility for users to join groups by browsing the page of users in their friends, subscribers or subscriptions network who are already part of the group. So in this section we are going to examine how new users are distributed over the different networks and find which form of linkage contributes significantly towards this regard and also find the distance of these new members from the existing group members.

Table 6.3 shows the percentage of new members reachable through various numbers of hops from the members who were already part of the group. The members who can be reached 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 6.3: Percentage of new members reachable through various numbers of hops from the group members

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%

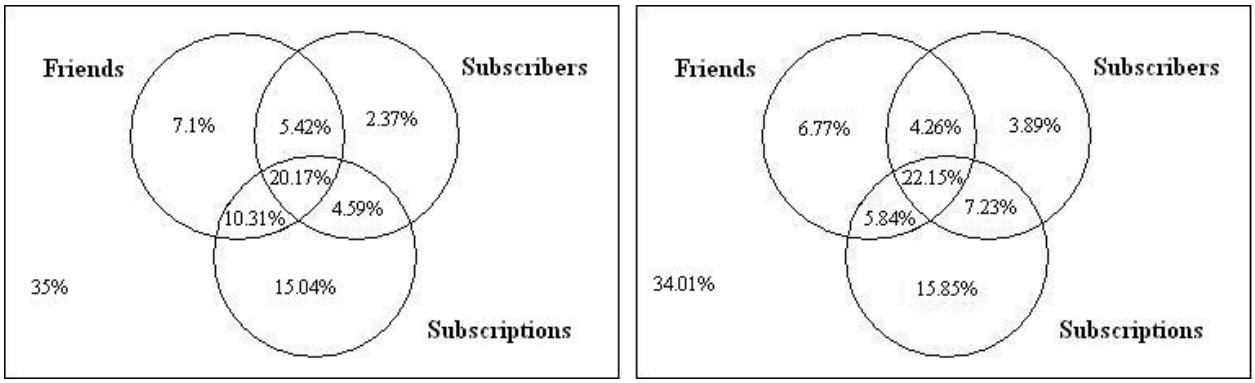

Figure 6.1: Distribution of links contributing to new members of the group (at the end of 7th month) in the Education (left) and Science & Technology (right) category

We clearly see that subscription dominates over all other linkages which indicates that majority of the new members had a subscription to at least one other group member which is an indicator that users are driven more by content than any other form of linkages. 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 the case vice versa. The graph below shows the number of hops over which the new members can be reached from the current group members. 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 Science & Technology and Education category respectively.

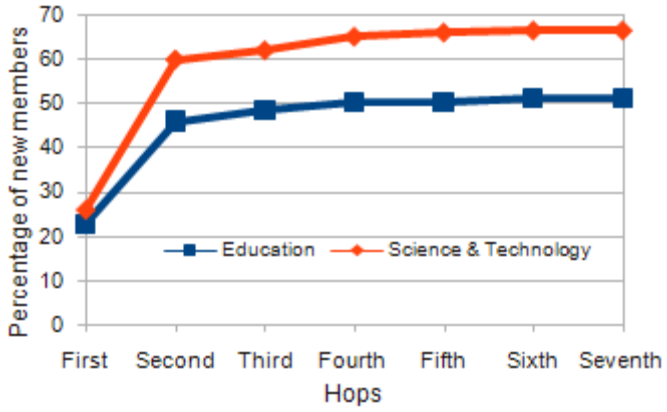

Figure 6.2: New members in the groups at the end of 5th month

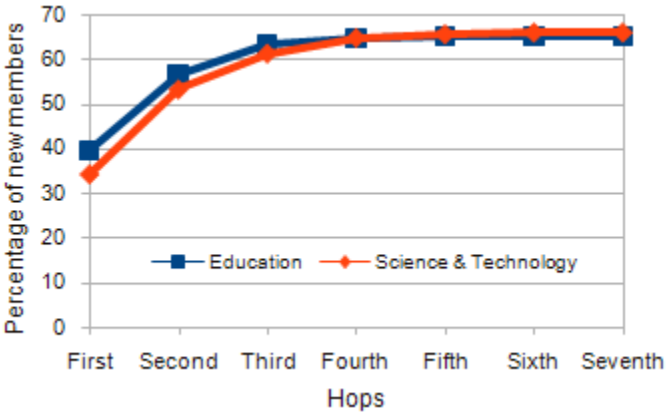

Figure 6.3: New members in the groups at the end of 7th month

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. The remaining 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.

Chapter 7

7. Conclusion and Future Directions

7.1. Conclusions

In this thesis, 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.

.

7.2. Future Directions

Aside from the obvious immediate direction of extending the analysis to other categories, there are several interesting directions in which this study may be extended: examining the groups' influences in content propagation over the general YouTube; 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; 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; better understanding the interplay between the dynamics of local networks (e.g. groups) and the global network; test whether group growth is consistent with existing models and if it is not consistent with existing models, develop better models to capture community growth.

8. References

[1]M. K. Reiter and A. D. Rubin, "Crowds: anonymity for Web transactions," ACM Trans. Inf. Syst. Secur., vol. 1, pp. 66-92, 1998.

[2]Z. Kou and C. Zhang, "Reply networks on a bulletin board system," Phys. Rev. E, vol. 67, p. 6, 2003 2003.

[3]G. W. F. a. S. L. a. C. L. Giles, Efficient identification of Web communities, 2000.

[4]M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences (PNAS), vol. 99, pp. 7821-7826, 2002.

[5]N. M. a. D. E. G. a. X. Llor, "Mining directed social network from message board," WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web, 2005.

[6]K.-I. G. a. Y.-H. E. a. H. J. a. B. K. a. D. Kim, "Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions," Physical Review E (Statistical, Nonlinear, and Soft Matter Physics, vol. 73, p. 8, 2006.

[7]R. K. a. J. N. a. A. Tomkins, Structure and evolution of online social networks. Philadelphia, PA, USA, 2006.

[8]D. G. a. J. K. a. P. Raghavan, Inferring Web communities from link topology. Pittsburgh, Pennsylvania, United States, 1998.

[9]Y.-R. L. a. Y. C. a. S. Z. a. H. S. a. B. L. Tseng, Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. Beijing, China, 2008.

[10]A. Mislove, et al., "Measurement and analysis of online social networks," in IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, ed. San Diego, California, USA: ACM, 2007, pp. 29-42.

[11]Y.-Y. Ahn, et al., "Analysis of topological characteristics of huge online social networking services," in WWW '07: Proceedings of the 16th international conference on World Wide Web, ed. Banff, Alberta, Canada: ACM, 2007, pp. 835-844.

[12]M. Wilson and C. Nicholas, "Topological analysis of an online social network for older adults," in SSM '08: Proceeding of the 2008 ACM workshop on Search in social media, ed. Napa Valley, California, USA: ACM, 2008, pp. 51-58.

[13]G. Robins, et al., "An introduction to exponential random graph (p*) models for social networks," Social Networks, vol. 29, pp. 173 - 191, 2007.

[14]P. S. a. M. Richardson, Yes, there is a correlation: - from social networks to personal behavior on the web. Beijing, China, 2008.

[15]V. c. c. G. o. m. a. A. K. a. V. L\'{o}pez, Statistical analysis of the social network and discussion threads in slashdot. Beijing, China, 2008.

[16]J. L. a. E. Horvitz, Planetary-scale views on a large instant-messaging network. Beijing, China, 2008.

[17]E. E. a. D. Lee, "On six degrees of separation in DBLP-DB and more," SIGMOD Rec., vol. 34, pp. 33--40, 2005.

[18]L. Backstrom, et al., "Group formation in large social networks: membership, growth, and evolution," in KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ed. Philadelphia, PA, USA: ACM, 2006, pp. 44-54.

[19]P. Yu, et al., "Social network analysis: YouTube," 2007.

[20]D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," Nature, vol. 393, pp. 440-442, June 1998.

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

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

[23]A. Java, et al., "Why we twitter: understanding microblogging usage and communities," in WebKDD/SNA-KDD '07, ed. San Jose, California, 2007, pp. 56-65.

[24]F. Benevenuto, et al., "Understanding video interactions in youtube," in 16th ACM international conference on Multimedia, ed. Vancouver, British Columbia, Canada, 2008, pp. 761-764.

[25]A. Clauset, et al., "Power-law distributions in empirical data," Feb 2009.

[26]A. Clauset, et al., "Finding community structure in very large networks," Aug 2004.

[27]J. a. K. Leskovec, Jon and Faloutsos, Christos, Graphs over time: densification laws, shrinking diameters and possible explanations. Chicago, Illinois, USA: ACM, 2005.

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

[29] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network," in WOSP '08: Proceedings of the first workshop on Online social networks. New York, NY, USA: ACM, 2008, pp. 25-30