2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10021128

Knowledge Graphs of the QAnon Twitter Network

Clay Adams[*††], Malvina Bozhidarova[†††], James Chen[‡††], Andrew Gao[§††], Zhengtong Liu[‡††],
J. Hunter Priniski[¶], Junyuan Lin[‖], Rishi Sonthalia[‡],
Andrea L. Bertozzi[‡] and P. Jeffrey Brantingham[**]

[*]Department of Mathematics, Harvey Mudd College, Claremont, CA, USA
[†]School of Mathematical Sciences, The University of Nottingham, University Park, Nottingham, UK
[‡]Department of Mathematics, University of California, Los Angeles, Los Angeles, CA, USA
[§]Department of Mathematics Stanford University, Stanford, CA, USA
[¶]Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA
[‖]Department of Mathematics, Loyola Marymount University, Los Angeles, CA, USA
[**]Department of Anthropology, University of California, Los Angeles, Los Angeles, CA, USA
[††] Equal Contribution

***Abstract*—Using Knowledge Graphs to understand noisy naturalistic data has gained significant prominence in recent years. In this paper, we apply Knowledge Graphs to a new dataset of tweets of an ideologically far-right Twitter network by sourcing tweet histories of users who discussed QAnon in the summer of 2018 [1]. We further develop a new method that arms topic models with relational information from Knowledge Graphs and apply the new technique to study this dataset. Our analysis shows that users do not form a monolithic belief or social network, but rather comprise many smaller interlinking communities which discuss unique key political events (e.g., the January 6th Capitol riots).**

I. Introduction

Knowledge Graphs are useful tools for representing and structuring information in large corpora of text. They are semantically rich, with nodes representing entities extracted from text documents and edges encoding how entities are related. Knowledge Graphs have been successfully applied to social media data such as tweets [2] and longer-formatted narrative structures such as scripts [3]. Well-known examples include the Google Knowledge Graph [4], the LinkedIn Knowledge Graph [5], and DBpedia [6], which have been successfully applied in practice.

In this paper, we couple Knowledge Graphs with social network analyses to understand the semantic content and social network structure of the QAnon conspiracy on Twitter. The QAnon conspiracy, which originated with postings on the anonymous message board 4chan in 2017 and gained widespread popularity in 2018, posits that a global cabal of liberal elites run a covert sex-trafficking ring that former President Donald Trump is destined to uncover and annihilate [1], [7], [8]. While QAnon's central narrative is radical, perhaps more problematic are the far-reaching peripheral narratives, which can have the effect of sowing active distrust in democratic processes [9] and in science [10]. Our work aims to improve the understanding of these interlocking narratives in relation to how they vary across a social network (e.g., how Knowledge Graphs vary across communities). We are also interested in how the network discusses key political entities and events (e.g., Donald Trump during the January 6th Capitol riots). We analyze data pulled from tweet histories of users previously identified in [1] to have tweeted about QAnon in 2018. We analyze their metadata by constructing Knowledge Graphs using topic modeling, sentiment analysis, time series analysis, and community detection. This process lets us determine topics (and their sentiments) of interest to the conspiratorial network, as well as the influence of key political events such as the election and the pandemic.

We find that the users do not form a monolithic super-cluster with identical beliefs, but rather involve networks of smaller communities with varying beliefs that discuss different events. We also observe that vaccines are discussed in close proximity to political entities which hold increased negative valence, suggesting that medical information about the COVID-19 vaccine may be associated with politicized information. This is a hallmark feature of anti-science misinformation, that likely targets pre-existing priors of distrust to shape COVID-19 skepticism [11]. Different communities discussing different political topics (e.g., conservative vs liberal politicians, news outlets) appear at different times and places in the network. We also see a spike of tweets following the January 6th Capitol riots, with various topics discussing the events, suggesting that these methods can be applied to study how real-world events trigger shifts in the discussion.

Our analysis has three major steps. The first step is to understand the data prior to constructing a Knowledge Graph. We analyze the geo-locations of the users, a hashtag co-occurrence network, and the volume of tweets over time. We present the results in Section II. Next, to better understand how topics are discussed across the social network in relation to the Knowledge Graph, we apply topic modeling and community

Email: ccadams@hmc.edu, drewgao@stanford.edu, malvina.bozhidarova@nottingham.ac.uk, junyuan.lin@lmu.edu, { jamesdc1, ericliu2023, priniski, rsonthal, bertozzi, branting}@ucla.edu. This work was partially supported by NSF grant DMS-2027277, and AFOSR MURI grant FA9550-22-1-0380. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Air Force Office of Scientific Research.

detection. This is detailed in Section III. Because of the noisy nature of the data, traditional topic modeling is virtually impossible. In the last section IV, we develope *KG-labeled topic modeling*, a novel method for producing labeled topics with Knowledge Graph relations. The *KG-labeled topic modeling* is better suited for uncovering topics in this extremely noisy dataset. We use this method to understand how different key events are discussed in communities across the social network.

Our paper contains three major contributions[1]:

1) a novel dataset of the relevant tweets of before and after the January 6th Capitol riots;
2) application of Knowledge Graphs to produce KG-labeled topics;
3) analysis of the dataset to get new insights into the community of Twitter users that discuss QAnon.

II. Data Collection and Visualization

A. Data Collection

Data collection followed a two-step process. In June 2018, [1] collected 800k tweets from 100k users by collecting tweets that contained at least one of the following strings: “#q”, “#qanon”, “qanon”. Then on August, 3rd 2021, we sampled the 200 most recent tweets for these users. The resulting data has $\sim 900k$ tweets from $\sim 10k$ users[2]. Since not every user had 200 tweets, we note the data has a median of 67 tweets per user and a mean of ~ 80 tweets per user with a minimum of 2 tweets per user. Our data, however, notably omits some important network properties and metadata, such as follower networks, like, retweet, and comment data.

B. Geolocation of Users

We explored the tweet metadata structure to get a sense of what type of analyses to run. We obtained a complete data set of the full tweet histories of the users, instead of just the most recent 200 tweets, and extracted the geo-location data using the geopy API [12] to convert from text to longitude/latitude pairs. The extracted geo-location data is plotted on the world map (Fig. 1).

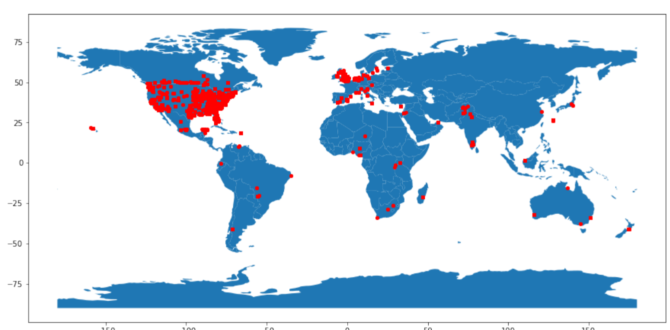


Fig. 1: Map of user locations. There is a concentration of users in the south of the United States with a significant amount of users located outside the US in our dataset.

[1]https://github.com/rsonthal/kg-twitter-reu

[2]Note the other 90k users no longer had Twitter accounts

C. Hashtag Co-occurrence

Hashtags allow users to label their tweets with a searchable tag. Oftentimes tweets also contain multiple hashtags. Thus, studying the co-occurrences of these hashtags can also illuminate how various topics are discussed, which can lead to follow-up analyses examining how ideas are conceptually linked as users either click between the co-occurring hashtags or by simply relating key semantic content (e.g., ‘Was forced to get the COVID-19 jab and now I’ve been ill for 3 months #Trump2020’).

We first computed the number of co-occurrences for all hashtags across 105747 tweets that utilize at least one hashtag, and we only consider hashtags with more than 100 uses in our data set to compute the Jaccard Index $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, where A and B are the sets of tweets that contain the two hashtags respectively. We then construct a network with the hashtags as nodes with an edge between pairs of nodes if they have a Jaccard Index of over 0.1. Edge weights are set equal to the Jaccard Index. We apply the Louvain algorithm [13], [14] to detect the communities and proxying topics, and we label the Louvain hashtag clusters with the hashtag having the highest closeness and betweenness centrality in the community [15]–[17]. The closeness centrality is defined as

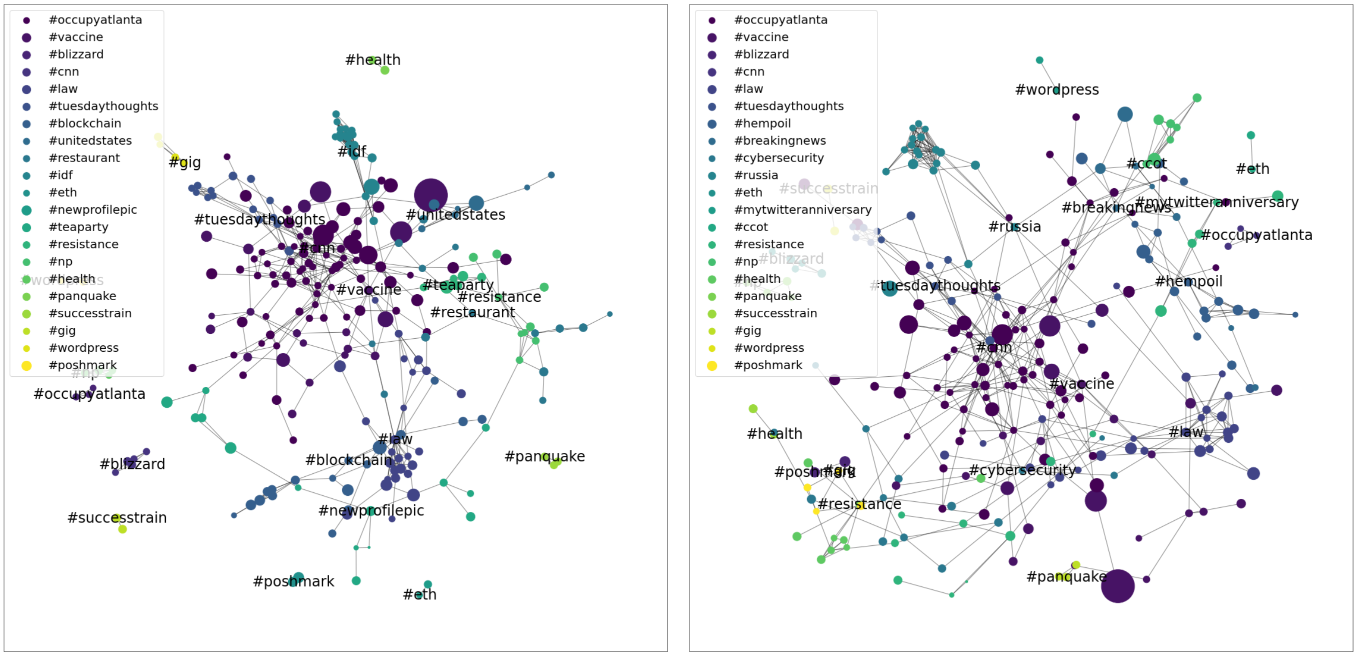
$$C(u) = \frac{k-1}{\sum_{v=1}^{k-1} \sigma(v,u)},$$

where k is the number of nodes within the community, and $\sigma(v,u)$ is the shortest path distance between two distinct nodes v and u in the node-set. The betweenness centrality is

$$C_B(v) = \sum_{\substack{s,t \in V \\ s \neq v \neq t}} \frac{\sigma(s,t \mid v)}{\sigma(s,t)},$$

where V is the set of nodes in the whole graph, and $\sigma(\cdot,\cdot \mid v)$ is the shortest path that passes through v.

Fig. 2 shows the resulting graph labeled with the highest closeness centrality and betweenness centrality in each community. While the hashtag with the highest betweenness and closeness centrality scores tend to be the same for each community, there are cases where our two notions of most central node differ, notably the community in which node #teaparty has the highest closeness centrality and #ccot (an acronym for Conservative Christians of Twitter) is the highest betweenness centrality node. In these cases, we observe that the highest closeness centrality hashtag is associated with a particular political movement whereas the highest betweenness centrality hashtag is broader and more likely to reach a wider audience. We notice that there is one large connected component consisting of politically-valenced hashtags, with smaller connected components to commericial activity such as selling clothing on #poshmark or discussing the Ethereum cryptography using #eth. These exploratory findings suggest that, as one would expect, political ideology is intertwined in a variety of topics in the network.



(a) Closeness Centrality

(b) Betweenness Centrality

Fig. 2: **(a)** Hashtag co-occurrence graphs with communities labeled by highest closeness centrality, and **(b)** communities labeled by highest betweenness centrality. The node size indicates the frequencies of each hashtag, the edges represent the co-occurrence relations, and the colors distinguish between different communities.

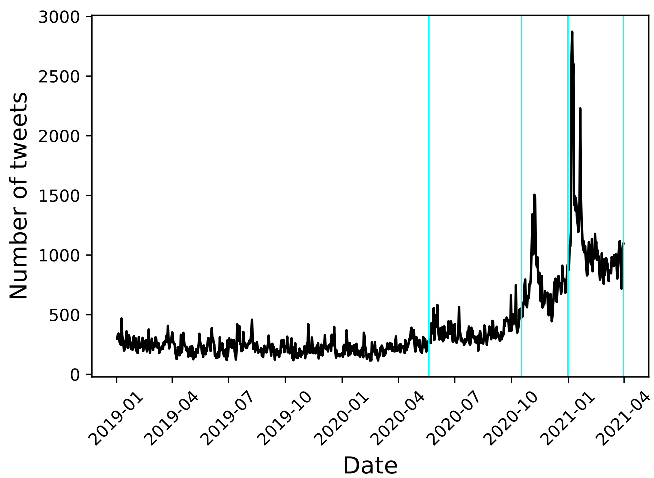


Fig. 3: The distribution of the number of tweets from Jan 2019 to April 2021 (in black) and the detected change points.

D. Change-point Detection

To better understand tweet dynamics, we ran change point detection on daily tweet volumes. Change point detection locates which time stamps in a data set exhibit statistically meaningful change. To detect change points, we use the Pruned Exact Linear Time (PELT) algorithm [18], which finds a given number p, as input, of change points $\tau_1, .., \tau_p$, by minimizing a cost function which measures changes.

Because the data was collected by backtracking the 200 most recent tweets of a set of users, the volume of tweets is imbalanced with respect to time. Therefore, we only consider the tweets from 1st January 2019 to 1st April 2021, a period in the middle of the whole time span where the volume of tweets is relatively steady over time. Fig. 3 illustrates the change points that the algorithm returns for this time period. The first change point is at the end of May 2020, which aligns with the murder of George Floyd on May 25, 2020. The second change point is around November 2020, aligning with the 2020 U.S. Presidential Election. The third change point, where we can see the highest peak in the number of tweets, is at the beginning of January 2021, corresponding to the January 6th Capitol riot. Previous work found that the police killing of George Floyd triggered similar spikes of discussion in left-leaning communities [19]. Future work could compare these two networks at this time point and see how discussion topic dynamics vary.

III. Knowledge Graph Construction

The Knowledge Graph has a two-fold structure: (1) a tweet metadata network augmented with features from topic models and sentiment analysis, (2) a social network aggregator that associates relations to the user. The general structure of the Knowledge Graph can be seen in Fig. 4. In this section, we detail the steps involved and the insights obtained by the Knowledge Graph construction.

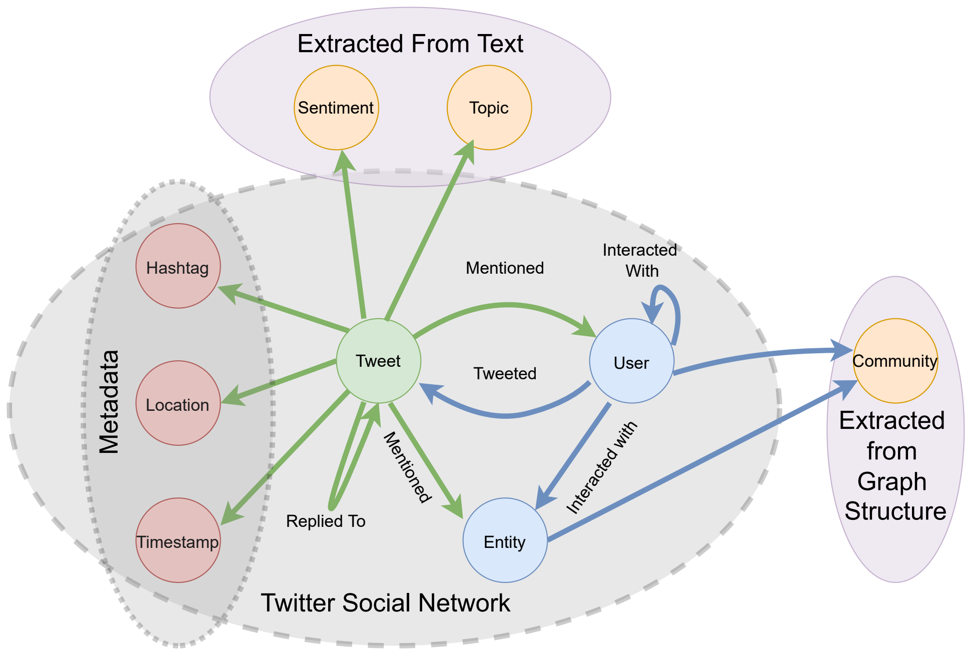


Fig. 4: Figure showing the Knowledge Graph structure. The location, hashtag, and timestamp are extracted directly from the tweet metadata. The topic and sentiment nodes are constructed from topic modeling and sentiment analysis. At the user level, the *hashtagged* and *mentioned* relations are aggregated to the *interacted_with* relation. Twitter accounts that are only involved in the *interacted_with* relation in the dataset are classified as entities. Also, users and entities are separated into communities on the social network.

A. Topic Modeling and Sentiment Analysis

First, we used topic modeling to determine which topics are of interest in our cleaned dataset. The large corpus of the QAnon tweets data set and the relatively short length of a tweet make it natural to choose a topic modeling method that clusters in the semantic space. Before analyzing the tweets, we clean them as follows:

1) Clean the text by removing the URLs, emojis, and other non-alphabet symbols. This text pre-processing procedure preserves the original structure of each sentence.
2) Filter the users and tweets to remove bots from the dataset. The procedure for this is described at the end of the section.

Next, BERTopic [20], a semantic-based topic modeling method, is employed to extract 50 topics from the dataset. Here, due to computational limitations, we only use 90% of the tweets chosen uniformly randomly. The BERTopic model first embeds the documents using the BERT model [21]. Next, it uses UMAP [22] to reduce the dimensionality of the embeddings. Then it clusters the tweets using DBSCAN algorithm [23], [24], which is a density-based hierarchical clustering algorithm. Finally, it creates the representations of each cluster based on the class-based TF-IDF (c-IF-IDF). The hierarchical clustering algorithm allows for straightforward visualization of the distance between different clusters in the embedding space.

Fig. 5 shows a subset of the topics extracted along with the hierarchical structure of the topics. The largest topic in terms of the number of tweets from BERTopic has “the election”, “voter fraud”, and “to vote” as its top three keywords. Hence we see that a large portion of the users are discussing voter fraud. We refer to this topic as the topic `election` in the following discussions. Hierarchical clustering also lets us determine which topics are closely related to other topics. As we can see from Fig. 5 the `election` topic is close to topics related to the Democratic and Republican parties on the hierarchical tree, and they both merge with a topic with representative words including “the media,” “fake news,” and “president trump.” Each topic is then added to the Knowledge Graph as a node. Topic nodes are connected to the user and tweet nodes. A tweet node is connected to a topic node via a *in_topic* relation if it was labeled with that topic and a user was connected to a topic node via a *discussed* relation if the user had a tweet that was labeled with that topic. Note due to the amount of noise in the dataset only 18.4% of tweets have topic labels.

Secondly, we perform sentiment analysis on the dataset. The sentiment is given by the sentiment score $i \in [-1, 1]$, where $|i|$ stands for sentiment intensity and the sign indicates positive or negative emotion. We assign the sentiment score of each tweet’s original text using [25]. Using this we can calculate the variance in the user’s sentiment. The sentiment is then added

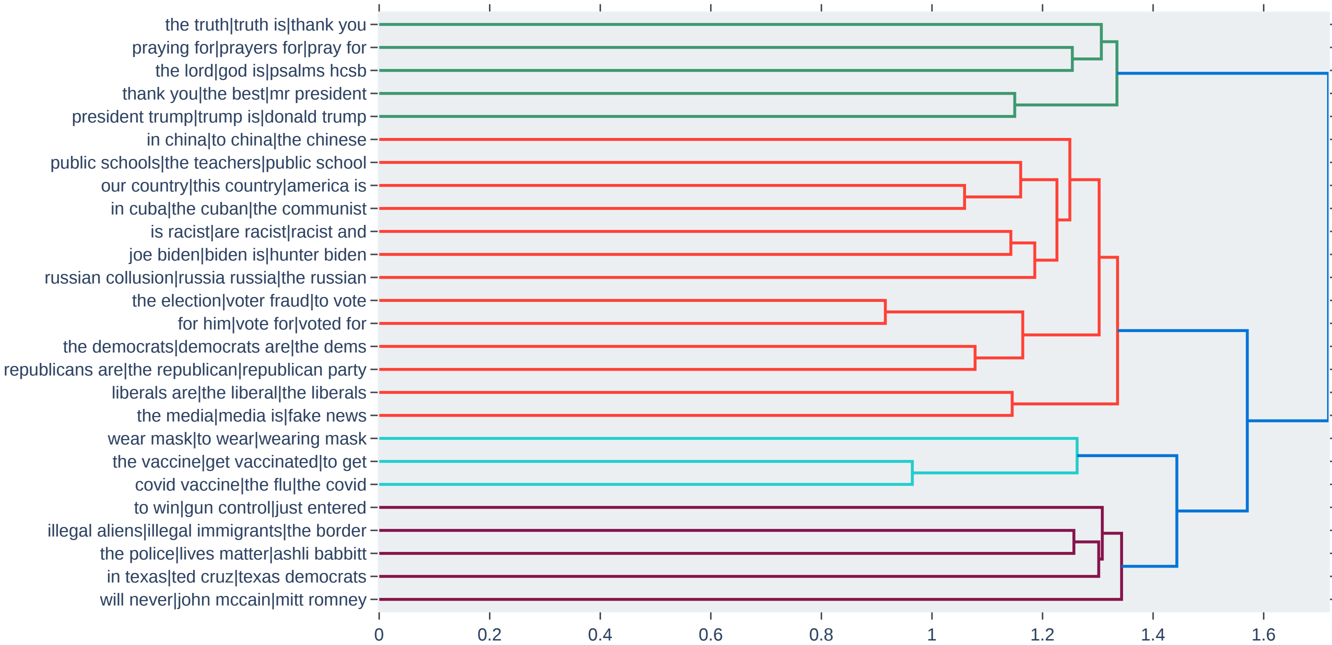


Fig. 5: Hierarchical clustering of topics from BERTopic (selected topics). Each topic is labeled with the top 3 two-grams or three-grams based on the c-TF-IDF. The x-axis shows the cosine distance between clusters.

to Knowledge Graph as a node and is connected to tweet nodes via the *feels* relation.

Bot Detection. We detect bots by looking at two statistics. The first is the sentiment variance for a user. Additionally, the tweets are embedded using the sentence-transformer model Sentence-T5 [26]. We then concatenate the embedding vectors of tweets for a user into a matrix Y and then calculated the order 1 coefficient of determination:

$$R_1^2 := \frac{\|\hat{Y}_1\|_F^2}{\|Y\|_F^2}.$$

Here $\hat{Y}_1$ is the best rank 1 approximation of Y. Note that the smaller R_1^2 the more diverse the tweets are.

We then cluster the tweets based on the two features using KMeans [27] and Gaussian Mixture Model and took the intersection of the smallest clusterings of these methods. The 884 bot users identified in this process have a mean sentiment variance of 8.62×10^{-7} and a mean sentence similarity of 0.969, compared to 0.147 and 0.772 of all users in the dataset. We removed the bot users in topic modeling and when building the user-level social network.

B. Dynamic Community Detection

To better understand the dynamics of the social network, we examine snapshots of the network at the end of each month from January 2020 to July 2021. The time difference is chosen such that the snapshot networks are feasible to analyze. We used the Louvain algorithm [13], [14] to detect communities at each timestamp and the Jaccard Index to track the communities over time. As Fig. 6 shows, there is one large community that persists throughout the studied period, while many smaller communities emerge near the end of the studied period. It is also interesting that one community splits into two at a certain time, and one branch grows larger while the other gradually dies.

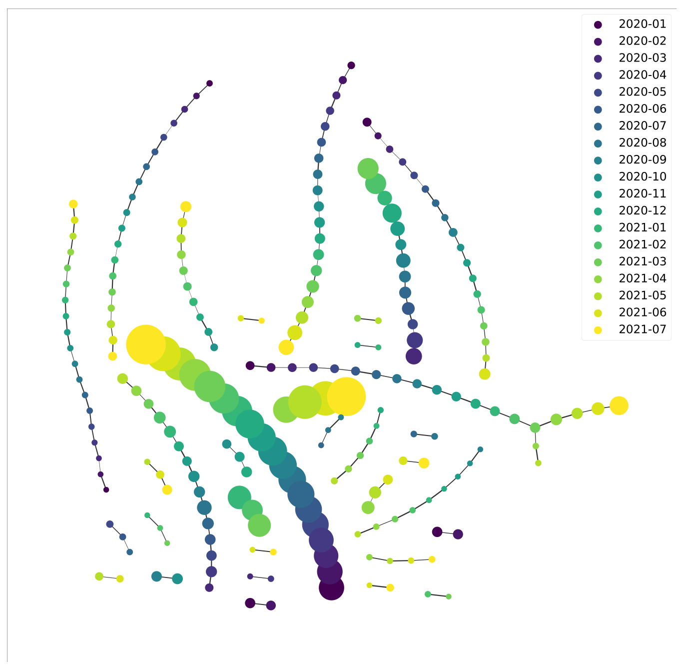


Fig. 6: Evolution of communities in the dynamic social network. Here the node size indicates the size of a community, and the color of a node indicates the time.

We define central entities as Twitter accounts tagged in

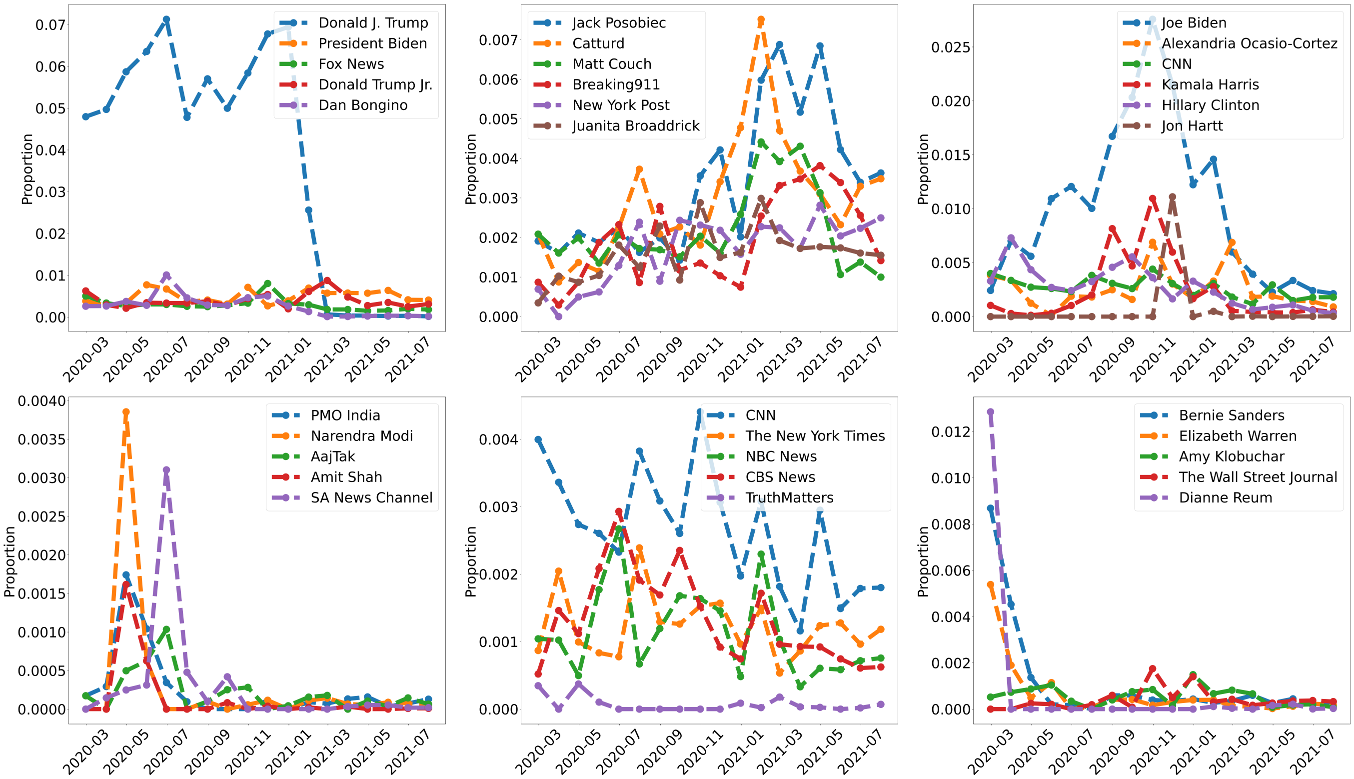


Fig. 7: Proportion of users that interact with the central entities of the selected communities in each time period.

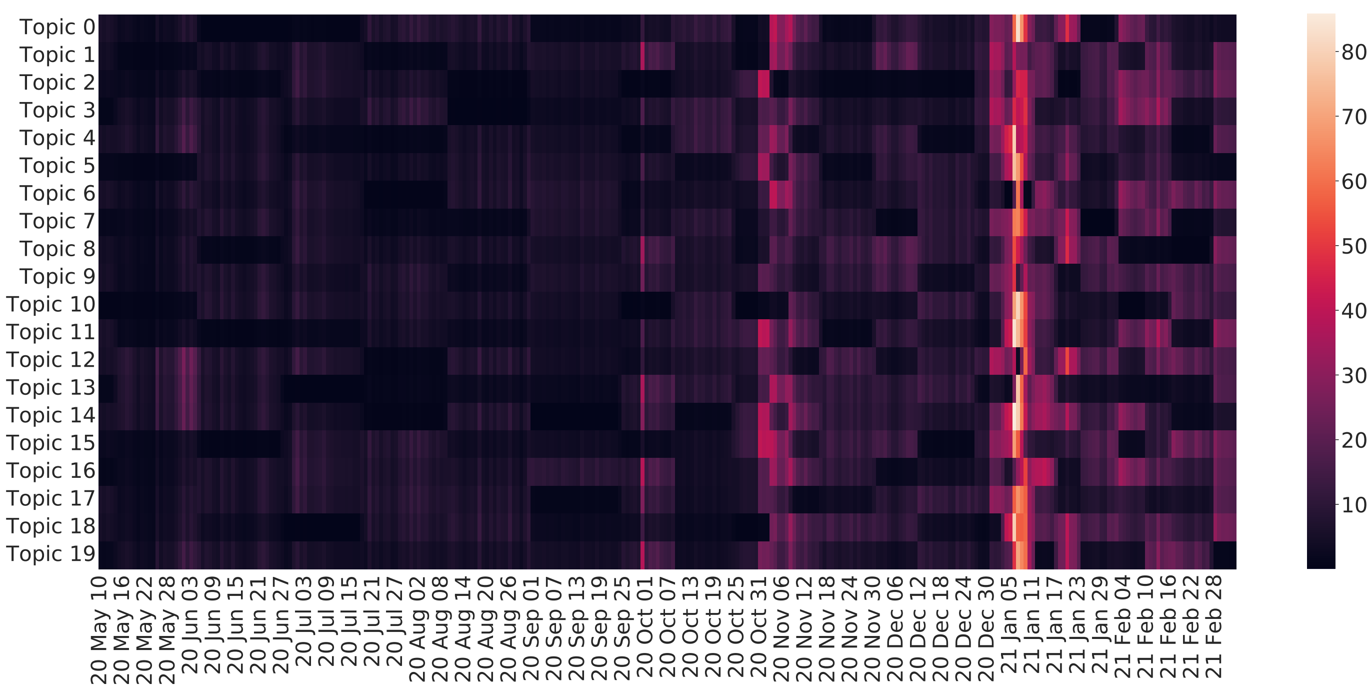


Fig. 8: Heatmap displaying the relative frequencies of topics with time on the largest community.

ID	Lifespan	Central Entities
1	Jan., 2020 – July, 2021	Donald J. Trump, President Biden, Fox News, Donald Trump Jr., Dan Bongino
2	Mar., 2021 – July, 2021	Jack Posobiec, Catturd, Matt Couch, Breaking911, New York Post, Juanita Broaddrick
7	Jan., 2021 – Mar. 2021	Joe Biden, Alexandria Ocasio-Cortez, CNN, Kamala Harris, Hillary Clinton, Jon Hartt
10	Jan., 2020 – July, 2021	PMO India, Narendra Modi, AdjTak, Amit Shah, SA News Channel
14	Oct., 2020 – Dec., 2020	CNN, The New York Times, NBC News, CBS News, Truth Matters
22	May, 2020 – July, 2020	Bernie Sanders, Elizabeth Warren, Amy Klobuchar, The Wall Street Journal, Dianne Reum

TABLE I: Selected dynamic communities with lifespan and central entities.

tweets that do not themselves have tweets in the data set. For example, Donald Trump and Joe Biden are both tagged or mentioned by users in the network, but do not have tweets in the network. Thus, central entities are simply subjects of online conversation. We hypothesize that central entities nevertheless play a important role in how nodes interact in the network. Therefore, we examine the closeness centrality of each node within certain communities [15], [17]. The nodes in each community are then sorted with their closeness centrality scores, and the top 10 users or entities are regarded as the central users in that community. In Table I, six selected communities with the central entities are shown, and the users are eliminated in the visualization for privacy purposes. The results show that community #1 has the most central entities "Donald J. Trump," "President Biden," and "Fox News," and this community remains large over time. We also identify a group related to India, which is centered around "PMO India" and "Narendra Modi" (community #10). Additionally, we see that the community #7, which emerges at the end of 2020 and disappears after February 2021, focuses on the entity "Joe Biden," "Alexandria Ocasio-Cortez," and "CNN." The central nodes of this group and the existence time of this community could potentially be related to the January 6th Capitol riots. Community #14, which focused on the mainstream news media, appeared in September 2020 and diminished in November 2020, which coincides with the time of the 2020 U.S. Presidential Election. This confirms that some QAnon-related Twitter accounts interact extensively with the news media during the election. Community #22, with central nodes "Bernie Sanders" and "Elizabeth Warren," only exists for a short time period after the Democratic party presidential candidates withdrew from the election.

Fig. 7 shwos the proportion of users that interact with the central entities in each month. This provides a clear view of the popularity of the central entities in different communities. From the plot of community #1, we see that nearly 7% of the users in the whole data set hashtagged or replied to Donald Trump in July 2020 and January 2021, while the proportion quickly drops to zero as Trump's account was suspended on Twitter. Also, we see that the community that focuses on entities like "Joe Biden," "Alexandria Ocasio-Cortez," and "CNN" in our dataset, or community #7, is only active during the election.

In addition to leveraging the results of previous topic models, we also used TF-IDF and NCPD [28]–[30] to gain a deeper understanding of the dynamics of the topics within these communities. Selected results from the topic modeling along with the topic-frequency heatmap are plotted below in Fig. 8. One striking feature of the heatmap of the four noticeable peaks on May 26th, 2020, September 29th, 2020, November 6th, 2020, and January 6th, 2021. By investigating some of the tweets sent during around these times, these peaks were found to be reactions to the killing of George Floyd, Donald Trump's "stand back and stand by" comment in the 1st Presidential Debate, the 2020 U.S. Presidential Election, and the January 6th Capitol riots respectively.

IV. Knowledge Graph Completion: Topic Modeling

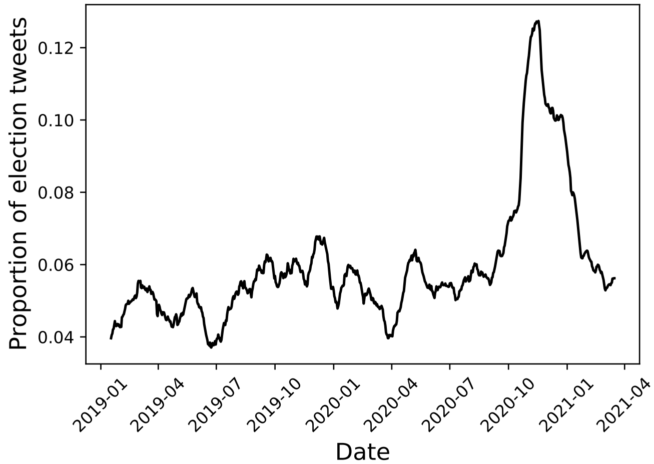
Due to the high degree of noise in our dataset, BERTopic only assigns topics to 18.4% of the tweets, leaving the rest with topic -1 or unlabeled. This motivates us to leverage the social relations provided by the Knowledge Graph to enrich the topic modeling results. Specifically, as shown in Fig. 4, since topics are nodes in our graph, we can think of topic modeling as link prediction (between tweets and topic nodes) or a Knowledge Graph completion problem. We then use the new topics to further analyze the data.

A. Graph Completion

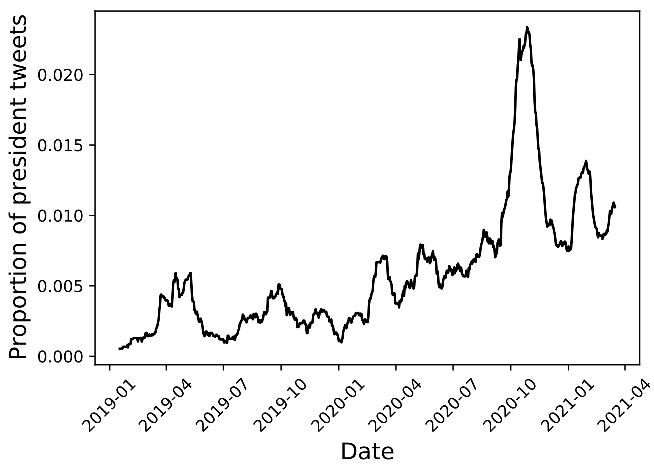
To do this, a graph neural network (GNN) was trained on the social network consisting of tweets, users, and entities. To train a GNN we need feature vectors for each node. We obtained features for each tweet as follows. First, we embedded the tweet using the sentence transformer model [31]. However, we have more than just semantic information from the text of the tweet. In particular, we also have temporal information. Hence, the embedding of the tweet is concatenated with an embedding of the timestamp which was computed using Time2Vec [32].

The embedding for a user consists of the embedding vector that is output by Deepwalk [33]. The Deepwalk algorithm was applied to the user-to-user network constructed from the *interacted_with* relation, and it uses short random walks to learn the representation of nodes on the graph. This vector is augmented using the Twitter account metadata including "favourites count", "followers count," "friends count," and "statuses count." Then, a heterogeneous GNN [34] based on the GATv2 [35] architecture was built for the task of topic labeling. The model is trained on existing topic labels.

The trained model exhibits predictive power on seven classes (topics) , where the class-specific F1-score and



(a) `Election` Topic



(b) `Joe Biden` Topic

Fig. 9: The proportion of the volume of tweets of topic `election` and topic `Joe Biden` in all tweets per day from January 2019 to April 2021.

accuracy on the testing set are both consistently higher than 0.7. For example, the averaged accuracy is 0.808 and F1-Score is 0.726 for the topic `election`. We then use the model to label all of the tweets and only kept the labels that corresponded to one of these seven topics. Using this method we increase the proportion of tweets with topic labels from 18.4% to 34.3%. For example, we now have 57855 tweets in the `election` topic, compared to 10559 tweets originally.

We also highlight that this method enables us to label tweets whose topic is clear given the context but cannot be inferred by the text alone. As an example, consider the tweets "Uniformed. Millions will regret getting it and people are already starting too. Don't get it to protect your kids!" and "No Ernie. It's designed to cause clots. Sinister." For both tweets, without the context of the COVID-19 pandemic, we cannot infer their topics. However, the model correctly assigned the two tweets to the `Vaccine` topic.

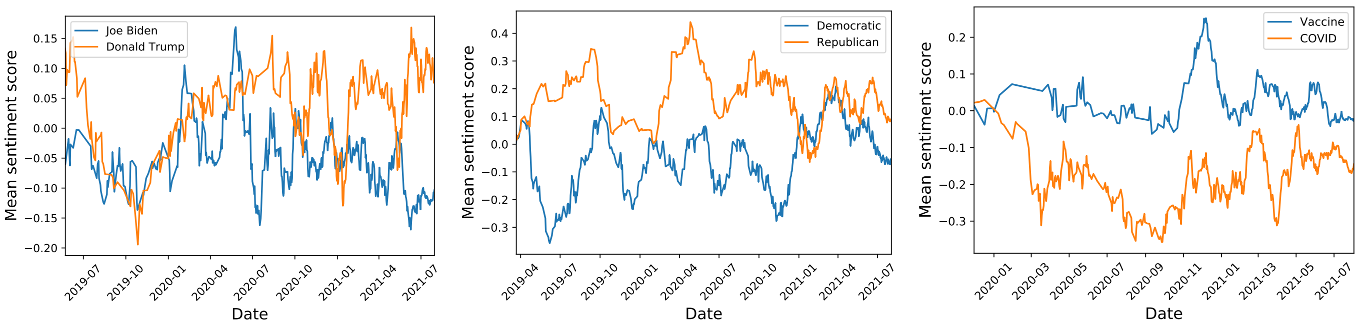
B. Time Series Analysis

With the topic modeling results and the sentiment labels on the tweets, we perform a set of time series analyses by considering the tweets as sequential data. Specifically, we investigate the tweet volumes of certain topics over time and compared the dynamic sentiments of related topics.

1) Topic Change: The time series analysis focuses on specific topics. We first clean the data by taking only the tweets that have nonempty cleaned text. For each day from January 2019 to April 2021, we count the proportion of the tweets discussing a particular topic. To reduce the noise in the data, we average the proportion for 30 days time windows. Fig. 9a and 9b illustrate the mean proportion of tweets of topic `election` and topic `Joe Biden`, respectively. Near the 2020 U.S. Presidential Election the proportion of tweets increases significantly for both topics. In addition, it can be seen that the proportion of tweets of topic `Joe Biden` increases right after the January 6th Capitol riot.

2) Sentiment Change: As day-by-day sentiment data is too noisy, we compute the mean sentiment for a 30-day period. In addition, we clean the data by taking only the tweets whose cleaned text is not empty. As shown in Fig. 10, the dynamic sentiments are compared between topic `Joe Biden` and topic `Donald Trump`, topic `Democratic` and topic `Republican`, and topic `Vaccine` and topic `COVID`, respectively. The time spans chosen are the overlapping periods where both topics have data. From Fig. 10a, we see that the sentiment increases for both topic `Joe Biden` and `Donald Trump` from October 2019, one year before the 2020 U.S. Presidential Election. The sentiments of two topics, however, diverge near July 2021, where the sentiment for topic `Donald Trump` becomes almost consistently higher than topic `Joe Biden`. For the comparison between topic `Democratic` and topic `Republican` in Fig. 10b, it can be seen that the sentiment for topic `Republican` stays positive while negative for topic `Democratic` throughout the time period, potentially indicating the popular political orientation in the QAnon community on Twitter. Lastly, the sentiment for topic `Vaccine` fluctuates around zero and has a clear peak at the time when the COVID-19 vaccine first became available in the U.S., but the sentiment for topic `COVID` is almost always negative. The contrast in sentiment for topic `Vaccine` and topic `COVID`, served as counter-evidence for the vaccine skepticism, suggesting that the voice against the COVID-19 vaccine might not be that significant in the QAnon community on Twitter.

The time series analysis showed that the rate of tweets in the QAnon network responded strongly to major real-world events. However, it is interesting that the rate of tweets did not appear to be affected by smaller-scale events which disproportionately affected the QAnon community. For instance, there was no discernible change in tweet rate following the taking down of 8chan on August 8th 2019, despite the fact that the forum was the primary platform on which the eponymous Q would post. We can also interpret our time series analysis results as a validation of the efficacy of our topic modeling,



(a) (b) (c)

Fig. 10: Sentiment change comparisons between related topics.

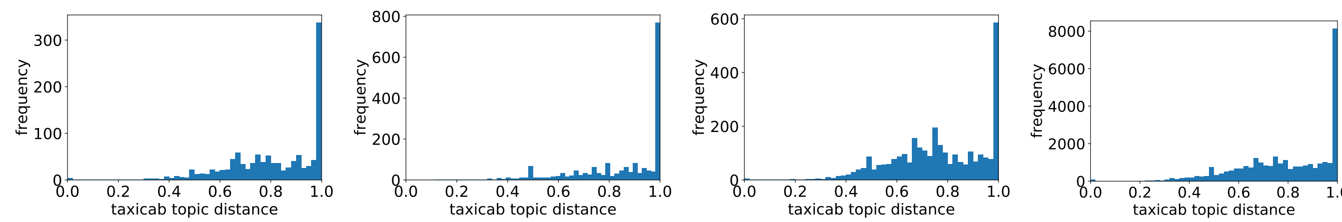


Fig. 11: Distribution of L_1 topic distances between communities.

as we can observe that the algorithmically identified topics, purported, feature keywords related to the election do indeed see an increase in tweet frequency around the time of the 2020 U.S. Presidential Election.

C. Community-specific Topic Modeling

To determine if different communities are discussing the same topics, we turn to a community-based topic modeling approach. To do this, we concentrate on a few selected months and study the distribution of L_1 distance between a topic frequency vector for all pairs of communities during these months. We investigate the topics in November 2020, January 2021, and April 2021. This is because we suspect that the 2020 U.S. Presidential Election and the January 6$^{\text{th}}$ Capitol riot might have a centralizing effect on discussions, and lead to the communities discussing similar topics, whereas April 2021 did not have prominent events, so it could serve as a contrasting example. These plots are shown chronologically (first three plots from the left). The last plot in Fig. 11 shows the sum of distributions of this form over all the months covered by our data, so it pictures the distribution of all L_1 distances between two communities occurring at the same time.

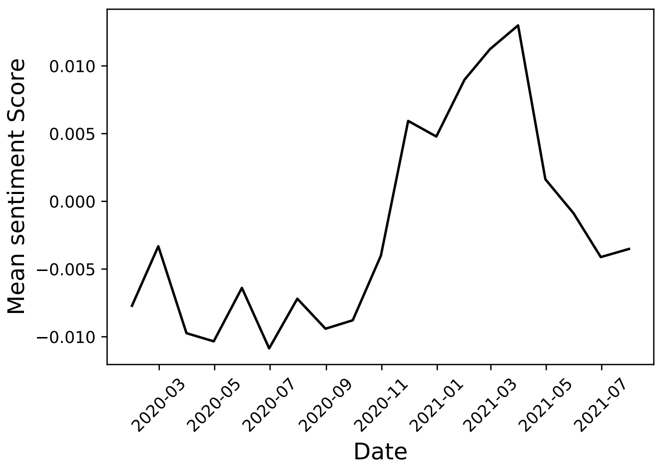
In all of the plots included in Fig. 11, the L_1 topic distances cluster near one, which could be interpreted as communities discussing different things. This figure indicates that even major events that should dominate discussion do not. Most communities tend to be discussing different topics. This result is surprising because it indicates that most of the discussion in these communities is internal in origin.

D. Community-specific Sentiment Change

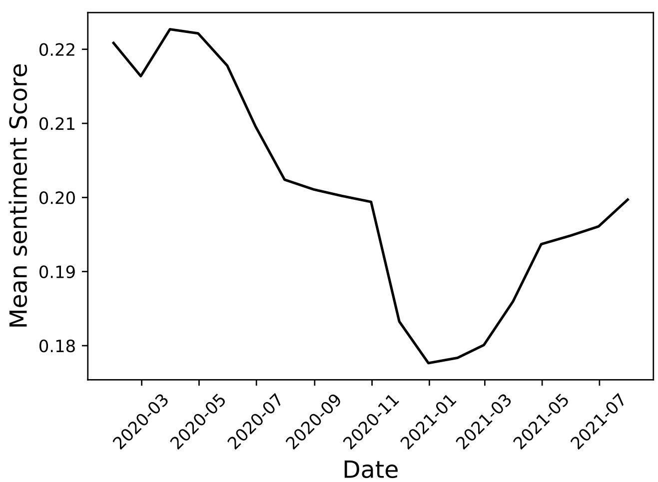
We also study how the sentiment changes over time for the different communities. Having community snapshots from the end of each month from January 2020 to July 2021, we collected all the tweets published by this community before the snapshot was taken. Using only the tweets for which the cleaned text is non-empty, we collect the sentiment score for each tweet. Then, having the sentiment score for each tweet before the snapshot was taken, we calculated the mean sentiment. Fig. 12 illustrates the mean sentiment change for 2 different communities: the largest community, where the most central is "Donald Trump", and community #10, where the central are Indian media. In Fig. 12a we can see that the mean sentiment score increases significantly after the 2020 U.S. Presidential Election to March 2021. In March 2021 Trump's account was banned from Twitter, which aligns with the sharp sentiment decrease. On the other hand, we can see the opposite behavior in Community #10 (Fig. 12b).

V. Conclusion

Despite our dataset missing a significant amount of key data, we still can derive meaningful results from it. For instance, Twitter accounts can be grouped into distinct communities centered around certain influential entities. Moreover, we find that the mainstream political topics are common discussion points within the QAnon community we studied. We generally noticed that the behavior of the QAnon Twitter community responded in relatively predictable ways to major real-world events, but seemed to not respond as strongly to events solely relevant to the QAnon community, such as new Q drops. These



(a) Community #1



(b) Community #10

Fig. 12: Mean sentiment change for select communities.

findings are consistent with an observed trend that the QAnon conspiracy is more mainstream [36].

REFERENCES

[1] J. H. Priniski, M. McClay, and K. J. Holyoak, "Rise of QAnon: A Mental Model of Good and Evil Stews in an Echochamber," 2021.

[2] D. Flocco, B. Palmer-Toy, R. Wang, H. Zhu, R. Sonthalia, J. Lin, A. Bertozzi, and P. J. Brantingham, "An Analysis of COVID-19 Knowledge Graph Construction and Applications," *IEEE Big Data*, pp. 2508–2512, 2021.

[3] M. Alaverdian, W. Gilroy, V. Kirgios, X. Li, C. Matuk, D. Mckenzie, T. Ruangkriengsin, A. Bertozzi, and P. J. Brantingham, "Who Killed Lilly Kane? A Case Study in Applying Knowledge Graphs to Crime Fiction." *IEEE Big Data*, pp. 2631–2640, 2020.

[4] A. Singhal, "Introducing the Knowledge Graph: Things, Not Strings," *Google Blog*, 2012.

[5] Q. He, B.-C. Chen, and D. Agarwal, "Building the Linkedin Knowledge Graph," Oct 2016.

[6] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Konotokostas, P. Mendes N., S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Cizer, "DBpedia - a Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia," *Semantic Web Journal*, 2015.

[7] W. C. Partin and A. E. Marwick, "The Construction of Alternative Facts: Dark Participation and Knowledge Production in the QAnon Conspiracy," *AoIR Selected Papers of Internet Research*, Oct. 2020.

[8] M. N. Hannah, "a conspiracy of data: Qanon, social media, and information visualization."

[9] B. E. Bond and R. Neville-Shepard, "The Rise of Presidential Eschatology: Conspiracy Theories, Religion, and the January 6th Insurrection," *American Behavioral Scientist*, 2021.

[10] J. Bodner, W. Welch, and I. Brodie, *COVID-19 Conspiracy Theories: QAnon, 5G, the New World Order and Other Viral Ideas*, 2020.

[11] J. H. Priniski and K. J. Holyoak, "A Darkening Spring: How Preexisting Distrust Shaped COVID-19 Skepticism," *PloS one*, vol. 17, no. 1, p. e0263191, 2022.

[12] "Geopy," 2021. [Online]. Available: https://github.com/geopy/geopy

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[14] G. Rossetti, L. Milli, and R. Cazabet, "CDLIB: A Python Library to Extract, Compare and Evaluate Communities from Complex Networks," *Applied Network Science*, 2019.

[15] L. C. Freeman, "Centrality in Social Networks i: Conceptual Clarification," *Social Networks*, 1979.

[16] U. Brandes, "A Faster Algorithm for Betweenness Centrality," *The Journal of Mathematical Sociology*, 2001.

[17] A. Hagberg, P. Swart, and D. Chult, "Exploring Network Structure, Dynamics, and Function Using Networkx," 2008.

[18] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal Detection of Changepoints with a Linear Computational Cost," *Journal of the American Statistical Association*, 2012.

[19] J. H. Priniski, N. Mokhberian, B. Harandizadeh, F. Morstatter, K. Lerman, H. Lu, and P. J. Brantingham, "Mapping Moral Valence of Tweets Following the Killing of George Floyd," 2021.

[20] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based TF-IDF procedure," 2022.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *ACL*, 2019.

[22] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," 2018.

[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *ICKDDM*, 1996.

[24] C. Malzer and M. Baum, "A Hybrid Approach To Hierarchical Density-based Cluster Selection," in *MFI*, 2020.

[25] C. Hutto and E. Gilbert, "Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *AAAI*, 2014.

[26] J. Ni, G. Hernandez Abrego, N. Constant, J. Ma, K. Hall, D. Cer, and Y. Yang, "Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models." Association for Computational Linguistics, 2022.

[27] D. Arthur and S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding." USA: Society for Industrial and Applied Mathematics, 2007.

[28] L. Kassab, A. Kryshchenko, H. Lyu, D. Molitor, D. Needell, and E. Rebrova, "On Nonnegative Matrix and Tensor Decompositions for COVID-19 Twitter Dynamics," 2020.

[29] T. Maehara, K. Hayashi, and K. ichi Kawarabayashi, "Expected Tensor Decomposition with Stochastic Gradient Descent," 2016.

[30] A. Garla, "NMF – A Visual Explainer and Python Implementation," 2021.

[31] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A Pre-trained Language Model for English Tweets." Association for Computational Linguistics, 2020.

[32] S. M. Kazemi, R. Goel, S. Eghbali, J. Ramanan, J. Sahota, S. Thakur, S. Wu, C. Smyth, P. Poupart, and M. Brubaker, "Time2Vec: Learning a Vector Representation of Time," 2019.

[33] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online Learning of Social Representations," in *SIGKDD*.

[34] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous Graph Transformer." Association for Computing Machinery, 2020.

[35] S. Brody, U. Alon, and E. Yahav, "How Attentive are Graph Attention Networks?" in *ICLR*, 2022.

[36] R. Kleinfeld, "Opinion | How Political Violence Went Mainstream on the Right," Jul 2022. [Online]. Available: https://www.politico.com/amp/news/magazine/2022/11/05/political-violence-mainstream-right-wing-00065297