Detecting and Tracking Polarized Communities in Temporal Networks

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Joe Spaeth

Spring, 2021.

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

John Hott, Department of Computer Science

Detecting and Tracking Polarized Communities in Temporal Networks

Joseph Spaeth
jss5ha@virginia.edu
University of Virginia
Charlottesville, VA, USA

Abstract

Community detection in temporal networks is a challenging task. Particularly difficult is determining how communities form and when they can be detected. This paper attempts to identify groups tweeting about mask usage during the COVID-19 pandemic between March 1st and May 31st of 2020. Snapshot based graphs were created via Networkx and analyzed using Gephi to find communities across the whole time period, as well as during every week long time slice [1]. This analysis found mixed results, but was successful in examining how communities change structure over time.

Keywords: graph theory, dynamic networks, temporal networks, community detection, temporal graphs, dynamic graphs

ACM Reference Format:
Joseph Spaeth. 2021. Detecting and Tracking Polarized Communities in Temporal Networks. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Temporal, or Dynamic, Networks are vital to understanding any phenomenon happening on social media. However, the addition of time series information often comes at a cost, including but not limited to larger file sizes, increased difficulty in computing various graph metrics, and increased computing time for those graph metrics.

In this paper, we present an analysis of community formation in a subset of Chen's COVID-19 Twitter dataset, which only contains the tweets containing the word mask, tweeted between March 1, 2020 and May 31, 2020[4]. The goal of this analysis is to compare the efficacy of community detection algorithms in a larger temporal network to the "snapshot" (subgraphs representing a given time-step) graphs which compose it. In doing so we attempt to determine if community formation can be tracked temporally, and if large communities which develop late in a graph's lifespan can be found earlier in its development. Specifically, we aim to determine if both pro- and anti-mask communities can be found in this period corresponding with the beginning of the COVID-19 pandemic and early discussion around the adoption of masks as a public health measure, and if those communities look like the polarized political communities found in other work[7],[11].

Conference'17, July 2017, Washington, DC, USA
2021. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Related Work

Others have examined community detection algorithms in large networks, temporal community detection, and the analysis of graph structure. This work builds upon and relies on the results of this work, by applying those algorithms to this our temporal data set, and examining the structure of the communities within the created graphs.

2.1 Community Detection in Large Networks

Work has already been done in the field of identifying communities in large networks, of particular relevance is the Louvain method from Blondel et al [2]. Other researchers have done community detection using algorithms like INFOMAP or K-means[3],[10]. K-means clustering is not ideal for this application (as the user specifies the value of K), but work has been done to automate the process of determining the existence and memebers of communities in large networks. We elected to use The Louvain method over INFOMAP and K-means because of Louvain's compatibility with the visualization software we used and that we did not wish to pre-define the number of communities we expected to find respectively.

2.2 Community Detection in Temporal Networks

Several others have examined how to best detect communities within temporal networks. He and Chen propose a snapshot-based method that applies the Louvain algorithm at each time step, as well as processing information about the graph structure from the prior time step in order to find communities at any given time step[6]. This methodology preserves temporal information, as well as minimizes CPU overhead by only looking back a single time step when performing community detection. Additionally, there has also been work done using network embedding in high dimensional vector spaces. Algorithms like those used by Chhetri et al are a very different and effective means of detecting communities in temporal networks[5]. Additionally, Zhao et al and Li et al have examined how to track changes to communities over time, and how nodes transition in and out of communities over time[8],[14]. This work is quite helpful in understanding community development over time. This is because they provide mechanisms to track community development, and how individual nodes will can go from one community to another over time.

2.3 Community Structure

Recuero et al analyze the roles of specific users in political networks on Twitter, and present a specific structure in networks surrounding political events[11]. This structure is a highly dense and highly modular structure where one community represents one side of an issue, and the other community represents the opposite side. In this case, Recuero examines the trial of Luiz Silva and the pro- and anti-Silva communities which formed during the trial. Other useful work on community structure includes Himelboim et al's work on classifying Twitter Topic Networks[7]. Himelboim found recurring patterns in Twitter networks, including the previously discussed high density high modularity structure in polarized political topics. This methodology however, was only applied to static graphs, and not to any temporal data.

3 Methods

3.1 Data Collection

We collected data from a subset of Chen's COVID-19 Twitter dataset[4]. In order to create this subset, we hydrated all tweets between March and June (non-inclusive) using Twarc, and inserted them into a MySQL relational database[12]. It is of note, that there are two gaps in the dataset. It is missing tweets for March 2, 2020, as well as May, 14, 2020 from 7:00-8:00 AM. Once this was complete, we iterated through all of the tweets, storing each tweet containing the word mask in a separate table, as well as using NLTK to assign a sentiment score to each tweet[9].

3.2 Graph Creation

Three kinds of graphs were created from this data. The graphs we created are either snapshot graphs representing any specific week, or temporal graphs made of each individual snapshot.

- User to User: these graphs connect users when they are retweeted or mentioned by each other. In addition to tracking which users are interacting with which, each node and each edge track the overall sentiment of this user's posts. Node sentiment tracks the sentiment of each post that user makes, edge sentiment tracks the sentiment of each post between the users an edge connects. Both of these measures are calculated by taking the arithmetic mean of all relevant sentiment scores for a given node or edge.
- User to Hashtag: these graphs connect users to any hashtags they include in their tweets. They are bipartite graphs, which serve as a means of identifying popular hashtags as well as who they are popular among. Edges in these graphs also contain a field which keeps track of the number of times a user tweets a tweet contain a given hashtag.
- Hashtag to Hashtag (Hashtag Co-occurrence): these graphs map hashtags to other hashtags which appear in the same tweets. These are the simplest of the three types of graphs, as they only connect co-occurring hashtags. Edges in these graphs also contain a data field acting as edge weight which simply tracks the number of tweets between the two nodes. An example of this can be seen in Figure 1.

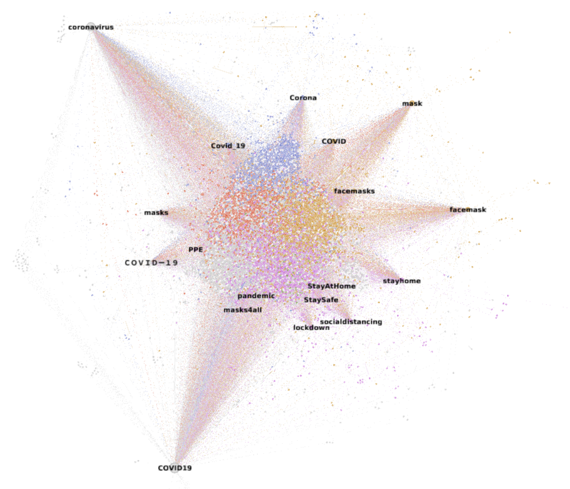

Figure 1. An example of a Hashtag Co-occurrence graph. This graph contains all communities across all time steps, and each community is demarcated with a different color.

These graphs were made using 3 python scripts. These scripts, as well as those used to create the database, can be accessed from this git repository. After making the graphs, the scripts then exported as `.gexf` files. The `.gexf` file format is an XML based format for encoding graph data used by Gephi[1]. Each individual `.gexf` file is a snapshot graph representing a given week. To turn the snapshots into temporal networks, each `.gexf` file's metadata was edited to indicate that they represented time-slice data and were applied the appropriate time stamp. Additionally, unconnected nodes in any of the graphs were removed when imported into Gephi[1]. This was to improve computation time when finding communities, as unconnected nodes will be sorted into their own communities, not substantially aiding analysis.

3.3 Graph Analysis and Visualization

Graphs were analyzed using the graph visualization utility Gephi[1]. It was used additionally to compute modularity measures, using Blondel et al's Louvain Method of community detection[2]. Communities found within the graph over the entire time period (March 1 through May 31) were then compared to the communities found at individual time slices. Those comparisons were primarily subjective, manually checking for significant overlap in the membership of any given community in both the relevant time slice graph, as well as the graph representing the entire time period. Community detection on the temporal graph was done by applying the Louvain algorithm on the graph across all time slices simultaneously.

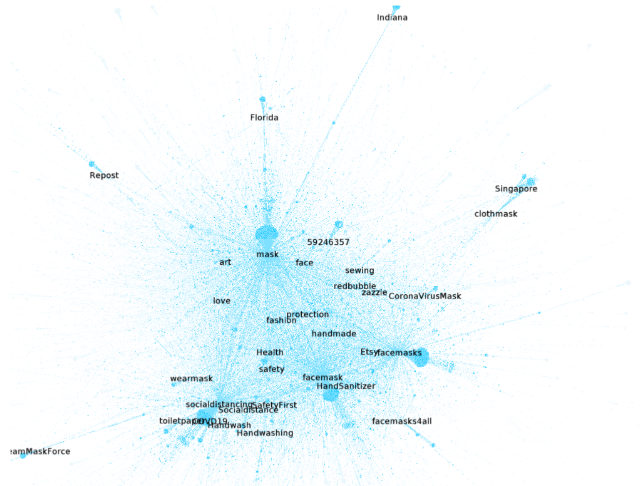

Figure 2. The mask selling community across all weeks

4 Results

It was found to be possible to find and track communities through time in the networks. The communities we expected to find we not always found however. Additionally, there were mixed results in finding expected structure within these networks.

4.1 Temporal Community Detection

When comparing communities at a given snapshot relative to the entire temporal network, we were able to find those communities existing both in the temporal network, as well as in individual time slices. Even when these communities were initially small and grew to a much larger size over time, they were still able to be found in individual snapshots. Figures 2, 3, 4, and 5 are an example of this. This shows 3 different versions of the same community from the temporal hashtag co-occurrence graph. This community shows a group of pro-mask hashtags and hashtags related to the sale and creation of masks. Figure 2 is the community across all time slices, Figure 3 is that same community at time slice 2, Figure 4 is that same community taken from the appropriate snapshot graph, and Figure 5 is that same community at time slice 9. The most interesting graph is Figure 4. It would seem that the co-occurring hashtags at this point in time are not as related to each other as the temporal graph would suggest, but upon closer inspection one will see that the relevant hashtags from Figures 2, 3, and 5 are all central in this graph. This is mostly a product of the graph being much smaller, and the community detection acting in a finer grained manner. Additionally, if one runs the Louvain algorithm on the subgraph in Figure 3 the results are similar to what is seen in Figure 4.

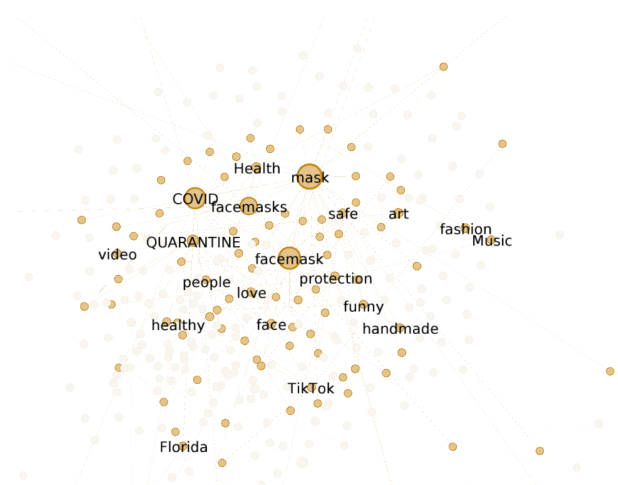

Figure 3. The mask selling graph at week 2. Many, though not all of the high degree nodes from the full community are present.

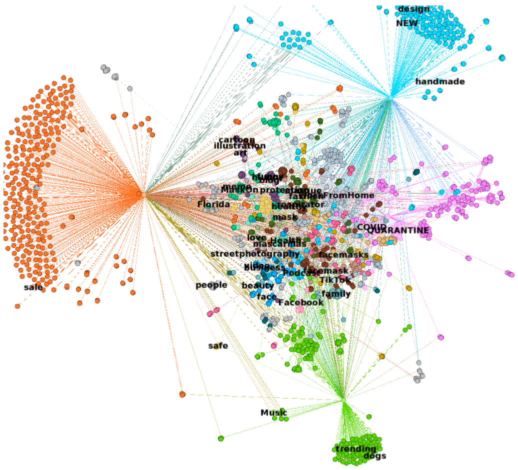

Figure 4. All communities at week 2. Communities are demarcated by color.

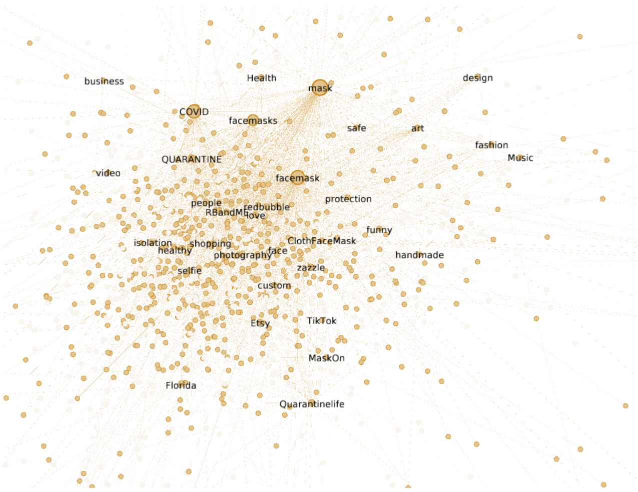

Figure 5. The mask selling graph at week 9. Most of the high degree nodes from the full community are present.

This is one example, among many, of communities which can be tracked temporally.

4.2 Expected Communities

Though many communities were found representing different groups, this analysis was not able to lead to the detection of both pro- and anti-mask communities within these networks. In the hashtag co-occurrence graphs as well as the hashtag to user graphs, multiple pro-mask communities can be found. However, no explicitly anti-mask communities were found. In the user to user network, the largest communities detected were a group centering around Indian news sources, a large community defined by users retweeting an American political figure (we cannot identify whom in this paper as it is a violation of the Twitter terms of service), a mix of liberal news sources and figures (a large portion of whom were tweeting pro-mask content at the time), a center right UK news network and UK citizens, and a group of African musicians. These communities can be seen in Figure 6. Within that figure, the pink nodes correspond to the liberal news community, the cyan nodes are the Indian news community, the tightly clustered turquoise community is the group of African musicians, and the tightly clustered orange community is the retweet community of the previously mentioned political figure.

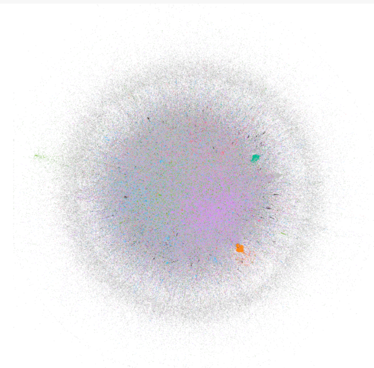

Figure 6. The user to user graph over the entire time period.

Though the results in the user to user graph were not particularly relevant in finding pro- and anti-mask communities, it did result in finding clearly defined communities surrounding particular topics. The hashtag to user and hashtag co-occurrence graphs however, were more fruitful in detecting pro-mask communities. Both of these networks found two pro-mask communities, one centered around the sale of masks, and another centered around the CDC, WHO, and other organizations promoting them as a public health measure. There were no explicitly anti-mask communities found in either of these networks using our methodology. There were two politically conservative communities centered around different topics. Within those networks are also communities primarily centered around hard news. They have no political leaning, and primarily are users interacting with far larger news networks. There is also a large group of EU related hashtags and users in those networks, made insular because they are primarily tweeting about their own countries.

4.3 Community Structure

Though the communities we expected to see were not found, our methodology was able to find politically polarized communities. These polarized communities can be seen in Figure 7, and are represented with a hashtag to user graph, which is analogous to the Twitter Topic Networks discussed by Himelboim[7]. This figure shows four communities, two of which are pro-mask and politically liberal, the other two are not anti-mask, but are both conservative. The pro-mask communities are represented by the blue and orange nodes. They have considerable crossover and are dense, but highly modular relative to the two other communities. The green and black communities are both also dense and modular, but have little overlap. This has to do with the subject matter of those communities, green is a mix of right wing news content from India and the United States, and black is a collection of anti-China/pro-Hong Kong hashtags and users. This shows a clear split along ideological lines amongst twitter users at this time, as well as addressing why the right-wing groups found do not map onto each other as cleanly as the liberal groups.

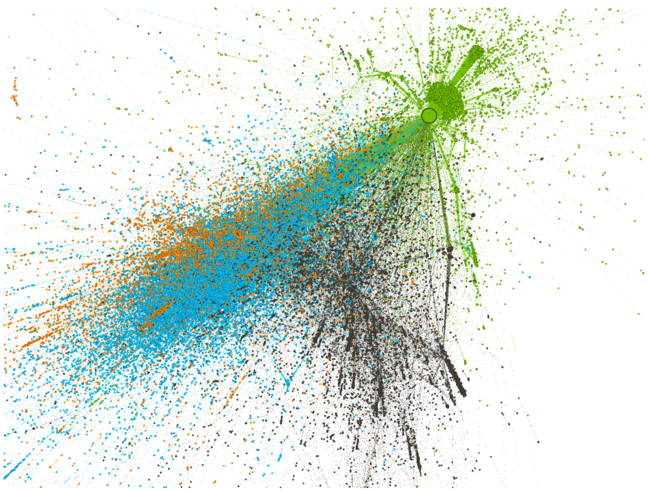

Figure 7. This graph contains 4 of the largest communities in the full temporal hashtag to user network.

Additionally, there is temporal behavior within the structure of these communities. Of particular note is the anti-China community. The anti-China community does not appear in the modular manner it does in the full time graph until week 4, becomes its largest at week 6, and then dissipates after week 8. Particularly, at this point the hashtag "Taiwan" and the users tweeting about it created significantly less overlap between it and the other communities at earlier times. Similarly, as time goes on, the community fades away almost entirely. By week 11 there are few nodes in this community still present in the graph. Figure 8 illustrates the temporal development of this community.

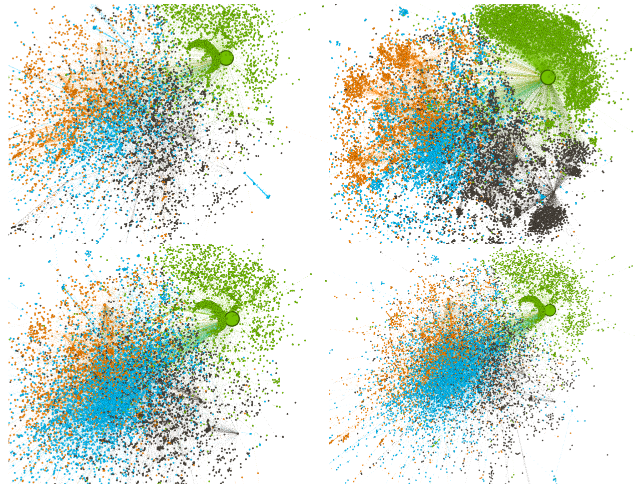

Figure 8. This image shows the development of 4 communities at weeks 3, 6, 8 and 11.

5 Further Study

There are a few avenues for further study, these include, but are not limited to, applying finer grained time-slices, applying this analysis to larger datasets (including more tweets and or analyzing a larger time period), changing the representation of the temporal networks, and applying different community detection algorithms. Finer grained time slices would potentially allow for the detection more better defined temporal behavior. Because the time slices applied are a week long, though temporal behavior is possible to notice, it does bring into question if smaller time slices would have helped to identify any additional time-based phenomena in the data. Larger datasets would potentially be able to find some of the communities we had expected to find, and the issues with not finding anti-mask communities could potentially be addressed by not looking at a subset of the Chen's dataset, but instead the entirety of it[4]. Similarly, examining tweets past the month of May could also address this issue.

Additionally different methods of temporal community detection could find different community structures, as well as rely more on the temporal aspects of the data being analyzed. This can either be done by using embedding techniques to make the graphs and perform analysis, or simply using different community detection algorithms such as the PCA community detection approaches, or simply iterating on Blondel's Louvain method[2],[13]. Any community detection measure which could be used to better represent and analyze the temporal behavior of these networks is vital to further study in this area. Particularly, I believe that He and Chen's method would be helpful to this analysis, and went unused due to limitations of the software library used to create the graphs.

6 Acknowledgements

This work would not have been possible without the help of my advisor, Robbie Hott. He has helped with this research since my third year, and I am immensely grateful to him for all of the help and ideas he has given me through the last summer and last two semesters. This research also would not have been possible without institutional support from the University of Virginia, as I would not have been able to build the database and run the scripts I've written for hours on end without their computing resources. Finally, I would like to explicitly thank Emily Chen, Kristina Lerman, and Emillio Ferrara, for making their COVID tweet dataset publicly available for the use of other researchers.

References

[1] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media* 3, 1 (Mar. 2009). https://ojs.aaai.org/index.php/ICWSM/article/view/13937

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (oct 2008), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

[3] Ludvig Bohlin, Daniel Edler, Andrea Lancichinetti, and Martin Rosvall. 2014. Community Detection and Visualization of Networks with the Map Equation Framework. In *Measuring Scholarly Impact*. Springer International Publishing, 3–34. https://doi.org/10.1007/978-3-319-10377-8_1

[4] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance* 6, 2 (2020), e19273.

[5] Sujit Rokka Chhetri, Palash Goyal, and Arquimedes Canedo. 2019. Tracking Temporal Evolution of Graphs using Non-Timestamped Data. (July 2019). arXiv:1907.02222 [cs.SI]

[6] Jialin He and Duanbing Chen. 2015. A fast algorithm for community detection in temporal network. *Physica A: Statistical Mechanics and its Applications* 429 (jul 2015), 87–94. https://doi.org/10.1016/j.physa.2015.02.069

[7] Itai Himelboim, Marc A. Smith, Lee Rainie, Ben Shneiderman, and Camila Espina. 2017. Classifying Twitter Topic-Networks Using Social Network Analysis. *Social Media Society* 3, 1 (jan 2017), 205630511769154. https://doi.org/10.1177/2056305117691545

[8] Tianpeng Li, Wenjun Wang, Xuxun Wu, Huaming Wu, Pengfei Jiao, and Yandong Yu. 2020. Exploring the transition behavior of nodes in temporal networks based on dynamic community detection. *Future Generation Computer Systems* 107 (jun 2020), 458–468. https://doi.org/10.1016/j.future.2020.02.006

[9] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. (May 2002). arXiv:cs/0205028 [cs.CL]

[10] Shie Mannor, Xin Jin, Jiawei Han, Xin Jin, Jiawei Han, Xin Jin, Jiawei Han, and Xinhua Zhang. 2011. K-Means Clustering. In *Encyclopedia of Machine Learning*. Springer US, 563–564. https://doi.org/10.1007/978-0-387-30164-8_425

[11] Raquel Recuero, Gabriela Zago, and Felipe Soares. 2019. Using Social Network Analysis and Social Capital to Identify User Roles on Polarized Political Conversations on Twitter. *Social Media Society* 5, 2 (apr 2019), 205630511984874. https://doi.org/10.1177/2056305119848745

[12] Ed Summers, Hugo Van Kemenade, Peter Binkley, Nick Ruest, Recrm, Stefano Costa, Eric Phetteplace, The Gitter Badger, Mx A. Matienzo, Lukas Blakk, Dan Chudnov, and Chad Nelson. 2015. Twarc: V0.3.4. https://doi.org/10.5281/ZENODO.31919

[13] Joseph H. Tien, Marisa C. Eisenberg, Sarah T. Cherng, and Mason A. Porter. 2020. Online reactions to the 2017 'Unite the right' rally in Charlottesville: measuring polarization in Twitter networks using media followership. *Applied Network Science* 5, 1 (jan 2020). https://doi.org/10.1007/s41109-019-0223-3

[14] Zhongying Zhao, Chao Li, Xuejian Zhang, Francisco Chiclana, and Enrique Herrera Viedma. 2019. An incremental method to detect communities in dynamic evolving social networks. *Knowledge-Based Systems* 163 (jan 2019), 404–415. https://doi.org/10.1016/j.knosys.2018.09.002