Sedition Hunters: A Quantitative Study of the Crowdsourced Investigation into the 2021 U.S. Capitol Attack

Tianjiao Yu
Virginia Tech
Blacksburg, VA, USA
tianjiao@vt.edu

Sukrit Venkatagiri*
University of Washington
Seattle, WA, USA
sukritv@uw.edu

Ismini Lourentzou
Virginia Tech
Blacksburg, VA, USA
ilourentzou@vt.edu

Kurt Luther
Virginia Tech
Arlington, VA, USA
kluther@vt.edu

ABSTRACT

Social media platforms have enabled extremists to organize violent events, such as the 2021 U.S. Capitol Attack. Simultaneously, these platforms enable professional investigators and amateur sleuths to collaboratively collect and identify imagery of suspects with the goal of holding them accountable for their actions. Through a case study of Sedition Hunters, a Twitter community whose goal is to identify individuals who participated in the 2021 U.S. Capitol Attack, we explore what are the main topics or targets of the community, who participates in the community, and how. Using topic modeling, we find that information sharing is the main focus of the community. We also note an increase in awareness of privacy concerns. Furthermore, using social network analysis, we show how some participants played important roles in the community. Finally, we discuss implications for the content and structure of online crowdsourced investigations.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

KEYWORDS

crowdsourcing, collective action, crowdsourced investigations, extremism, topic modeling, Capitol Riot, social network analysis

ACM Reference Format:
Tianjiao Yu, Sukrit Venkatagiri, Ismini Lourentzou, and Kurt Luther. 2023. Sedition Hunters: A Quantitative Study of the Crowdsourced Investigation into the 2021 U.S. Capitol Attack. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543507.3583514

*A portion of this work was completed while this author was at Virginia Tech.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583514

1 INTRODUCTION

On January 6, 2021, the U.S. Capitol was violently attacked by a group of more than 2,000 Donald Trump supporters aiming to nullify the formalization of Joe Biden's presidential victory. This was the most severe attack on the U.S. Capitol since the War of 1812, causing several deaths, hundreds of injuries, and at least $30 million in property damage and related security expenses [16, 17]. It is also arguably "the most documented crime in U.S. history" [10] as there is evidence collected not only from surveillance infrastructure but also social media where hundreds of rioters were live streaming and uploading posts during and after the event [11]. Following the Capitol Attack, the U.S. Department of Justice (DOJ) and Federal Bureau of Investigation (FBI) began investigations that have led to over 840 arrests and 185 criminal sentences (as of June 2022) [41]. However, hundreds of suspects remain unidentified due to the large scale of the event, prompting law enforcement officials from the FBI and the Washington, D.C. police department to seek help from the public [6].

Among the members of the public assisting in suspect identification, a Twitter community quickly formed around the hashtag `#SeditionHunters`. A related account, `@SeditionHunters`, was created on January 12, 2020, and gathered more than 66,000 followers. This movement of online sleuthing or web-sleuthing [37, 56] harnessed the power of collective knowledge and resources to assist the police in their criminal investigation. They described themselves as Open-Source Intelligence (OSINT) investigators working together to identify suspects who allegedly committed crimes in the January 6 Capitol Attack [2]. The Sedition Hunters community is notable for their significant contributions to multiple successful arrests [31] and legal proceedings [4, 5]. Over a year later, the `#SeditionHunters` community remains active, in contrast to most similar gatherings [47].

Much discussion has been made about the potential value of crowdsourced data contributions to law enforcement [38, 44]. The investigation of the Vancouver Hockey Riots of 2011 is an early example illustrating how the Internet offers a venue for the public and law enforcement officials to achieve a shared objective through pooled intelligence and resources [46]. Despite the growing value of crowdsourced investigations, they have also raised serious concerns. For example, as police actively incorporated the public's help to provide information about the suspects who planted bombs at

the 2013 Boston Marathon, the crowd yearned to be more than information suppliers. They sought to participate in the crime-solving process, which ultimately led to four misidentifications and harmful consequences for the victims [37].

Analyzing patterns of public participation in online cyber-policing gives insights into when and how collaborative efforts can bring about positive or negative impacts during the so-called Information Era of policing [27]. It also provides insights into how crowdsourced investigations can be better harnessed for efficiency and effectiveness. Given the most researched large-scale crowdsourced criminal investigation event, the 2013 Boston Marathon Bombing, happened almost a decade ago, Sedition Hunters provide us a unique opportunity to investigate how such a community deals with the ever-growing overabundance of data [19], how recent technology (e.g., Face recognition) can be utilized for better output, and how the community is organized and regulated for improved effectiveness.

In this paper, we use quantitative approaches, namely topic modeling and social network analysis, to conduct an analysis of more than 300,000 posts and 65,282 unique users related to the Sedition Hunters community, spanning exactly one year from the Capitol Attack. We find that the content of most posts focused on information sharing, which included details about suspects and related news and updates. Topic analysis also reveals the tracking tools the community used for the investigation process such as hashtags, spreadsheets, and self-developed websites. Social network analysis shows that, as most participants did not engage comprehensively, a few dominant accounts are responsible for multiple roles of gathering, processing, and distributing information. We conclude by discussing the factors that are unique to the Sedition Hunters community and contributed to its success.

2 RELATED WORK

With the prevalence of portable computing devices and the desire to document daily activities, modern society empowers people to provide huge amounts of persistent, searchable, and remotely accessible archives of formerly private information and events [32]. The openness and interactivity of social media also encourage crowdsourced investigations, which use the knowledge, work, and content of online communities as a vastly extended surveillance network [54]. Simultaneously, the entire format of policing has advanced into the Information Era in which the crime-fighting process has become data-driven, intelligence-led, and technologically mediated [27]. Law enforcement officials are considered more as "knowledge workers" who now have more responsibility for collecting and processing information [23]. Our analysis of Sedition Hunters highlights the complex and nuanced role the public sometimes seeks to play in online crime investigation. Further, we see that the Sedition Hunters community is able to self-organize regulatory and other strategies by which they can direct themselves in ways that generate useful information more efficiently while minimizing negative impacts such as misidentification.

Previous studies posited three primary investigative models for solving crimes. Trottier [18] studied how traditional top-down policing by investigative agencies and bottom-up policing by crowdsourced users converge and co-exist on social media. [52] studied expert-led crowdsourced investigations, a hybrid of top-down and bottom-up criminal investigation models, in the hope of balancing the tension between experts and crowds to achieve greater results. Sedition Hunters is different from previous investigative models because, while most participants of the Sedition Hunters are amateurs, some members claim they have expertise or previous experience in OSINT [42]. In other words, the experts are part of the crowd. The Sedition Hunters community was able to self-regulate the interactions between experienced members and novices, in which they created relatively standardized procedures for the investigation, such as creating composite suspect images, releasing information with the #doyouknow hashtag, and building progress-tracking and coordination websites.

Scholars use terms such as digilantism [37] and web-sleuthing [56] to describe cases where members of the public take part in online investigations. Tapia and Lalone used sentiment analysis to examine moments during crowdsourced crisis investigation when the public became either more positively or negatively inclined toward the acts of the web-sleuthing participants, reflecting social approval or disapproval of such actions [50]. Marx also qualitatively examined opportunities and risks of crowdsourced efforts and implications for issues of justice given the new technologies [33]. Compared with these studies, we use quantitative methods to explore the content and participants of this specific online crowdsourced community, the Sedition Hunters.

Partially due to the limited scale of most crowdsourced crime investigations and their dispersed nature, prior work has focused on how they play a role in criminology in a broader sense. Sedition Hunters provided an opportunity to study the community itself and analyze how such a large-scale crowdsourced investigation community successfully delivers useful information. Most related to our work is the study of Venkatagiri et al. who conducted a mixed-methods analysis of the Sedition Hunters community. They presented preliminary results regarding who are the sedition hunters and what did they do to produce successful results [53]. Our research extends this study in an in-depth, quantitative approach with topic modeling and social network analysis, along with a focused manual qualitative examination to further investigate how and why the Sedition Hunters do what they do to produce successful results. This paper addresses the following research questions:

- **RQ1**: What are the discussions within and around the Sedition Hunters community? What are the main topics or targets of #SH?
- **RQ2**: Who participates in the Sedition Hunters community and how? What are the different roles in the SH community? What methods have they used?

3 METHODS

This section introduces our methods, which focus on assessing what topics and questions gathered the most attention, and identifying different roles within the Sedition Hunters community.

Data. Using the Twitter API 2.0 search endpoints, we used the hashtag `#Seditionhunters` as query to collect all the tweets from 2021 January 12 (when the first usage appeared) to 2022 January 12 (one year later). As the Capitol Attack and discussions around it focused on the U.S., we only store tweets in English and their associated metadata. There were 321,662 unique tweets (25,324

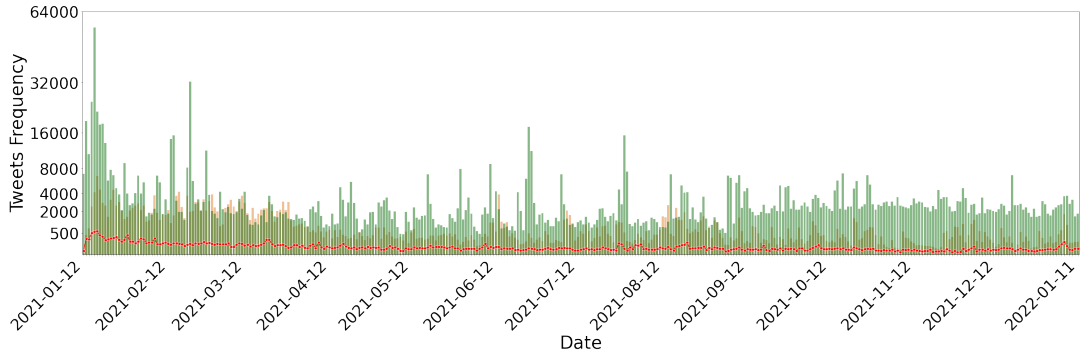


Figure 1: Number of tweets posted each day from 2021 Jan 12 to 2022 Jan 22. Each column represents one day. The red line represents the original tweet frequency, the orange bars represent the original tweets with retweets frequency, and the green bars represent the reply frequency. For better representation, the y-axis is scaled $x^{1/2}$.

if excluding retweets) over the year, with 65,282 users participating. From all collected tweets, only a fraction (7.87%) are original tweets or quote tweets where new content is presented. The rest are retweets. However, the Twitter search endpoint ignores tweet replies, an important dataset consisting of many original discussions and conversations. Therefore, we further collected all replies to tweets using conversation IDs provided by the Twitter API. There are 8,317 tweets that have replies, totaling 1,173,870 replies (including replies of replies). We collected the data in 12 monthly intervals starting from the 12^{th} day of each month to the 11^{th} day of the following month. Tweet frequency is shown in Figure 1.

Topic Model. Given the large size of the dataset, manual text analysis is infeasible. Thus, we use a topic modeling algorithm to analyze what discussion took place within the `#SeditionHunters` community. BERTopic is an algorithm that leverages BERT embeddings and dense-based clustering to overcome the limitations of other methods such as LDA. In particular, LDA, as one of the most widely used methods, assumes that documents are created from words that belong to different topics. This is usually not the case with Twitter data in which one tweet is normally about one specific topic. Also, LDA requires prior knowledge of topic numbers, which is hard to conclude in this scenario, and it makes the bag-of-words assumption, ignoring the context of words [13, 20]. In BERTopic, the assumption of each data entry is talking about only one topic is especially convenient in that it allows us easily trace back each topic and look at the original texts of the topic. BERTopic first transforms each document into a dense vector so that documents within close spatial proximity can be grouped using the HDBSCAN [34] clustering algorithm. Then, topics were extracted and described by class-based TF-IDF. In addition, BERTopic uses Uniform Manifold Approximation and Projection (UMAP) [35] for dimension reduction in that it often deals with high dimensional vectors. A detailed explanation is available in the Appendix.

In our study, we first filter out the retweets, since their textual content is simply repeating the original tweet. Then, we remove the usernames and hashtags from the tweet. Next, we lemmatize the words in the remaining tweets and remove the duplicate after they are processed. Finally, we pass this subset to BERTopic. For better interpretability, we fine-tune BERTopic so that the number of topics is limited to a reasonable range. Given the scale of the data, the number of topics can be reduced to around 100 while still having interpretable results. After we acquire the result, we use the top-10 most weighted topics as a guide to categorize and present other less weighted topics. We manually examine at least 100 tweets for presented topics to support our statements.

Network Analysis In addition to the text analysis with topic modeling, we employ social network analysis (SNA) to investigate the structure of the community, i.e., who participated in the `#SeditionHunters` community, and how they participated. As mentioned, retweet activity is dominant. Therefore, we first build the retweet network where each node in the graph represents a Twitter account and directed edges between nodes represent retweets. We additionally construct a mention network where each node represents a Twitter account and directed edges represent mentioning activity. The mention network is complementary to the retweet network; a retweet activity shows where the information comes from, and the mention activity shows where the information goes.

SNA offers numerous metrics that can show the importance of each node, and these metrics also reflect the social capital of the nodes [15]. We make use of three basic metrics: in-degree, out-degree, and modularity. In-degree refers to the number of connections received in a directed network. For the retweet network, high in-degree means the user is retweeting many other users, which implies high participation. A node with a high in-degree in the mention network is a node where many other users are directing information towards it. Out-degree is the opposite of in-degree and measures the number of connections sent in a directed network. In the retweet network, a node with a high out-degree means the respective node received many retweets. Such users usually have high visibility since the node influenced the spread of the information. Finally, the likelihood of forming clusters or groupings of closely connected nodes in a network is measured by modularity. Higher modularity means connections within a group are denser and links to other groups are less dense [36]. When calculating modularity using Gephi [8], we also compute the number of modules. A high number of modules usually implies a more disconnected network where informational flow might be negatively affected [55].

The original dataset is processed into networks and then visualized using the Force Atlas 2 layout in Gephi [9]. We analyze the community and list the important users according to these metrics

(see Tables 1 and 2). Finally, we examine the Twitter account page of important users to categorize them.

4 FINDINGS

As shown in Figure 1, the community quickly generated much attention in the first month, with many discussions. The retweet counts remained at a high level for the first three months. They gradually decreased with small bursts in June, August, December, and the following January. The reply frequency has bursts across the year. However, unlike retweets, as the frequency decreased during the first 9 months, it increased around September 12 and remained at a relatively high level afterward.

4.1 What are the main topics or targets of #SH?

Using BERTopic, we extract 119 topics from the original tweets and 57 topics from the replies. These topics reveal several main categories describing how users participated in the `#SeditionHunters` discussion. First, a majority of the tweets are related to asking for help and promoting information sharing. Second, there are some topics related to how this community keeps track of progress. Third, there are discussions about news, events, and updates that are not necessarily related to the online sleuthing process, including many participants simply tweeting as a means of self-expression.

Topics from the Tweets: Figure 2 shows the top 10 topics out of the 119 topics from tweets in terms of their weight. Many heavily weighted topics are about information sharing and asking for information sharing. For example, Topic 0 has keywords such as "please contact fbi", "fbi afo" ("afo" refers to target labels assigned by the FBI), "please share across", and "across platform". An example tweet of Topic 0 is shown in Figure 3. Similar tweets with different target labels are a major component of the #SeditionHunters community. We also find other topics with similar semantic meanings but phrased differently, such as Topic 18 with keywords "track tag", "hey see", "recognize contact directly" and Topic 28 with keywords "let identify", "need identify", "recognize", and "anyone identify".

In Topic 0, the specified targets are often independent of each other. In contrast, some topics cluster tweets that have specific targets related to a single theme. Topic 4, for example, has keywords such as "proud boy", "march proud boys", "joe biggs", and "enrique tarrio". These keywords are all related to the Proud Boys, a right-wing extremist group that has taken part in multiple acts of violence and intimidation. Out of around 110 active members, more than 30 Proud Boys have been identified in connection with the Capitol Attack [1]. Some tweets labeled as Topic 4 are asking for more information about specific members of the Proud Boys, similar to the tweets in Topic 0. Others are news and updates related to the Proud Boys, such as the leader of the Proud Boys has been charged with conspiracy in the January 6 attack, and another leading figure of the Proud Boys who participated in the Capitol Attack and was consequently arrested later.

Topic 5 with keywords "tunnel entrance", "tunnel video", "inside tunnel", "upper western terrace", and "tunnel fight" is another single-theme topic. This topic describes the 3-hour-long fight that happened around the west terrace and inside the western tunnel entrance of the Capitol where rioters fought the police blockade and tried to gain entry to the Capitol. Comparing the tweets of Topic 0 and Topic 5, the first contains tweets asking for images of specific targets with publicized information. As we can see in Figure 3, the target is already on the FBI's wanted list with the FBI-generated suspect ID 244-AFO. However, Topic 5 has more original content such as capturing images and creating hashtags of individuals who were involved in the tunnel fight. As the whole tunnel was blocked by the police, a picture of an individual entering or leaving the tunnel can be used to accuse the person of fighting the police officers inside the tunnel. Such information is usually passed to the FBI and contributes to the formal criminal investigation process.

Besides distributing information about potential suspects, related updates regarding legal proceedings are often shared as well. Some of the tweets in Topic 6 with keywords "arrest yet", "charge", "arrest today", together with Topic 61's keywords "update arrest", "link update arrest", "link update", show that there are constant updates on the sleuthing process.

The topic model also shows hints on how `#SeditionHunters` keep track of the progress. Topic 1 with keywords "new hashtag", "videos", "hashtags", "tag", "find", and "archive", shows that hashtags are one way for the entire community to track specific targets. As mentioned above, some of these hashtags come from law enforcement officials' suspect IDs, while others are newly generated by Sedition Hunters. In addition to hashtags, Topic 32 with keywords such as "spreadsheet", "data", "add spreadsheet", "info sheet" reveals that the community uses shared Google Docs spreadsheets to keep track of progress. After examining the corresponding tweets and spreadsheets, we find that the usage of the spreadsheets is multi-purpose. Some spreadsheets are used to keep track of multiple targets. Each entry in such spreadsheets is about one specific target and its related information such as legal status, corresponding hashtag, appearance notes, etc. Other spreadsheets might be used to keep track of one specific video as each entry is an important frame of the video with its related information.

Another way for Sedition Hunters to track and organize their progress is by building suspect profile web pages. Topic 7 has keywords such as "know link", "know", "link more info". Back tracing the tweets and the links from Topic 7, we find several community-built websites. All of them have a listing page that tracks all targeted individuals. Although presented in different styles, these listing pages are often associated with corresponding individual profile pages which provide sleuthing progress updates. For example, the official website of `#SeditionHunters` has a "Perp sheet" where all avatars of suspects are listed (see Figure 4a). Each avatar leads to its corresponding profile page which has details such as pictures or videos of the suspect committing crimes during the Capitol Attack and corresponding legal documents. Other websites incorporate interactive tools for better usage. For example, *jan6attack.com* provides detailed filters in which the user can search for a person of interest by headgear types, face covering colors, top wear brands, etc. Similar to spreadsheets, these websites are not limited to tracking suspects. The website *jan6evidence.com* pinned over 8000 videos on an interactive map in chronological order (see Figure 4b). Moreover, *capitolmap.com* has an interactive face recognition tool that takes a portrait photo and returns similar faces so that the user can easily "compare faces" based on the input.

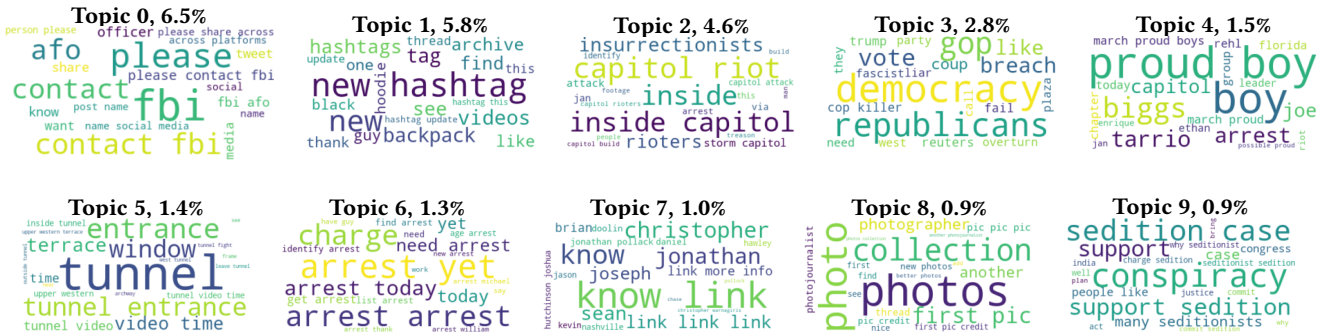
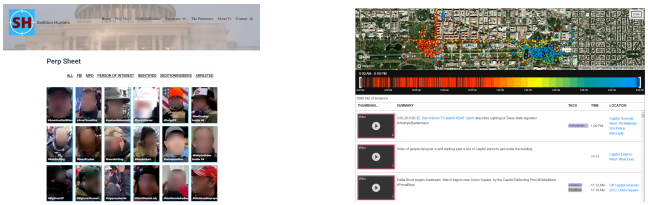


Figure 2: Keywords and weights for the top 10 most weighted topics

Figure 3: Example tweet for Topic 0 (information sharing)



(a) "Perp Sheet" page of the official Sedition Hunters website

(b) Interactive map locating different evidence videos in chronological order

Figure 4: Related websites for Sedition Hunters

Parallel to the topics related to the online investigation, many tweets commented on discussions about trending events. For example, Topic 3 is about the discussion of "democracy", "republicans", "GOP", and "vote". The discussion may relate to the January 6 attack, but they are not directly associated with the sleuthing process. Furthermore, some topics show that some tweets are self-expressions with no clear intention of making conversation or contributing to the investigation process. For instance, Topic 13 with keywords "yeah baby", "lol", "shit", "wtf wtf"; Topic 24 with keywords "great jobs", " heroes thanks", "work good"; and a more negative Topic 60 with keywords "loser", "douchebag" and "stfu".

In addition to the three main categories mentioned above, as [53] showed, in order to avoid ethical issues such as exposing private information, many "DoYouKnow" composites created by the Sedition Hunters explicitly instructed community members to refrain from publicly naming suspects and instead submit tips directly to the FBI. We trace back to some of the original tweets using keywords such as "contact fbi" in Topic 0 and "recognize contact directly" and "directly" in Topic 18. We found similar textual cues, suggesting that the community is willing to learn from their own mistakes [28] and previous events [29] so that they can minimize negative impacts caused by errors such as misidentification.

Topics from replies: The large number of replies led to 2165 topic clusters. For a deeper investigation, we increased the n-neighbors of the UMAP from 10 to 100. Accordingly, we increased the minimal topic size of HDBScan from 10 to 100. This setup reduced the final number of topics to 448, which are either discussions of trending events or self-expression. For event discussion topics, most of them do not contribute to the investigation process. For example, Topic 1 is talking about "joe biden", "obama" and "worst president". Topic 3 is talking about "impeachments" and "twice impeach" of the "president". Some event discussion topics are not related to the Capitol Attack or Sedition Hunters at all. For example, Topic 5 is mainly about "taliban" and "afghanistan". Topic 8 is about the COVID with keywords such as "vaccines", "covid", and "wear mask". And Topic 9 is discussing "debt" and "tax cut". Additionally, many replies are used as a means of self-expression: Topic 4 has keywords such as "stupid", "know", "idiots", "happen", and "dumbest".

Outliers: 48.3% of tweets are classified as outliers even though the n-neighbors of the UMAP and the minimal topic size of HDBScan were set to a small number (10), which means the textual content of the community is sparse. The topic results of the replies show similar trend as 51.8% of the replies are outliers. The content of the replies is less regulated compared with the tweets. This is expected as discussions can easily diverge in the reply section. Also, most replies did not contribute to the online investigation process.

The textual content does not characterize the uniqueness of this community. If we take tweets, retweets, and replies all into account, we find that information sharing and distributing tweets, especially with specified targets, are the main focus of this community. This aligns with one of the community's self-defined goals, which is creating "Do You Know" posters and composites and asking the public to share them on social media platforms [2]. Among the distributed targets, there is no single target that stands out. However, some targets together lead to a converged theme such as the "tunnel fight" and the "Proud Boys". Additionally, we notice the

increased awareness of protecting private information, as we see explicit efforts of the community where they request people to not use real names and contact law officials directly for providing tips.

4.2 Who participates in #SH and how?

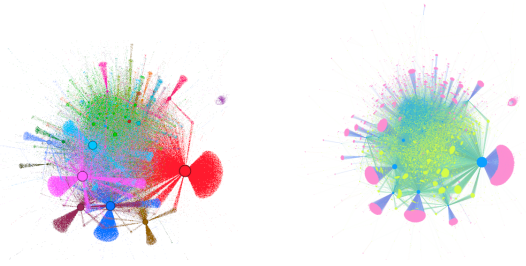
We use SNA to identify different roles in the #SH community. We transform the original dataset into two types of networks, retweet and mention networks. In retweet networks, each node in the graph represents a Twitter account and the directed edges represent retweet activities. An edge from node A to node B represents B retweeting A. In mention networks, nodes represent Twitter accounts and edges represent mention activities where an edge from node A to node B means that A posted a tweet mentioning B.

Retweet Network: There are 63,859 nodes and 126,353 edges. The graph modality, 0.473, is relatively high given the range of 0 to 1. We colored the top 10 out of 107 groups according to the modality as shown in Figure 5 (a). Besides the central dense area, many groups are centered around one or two high-degree nodes. For example, the nodes in the red group (23.24% of the total number of nodes) almost all directly connect to the central node "SeditionHunters".

To better identify different roles in the community, we manually group users according to their degrees as shown in Figure 5 (b). The first group is the pink nodes with a degree of 1. More than half of the nodes (62.73%) belong to this group. They are almost exclusively in periphery positions. These nodes were pushed outward and converged as clusters because they have retweet relations with only one other node. Note that a user could retweet a second user many times without increasing the first user's degree. A degree of 1 does not necessarily imply an inactive user. Instead, it means the user did not engage comprehensively with more users in the dataset. The second group is yellow nodes with degree 2–10, accounting for 35.69% of nodes. Compared to the first group, these nodes are positioned more in the center area. We also see some clusters in this group. They are formed when the nodes within are retweeting from the same nodes. In addition to the clusters, we see a more distributed pattern, implying that users may be more engaged. The third group is nodes in blue with degree more than 10. Only 1.58% of total nodes have degree more than 10. However, these nodes are the backbone of the `#SeditionHunters` community. Nodes in this group are dispersed and often are the source of the clusters. These accounts either posted original content worthy to be retweeted many times, or it is already an influential account that brings more exposure to the content, leading to more retweets.

As the degree increases, we can then focus on the differences between in-degree and out-degree. The in-degree interval of the retweet network falls between 0 and 4,487. The out-degree is between 0 and 30,429. Figure 6 shows the high in- and out-degree users in the retweet network. The larger size and greener hue means the in- or out-degree is higher. Notice that there there is only one node that has both high in-degree and out-degree.

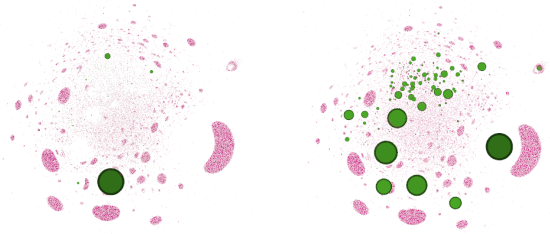
To understand the different roles that users played in the large `#SeditionHunters` community, we further analyze the top 10 users with higher in-degree and out-degree (see Table 1). As suggested above, users with higher degrees are essential to the functioning of the community. We investigate the follower account along with the degrees with the assumption that users with more followers are able to distribute information more effectively.



(a) Grouped by modality

(b) Grouped by degree, Pink: degree 1, Yellow: degree 2 to 10, Blue: degree more than 10

Figure 5: Retweet networks grouped by modality or degree



(a) In-degree

(b) Out-degree

Figure 6: Retweet network node size proportional to in-degree and out-degree

For higher out-degree accounts, `@SeditionHunters` is the official account of the community. Accounts such as`@capitolhunters`, `@seditiontrack`, and `@DomesticTerror2` are created specifically for `#SeditionHunters` shortly after January 6, 2021. `@ryanjreilly` and `@nate_thayer` are journalists. Others are normal Twitter accounts without obvious Sedition Hunters-related identifications such as usernames, profile pictures, or profile descriptions. However, by looking at the most 100 recent tweets, we find that tweets of `@Anonymized-4` and `@Anonymized-6` are exclusively about the Capitol Attack and Sedition Hunters.

Compared with top out-degree accounts, most of the higher in-degree accounts are normal Twitter users. Besides `@ryanjreilly` and `@DomesticTerror2`, we already see in the out-degree list, many users are passionate about politics. For example, `@patriottakes` is described as an account of researchers exposing right-wing extremism, and `@Anonymized-5` self-identifies as a political activist. The in-degree range is considerably lower than the out-degree. The highest in-degree is 4,487, then it quickly decreases to a few hundred. In-degree indicates the number of retweets one user gathers, which might be retweeted by their own followers. Thus, these nodes can be considered information transmitters. More green nodes in the out-degree graph together with the large gap between in-degree and out-degree confirmed the idea that the main goal of the community is sharing and distributing information.

The analysis of Table 1 shows that the number of followers does not reflect how important the account is. There are accounts that have few followers but high in- or out-degrees, and vice versa. It is also noticeable that the top out-degree nodes rarely retweet others. The highest out-degree account, @SeditionHunters, never retweets other users. This suggests that they act more as an authority or information generator. The one node with large in-degree and out-degree is `@ryanjreilly`, an NBC Justice News reporter.

Top in-degree users				Top out-degree users			
Usernames	**Out-degree**	**In-degree**	**Follower counts**	**Usernames**	**Out-degree**	**In-degree**	**Follower counts**
@ryanjreilly	10,341	4,487	167,835	@SeditionHunters	30,429	0	64,310
@patriottakes	568	941	449,980	@capitiolhunters	14,311	6	34,763
@Anonymized-1	501	513	255	@ryanjreilly	10,341	4,487	167,835
@DomesticTerror2	5,670	352	583	@seditiontrack	8,886	181	70,196
@Anonymized-2	217	274	850	@DomesticTerror2	5,670	352	583
@nycjim	126	258	221,756	@Anonymized-3	3,301	108	324
@SlickRockWeb	1,173	249	6,951	@nate_thayer	2,252	30	6,666
@SeditionSleuth	88	209	495	@Anonymized-4	2,143	10	8,923
@Anonymized-5	154	197	31,480	@Anonymized-6	1,832	23	709
@Anonymized-7	486	188	545	@Anonymized-8	1,679	22	305

Table 1: Users with top in- and out-degree in retweet network (full usernames are anonymized except for high-profile accounts)

Figure 7: Mention Network grouped by modality

Mention Network: There are 70,317 nodes and 169,623 edges. As there are slightly more nodes and edges, the total of 734 detected communities is significantly higher than the retweet network (107 detected communities). However, the modality, 0.474, is very close to the retweet network (0.473), meaning the intensity of the connection within each group is almost identical to the retweet network, but it is more disconnected.

The mention network also has many clusters formed around the dominant accounts (see Figure 7). Unlike retweeted users, mentioned users do not need any related content to be mentioned. Thus, the mention network contains 0-degree accounts and they are pushed to the periphery forming the circle around the graph. The degree distribution also shows a similar trend in which more than half of the users (55.89%) have a degree of 1. 39.99% of users have degree of 2 to 10, and 3.31% of users have degree of more than 10. As mentioned above, the mention dataset now contains 0-degree users (0.81%). In the context of Twitter communication, the mention degree tends to be higher than the retweet degree since users can mention multiple names in one tweet, but one specific tweet can only be retweeted once. This partially explains why there are more users with a degree of more than 1.

Table 2 depicts the top 10 users with in-degree and out-degree. We can see that the scale of the in-degree is remarkably larger than the out-degree. Users with high in-degree are pivotal because many other users expose and direct the information to them. Nine of the top 10 in-degree users are also on the list of top 10 out-degree users in the retweet network, which emphasizes their importance. Many high out-degree users in the retweet network are also in the top 10 in-degree list of the mention network. In addition to the users already introduced, we see two other noteworthy accounts: `@MacfarlaneNews`, the CBS News congressional correspondent, and `@January6thCmte`, the official account of the Select Committee to investigate the January 6th attack.

5 DISCUSSION

The Sedition Hunters community was formed on Twitter in response to the FBI seeking the public's help in identifying individuals who engaged in unlawful activities during the U.S. Capitol riot [6]. As the main account `@SeditionHunters` quickly amassed many followers, the FBI maintained limited interactions with the crowd except by constantly publishing new suspects and requesting more information about the posted suspects. Similar constrained behavior was shown in the previous web-sleuthing of the Boston Marathon Bombing [50]. Sedition Hunters would sift through a large amount of data and send related information back to officials or simply publish them on the main Twitter account. This task-focused communication led to notable successes, as numerous arrests and legal proceedings used the investigation results from the community [4, 5]. Previous studies showed that the lack of interaction between investigation officials and the crowd posed a major problem [50, 52]. However, the Sedition Hunters community shows that the value of crowdsourced investigation can be realized even with constrained interaction or communication between law enforcement and the crowd.

From Topic Modeling: To better understand this Twitter community, our first step beyond the preliminary examination was performing a content analysis of its tweets. According to the topic modeling results, the majority of the tweets were about promoting information sharing. Despite a small number of tweets sharing related news and investigation updates, most tweets were trying to expose the specified suspects more broadly by tweeting or retweeting related information, such as the "DoYouKnow" composites.

In the 2013 Boston Marathon Bombing, police asked for help identifying two potential suspects. In contrast, the Sedition Hunters community is constantly investigating dozens of suspects. The results show the importance of keeping track of the overall investigation, especially when numerous suspects were under investigation in parallel and with a such large amount of data. We see the hashtag function of Twitter was useful as it allows targets and topics to be easily documented. For example, in addition to the FBI-generated IDs (e.g., tags with prefix or suffix AFO), the community often creates hashtags of individuals based on their appearances (e.g., `#lordlonghair` in Figure 3). However, as the investigation progressed, we see many tweets with a long list of hashtags indicating different descriptions of the same target. Many targets had two

Top in-degree users				Top out-degree users			
Usernames	**Out-degree**	**In-degree**	**Follower counts**	**Usernames**	**Out-degree**	**In-degree**	**Follower counts**
@SeditionHunters	31,590	0	64,310	@mAnonymized-1	1,546	640	255
@capitiolhunters	15,672	3	34,763	@MacFarlaneNews	822	467	271,076
@ryanjreilly	15,038	15	167,835	@January6thCmte	169	467	594,111
@seditiontrack	8,966	207	70,196	@Anonymized-2	281	352	850
@DomesticTerror2	5,939	119	583	@SeditionSleuth	103	276	495
@Anonymized-3	3,332	118	324	@SlickRockWeb	1205	270	6,951
@FBI	2742	59	3,484,047	@Anonymized-7	575	223	545
@Anonymized-4	2,388	10	8,923	@seditiontrack	8,966	207	70,196
@nate_thayer	2,261	39	6,666	@Anonymized-5	154	199	31,480
@Anonymized-6	2,070	33	709	@Anonymized-9	120	190	285,942

Table 2: Users with top in- out-degree in mention network (full usernames are anonymized except for high-profile accounts)

identification hashtags, with one created by the FBI and the other created by the crowd. This may introduce inefficiency when listed randomly at the end of a tweet.

Design Recommendation 1: Prior research [57] shows hashtags can convey complex meanings and shape the crowd's reaction. Hashtags as documentation tool can be improved with proper curation, such as eliminating co-references and providing structural relations [49].

Design Recommendation 2: We see self-restriction efforts such as intentionally not naming suspects in public. However, there is no explicit monitoring or prevention mechanism. Mistakes with terrible consequences (i.e., [29]) can easily happen again. Moreover, [48] shows that strategic information operations are a critical concern for CSCW researchers. We see that the entire community participates around the few influential accounts, making it fertile ground for strategic information operations such as disinformation.

In addition to hashtags, our results revealed how the crowd uses various platform-independent methods to facilitate their investigation. For example, apart from spreadsheets (similar to other crowdsourcing events [14, 25]), several standalone websites have been built to systematically track the investigation process. As mentioned previously, these websites show that the crowd can incorporate advanced technology such as facial recognition tools into their investigation and create convenient yet powerful interfaces to organize and visualize the entire event. Although the author of the facial recognition website limited the functionality to avoid linking the faces with named identities, considering it may cause serious misidentification, it shows that such advanced technologies are becoming more accessible to crowd workers and they are now able to harness such tools actively or passively [3, 31].

Design Recommendation 3: We also note that almost all websites had a profile page listing the investigation progress of suspects. This seemingly universal desire for progress tracking revealed the lack of such features on online social media platforms and emphasizes the importance of information visibility.

From Social Network Analysis: We used SNA to identify different roles in the Sedition Hunters community. Prior work has shown three types of investigative models: top-down [e.g., 7, 40], bottom-up [e.g., 22, 26], and hybrid [e.g., 52]. The Sedition Hunters community is similar to top-down law enforcement-led investigations with a few key differences. First, the crowd workers are led by a few emergent leaders within the community rather than law enforcement experts. Despite the FBI hashtags appearing in some tweets, almost all tweets are in direct relation to the main `@SeditionHunters` account. The connection between crowd workers and the few dominant accounts is closer compared to law enforcement investigations. Second, the small number of influential accounts such as `@SeditionHunter` and `@seditiontracks` can be viewed as information processors that filter information from others and create composite images. This task is usually performed by law enforcement [52]. Third, unlike traditional law enforcement-led models where information flows unidirectionally from the public to law enforcement, in Sedition Hunters, information mainly flows from influential accounts. As influential accounts act as information processors through at-mentions, they also act as information generators by instigating a significant amount of retweets that increase the spread of information.

Design Recommendation 4: Offline crowdsourced investigations can benefit from redistributing leadership as it divides the work and facilitates better feedback from crowdworkers [30, 52]. We see that the success of the Sedition Hunters community also partially relies on having multiple leading accounts. These accounts dramatically increase information visibility and open more points of contact.

Sensemaking tasks consist of information gathering, creating information representations, developing insights from the representation, and creating knowledge with the insights [39]. The efforts of Sedition Hunters can be viewed as a large-scale collective sensemaking effort where the main goal is to gather information. However, this parallelization of search tasks may lead to individuals having incomplete access to information [45]. Sedition Hunters addressed this concern by incorporating the information-sharing task into their influential accounts. This includes specifying targets, sharing progress on different suspects, and other updates. We argue this is a reason why the community achieved successful results, despite being dispersed, with different groups of low-participation users gathering around several leading accounts.

6 CONCLUSION

In this work, we examine the behavior and structure of the Sedition Hunters community. We uncover three categories of discussions and find that a small number of members played important roles in the community. We also discuss the unique characteristics of the Sedition Hunters community and how they conduct investigations. Crowdsourced intelligence can be a powerful tool for solving real-world investigations, but scrutiny over their methods can highlight how to better harness the power of crowds.

ACKNOWLEDGMENTS

We wish to thank Vikram Mohanty for initial data collection. This work was supported by NSF IIS-1651969.

REFERENCES

[1] 2021. Pride & Prejudice: The Violent Evolution of the Proud Boys. https://ctc.westpoint.edu/pride-prejudice-the-violent-evolution-of-the-proud-boys/
[2] 2021. Sedition Hunters - About Us. https://seditionhunters.org/aboutus/
[3] 2021. This Site Published Every Face From Parler's Capitol Riot Videos. https://www.wired.com/story/faces-of-the-riot-capitol-insurrection-facial-recognition/
[4] 2021. United States of America v. Chase Kevin Allen. www.justice.gov/usao-dc/case-multi-defendant/file/1408341/download
[5] 2021. United States of America v. David Nicholas Dempsey. www.justice.gov/usao-dc/case-multi-defendant/file/1428151/download
[6] 2021. U.S. Capitol Violence — FBI Most Wanted. www.fbi.gov/wanted/capitol-violence
[7] Joelle Alcaidinho, Larry Freil, Taylor Kelly, Kayla Marland, Chunhui Wu, Bradley Wittenbrook, Giancarlo Valentin, and Melody Jackson. 2017. Mobile collaboration for human and canine police explosive detection teams. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 925–933.
[8] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media*, Vol. 3. 361–362.
[9] 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), 361–362. https://ojs.aaai.org/index.php/ICWSM/article/view/13937
[10] Vera Bergengruen and W.J. Hennigan. 2021. The Capitol Attack Was the Most Documented Crime in History. Will That Ensure Justice? https://time.com/5953486/january-capitol-attack-investigation
[11] Vera Bergengruen and W.J. Hennigan. 2021. The Capitol Attack Was the Most Documented Crime in History. Will That Ensure Justice? *Time* (April 2021). time.com/5953486/january-capitol-attack-investigation/
[12] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When Is "Nearest Neighbor" Meaningful?. In *Database Theory — ICDT'99*, Catriel Beeri and Peter Buneman (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 217–235.
[13] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[14] Hansen C Bow, Jonathan R Dattilo, Andrea M Jonas, and Christoph U Lehmann. 2013. A crowdsourcing model for creating preclinical medical education study tools. *Academic Medicine* 88, 6 (2013), 766–770.
[15] R.S. Burt. 1992. *Structural Holes: The Social Structure of Competition*. Harvard University Press. https://books.google.com/books?id=_gjtAAAAMAAJ
[16] Bill Chappell. 2021. Architect Of The Capitol Outlines $30 Million In Damages From Pro-Trump Riot. https://www.npr.org/sections/insurrection-at-the-capitol/2021/02/24/970977612/architect-of-the-capitol-outlines-30-million-in-damages-from-pro-trump-riot
[17] Emily Cochrane and Luke Broadwater. 2021. Capitol Riot Costs Will Exceed $30 Million, Official Tells Congress - The New York Times. *New York Times* (Feb. 2021). www.nytimes.com/2021/02/24/us/politics/capitol-riot-damage.html
[18] Trottier Daniel. 2014. Police and user-led investigations on social media. *Journal of law, information and science* 23, 1 (2014), 75–96.
[19] Steve Doig. 2022. It is difficult, if not impossible, to estimate the size of the crowd that stormed Capitol Hill. https://theconversation.com/it-is-difficult-if-not-impossible-to-estimate-the-size-of-the-crowd-that-stormed-capitol-hill-152889
[20] Roman Egger and Joanne Yu. 2021. Identifying hidden semantic structures in Instagram data: a topic modelling comparison. *Tourism Review* (2021).
[21] Roman Egger and Joanne Yu. 2022. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology* 7 (05 2022). https://doi.org/10.3389/fsoc.2022.886498
[22] Sheena L Erete. 2015. Engaging around neighborhood issues: How online communication affects offline behavior. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1590–1601.
[23] Richard V Ericson and Kevin D Haggerty. 1997. Policing the risk society. (1997).
[24] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *kdd*, Vol. 96. 226–231.
[25] Ramesh Guntha, Sethuraman N Rao, and Avinash Shivdas. 2020. Lessons learned from deploying crowdsourced technology for disaster relief during Kerala floods. *Procedia Computer Science* 171 (2020), 2410–2419.
[26] Aarti Israni, Sheena Erete, and Che L Smith. 2017. Snitches, trolls, and social norms: Unpacking perceptions of social media use for crime prevention. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1193–1209.
[27] George L Kelling and Mark Harrison Moore. 1989. *The evolving strategy of policing*. Number 4. US Department of Justice, Office of Justice Programs, National Institute of Justice.
[28] Mark Konkol. 2021. Trolls Wrongly Accused Retired Firefighter Of Capitol Riot Murder. *Chicago, IL Patch* (Jan. 2021). patch.com/illinois/chicago/trolls-wrongly-accused-retired-firefighter-capitol-riot-murder
[29] Traci G. Lee. 2015. The Real Story of Sunil Tripathi, the Boston Bomber Who Wasn't. https://www.nbcnews.com/news/asian-america/wrongly-accused-boston-bombing-sunil-tripathys-story-now-being-told-n373141
[30] Kurt Luther, Casey Fiesler, and Amy Bruckman. 2013. Redistributing leadership in online creative collaboration. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1007–1022.
[31] Tim Mak. 2021. The FBI Keeps Using Clues From Volunteer Sleuths To Find The Jan. 6 Capitol Rioters. https://www.npr.org/2021/08/18/1028527768/the-fbi-keeps-using-clues-from-volunteer-sleuths-to-find-the-jan-6-capitol-riote
[32] Alice Marwick. 2012. The public domain: Surveillance in everyday life. *Surveillance & Society* 9, 4 (2012), 378–393.
[33] Gary T Marx. 2013. The public as partner? Technology can make us auxiliaries as well as vigilantes. *IEEE Security & Privacy* 11, 5 (2013), 56–61.
[34] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.
[35] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
[36] M. E. J. Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103, 23 (2006), 8577–8582. https://doi.org/10.1073/pnas.0601602103 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0601602103
[37] Johnny Nhan, Laura Huey, and Ryan Broll. 2017. Digilantism: An analysis of crowdsourcing and the Boston marathon bombings. *The British journal of criminology* 57, 2 (2017), 341–361.
[38] Leysia Palen and Sarah Vieweg. 2008. The emergence of online widescale interaction in unexpected events: assistance, alliance & retreat. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 117–126.
[39] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.
[40] Ronald Poelman, Oytun Akman, Stephan Lukosch, and Pieter Jonker. 2012. As if being there: mediated reality for crime scene investigation. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 1267–1276.
[41] Nik Popli and Julia Zorthian. 2022. What Happened to the Jan. 6 Insurrectionists Arrested Since the Capitol Riot. https://time.com/6133336/jan-6-capitol-riot-arrests-sentences/
[42] Ryan J. Reilly. 2021. 'Sedition Hunters': Meet The Online Sleuths Aiding The FBI's Capitol Manhunt. https://www.huffpost.com/entry/sedition-hunters-fbi-capitol-attack-manhunt-online-sleuths_n_60479dd7c5b653040034f749
[43] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
[44] Caroline Rizza, Ângela Guimarães Pereira, and Paula Curvelo. 2014. "Do-it-yourself justice": considerations of social media use in a crisis situation: the case of the 2011 Vancouver riots. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 6, 4 (2014), 42–59.
[45] Daniel M Russell, Gregorio Convertino, Aniket Kittur, Peter Pirolli, and Elizabeth Anne Watkins. 2018. Sensemaking in a senseless world: 2018 workshop abstract. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–7.
[46] Christopher J Schneider and Daniel Trottier. 2013. Social media and the 2011 Vancouver riot. In *40th anniversary of studies in symbolic interaction*. Emerald Group Publishing Limited.
[47] Joshua Smallridge, Philip Wagner, and Justin N Crowl. 2016. Understanding cyber-vigilantism: A conceptual framework. *Journal of Theoretical & Philosophical Criminology* 8, 1 (2016).
[48] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.
[49] Kate Starbird and Jeannie A Stamberger. 2010. Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting.. In *ISCRAM*.
[50] Andrea H Tapia and Nicolas J LaLone. 2014. Crowdsourcing investigations: Crowd participation in identifying the bomb and bomber from the Boston marathon bombing. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 6, 4 (2014), 60–75.

[51] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[52] Sukrit Venkatagiri, Aakash Gautam, and Kurt Luther. 2021. Crowdsolve: Managing tensions in an expert-led crowdsourced investigation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–30.
[53] Sukrit Venkatagiri, Tianjiao Yu, Vikram Mohanty, and Kurt Luther. 2021. Sedition Hunters: Countering Extremism through Collective Action. In *CSCW 2021 Workshop on Addressing Challenges and Opportunities in Online Extremism Research: An Interdisciplinary Perspective*.
[54] James P Walsh and Christopher O'Connor. 2019. Social media and policing: A review of recent research. *Sociology compass* 13, 1 (2019), e12648.
[55] Magdalena Wojcieszak and Diana Mutz. 2009. Online Groups and Political Discourse: Do Online Discussion Spaces Facilitate Exposure to Political Disagreement? *Journal of Communication* 59 (03 2009), 40 – 56. https://doi.org/10.1111/j.1460-2466.2008.01403.x
[56] Elizabeth Yardley, Adam George Thomas Lynes, David Wilson, and Emma Kelly. 2018. What's the deal with 'websleuthing'? News media representations of amateur detectives in networked spaces. *Crime, Media, Culture* 14, 1 (2018), 81–109.
[57] Michele Zappavigna. 2015. Searchable talk: The linguistic functions of hashtags. *Social Semiotics* 25, 3 (2015), 274–291.

A APPENDIX

BERTopic: There are four main components to the BERTopic. First, using Sentence-BERT [43], each document is transformed into a dense vector representation. These vectors make it possible to locate semantically similar words, sentences, or documents within close spatial proximity [21]. The default all-MiniLM-L6-v2 embedding from the Sentence-Transformers package is an all-around model tuned for many use cases. However, other language models can be used for embedding. For example, BERTweet is a good alternative, trained exclusively on tweets. We chose to use the default all-MiniLM-L6-v2 embedding since it produced a lower topic coherence score (-6.06 vs. -6.76).

As data increases in dimensionality, the algorithm requires higher computational power, and more importantly, spatial locality becomes ill-defined [12]. Thus, the next component reduces the dimensions of the embeddings. BERTopic uses Uniform Manifold Approximation and Projection (UMAP) [35] for dimension reduction. Compared with other popular methods such as t-SNE[51], UMAP is able to leverage both local structure and global structure and be explicitly controlled by user-defined parameters. It is also faster as it creates estimations of the high dimensional graph instead of measuring every point. The core idea of UMAP is to construct a high-dimensional graph representation of the data and then optimize a low-dimensional graph that is as structurally similar as possible. In order to build the high-dimensional graph, UMAP uses simplices, which enable us to build a high-dimensional graph robustly. Starting from extending the radius of each point, UMAP connects points where the radii overlap. Choosing the radius is critical as a small radius can lead to trivial isolated clusters, while a large radius will connect everything together. UMAP overcomes this by choosing the radius locally. However, as the distance within the radius increases, the probability of connecting the original points with other points decreases [35].

The third component is hierarchical density-based spatial clustering of applications with noise (HDBSCAN [34]), which groups the reduced embeddings. It is an extension of DBSCAN [24] which hierarchically groups clusters by different densities. It uses a soft-clustering approach, allowing noise to be modeled as outliers.

Finally, topics are extracted with class-based TF-IDF (c-TF-IDF). The idea behind the original TF-IDF is to compare the importance of words between documents by computing the frequency of a word in a given document and also how prevalent the word is given all other documents. c-TF-IDF proposes to treat all documents in a single class as one single document. Important words are computed by measuring how frequent a word is in a class and how prevalent the word is in all different classes. The result would be an importance score for words within a class instead of a document. If we extract the most important words in a cluster, they are the description of a topic.