Structure and Network in the YouTube Core

John C. Paolillo, Informatics and SLIS, Indiana University
paolillo@indiana.edu

Abstract
In this paper, we present results of an empirical investigation into the social structure of YouTube, addressing friend relations and their correlation with tags applied to uploaded videos. Results indicate that YouTube producers are strongly linked to others producing similar content. Furthermore, there is a socially cohesive core of producers of mixed content, with smaller cohesive groups around Korean music video and anime music videos. Thus, social interaction on YouTube appears to be structured in ways similar to other social networking sites, but with greater semantic coherence around content. These results are explained in terms of the relationship of video producers to the tagging of uploaded content on the site.

1. Introduction

Since being launched in December 2005, YouTube has unexpectedly emerged as a major player in video distribution. Beginning as a "personal video sharing service" [17], it has become a multi-billion dollar business, generating advertising revenues for Google and fears of displacement for traditional producers of video. In 2006, when YouTube sold to Google for $1.5 million, the site boasted 100 million views and 65,000 video uploads per day; no exact figures are available today, but the site's popularity and influence has only increased. Popular YouTube video links are exchanged in email, on weblogs, and are even criticized in mainstream media. Popular Internet videos are archived on YouTube, alongside clips from television and cable broadcasts. Some mainstream media companies have responded by posting their own content on YouTube, while others such as Viacom have sued over copyright infringement. YouTube has even acquired its own stars, such as actress Jessica Rose who plays lonelygirl15, and Gary Brolsma of the Numa Numa dance.

In spite of all of this activity, little is known about ordinary people's use of YouTube, or even what characterizes its offerings. Most of the available information comes from mass media or marketing perspectives and addresses its most visible controversies and/or legal issues (deceptive content, copyright infringement, pornography, privacy). Little has been done to characterize the predominant uses of YouTube from a systematic or scholarly perspective. The site itself aggregates only the most basic usage statistics (raw counts of views, comments, etc.), and there is no clear picture of how people use YouTube and why. It is important to address these issues if we are to understand what sort of influence YouTube has on media consumption and production, or what sort of niche it occupies in the media ecology.

YouTube's primary features are the ability to upload and deliver video clips of any reasonable length. Video is accepted in most standard formats and converted to low-resolution Flash (swf) for delivery. Any user with a web browser can view YouTube videos, but users must create a user account ("channel") to upload videos. This account provides a profile page that serves as an index to the user's uploaded videos, and on which users may optionally disclose personal details, or "subscribe" to other users' videos and "friend" other users; these details are then displayed in their profile pages. Users may also comment on other users, or more commonly, on specific videos. These comments are also displayed in the relevant pages. Finally, YouTube offers community "groups" which users may join to declare particular interests. Groups provide a way to serialize video content as well as offering a text message interface similar to discussion boards or Usenet.

YouTube is thus a social networking site, with the added feature of hosting video content. Online social networks are often characterized by a core-periphery social network structure [2, 15], in which central participants disproportionately influence social interaction and the devel-

1530-1605/08 $25.00 © 2008 IEEE

opment of content. Core-periphery structures have been observed in a broad range of such networks, including email discussion lists [7], Usenet [9], IRC [10], weblogs [4, 11], and online learning environments [8].

In a few such cases, research also addresses the relation of content to the social organization of the network. In an extensive study of LiveJournal profiles [11], users' interests were found to be only weakly correlated with their friends. Over the span of a year, the correlation appeared to decrease, even while groups of shared friends and shared interests strengthened. Likewise, a two-month sample of tagged video from the social bookmarking site del.icio.us revealed clusters of individuals using overlapping sets of tags for specific online video links, but little connection or sharing of tags among even related links [12]. Hence, there is no necessary relationship between content (interests, tagged videos) and the social structure of the network. In contrast with this, on the amateur Flash exchange website Newgrounds.com, seven distinct social groups in the network of authors were shown to correlate closely to the production of distinct genres and sub-genres of Flash [13]. Since Newgrounds users rate one another's Flash and compete for popularity, authors who share artistic styles and/or communicative goals benefit from providing one another support against competing groups. This circumstance results in a stronger association of social network and content.

On YouTube, producers of video operate in at least two distinct modes that bear on the relation of social network to content. In the first mode, users upload their own original creations, like the amateur Flash authors of Newgrounds. Such users might potentially form groups for mutual support, cultivating particular subject matter or tastes over others. Central or popular users have a special role in this circumstance, in that they serve as points of reference for genre emergence by providing widely emulated examples [13]. In the second mode of production users upload video obtained from outside sources, typically broadcast or cable television. Such users essentially "forward" information to the YouTube audience at large, performing a filtering function. Their orientation to content is potentially more topical and event oriented and less semantically coherent, as found in the bookmarking of video on del.icio.us.

Whichever practice characterizes the prevailing mode of YouTube video posting, we should be able to find indications of it in the social network structure of the site and semantic coherence found among the posted content. Should multiple cores exist, they might represent competing or even antagonistic social groups, whose interactional dynamics shape both the social environment and the forms of content available. Hence, examining the social network structure of YouTube is an important first step to understanding what sort of a media and social space it represents.

2. Data

Identification of the relationship between social network and content requires collection of a large sample of user profiles and relations. In social network analysis the chief method for sampling large networks is "snowball sampling", in which relations among individuals are followed iteratively to identify new members of the network. In part this is because the bounds of a large network are difficult to determine in advance. Following the network ties we are interested in has the advantage of discovering new individuals at each step that are potentially relevant to the analysis. The chief disadvantage is that snowball sampling biases the sample toward well-connected individuals, and it is possible to entirely miss components of the network that are not connected to the starting point of sampling. These problems can be mitigated if multiple starting points are used.

The social network analysis conducted here is based on a crawl of YouTube's user profiles, following two of the site's social networking features, friends and comments. For the analysis of content, the video references themselves were processed to extract the author-provided tags (uncontrolled keywords). Authors were characterized by the aggregate of all the tags they have used on uploaded videos. In this way, we can associate the social network positions of authors with the type of content they produce. Analysis of the content of actual videos on YouTube is left to future research. The data and its collection are described in the remainder of this section.

2.1. Entry points

To seed the crawl, a large number of entry points needed to be identified. Ideally, these should not be biased toward any particular subject matter or genre. Random sampling of YouTube is not possible, since there is no publicly available comprehensive list of videos or authors. Hence an alternative strategy for identifying starting points was required. This issue was addressed using a specially-tuned Google search. For the period of one month (mid-February to mid-March, 2007), I ran a Google Blog Search looking for references to youtube.com in weblog postings. The search was implemented as a Google Alert emailed to my email address daily, listing the top ten search results. The search result pages are easily visited and processed using standard tools (e.g. wget, grep) to reveal the YouTube references. Through this method we obtained 216 starting videos for the crawl.

There are certain disadvantages of this method. First, as the Google search algorithm is proprietary, it is impossible to know how the results are ranked, and consequently, what ends up in the top-ten results. Presumably, link in-degree is a large component of the ranking, but it is unclear what other features may be in use, or just how different the blog pages are with respect to in-degree. Second, by using weblogs to provide the entry points to YouTube, we are subject to whatever biases exist among weblog authors, or whatever motivations they may have for linking to YouTube videos. Consequently, the popularity or significance of the video pages used as starting points is unclear, although it is expected that they will tend to be higher in popularity than average YouTube videos. Hence, a sample starting points collected this way will not likely be representative of YouTube videos more generally, and the resulting crawl is potentially biased by this approach.

A couple of factors mitigate these concerns. First, by using a large number of starting points, we maximize our chances of finding any large components, even if some are unconnected to the main core. Second, since our intent is to characterize the social structure of YouTube, in particular its core, popular videos, as would likely be found in popular blog posts, will get us to the core or cores faster. Moreover, popular blog posts tend to be ones linked to by other blogs (as opposed to non-blog sites), and given the topical and current-time orientation of blogs, this should assist in identifying videos that were popular at the time of the search. Hence, this search should suffice to take us to core members of the YouTube social network in a small number of steps.

2.2. Crawl

To crawl YouTube, we wrote a custom program to both fetch and parse the files using the YouTube REST developer API. SWI-Prolog was used for this purpose, as it has good libraries for both HTTP requests and SGML/XML parsing, and because Prolog simplifies the implementation of complex heuristic search algorithms, should they be needed at a later point.

Our crawler operates iteratively in two phases. In the first phase, the crawler identifies the video details from the video references we have. From these we identify the author of each video, and any comments on the video by other YouTube users. In the second phase, we identify all of the videos produced by an author, and all listed friends. This phase yields new video and user references which are then followed, iteratively. References that have been followed on a given iteration are remembered so that they will not be followed in future iterations. Table 1 illustrates the operation of the crawler in terms of the number of new references produced at each iteration. Only data up to the seventh iteration was used for the subsequent SNA, because of data size considerations. The fan-out (the ratio of new references to previously known ones) varies a great deal from iteration to iteration, and in the eighth iteration, falls off considerably. This suggests that most of the references revealed at iteration 8 point back to users of videos already discovered, and hence, the core or cores of the network have already been reached.

Table 1. Number of new video+user references in each crawl iteration.

Iteration	no. refs	fan-out
1	216	—
2	973	4.50
3	4216	4.33
4	22171	5.16
5	51730	2.33
6	172331	3.33
7	857456	4.98
(not used) 8	1073082	1.25

The YouTube REST API has some unfortunate features that impede the crawling process.[1] Friends, video and comment lists are all truncated (to 20, 20, and 11 items, respectively). This makes it hard to discover all of the friends of the more popular users, and impossible to discover all of the comments on a given video, and all of the videos produced by a prolific user (this information is viewable on the HTML pages, which are harder to crawl and parse). Conse-

[1] Since this crawl was conducted, changes in the YouTube API have removed some of the limitations described here.

quently, our use of comments is strictly to discover other users we might not have seen before, and the social network analysis focuses exclusively on the friends network.

2.3. Processing

User and video information identified by the crawler were saved internally as Prolog facts and output for later use into a Prolog source file. Information was retained for the user profiles, user videos, tags, comments, and descriptions. To prepare the data for social network analysis, we re-formatted the facts as tab-delimited text files which were then imported into R. As R and Prolog use somewhat different external representations of text, some minor data cleaning was necessary at this stage. The chief issues concerned user comments with Unicode characters that needed to be quoted for import into R. In the 2.7 million facts generated, there were less than a dozen such cases.

Two data files were created. One contained exhaustive information on all of the videos identified, the other contained author-to-author ties from friends ties. There were 9,948 videos, from 1,070 different authors, out of 82,185 total users identified; 148,235 friends ties were identified, but since this relationship is reciprocal, the actual number of edges is 136,797, or 273,594 if represented as directed ties.

3. Analysis

The data were subjected to two forms of analysis. First, the tags were extracted from the video data file and aggregated according to author. Clusters of authors and associated keywords were identified through vector-space projection and hierarchical cluster analysis. Second, after diagnostic evaluation of the friends network data, a pair of sociograms were generated representing the social connections among the 1,070 video authors. Cluster assignments of the authors were reproduced in these plots, as well as degree information, to assist their interpretation.

3.1. Keywords

Keywords were rolled to lower case before processing. Authors used a total of 20,914 distinct keywords, of which 5,313 were used more than once. Authors not using at least one of these keywords were dropped, leaving 1,022 authors and 5,313 keywords in the vector space. The data were log transformed, row and column z-score normalized and submitted to Principal Components Analysis. The resulting projections of authors in the reduced (20-dimension) vector space were then further clustered hierarchically using a combination of Euclidean distance and Ward's clustering method. A cut with nine clusters was chosen as reasonably representing the groupings of authors.

Table 2. Author groups and associated keywords from the YouTube crawl (keywords in italics).

Group	No.	Keyword/Descr
1	0 +	
Red	623 –	Celebrities, pop music, pornography, sex
2 Orange	20 +	*360 cloud enix fantasy final game hearts iraq kingdom namco nintendo playstation ps2 ps3 sora square test war wii xbox*
	687 –	R&B music, Celtic, Spanish, skate sports
3	3 +	*acoustic guitar live*
Yel-Gr	705 –	Anime character names (Bleach, Naruto series), video games, Star Wars
4	0 +	
Green	2263 –	Miscelaneous (unclear semantic relations)
5 Cyan	97 +	*Islam, christianity, linkin park, Dragon Ball GTand fullmetal alchemist anime*
	36 –	*747 911 airbus airplane airport amsterdam beach boeing bucknakeddragking china cover drag ds electronic flight flying guilty hank japan jazz jet king landing male philadelphia plane planes progressive runway schiphol senior theory torture tower wtc xii*
6 Sky Blue	11 +	*cat crazy cute dancing fun funny horse lol naked silly stupid*
	557 –	Political current events, hard rock and metal artists
7 Blue	135 +	Rap, hip-hop, freestyle artist names
	0 –	
8 B-Violet	79 +	Anime music video, anime character names (Naruto series)
	2 –	*magic trick*
9	101 +	Korean pop artist names
R-Violet	0 –	

To characterize the authors according to their selected keywords, the 5,313 projections of the words into the reduced vector space were correlated with the centroids of the nine author clusters, and assigned to whichever had the highest squared correlation. This separated the words into fourteen groups either associated with or avoided by specific author clusters. This permitted author clusters to be characterized according to the content of the videos produced.

As illustrated in Table 2, most of the nine author clusters are readily characterizable in some terms, and a few represent important genres of Internet videos. Only clusters 1 and 4 are not readily characterizable, having nothing in the way of associated keywords, and relatively large numbers of avoided keywords. Cluster 2 represents authors of videos about (console) video games, and is complementary to R&B, Celtic, Spanish and skate sports; cluster 3 addresses live acoustic guitar music, and avoids anime character names, video game and Star Wars references; cluster 5 refers to Islam, Christianity, and one group of anime characters (Fullmetal Alchemist), while avoiding 9/11 and aviation references; cluster 6 authors produce humorous videos while avoiding current events and hard rock; cluster 7 authors post rap and hip-hop videos; cluster 8 authors produce another set of anime music videos; and cluster 9 authors post Korean pop music performance videos (from artists Shin Hyesung, Hwang Yoon-Suk, Park Hwayobi, etc.).

Among the genres of Internet videos, anime music videos figure importantly in distinguishing several author clusters. Based on selective viewing of sample video content in this category, these videos characteristically feature edited clips of anime videos synchronized to popular music, sometimes but not exclusively sung in Japanese. From the character names that occur, the Naruto series generates the most references (the characters in cluster 8 come strictly from that series), although the availability of variant spellings (particularly for romantic pairings of characters: narutoxsakura, naruxsaku, narusaku, sakuraxnaruto, sakuxnaru, etc.) sometimes cause equivalent character names to be assigned to different clusters. Bleach, presumably a less popular series, appears negatively correlated with cluster 3.

Anime music video is highly prominent on YouTube: the third most popular YouTube group with 9,762 members is devoted to the anime music video form. Similarly, the prevalence of acoustic guitar is not surprising given that the most popular YouTube group, with 13,119 members, is devoted to guitar music. This cluster is also negatively associated with the Naruto and Bleach anime music video characters. Hence, it appears that without consulting the YouTube group memberships directly, our crawl has identified a major axis differentiating popular content types on YouTube. Humor is a similarly popular genre of Internet video, with its own YouTube groups as well, and cluster 6 appears to represent this category of video as well.

Notably absent from the keywords characterizing the video references found are terms identifying content currently in copyright dispute (e.g. Viacom's properties The Daily Show, John Stewart, Stephen Colbert, South Park, Laguna Beach, Sponge Bob, etc.), in spite of the potentially large YouTube audience for them. Possible reasons for their absence include policing by YouTube/Google, or merely a lack of social connection between people uploading such videos and the larger YouTube network. Such could be the case, for example, if a production of such illicitly copied videos were done in a one-off fashion. Other emblematic YouTube videos, such as the lonelygirl15 series, or the New Numa by Gary Brolsma, are also absent from mention in our sample. The reasons for this absence deserve fuller investigation in a future study.

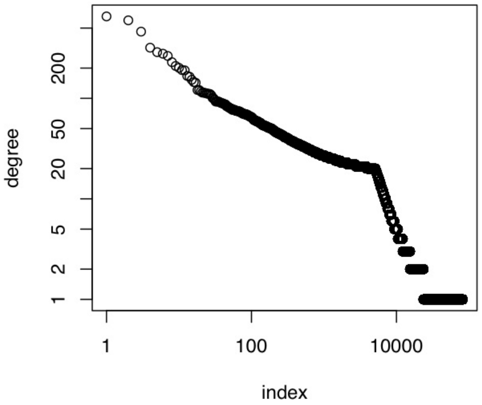

Figure 1. YouTube Friends degree distribution.

3.2. Social Network Analysis

Social network analysis was conducted using the R sna package [1], as well as general features of R [14]. The first step in this analysis was a degree distribution evaluation of the friends relation data, to determine if it has the expected power-law distribution. If it does not, this would question how representative the data are of You-

Tube's social network. This degree distribution is represented in Figure 1.

The first observation that can be made about the degree distribution is that it appears to represent two distinct power laws, with a phase break at 20. This is easily explained by the truncation of the friends lists to 20 items, as reported by the YouTube REST API. Hence, the portion of the graph below 20 links primarily represents out-degree information from user IDs that have not recurred in the data. Those above 20 up to the maximum of 653 represent more closely the actual degree distribution of friends on YouTube. A consequence of this truncation is that we cannot determine the average or median number of YouTube friends. At the same time, the range of the distribution compares closely to that of other social networking sites like LiveJournal [5, 11]. Hence, YouTube appears to function like other social networking sites, even though its primary function is the sharing of video.

When friends links among the 1,022 authors are aggregated according to the author clusters, we can draw a reduced sociogram suggesting the relative degree of connection among the authors of different categories of video. This sociogram is shown in Figure 2. Node size in this diagram represents the number of members of each group, log scaled, and edge weight represents the number of ties between clusters, corrected for cluster size. Different thresholds were examined to reveal the different levels of connectivity between the clusters. For clarity, self-ties within the clusters are omitted. All the clusters show very strong self-linkage, meaning that they are socially cohesive groups. Anime music video and Korean pop video producers have the strongest self-ties of all the clusters, followed by Religion and Rap. Hence they may represent additional cores, and further sampling might need to focus on those clusters for further clarification.

Outside of self-linkage, the strongest connections in this network are found among the triangle at the top (Misc 1, Misc 2 and Religion), and the two triangles at the bottom, which resolves to the chain, Guitar-Humor-Video Game-Rap, if the threshold is raised. The most weakly connected clusters are the Anime and Korean clusters, in spite of the fact that the cluster labeled Religion has anime music video references in it. This suggests that producers of anime music videos and Korean pop-music videos occupy more peripheral positions in the YouTube social network, while nonetheless having internal social cohesion. At the same time, the social core, with authors in the clusters 1–7, itself is divided somewhat into one aggregate, containing the two clusters of miscellaneous videos, the current events/religion/Fullmetal Alchemist anime music videos on the one hand, and another aggregate with guitar music, humor, video game and rap video producers on the other. This could represent a structural division in the core, and hence bears closer scrutiny.

To examine the core structure more closely, we constructed another sociogram with the individual authors disaggregated, but color-coded according to their cluster memberships. Node size was used to represent node degree, log-scaled. This sociogram is presented in Figure 3. Ties in Figure 3 are unweighted, whereas in Figure 2 they are weighted by both number of ties and corrected for group size.

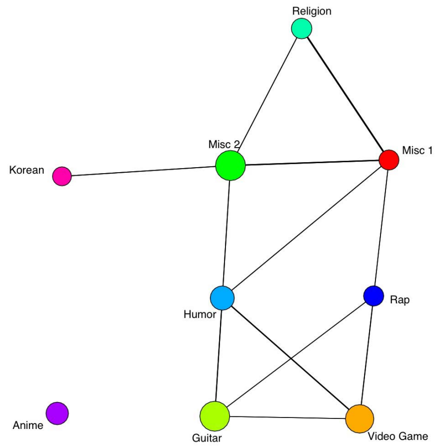

Figure 2. Reduced sociogram of friends relations among the different author clusters.

The relations observable in Figure 3 are only partly anticipated by those made regarding Figure 2. There is apparently a single core, with mixed characteristics from most of the user clusters, although dominated by the Video Game (yellow) Guitar (yellow-green) and Misc 2 (green) clusters. Anime music video (blue-violet) and Korean music videos (red-violet) do form two small but relatively connected groups, pushed to the periphery of the network. The Korean group is tightly cohesive, suggesting it is indeed a small but distinct core. The anime music video group is somewhat less cohesive, consisting of three somewhat loosely connected groups that are internally more cohesive. Finally, there are a large number of unconnected components with members from a range of different clusters.

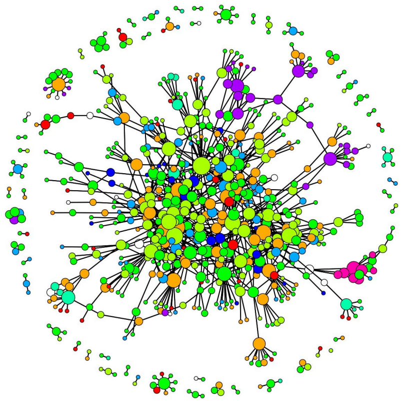

Figure 3. Sociogram of video authors' friends.

It should be remembered that the links represented here are only the direct links among the authors in the sample. The vast majority of YouTube users apparently do not upload any videos (and probably an even larger number of YouTube users do not have accounts). For reasons of data size, it is difficult to handle the entire set of users as individuals. Yet many of the non-producing users have numbers of friends equal to or greater than those of video producers, hence they are clearly important in the social structuring of the site.

To get a better indication of the role of non-producing users, all users with degree greater than 20 (the cutoff in the YouTube API) were assembled into a sociomatrix. This dataset includes 3777 users, 302 of whom are authors in Figure 3. Since this sociomatrix is still too large to be readily plotted, it was submitted to principal components analysis and hierarchical cluster analysis. Non-producing users were then aggregated into clusters, based on hierarchical cluster analysis using euclidean distance and Ward's method; authors were left disaggregated so that their content cluster assignments could be color-coded in the plot as before. The resulting sociogram is shown in Figure 4, in which node size is weighted according to log in-degree, and tie strength is weighted according to group size.

A number of observations can be made from Figure 4. First, the Naruto anime cluster and the Korean music video cluster remain tightly organized and distinct from the remaining core, much as in Figure 3. Separation from the main core is not as great, but this is evidently due to the highly-connected groups of non-producers. At the same time, clusters of non-producers are more closely associated with either main core producers, Naruto anime producers, or Korean music video producers. The main core is characterized by the guitar, video game, humor and rap categories of producers, those in the miscellaneous 2 category are uniformly spread throughout the sociogram, and the miscellaneous 1 and religion categories are only weakly present, being pushed outside the main core.

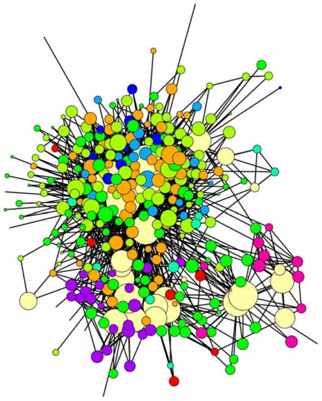

Figure 4. Sociogram of all users with degree greater than 20 (non-producing users are aggregated into clusters, which are cream-colored).

Hence, when considering centrality according to in-degree, the core-most social organization of YouTube is composed of producers and non-producers corresponding to the lower four nodes of Figure 2. The upper three nodes are indeed more peripheral, while the Korean and anime music video producers collect a substantial concentration of interest equivalent to smaller but competing social cores.

4. Discussion

Among the authors of YouTube videos, there is apparently a single main core, whose content is mixed from among the most popular genres on YouTube, as represented by author-applied uncontrolled keywords. Guitar, having the largest and apparently most active YouTube group, and humor, a characteristic Internet video genre, figure heavily in the core, alongside video game-related movies and rap music. Anime music videos and Korean music videos represent important genres with their own small cores, less connected to the central core. Other categories of content exist, but are less central to the social organization of YouTube's core. These results are reflected in all three sociograms, such that socially coherent activity (around friending) is strongly coherent semantically (around tagging), and the same patterns can be observed among producers with or without non-producers.

It is possible that the picture of the YouTube core we have obtained is biased by the starting points we employed. Since we obtained weblog references via search engine queries, they might be biased toward English-language weblogs, for example, and hence the authors favored by international YouTube users could be under-represented with Korean music videos being the only real example. This could be addressed by conducting similar YouTube crawls from other sets of starting points, and comparing the video author lists obtained. Should different results be found, this would suggest that YouTube has multiple cores, rather than a single core, with the core discovered here being just one, and that substantial distances separate the different cores. Two places to focus crawling that would help ascertain this would be among the Korean and anime music video producers, since they already represent cores separated somewhat from the main core. Also relevant would be the religion cluster (which includes references to another anime series, Fullmetal Alchemist) and the two miscellaneous clusters. These three clusters are presently assigned to peripheral positions in our analysis; if they instead represent additional, less-connected cores, this could be readily confirmed by specifically targeting them as starting points in a new crawl.

YouTube's social structure appears to have a typical core-periphery structure, similar to that observed in other social media spaces like LiveJournal [11] and Newgrounds [13]. Since YouTube video content is distinguished by tags, similar to the shared bookmarks of del.icio.us [12] and the user interests in LiveJournal profiles [11], one might expect to find little relationship between types of media on the one hand and the social organization of the site on the other. Yet quite the opposite is true: unlike LiveJournal and del.icio.us, social contacts of authors are characterized by strong cohesion around coherent semantic clusters of content. In this respect, YouTube is rather more like Newgrounds, in that both show signs that socially connected groups of users appear to cultivate specific kinds of content. On YouTube, we also see that this tendency also extends to audiences, as illustrated by the positions of the non-producing user clusters in Figure 4.

The reason for this difference among tagging systems may stem from the nature of the relationship expressed by tagging. When users tag videos on del.icio.us, they need not have any other users in mind, as their purpose is largely to assist themselves in finding the content later. Hence, tags express a relation of the user to the resource, and this can vary tremendously depending on the nature of the resource tagged. In LiveJournal profiles, interest tags are intended to express something about the nature of the individual whose profile it is, possibly for discovery by other users. Hence, users may converge on certain conventional meanings for tags, but some extrinsic norming process is needed to cause people with similar interests to become socially connected (this may also be antithetical to the ethos of LiveJournal). On YouTube, only videos are tagged, and they are only tagged by their authors. Hence, tagging on YouTube primarily exposes the content for discovery by other users, and convergence around conventional meanings can be expected.[2] Moreover, video producers are involved in a creative process, much like the Flash authors of Newgrounds, which foregrounds the need to cultivate audiences. Hence video producers benefit from social connections that provide mutual support, such as friending users who produce similar content.

General observations about the production and consumption of user-generated content can also be made from our analysis. A scant thousand of the full set of 82 thousand users identified are actually producers of video. This suggests that, in spite of any expectations to the contrary, YouTube's permissive environment for posting and distributing video do not fundamentally change the way video is produced and consumed. Video production, even of the most rudimentary type (e.g. recorded from mass media sources and uploaded), is a craft engaged in by a small minority of users. Hence, it is important to investigate in future research the relation between producers and consumers of video on YouTube, and what makes a user become a producer of video.

If social network relations play a role in this question, some attention will need to be directed to managing the scale of the YouTube user base. We presently do not have the means to investi-

[2] Similar differences among tagging systems are discussed in [6], where a related taxonomy of tagging motivations is developed.

gate the existence of the YouTube core in terms of individual users more generally, as the sociomatrix with 82 thousand nodes requires prohibitive storage allocations (in excess of 56 GB). Even thresholding the network to include nodes of degree 20 or greater resulted in a sociomatrix of approximately 119 MB, which is still cumbersome to work with.

It is likely that any increase in the number of nodes considered, whether from additional authors found in a new level of crawl or from ordinary users, would increase the connectivity of the observed core. Consequently, extra care needs to be taken in the analysis to ensure that social structure which is present is actually revealed. For example, plotting the full set of 3,777 users of Figure 4 without the additional step of aggregating non-producers into clusters results in an unrevealing mass of undifferentiated points. Hence, other means, such as ERG or *p** modeling [3, 15, 16], are necessary if the role of the author-keyword clusters in structuring the social network is to be investigated more closely.

5. Conclusion

From our investigation, YouTube does appear to have a social core among authors. In many respects, YouTube functions like other social networking sites, both in terms of its degree distribution and internal structure. Types of videos are found to be distinguished to some extent by author-applied uncontrolled keywords (tags), and these tags in some cases identify cohesive subgroups of authors exchanging similar content. This result sets YouTube apart from bookmarking/filtering sites like del.icio.us, and makes it more like the creative Flash portal Newgrounds.

In interpreting these results, it is important to recognize that friending is not the only relationship that structures YouTube interaction. Commenting is also another important social affordance, of YouTube, whose complete character must be left to future study. An awkward characteristic of commenting is that it is asymmetric, and since the YouTube REST API truncates the comment list to 11 comments, it is impossible to identify all of the users commenting on a popular video.[3] Consequently, we are unable to compare commenting with friending to know if it represents a fundamentally different relationship or if the two correlate somehow in the YouTube social dynamic.

Also on account of the YouTube REST API, we could not investigate the full range of videos posted by authors. This deficit includes the keywords assigned to each video. These additional videos need to be identified by visiting the HTML pages directly, and, because of HTML pagination issues, and issues with screen-scraping the HTML and the accompanying lower quality of the data, this was not done for the present study. Hence, it is possible that some authors are incompletely characterized in terms of their keyword usage.

When both of these things are taken into account it is likely that an apparently even denser core structure will be revealed. Hence, future research should employ advanced statistical modeling techniques such as ERG/p* models to address questions of genre in relation to social connectivity. At the same time, our research indicates a cohesive region of social interconnection among YouTube users producing videos, and an organization in which at least certain kinds of content are cultivated by specific social groups. In the process, several types of popular YouTube content have been identified. The social structuring of the YouTube video authors and their relation to different types of content is clearly rich enough to merit much further attention.

6. References

[1] Butts, C.T. 2006. The SNA Package for R. Available at http://erzuli.ss.uci.edu/R.stuff/; http://cran.r-project.org/doc/packages/sna.pdf

[2] Degenne, A.; and Forsé, M. 1999. Introducing Social Networks. Thousand Oaks, CA: Sage.

[3] Frank, O.; and Strauss, D. 1986. Markov graphs. *Journal of the American Statistical Association*, 81, 832-842.

[4] Herring, S.C.; Paolillo, J.C.; Kouper, I.; Scheidt, L.A.; Tyworth, M.; Welsch, P.; and Wright, E.L. 2005. Conversations in the blogosphere: An analysis from the bottom up. *Proceedings of the 38th Hawaii International Conference on System Sciences*. Los Alamitos: IEEE Publications.

[5] Herring, S.C.; Paolillo, J.C.; Ramos-Vielba, I.; Kouper, I.; Wright, E.; Stoerger, S.; Scheidt, L.A.; and Clark, B. 2007. Language networks on LiveJournal. *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Computer Society.

[3] This restriction was removed in the YouTube API released in August, 2007, but I have not had the opportunity to experiment with it as of this writing.

[6] Marlow, C.; Naaman, M.; Boyd, D.; and Davis, M. 2006. Position paper, tagging, taxonomy, flickr, article, toread. Paper presented at the Collaborative Web Tagging Workshop, 15th International World Wide Web Conference, May 2006.

[7] Newman, M. E. J. and Forrest, Stephanie and Balthrop, Justin, 2002. Email networks and the spread of computer viruses. *Phys. Rev. E*, 66.3 DOI: 10.1103/PhysRevE.66.035101

[8] Palonen, T.; and Hakkarainen, K. 2000. Patterns of Interaction in Computer-Supported Learning: A Social Network Analysis. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Fourth International Conference of the Learning Sciences*, 334-339. Mahwah, NJ: Erlbaum.

[9] Paolillo, J.C. 2000. Visualizing Usenet: A factor analytic approach. *Proceedings of the 33rd Hawaii International Conference on Systems Sciences*. Los Alamitos, CA: Institute of Electrical and Electronics Engineers (IEEE) Computer Society.

[10] Paolillo, J.C. 2001. Language variation in the virtual speech community: A social network approach to Internet Relay Chat. *Journal of Sociolinguistics*. Oxford: Basil Blackwell.

[11] Paolillo, J.C. Mercure, S.G.; and Wright, E.L. 2005. The social semantics of LiveJournal FOAF: Structure and change from 2004 to 2005. *Proceedings of the International Semantic Web Conference, 2005 Workshop on Semantic Network Analysis*. Aachen: Sun-Cite Central Europe (CEUR).

[12] Paolillo, J.C.; and Penumarthy, S. 2007. The Social Structure of Tagging Internet Video on del.icio.us. *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Computer Society.

[13] Paolillo, J.C.; Warren, J.; and Kunz, B. 2007. Social Network and genre emergence in amateur Flash multimedia. *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Computer Society.

[14] R Development Core Team. 2007. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

[15] Wasserman, S.; and Faust, K. 1994. Social Network Analysis: Methods and Applications. Cambridge, UK: Cambridge University Press.

[16] Wasserman, S.; and Pattison, P. 1996. Logit models and logistic regressions for social network I: An introduction to Markov graphs and *p**. *Psychometrika*, 61.3, 401-425.

[17] YouTube. 2007. YouTube Fact Sheet. http://youtube.com/t/fact_sheet