A Network View of Social Media Platform History: Social Structure, Dynamics and Content on YouTube

John C. Paolillo
Indiana University Bloomington
paolillo@indiana.edu

Sharad Ghule
Indiana University Bloomington
sharadsghule@gmail.com

Brian P. Harper
Indiana University Bloomington
bpharper@indiana.edu

Abstract

Social media sites are prone to change from many internal and external causes, yet it is difficult to directly explore their histories in terms of the content itself. Search and browsing features are biased toward new and paid content, archives are difficult to navigate systematically, and their scale makes any observations challenging to contextualize. Here, we present results of an ongoing study of YouTube's history (currently with more than 76 million videos) using a combination of iterative browsing, network crawling and clustering within and across time periods. Through this method, we are able to identify historical patterns in YouTube's content related to internal and external events. Our approach thus illustrates an adaptation of network analysis for understanding the content histories of social media platforms.

1. Introduction

Currently, YouTube is at a crossroads: YouTube's dominance in online video is now challenged by Amazon, Facebook, Hulu, Netflix and Twitch. YouTube's visibility has exposed it to regulatory scrutiny and advertiser protests, threatening revenue. In response, YouTube has changed its advertising algorithms and upset the economic viability of many channels, alienating channel owners. Any of these conditions could induce large changes on the site, shaping its content or what we can access of it.

We therefore need a history that would chronicle the emergence and influence of the platform's dominant genres and content types since 2005, ideally indexed to changes in the platform's features and incentives as well as external world and media events. YouTube has archival properties, however, and the YouTube public data API reflects the historical character of the site through the publication dates of video and channel metadata. Channels and their videos are also structured as a network, via relations such as liking and favoriting videos. Can this information be used to further illuminate the history of the site?

Our answer to this question is yes, based on a network analysis in which the publication dates of videos are used to segment the YouTube network into a sequence of time slices, covering its entire history from May 2005 to December 2016. This analysis reveals the evolution of a range of different genres of content, which can be read in terms of responses to historical events and platform changes. This work provides a potentially important frame for the interpretation of past and current studies of YouTube content.

2. Literature Review

From its initial pre-launch public availability in 2005, YouTube rapidly became the dominant platform for the distribution of online video. This 12-year history has been unstable, punctuated by technical changes to the platform, purchase by Google, introduction of advertising, international expansion, for example. External events have also had effects: large user migrations, political events, copyright lawsuits, changes in national and international regulation of internet technology, major studio participation in YouTube, and the US presidential elections have all been felt in different ways by YouTube users.

Empirical research insufficiently contextualizes YouTube's content and its evolution. Early attempts at a global-scale analysis of YouTube's content exist [1], but they are either small in comparison to its actual scale at the time [2], are based on specific events [3], or they do little to address the nature of the content or how it might relate to platform features [4, 5]. A representative compilation of early research on YouTube is *The YouTube Reader* [6]. Early histories of the platform exist [7], but numerous changes in the site have obscured the relationships among YouTube's features, users, content and external events.

Other YouTube research has addressed YouTube's politics as a platform [8], the recommender system [9, 10, 11], social network effects on content propagation [3, 12, 13], the features of memes [14], multichannel

URI: https://hdl.handle.net/10125/59701
ISBN: 978-0-9981331-2-6

networks [15], and even specific genres of content [16]. These pieces often exist in isolation of YouTube's development over time, as can be seen in the contradictory findings at different times regarding the popularity of longer videos [17, 18].

An important contextual component missing from the discussion of YouTube is the role of mutual support among channels in the cultivation of its genres. YouTube's liked and favorited video playlists offer one record of such support, which also flows and ebbs over time, as channels become active or dormant. Such social processes have been shown to be instrumental in genre emergence [19], and a network analysis offers one approach for revealing them [20]. Time in network analyses, however, has no standardized treatment. We therefore ask: how can we use the network of likes and favorites among channels to read a history of genre evolution on YouTube?

3. Method

The method employed in this study has three main components: (i) construction of a sample using browsing and crawling and the Google/YouTube public data API, (ii) extraction of time-located network samples and clustering them, and (iii) organizing and interpreting the timeline of network clusters. Each of these corresponded to three distinct phases of research, discussed in turn below.

3.1 Sampling YouTube

YouTube is large and unwieldy, and its complete data are accessible only within Google. Data for individual videos are exposed only through search and browsing functions that are subject to unknown biases (e.g., sponsored search features and the video recommendation algorithm) and cannot be sampled in a truly random manner, and we must resort to crawling a large sample. Problematically, crawled samples miss unconnected components. Consequently, a diversified strategy for sampling is necessary, relying on searching and browsing to identify starting points for crawling, and iterative phases of both activities.

The initial sample for this study was based on a collection of YouTube channel IDs identified for a project on conspiracy theory videos in 2015, using searching and browsing strategies. A script written for the Firefox Greasemonkey plugin was used to collect channel IDs into a PostgreSQL database directly while browsing. In addition, the script reports whether the channel for the current page was already recorded in the database. YouTube search was used to initiate browsing, and browsing strategies were developed so as to rapidly gather distinct channel IDs. On a typical video page, the first video listed on the right bar often comes from the same channel, and the second is an advertisement. Videos from the third on come from a range of channels: the same channel, related channels and "recommended" channels. The last of these are fed by a YouTube algorithm that references a user's viewing history; typically these have already been visited. We therefore focused attention on videos after the first two with unfamiliar channel names, using the thumbnails and titles to help recognize if a particular video had already been seen. When the initial project was broadened beyond conspiracy theories, the same strategies were employed, merely using different YouTube searches from which to begin browsing. Channel IDs from browsing became the seed set for a crawl collected through the YouTube public data API. Each channel is associated with three playlists: uploads, likes and favorites. The first is merely the list of the videos uploaded by the channel; the second and third represent videos that users have identified as ones they like or favorite, using YouTube's interface features. Typically, these videos are ones produced by other channels (though they need not be). YouTube's recommendations are generated partly from videos that are co-liked or co-favorited with the video being watched. Hence, crawling these two playlists to obtain the video information and that of their associated channels tends to expand the set of channels observed while mirroring YouTube's video recommendations.

Unfortunately, crawling via the API has limitations. It does not list channels that liked or favorited a particular video, so we must always identify channels first. This requires that all our channels post videos, when many do not. Relations to such profiles could be crawled through the comments feature, but this would expand the data collection beyond the capabilities of our current system architecture. Similarly, channel subscriptions are treated as private by the API, and for non-posting channels, likes and favorites can also be made private. Without appropriate searching and browsing strategies it is likely that sections of the network would be missed, especially less popular channels. For this reason, the searching/browsing and crawling processes were repeated several times from July 2015 to March 2017, ending with a sample of 76,081,372 videos and 549,383 channels.

The resulting database contains metadata for a small but popular and highly connected fraction of the total activity on YouTube. Although our sampling began with the conspiracy theory channels, these are a

small proportion of the final network, which is otherwise dominated by entertainment content (below).

3.2 The Network Over Time

Our network analysis of YouTube is based on the structure induced by likes and favorites; we treat these as indicating directed links between channels, i.e., a channel has a (directed) link to another channel as strong as the number of times the first channel likes or favorites videos uploaded by the second. We treat likes and favorites as equivalent because the two relations are strongly correlated [4]. Likes and favorites also tend to occur in a short window of time after a video is released [4]. For this reason, we use the video publication date of the liked/favorited video as a proxy for historically dating the relationship.

Using 3-month intervals over the video publication dates as a moving window in which to examine connectivity of channels, we segmented the network into 141 samples, starting from April 2005, shifting the window by one month for each sample, and ending with the December 2016 sample. To keep our networks within a size that we could process, we used a threshold of a minimum of 10 likes/favorites from one channel to another within any given sample to include a link in the network.

3.3 Clustering

There are many approaches to clustering networks [21]; here, we employ the Louvain method of [22]. This algorithm performs well for large networks, especially with a high clustering coefficient and a fat-tailed degree distribution, as occurs in the YouTube network [4]. It performs an agglomerative hierarchical clustering in which a node is assigned to a cluster if doing so maximizes the modularity of the network, continuing until either a single node remains or modularity cannot be increased further. Modularity clustering is not perfect: it sometimes infers non-existing relationships between clusters based on weak false positive links [21], and tends to give large numbers of clusters in sparse networks. Nonetheless, it works well for detecting well-defined but small clusters in very large networks, as we expect to be the case with YouTube. Since crawling biases our samples toward connectivity, we anticipate some issues with interpretability in larger clusters.

Compatible implementations of this algorithm exist in Gephi [23] as "modularity class", and in the igraph package [24] as the function cluster_louvain(). For clustering the samples, we use the implementation in R [25]. This results in anywhere from 1 to 3747 clusters for each sample depending on its size and overall connectivity. Clusters are identified by arbitrary ID numbers and the only means for identifying them across different samples is through their aggregate common memberships, which we obtain by cross-tabulating clusters from successive pairs of samples. This is identical to a treating the entire sample as a network of clusters, in which links represent shared membership between clusters across different samples.

For convenience, the cluster comparison network was imported in its entirety into Gephi as a directed network with no minimum value for a link. Using Gephi's modularity class, we assigned each of the clusters within samples to new cross-sample clusters. Clusters with substantial overlap or that regularly exchange members fall together into a single new cluster assignment; clusters whose membership largely excludes those of another cluster over time appear in distinct new clusters, thereby identifying clusters with stable yet evolving membership across time. The success of this approach depends on the suitability of the threshold for the initial samples, the size of the moving sample window, the frequency of the samples, and the stability of class membership over time, so changes in these values would yield different results.

Gephi also provides network layouts; a suitable layout for this data should be able to find a linear structure or structures, showing the evolution and relative closeness of different content clusters. We used two force-directed layouts: the Yifan Hu layout for rapidly finding the global structure, and Force Atlas to verify that the observed structures were not peculiar to Yifan Hu.

The resulting layout appears as Figure 1, in which we find a single linear structure whose two large bends and single sharp elbow correspond to gradual and sharp changes in cluster membership, respectively. The layout has been rotated so that clusters from the earliest samples are on the left, and tracing along the main connected path takes one through more recent samples, to the final sample on the far right. Nodes in Figure 1 represent the sample modularity classes, with color indicating the cluster a node belongs to and size its number of members.

The largest 24 of the clusters (out of 14978 total) account for 75.9% of the network's nodes, with the next largest containing only 0.1%. Individual clusters are rendered in Figure 2, so that their lifespans can be more readily recognized, alongside their relative sizes and general type of content (this indicates in which subsection it will be discussed below). The largest nodes group around a central path, with fine filaments

representing the paths through the smaller classes and clusters extending outward from it on either side, including smaller clusters not shown in Figure 2, as the full layout exceeds the margins of the image. To clarify the cluster timelines, we produced Figure 3, in which each cluster is represented by a horizontal bar spanning the x-axis from its beginning point to its endpoint. Scanning vertically in Figure 3 indicates which clusters overlap at specific times.

To facilitate cluster interpretation, we created a web interface that provided summaries of the number of videos in each cluster for each month of the sample, along with a listing of ten videos from each of the 100 most connected channels in the cluster and active links to the videos and channels on YouTube. All three co-authors explored the full complement of clusters through this interface, meeting together to discuss and reconsider their interpretations.

4. Interpretation of the network

A few observations can be made from Figures 1 and 2 directly. First, there is a single central core to YouTube's network with varied content, as reported in [2]; it is stable over YouTube's history, although its composition changes. Many filaments diverge from the core, carrying channels toward or away from it, but these account for only a quarter of the observed network. In other words, YouTube's content is not strongly segmented due to, e.g., language markets, political polarization or content, as might have been expected. Such a pattern would appear as multiple, disentangled paths in the network, arising from clusters whose exchange of members with the core clusters is less frequent. We turn now to the specific patterns of content within the clusters that can be observed.

4.1 The Early Years of YouTube

Clusters 387, 629, and 909 represent the first stages of the development of YouTube's content. Cluster 387 arises in October 2005, just 8 months after YouTube became public; it contains mostly music-based channels, typically songs made by YouTube users or remixes of popular songs. This cluster also contains viral videos (for example, "Charlie bit my finger - again," "Evolution of Dance," and "Sneezing Baby Panda"), indicating their importance in YouTube's early history (they are otherwise infrequent). Cluster 387's content is predominantly entertainment, suggesting that the platform served a limited function in its early phase. Clusters 629 and 909 branch off from 387, maintaining continuity in both having music channels.

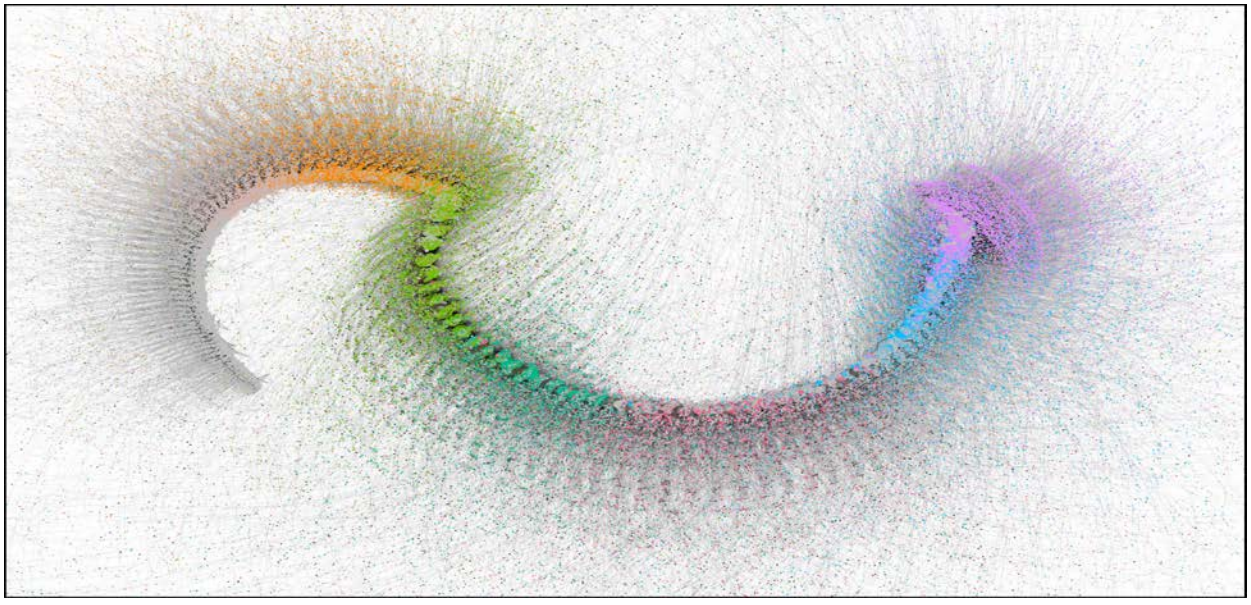

Figure 1. Final layout of network of shared membership in modularity classes of 141 sample networks, based on 3-month samples of YouTube channel-to-channel likes and favorites spaced at overlapping one-month intervals.

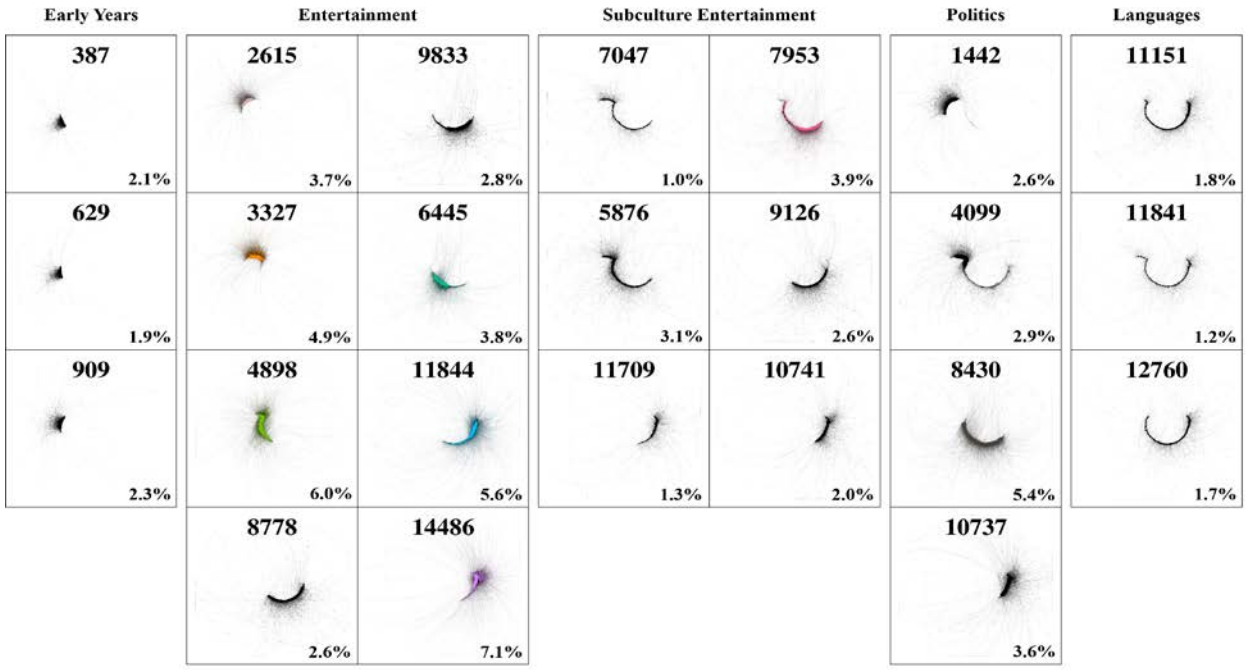

Figure 2. Clusters of modularity classes in the network of Figure 1, grouped in columns by type of content. Size of each cluster as percentage of modularity classes of the cluster comparison network is given in the lower right corner of each panel.

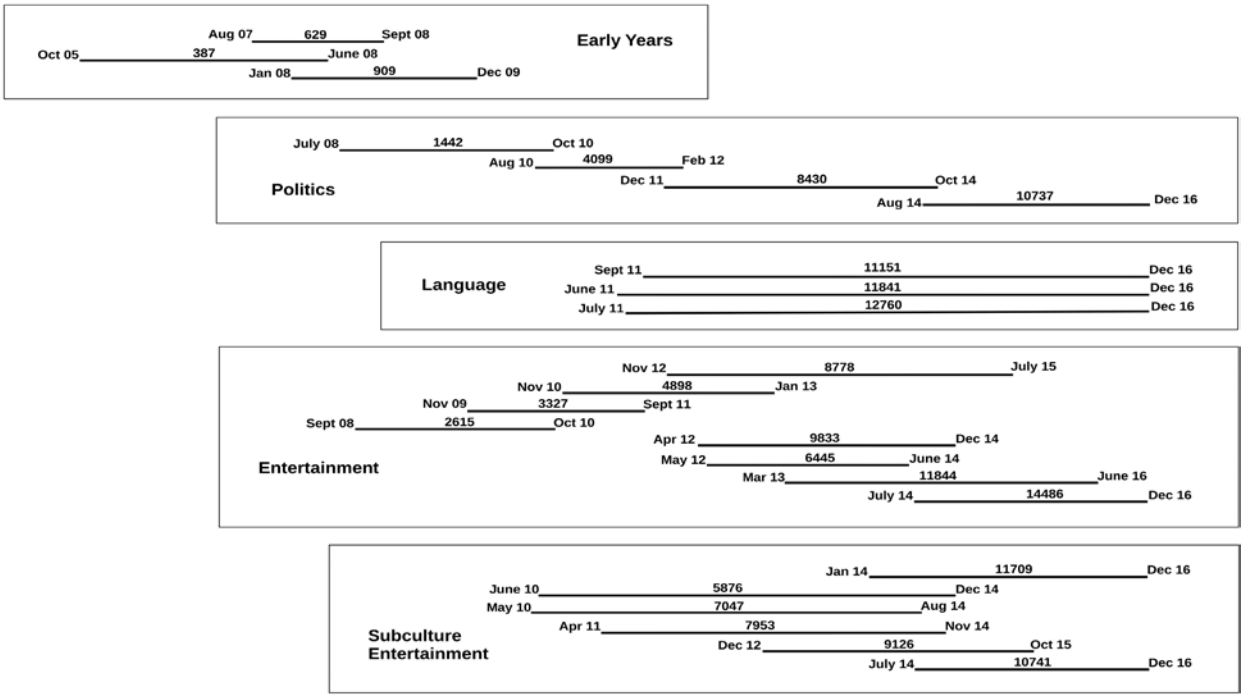

Figure 3. Temporal relationships of clusters in Figure 1, grouped by type of content. The left-right location and extent of a bar indicates the time period occupied by a cluster; clusters that overlap vertically occur at the same time.

Cluster 629 features more channels with hip-hop and non-pop music genres, alongside other early news channels such as the BBC News channel with material from its regular broadcasts, and The Young Turks, a news/commentary channel created specifically for Internet-based consumption. Both 387 and 629 end while increasingly accruing content related to the 2008 US presidential election. The official channel for the Barack Obama campaign especially gains prominence in summer 2008 in cluster 629, but both clusters feature electoral content.

Cluster 909 features music, comedy (especially for young males), and early YouTube-specific channels like Machinima, College Humor, and Rooster Teeth. It persists later than 387 and 629, possibly owing to its lack of political content from the 2008 election. Like cluster 629, broadcast media sources appear in 909, but while the media in 629 was political, 909 contains entertainment, such as the BBC cars and comedy show Top Gear. Cluster 909's last months overlap with the launch of Vevo channels, which likely served to disrupt the music channels present in this cluster.

4.2 The Two Main Streams

The differences in 629 and 909 become amplified after this point, yielding two distinct streams of subsequent clusters lasting through the rest of the timeline: one with primarily entertainment content and the other with political content. The entertainment stream, originating from cluster 909, is larger and more fluid, with some members separating into specific subculture topic clusters and rejoining popular entertainment clusters later. The politics stream, partially descended from 629, is more stable, and while it possesses multiple clusters within it, they have strong temporal continuity from cluster 629 onwards. Clusters outside these two streams represent language markets outside the dominant English-language market on YouTube.

4.3 The Entertainment Stream

The Entertainment stream is broad, with the three largest clusters in the network. Its size and diversity make it somewhat difficult to characterize, but several consistent elements recur: music, video games, and comedy. This stream begins in late 2009 with clusters 3327 and 2615, around the time Viacom's settlement with Google that resulted in the Content ID system. Google simultaneously changed its rating system from a five-star rating system to merely reporting likes. Both changes arguably affect the connectivity among channels, and hence may play some role in explaining the emergence of two distinct clusters. Cluster 2615 is dominated by copyright-protected popular music in the form of Vevo channels, i.e., predominantly pop and country music. There are occasional parkour and skateboarding channels, whose videos often contain musical backing similar to Vevo's pop music, though they are clearly not professionally mixed. Cluster 3327 contains music and content created within the YouTube platform itself, outside of any arrangement with Vevo. Its music comes predominantly from electronic genres such as dubstep and techno. Cluster 3327 also shows some of the early video bloggers ("vloggers").

Cluster 4898 wraps around the elbow in our data representation, and both clusters 2615 and 3327 feed into it, though not at the same time. This period appears to correspond to the shift in YouTube's definition of a view to include watch time in March, 2012. Prior to the elbow and the sudden decline of cluster 3327, 4898 is characterized by a mixture of video games and music content. The gaming content is not centered clearly on a specific game or genre, and the music content is partially Vevo channels formerly in cluster 2615 and partially variations of pop music. Lindsey Stirling's channel exemplifies this music; at this time she produced violin pieces inspired by video game music. After channels from cluster 3327 merge into 4898, it becomes dominated by personality-driven YouTubers: GennaMarbles, PewDiePie, CaptainSparkles, among others. These channels combine some contemporary topics, lifestyle, gaming, fashion, etc., with comedy focusing on the host's personality, and become central to the entertainment stream from this period onwards.

Rather than remaining a new large cluster, 4898 splinters into multiple clusters: 6445, 9833, and 8778. These show similar divisions as 2615 and 3327 did in that while all feature popular culture, 8778, 9833, and 2615 feature content produced for non-YouTube audiences (e.g., broadcast and cable television), while 6445 and 3327 feature content produced specifically for YouTube. Cluster 6445, like 4898 before it, features personality-driven YouTubers, primarily in gaming and fashion. Cluster 8778 features predominantly Vevo and other music channels, alongside late night comedy show clips. Cluster 9833 also features late night comedy channels like The Tonight Show Starring Jimmy Fallon, although much of the cluster focuses on reviews and critique of movies. Sketch comedy channels are also present. The presence of musical guests on late night comedy shows

helps to explain the presence of these shows in both clusters.

These three clusters feed into the two clusters of 11844 and 14486 at the end of our sample period. Much like cluster 4898, which straddled a major change in our network, 11844 begins with general gaming content, primarily of Let's Plays of varied games, though Minecraft makes a strong and consistent showing. As cluster 6445 ends, many of the channels from that cluster shift into 11844, leading to the personality-driven channels dominating it. Cluster 14486 began as a combination of fitness and lifestyle channels; at this time it also contains many popular videos in Italian, German, and Czech, though often with parts of their titles in English or subtitles in English. Later, language-specific clusters form around some of these channels.

By mid-2015, after clusters 9833 and 8778 end and their constituent channels merge into 11844 and 14486, it becomes increasingly difficult to accurately describe 11844 and 14486, with both clusters becoming so broad in scope that, while they are still clearly entertainment-focused, further characterization is difficult. This trend increases following the end of cluster 11844 in mid-2016 as most of the entertainment channel fall into 14486. One possible explanation for the difficulties in the period after mid-2015 is the candidacy announcements of the 2016 US Presidential Election, whose use of late-night comedy TV as a platform may have destabilized the entertainment clusters. A second is that the implementation of YouTube's paid subscription service YouTube Red may have disrupted some clusters by putting interactions behind a paywall where they are no longer observable. A third potential explanation is that the recency effect in likes and favorites distorts the network connectivity at the end of our sample period, closer to the point of our actual observations. The precise reason cannot be known right now, but we note that the clustering may be lower quality toward the end of our sample period.

4.4 Subcultural Entertainment

In addition to the main entertainment structure, there are several offshoot clusters of entertainment videos throughout the network. These clusters usually exist for brief periods of time, after which the channels involved rejoin one of the larger clusters of the entertainment stream. This is not universally true, and some channels in subcultural clusters regularly fluctuate between the main cluster and the subcultural cluster. Some are also more naturally associated with the politics stream rather than entertainment.

Cluster 7047 is one such subculture cluster, and lasts for an unusually long time for this network where the constituent channels join either clusters 11844 or 14486. Originally having short skateboarding and parkour videos, longer videos made using GoPro or other active point-of-view cameras start to show up in 7047, allowing an "urban spelunking" genre in which the content creators explore rooftops, climb tall buildings, and generally trespass, often without safety equipment. Skateboarding's importance diminishes after the emergence of the GoPro videos.

Another subcultural cluster, 5876, features plane flights and train operation as its primary material. Its focus is clearly technical, especially the inner workings of machines. At different points, this cluster attracts channels from the Maker movement (electronic gadget and computer hackers) and car tinkerers. It transitions into cluster 11709 with relatively little difference in the channels of interest, although at that point drones emerge as a video platform. Cluster 11709 remains distinct from the rest of the entertainment cluster until the end of our data collection rather than joining cluster 11844 or 14486.

Cluster 7953 is dominated by hyper-masculine channels aimed at high school and college-aged males. The videos in this cluster include sports, male heteronormative dating advice, pranks, and videos of gameplay in Call of Duty. More so than among the other clusters, videos of these channels tend to feature video thumbnails with attractive, scantily-clad women on them. This cluster could be characterized as appealing to "Bro" culture. A different male-audience cluster is 9126, featuring cars, guns, and gadgets, which partially overlaps with transportation clusters 5876 and 11709, although 9126 is favors experiences of speed and danger with cars more than their technical workings. This cluster also includes gun enthusiasts and maker channels. Many videos in 9126 demonstrate some technology a channel is dedicated to, thereby serving an advertising function. Members of this cluster arrive from both the entertainment and political (see below) streams, and some channels display salient political attitudes alongside the subjects of their technical interests, a stance less common in other entertainment clusters.

One subcultural cluster in our network is not predominantly masculine in character, namely cluster 10741. This cluster features relaxing content, primarily soft music, nature sounds, and videos of animals. As the content appears to be based more on a mood than on a particular language, this cluster also has some

crossover with non-English languages, especially Polish and German.

4.5 The Politics Stream

A second major stream of channels is one in which the content is primarily political. This stream is relatively consistent across the timeline, with four contiguous clusters spanning the period from the end of cluster 629 to the end of our sample. Cluster 629, while containing more news and political content than 909, still contained elements of popular culture like hip-hop music. Cluster 1442 does not share these elements, and is defined by its juxtaposition of news and conspiracy theory channels. These conspiracy theory channels promulgate a wide variety of different types of conspiracy theories (e.g. those involving extra-terrestrials, Illuminati/Freemasons, anti-Semitic themes, and Christian apocalyptic end-times), though they tend toward the conservative end of the political spectrum. Channels supporting atheism also appear in this stream.

Clusters 1442, 4099, 8430, and 10737 form a continuous chain, with clean breaks between the clusters. The shifts occur during the major events of the network graph: the 2008 curve, the 2012 elbow, and the disrupted period of mid-2015, indicating a connection to the rest of the environment not shared by the specific language-market clusters (below), which remain intact through these important events in the network. While having a preponderance of conservative channels, the political stream also has some clearly progressive and liberal channels, such as The Young Turks, hosted by Cenk Uygur. This channel has a relatively continuous history from early clusters like 1442 and 4099, although it occasionally crosses over into the entertainment stream at later points.

Also continuously present in this stream is the channel RT, the flagship channel of the Russian government sponsored news network of the same name that also includes RT America and RT UK alongside other content from the same network. This stream does not include Russian language versions of RT, which are located in the Russian language cluster, discussed below. RT and its sub-channels are highly integrated into the network of these clusters, with RT achieving very high numbers of likes/favorites throughout the stream. Also in this stream, and similar in form and media structure to RT is Al Jazeera, the Qatari government-sponsored news organization, with Al Jazeera English being their most successful channel on English-language YouTube.

Broadcast and cable television news media YouTube channels appear in this stream, including cable news channels ABC, CBC, CBS, CNN, Fox News, MSNBC, and NBC, several local affiliate channels of these networks. State-supported networks BBC and CBC also appear. However, most of these television media outlets arrive in this stream later and are not as well integrated into the network as the media entities discussed previously. While they spend more time in the clusters of the politics stream than elsewhere, they often cross over into other clusters more focused on entertainment. Despite the popularity of their content outside of YouTube, their likes/favorites pale in comparison to those of RT. It is likely that this represents a deliberate strategy of linking on the part of RT (and a reciprocal lack of that strategy on the part of other news organizations) rather than an absolute index of popularity. In other words, different organizations differ in their employment of search engine optimization techniques, and this influences the likes/favorites we have observed.

4.6 Specific Language Markets

In addition to these two main streams, we identify several major clusters of YouTube channels specific to different language markets. Clusters 11151, 11841, and 12760 represent Spanish, Brazilian Portuguese, and Russian language channels respectively, and are remarkably similar in their positions in the network. All three start in mid-2011 and persist until the end of our collection in 2016; these three are the longest-lived clusters in our network. While the Brazilian Portuguese and Russian clusters are heavily identified with their country of origin, the Spanish cluster, 11151, is somewhat more complex: many of the popular channels in the cluster are from Spain, but channels from Latin American countries are also present.

The Spanish and Brazilian Portuguese clusters appear to be most similar to the entertainment stream of the larger English-language YouTube environment: video gaming and music content dominate the large nodes in this cluster, with interspersed news and political content. The Russian cluster 12760, in contrast, appears closer to the political stream in content. While there are large channels producing gaming and music content, the frequency of political content is higher, buoyed by the constant presence of the Russian-language RT.

The specific language market clusters are not totally isolated from English language YouTube, as seen during February 2013, the month of the Chelyabinsk meteor. Images of the meteor were

recorded by cameras in people's cars and widely shared on YouTube, inviting a sudden surge of activity towards cluster 12760 from the rest of the network, and many news channels from English-language YouTube like the BBC cross into the Russian cluster for that month.

It is likely that there are other language clusters present within our network which fell below the 1% size threshold that we used to determine which clusters to explore qualitatively. However, we also noted Italian, Polish, and German language channels within the entertainment stream (especially within music-related clusters), so it is also possible that some non-English videos collect likes/favorites that link them more to the English-language YouTube environment than those of other language markets.

5. Conclusions

We have investigated the history of content on YouTube as a temporally changing network of channel-channel relationships expressed in the likes/favorites playlists of channel owners. The history is read using publication dates of the videos. From this, we have shown that YouTube has a strongly-connected core which is nonetheless differentiated throughout its history by certain prominent kinds of content: late-night comedy, popular music, political news, cars and trains, etc. These clusters exchange members on a relatively regular basis.

We furthermore observe that certain platform changes and external events appear to have an effect on the connectivity of the network and the nature of its content. For example, the change to watch-time diminished the popularity of independent comedy sketches, but new genres of videos like Let's Plays could then form. New genres of video also appear after the relaxation of video length restrictions in 2010, and after the introduction of point-of-view cameras such as the GoPro. Previously-existing clusters featuring hobby and sport-related content are reshaped to accommodate the new content types. Similarly, large external events such as the Chelyabinsk meteor in 2013 and the US presidential elections in 2008, 2012 and 2016, may be felt throughout the network, sometimes triggering major reorganizations, most evident here in the political stream of clusters.

A few caveats should be acknowledged. First, our network of YouTube, though large, is an incomplete picture of the platform and its evolution. It is highly challenging to characterize: many of our clusters represent the activity from tens of thousands of channels each with hundreds of hours of videos. Thousands of smaller clusters remain uncharacterized; among these may be some that are important to another view of YouTube or its history that cannot be accessed with this method alone.

Second, as with any network study, the view of YouTube we have produced depends in unknown ways on the sample. Stopping at December 2016 permitted the entire 2016 presidential cycle to be included, but the largest cluster coincides with this same period, meaning that it might be distorted by the 2016 election. Alternatively, this cluster is simply a recency effect of the likes/favorites having been sampled in this time.

Similarly, we cannot know what biasing effects the choice of seed channels might have had, although we expect such effects to be mitigated by large number of starting points. Yet another problem facing the network analysis is the effect of losses, especially when channels or videos are deleted by YouTube for violating copyright due to the Content ID system instituted after Viacom's 2007 lawsuit of YouTube. For any deleted videos, their contribution to the connectivity of the network, can now never be accessed. Losses due to decay and censorship impair all of our historical understandings, and YouTube's possession of a feature designed to delete infringing content adds to this problem. Remedies for these issues are possible, but not available to most researchers. For example, Google and YouTube possess complete knowledge of the YouTube network, including perhaps some of the censorship and losses; with such data, a more complete story could be told.

Nonetheless, it should be clear that paradoxes such as the conflicting interpretations of the value of longer videos [18, 17] can readily arise in the absence of a proper historical context. In this specific case, the two different observations were made at times before and after, respectively, YouTube implemented the watch-time feature incentivizing longer videos, affecting multiple genres. Similarly, the periodic effect of external events is palpable in the political stream on YouTube. In these and other ways, the present study serves to illustrate the potential utility of a network history in directing and developing historical interpretations of a social media platform.

6. References

[1] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: a view from the edge", In: Proceedings of the 7th ACM SIGCOMM, ACM Press, New York, NY, 2007, pg. 15-28.

[2] J Paolillo, “Structure and network in the YouTube core”, In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences, Los Alamitos, CA, 2008.
[3] J. Klausen, E. Barbieri, A. Reichlin-Melnick and A. Zelin, “The YouTube Jihadists: A social network analysis of Al-Muhajiroun’s propaganda campaign”, Perspectives on Terrorism, 2012, 6(1).
[4] X. Cheng, C. Dale and J. Liu, “Statistics and social network of youtube videos.” In: IWQoS 2008, 16th International Workshop on Quality of Service. IEEE Computer Society, Los Alamitos, CA, 2008, pg. 229-238.
[5] D. Halpern, and J. Gibbs, (2013) “Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression”, Computers in Human Behavior, 2013, 29(3): 1159-1168.
[6] Snickars, P., and P. Vonderau (eds), *The YouTube Reader*, National Library of Sweden, Stockholm, Sweden, 2009.
[7] Burgess, J., and J. Green, *YouTube: Digital Media and Society Series*, Polity Press, Malden, USA, 2009.
[8] T. Gillespie, “The politics of ‘platforms’”, New Media and Society, 2010, 12(3): 347-364.
[9] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston and D. Sampath “The YouTube video recommendation system,” In: Proceedings of the fourth ACM conference on Recommender systems, ACM, 2010, pg. 293-296.
[10] H. Yoganarasimhan “Impact of social network structure on content propagation: A study using YouTube data”, Quantitative Marketing and Economics, 2012, 10(1): 111-150.
[11] R, Zhou, S, Khemmarat and L, Gao, “The impact of YouTube recommendation system on video views”, In: Proceedings of the 10th ACM SIGCOMM, ACM Press, New York, NY, 2010, pg. 404-410.
[12] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro “How useful are your comments?: analyzing and predicting YouTube comments and comment ratings”, In: Proceedings of the 19th International Conference on the World Wide Web, ACM Press, New York, NY, 2010, pg. 891-900.
[13] A. Susarla, J.H. Oh, and Y. Tan, “Social networks and the diffusion of user-generated content: Evidence from YouTube”, Information Systems Research, 2012, 23(1): pg. 23-41.
[14] L. Shifman, “An anatomy of a YouTube meme”, New Media and Society, 2012, 14(2): pg. 187-203.
[15] S. Cunningham, D. Craig, and J. Silver. “YouTube, multichannel networks and the accelerated evolution of the new screen ecology”, Convergence, 22(4), 2016, pg. 376-391.
[16] D. Craig, and S. Cunningham, “Toy unboxing: living in a(n unregulated) material world”, Media International Australia, 163(1), 2017, pg. 77-86.
[17] D. Welbourne, and W. Grant, “Science communication on YouTube: Factors that affect channel and video popularity,” Public Understanding of Science, 2016, 25(6), pg.706-718.
[18] G. Chatzopoulou, C. Sheng, and M. Faloutsos, 2010, “A first step towards understanding popularity in YouTube,” In: INFOCOM IEEE Conference on Computer Communications Workshops, 2010 (pp. 1-6).
[19] Yates, J. and Orlikowski, W.J., 1992. Genres of organizational communication: A structurational approach to studying communication and media. *Academy of management review*, *17*(2), pp.299-326.
[20] Paolillo, J.C., Warren, J. and Kunz, B., 2007, January. Social network and genre emergence in amateur flash multimedia. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on* (pp. 70-70). IEEE.
[21] S. Fortunato, “Community detection in graphs”, Physics Reports, 486(3-5), 2010, pp. 75-174.
[22] J.L. Guillaume, R. Lambiotte, E. Lefebvre and V.D. Blondel, “Fast unfolding of communities in large networks”, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), pp. 10008.
[23] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media, 2009.
[24] G. Csardi and T. Nepusz, The igraph software package for complex network research, InterJournal, Complex Systems 1695. http://igraph.org, 2006.
[25] R Core Team, “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria.