Modeling Network Community Evolution in YouTube Comment Posting

Richard Allain[2], Ralucca Gera[1], Daniel Hall[2], Mark Raffetto[2]

[1] Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA
[2] Department of Operations Research, Naval Postgraduate School, Monterey, CA

Abstract. The present ready availability of time-tagged data enables new perspectives through which to study interactions within online social media. We empirically model the temporal development of both the largest component and largest community within the component. A power law relationship mapping community order and time explains a high percent of observed variance. The relationship is robust across a range of social media data sets with widely varying characteristics and suggests a deeper yet-underrepresented similarity between component and community.

Keywords: network topology, distance between graphs (similarity), network comparison metrics, distance in a network, communities, clustering, online social media, temporal analysis.

1 Introduction

Ever-improving information technologies continue to increase the level of interconnectedness and speed of messaging between otherwise-disparate people. Enabled by these advances, new methods of interaction emerge in the form of online social media. A better understanding of online social media may offer insights for many lines of study including sociology [24], recommendation systems [16], and targeted marketing and advertising [9].

Network science provides researchers a family of analytical tools to examine online social media. By representing the interactions within online social media as a network, the density of interpersonal interactions can be examined. Therefore the data can be studied at the microscale structure by looking at the nodes, at the macroscale structure through the analysis of the network overall, or at the mesoscale structure such as analysis of groups of nodes.

Individuals in a social network can be partitioned into groups by measuring the density of interactions with other individuals and aggregating those more tightly connected. A type of such groups can be defined as communities, which are densely connected individuals with sparse connections to the outside group members. These groups within online social media change in time both in membership and in relations between its members. Gaining insights into these online social media communities through a static model can thereby be problematic as we do not obtain the data on the relationships that appear or that degrade in

time. We focus our study on temporal online social media data to gain insights into community evolution in time, with a focus on the largest community.

Fortunato [6] mentions that the difficulties of community detection is in the agreement on the definition of communities in network. Radicchi [19] defined community as a general concept as a sub-network or subset of vertices with more internal edges to the community compared to edges between communities. This is the definition we will be using in the current research.

Most of the more popular community detection algorithms in network science are modularity based. Newman [14] uses the concepts of modularity to partition the network into communities and to measure the goodness of this partition by comparing it to a partition of a similar size (node and edge count) random network. That is, the modularity is the result of summing $a_{ij} - \frac{k_i k_j}{2m}$ for all pairs (i, j) in the same community.

Many modularity based algorithms for community detection exist; Louvain and Girvan-Newman methods are popular examples [2], [15]. While techniques to optimize these algorithms continue to grow, examining community growth and formation in time adds a new dimension to the study of online social media and the potential for additional insight. Current research on the temporal development of communities in real-world online social networks centers on the relationships and optimal techniques of identifying communities [5], [8], [10]. Several studies examining epidemics and financial trends focused on a temporal component within the dynamic network [1], [21]. These have been extended by evolutionary clustering and community detection approaches designed for dynamic social networks have been considered by [3], [4], [12], [25] and [23]. These papers analyze the evolution of communities, where communities are given by k-means, hierarchical, or spectral clustering.

The current study develops an empirical model identified with the change of social online networks in time, postulating a relationship between the membership of the largest modularity based community and the square root of elapsed time. This might be reminiscent of how the size of the Giant Component grows by the square root as edges are added to a fixed number of nodes [13]. However, our study also considers some graph data sets that represent change by varying both the number of vertices and edges over time.

The remaining of the paper is organized as follows. In Section 2 we introduce the data we have studied. In Section 3 we introduce our methodology, whose results and analysis on the three data sets are presented in Section 4. We then present the conclusion of our paper in Section 5.

2 The Data

The dynamic social network data studied herein was collected from YouTube video comment streams posted below videos of interest. Using netlytic.org applications, activity within the comment stream is modelled by a dynamic graph [7]. Vertices represent individual accounts posting comments within the video's comment stream (names). A directed edge is placed between two vertices when a

comment mentions another individual account within the post (name-mentions-name). Within the YouTube video comment stream, posts are persistent. Thereby, nodes and edges, once added, never disappear. Social interactions are forever ongoing once begun in this context.

We examined five disparate YouTube video comment streams to infer our mathematical model. Each data collection began and ended at different periods of time after the respective video was initially posted. The time elapsed between video posting and the initiation of data collection ranged from several minutes to several days. Additionally, we collected video comment streams containing different amounts of comments.

We examined five videos as shown below Case 1: SKRILLEX (SKX) video linked here $v = aAv8AtxuF8s\&lc$, Case 2: PewDiePie (PDP1) video also linked using $v = QA7W4XNXDzw$, Case 3: PewDiePie (PDP2) video linked using $v = xQ2WrglmsJk$, Case 4: Lady GagaVEVO (LGV) video linked using $v = 3aw\_sZvauCw$, and Case 5: RihannaVEVO (RIV) video linked using $v = SQWqksdE4gw$. A summary of the network information of the five data sets are shown in Table 1 for each data set.

Posted between 8 January and 24 February 2016, the videos originate from four different YouTube channels: SKRILLEX [22], PewDiePie [17], [18], Lady Gaga VEVO [11], and Rihannna VEVO [20].

Data set	Number of Nodes	Number of Edges	Clustering	Number of Time Slices
SKX	2049	3552	0.214	211
PDP(1)	2063	2426	0.088	151
PDP(2)	913	1068	0.092	212
LGV	483	659	0.145	173
RIV	2034	3297	0.205	1025

Table 1. Network Data Description

In two of the videos examined, namely SKX and PDP(1), new names and new name-mentions-name interactions are noted. In the other three of the videos considered, namely PDP(2), LGV and RIV, no new names are added, only new name-mentions-name interactions appear, an artifact of the data we collected.

The total length of time over which we collected the data for each video posting stream varied between 4 hours and 10 days. The significant variety in video topic, time period, and number of comments provides robustness to our empirical findings.

3 Mathematical Formulation

We examine $name - mentions - name$ directed social media networks. Within these networks, we represent the dynamic social interactions as a time sequence

of static graphs $G_1, G_2, ..., G_i$ where $G_k = (V_k, E_k)$ represents all recorded interactions up to and including time $k(1 \leq k \leq i)$. Our network interactions are thus persistent. Names are modelled by vertices and name-mentions-name interactions are modelled by directed edges. Using this sequence of graph data representations, we seek to gain insight into the evolution of the largest community within the largest component of dynamic social networks.

Once instantiated as dynamic graph data, we slice the data into a time-sequence of static graphs $G_1, G_2, ..., G_i$. All time slices within a sequence are of equal length, but vary in specific duration depending on the video examined. Slices are made to separate the data into the shortest intervals possible, given the precision of the data's time stamps. Time slices range from one second to one hour across the five videos analyzed.

Next, we isolate the largest component identified through multiplexing the graph data in time. It is important to reemphasize that because the comments are persistent once posted, the last time step examined is equivalent to a time-multiplex of the sequence of static graphs $G_1, G_2, ..., G_i$. We found that $20-30\%$ percent of the names included within the video comment stream made no mention of another name and thereby never participated in forming a discernible social community within the macro-community provided by the total membership of the video comment stream. These disconnected nodes were not used in further analysis.

Limited to the largest component, we measure n'_k, the order of the component in each period $k \geq 1$, and then apply each of the Louvain and Girvan-Newman static community detection algorithms to each graph G_k. After examining various characteristics of the resultant communities, we fit a mapping of 1) time period to the order of the largest component, n'_k; and 2) time period to the order of largest community, n''_k. Using the coefficient of determination (R^2) we assess the goodness of fit for each model. The result is a mathematical model of the largest component and largest community's evolution in time, respectively. We find that the models $n'_k \propto \sqrt{time}$ and $n''_k \propto \sqrt{time}$ explain greater than 90 percent of the variance in each of five data sets we examine.

4 Results and Analysis

In this section we present our results of the five data sets. We introduce a more detailed analysis of one case. We omit the others as they are similar, rather present the summary in Subsection 4.2.

4.1 Case 1: Illustrative Case

We first analyze data collected from the video comment stream of a music video, $v = aAv8AtxuF8s\&lc$, originally posted by an electronic music artist called Skrillex [22]. The video was posted to the Skrillex YouTube channel (SKX) on January 8, 2016 and as of March 4, 2016 received 3, 328, 879 views, 103, 041 likes and 9, 862 comments [22]. We examine a ten-day period of the video's comment

stream, including posts between 8 and 18 January 2016. The dataset collection began approximately four hours after the original posting of the video.

A dynamic graph with $2,050$ vertices and $3,550$ edges results from the comment stream. Multiplexing the graph in time, we extract the largest component. The component has 977 vertices and $2,446$ edges. In Figure 1 we show the SKX network and its largest component. Similar pictures are obtained for the other four networks.

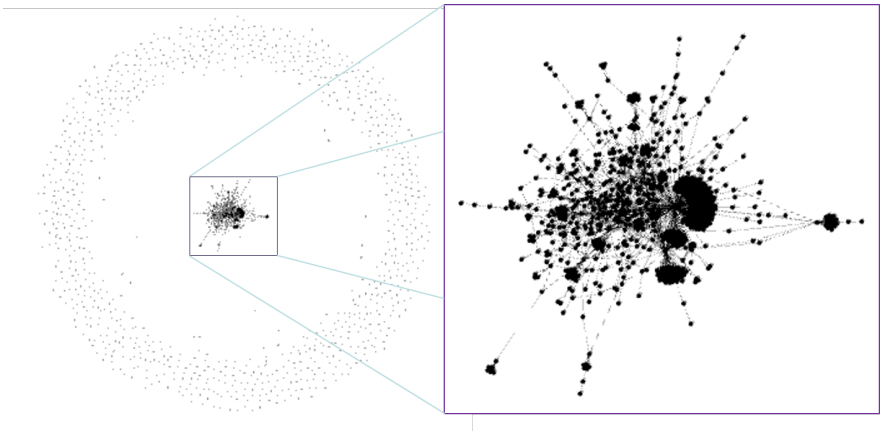


Fig. 1. The Network and the Largest Component in *SKRILLEX*

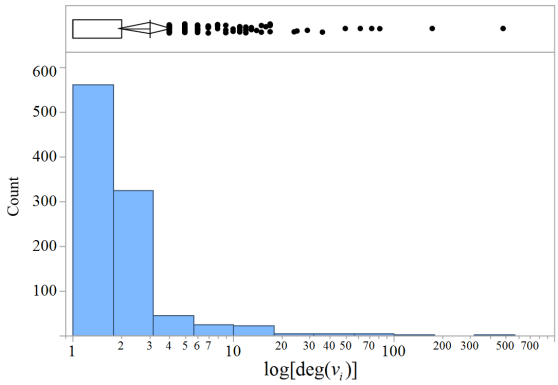
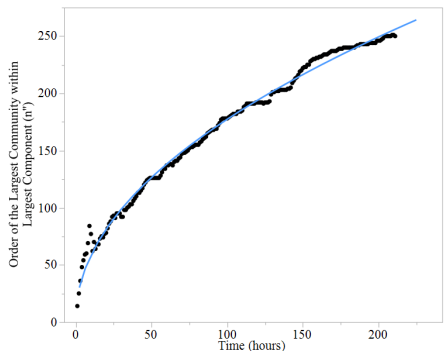


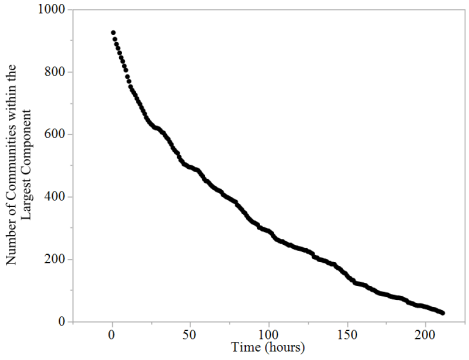
Fig. 2. The Degree Distribution of the Largest Component within *SKRILLEX*

Next, we divide the dynamic data for the largest component into 211 time-sequential static graphs $G_1, G_2, ...G_{211}$. Each graph G_i includes one hour of

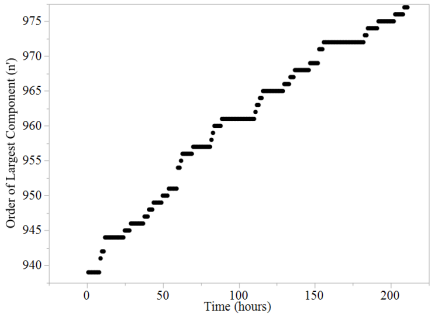
new posts and all past posts still present in the video comment stream. Both Louvain and Girvan-Newman static community detection algorithms are used to measure n_i'', the order of the largest community within the largest component in each static graph G_i. Using JMP software, we conduct a statistical analysis, fitting n_i'' against the time period i. The bivariate fit with the highest percent of variance explained is $n'' = c + a\sqrt{time}$ where $c = -14.433, a = 6.728$, and $time$ is the total elapsed time. This fit explains approximately 97.5% of the variance in n'', Figure 3(a). While this relationship might seem initially obvious, it became less so when we observed that the number of communities within the largest component changes from 925 at $i = 0$, to 27 at $i = 211$ as the communities coalesce and evolve, Figure 3(b-d).



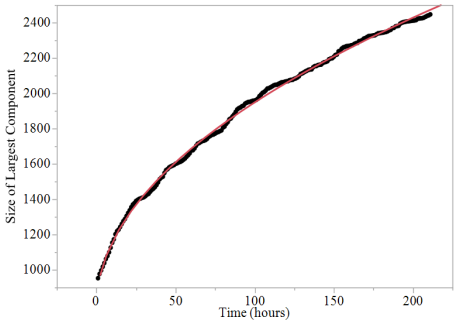
(a) Order of the Largest Community within the Largest Component (n'') over Time



(b) Number of Communities within the Largest Component over Time



(c) Order of the Largest Component over Time



(d) Size of the Largest Component over Time

Fig. 3. Illustrative Case: The Skrillex Network Attributes over Time

4.2 Additional Cases

Our examination includes data from four additional YouTube video comment streams as presented in Section 2. Our data sets contain between approximately 100 and 1000 posts within just the largest component of each data set's multiplexed network.

We apply the same statistical analysis and bivariate fit as described in the illustrative case. Our results are displayed in Table 2. The total number of communities within the largest component found in the last time slice is also displayed for careful consideration. In no case studied did the largest community include all vertices within the largest component. PewDiePie video 2 displayed only 18 communities in the last time slice, the smallest number of any case. In our data the membership of the largest community is always a subset of the largest component's membership.

Dataset	Order of Largest Component	Size of Largest Component	Number of Communities	c	a	$R^2(n'')$	$R^2(n')$
SKX	977	2446	27	3.492	17.380	0.99	0.98
PDP(1)	679	1459	37	-14.433	6.728	0.98	0.97
PDP(2)	200	376	18	-3.745	2.398	0.96	N/A
LGV	142	348	20	-9.103	4.219	0.99	N/A
RIV	892	2230	38	-1.434	3.622	0.93	N/A

Table 2. Comparison of n'' and n' model fit by case, with $n'' = c + a\sqrt{time}$. In cases PDP(2), LGV, and RIV, the number of vertices in the largest component, n', does not change during the period considered.

5 Conclusions and Further Studies

Our analysis suggests that a power law, the square root of elapsed time, models both the growth of the largest community and largest component in some types of online social media. This model consistently explained a high level of variance– greater than 93 percent–in all cases examined. These cases included social media videos that varied significantly in topic, time period, and size.

It is not lost on us that the discovery of a power law relationship is not necessarily unique. However, it is novel that we observe the largest community within the largest component growing by exactly the same power, without regard to the origin of the data. This power law model is robust, accurate even in data where the number of vertices remained constant. Such similarity across the range of online social media video sources examined and extreme change in community membership within each data set brings pause. We believe it provides the powerful suggestion that a deeper more fundamental characterization of graph data

might show a significant, yet-to-be defined relationship between the largest component in a graph and the largest community within that component. Perhaps an important phenomenon of social media evolution is under-represented by current graph-theoretic definitions and models. Careful consideration and study should be given to this suggestion.

Further analysis in other types of social media and across a larger number of data sets should be completed to continue exploring the extensibility of our claim.

6 Acknowledgments

The authors thank the DOD for partially sponsoring the current research.

References

[1] Marya Bazzi, Mason A Porter, Stacy Williams, Mark McDonald, Daniel J Fenn, and Sam D Howison. Community detection in temporal multilayer networks, with an application to correlation networks. Multiscale Modeling & Simulation, 14(1):1–41, 2016.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10):P10008, 2008.

[3] Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L Tseng. On evolutionary spectral clustering. ACM Transactions on Knowledge Discovery from Data (TKDD), 3(4):17, 2009.

[4] Tanja Falkowski, Jorg Bartelheimer, and Myra Spiliopoulou. Mining and visualizing the evolution of subgroups in social networks. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pages 52–58. IEEE Computer Society, 2006.

[5] Hossein Fani. Temporal formation and evolution of online communities. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, pages 717–717. ACM, 2016.

[6] Santo Fortunato and Claudio Castellano. Community structure in graphs. In Computational Complexity, pages 490–512. Springer, 2012.

[7] A. Gruzd. Netlytic: Software for Automated Text and Social Network Analysis. `http://netlytic.org`, January 2016. [Software Tool].

[8] Thanh Ho and Phuc Do. Discovering communities of users on social networks based on topic model combined with kohonen network. In Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on, pages 268–273. IEEE, 2015.

[9] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 137–146. ACM, 2003.

[10] David Burth Kurka, Alan Godoy, and Fernando J Von Zuben. Online social network analysis: A survey of research applications in computer science. arXiv preprint arXiv:1504.05655, 2015.

[11] LadyGagaVEVO. Lady Gaga- David Bowie Tribute by Lady Gaga From The 58th GRAMMYs. https://www.youtube.com/watch?v=3aw$_$sZvauCw, February 2016. [YouTube].

[12] Yu-Ru Lin, Yun Chi, Shenghuo Zhu, Hari Sundaram, and Belle L Tseng. Analyzing communities and their evolutions in dynamic social networks. ACM Transactions on Knowledge Discovery from Data (TKDD), 3(2):8, 2009.

[13] Mark EJ Newman. The structure and function of complex networks. SIAM review, 45(2):167–256, 2003.

[14] Mark EJ Newman. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23):8577–8582, 2006.

[15] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. Physical review E, 69(2):026113, 2004.

[16] Jordi Palau, Miquel Montaner, Beatriz López, and Josep Lluís De La Rosa. Collaboration analysis in recommender systems using social networks. In Cooperative Information Agents VIII, pages 137–151. Springer, 2004.

[17] PewDiePie. Just a Prank Bro (Animated). https://www.youtube.com/watch?v=xQ2WrglmsJk, February 2016. [YouTube].

[18] PewDiePie. PEDOBEAR (Gone Sexual). https://www.youtube.com/watch?v\=QA7W4XNXDzw, January 2016. [YouTube].

[19] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of the United States of America, 101(9):2658–2663, 2004.

[20] RihannaVEVO. Rihanna - Work - Live at The BRIT Awards 2016 ft. Drake. https://www.youtube.com/watch?v=SQWqksdE4gw, February 2016. [YouTube].

[21] Marta Sarzynska, Elizabeth A Leicht, Gerardo Chowell, and Mason A Porter. Null models for community detection in spatially embedded, temporal networks. Journal of Complex Networks, page cnv027, 2015.

[22] Skrillex. Torro Torro - Make A Move (Skrillex Remix). https://www.youtube.com/watch?v=aAv8AtxuF8s, January 2016. [YouTube].

[23] Chayant Tantipathananandh, Tanya Berger-Wolf, and David Kempe. A framework for community identification in dynamic social networks. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 717–726. ACM, 2007.

[24] Stanley Wasserman and Katherine Faust. Social network analysis: Methods and applications, volume 8. Cambridge university press, 1994.

[25] Kevin S Xu, Mark Kliger, and Alfred O Hero Iii. Adaptive evolutionary clustering. Data Mining and Knowledge Discovery, 28(2):304–336, 2014.