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

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.

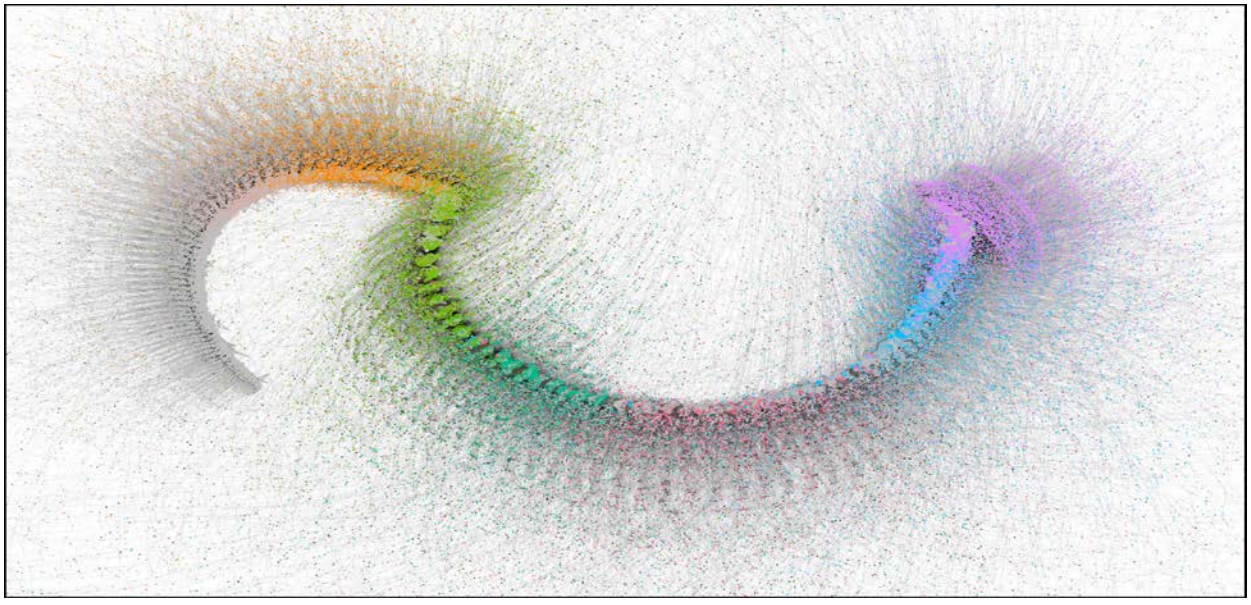

Figure 1. Final layout of network of shared membership in modularity classes of 141 sample networks, based on 3-month samples of YouTube channel-to-channel likes and favorites spaced at overlapping one-month intervals.

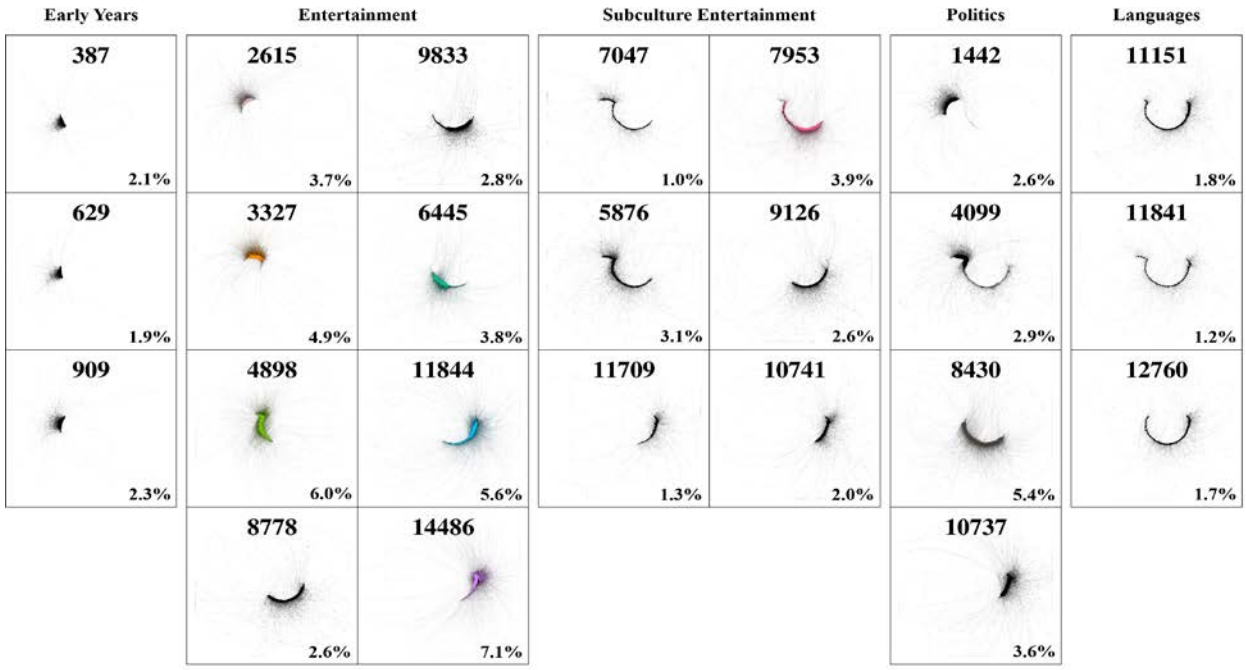

Figure 2. Clusters of modularity classes in the network of Figure 1, grouped in columns by type of content. Size of each cluster as percentage of modularity classes of the cluster comparison network is given in the lower right corner of each panel.

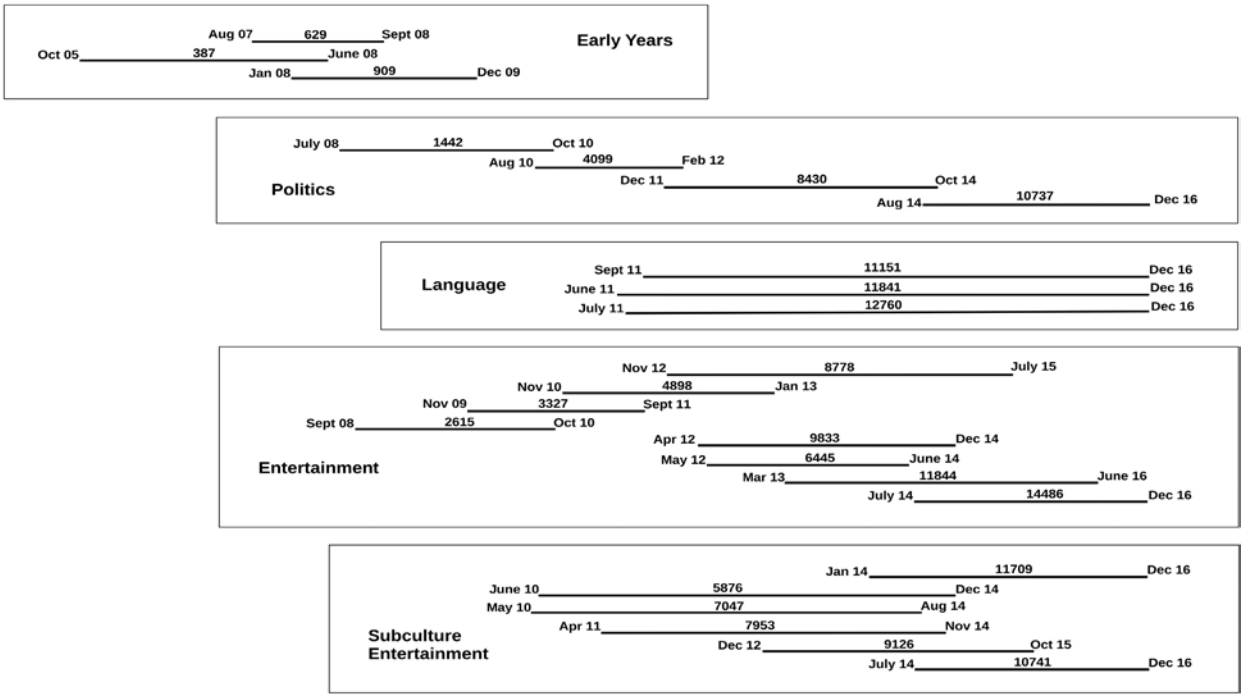

Figure 3. Temporal relationships of clusters in Figure 1, grouped by type of content. The left-right location and extent of a bar indicates the time period occupied by a cluster; clusters that overlap vertically occur at the same time.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.