Crowdsourced Archiving of the January 6th US Capitol Insurrection: An r/DataHoarders Case Study

Edward Miezio Chapman
Marquette University

CROWDSOURCED ARCHIVING OF THE JANUARY 6TH US CAPITOL INSURRECTION: AN R/DATAHOARDERS CASE STUDY

by
Edward Chapman

A Thesis submitted to the Faculty of the Graduate School,
Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Master of Science in Computing

Milwaukee, Wisconsin
December 2021

ABSTRACT
CROWDSOURCED ARCHIVING OF THE JANUARY 6TH US CAPITOL INSURRECTION: AN R/DATAHOARDERS CASE STUDY

Edward Chapman

Marquette University, 2021

The crowdsourced archiving that occurred in the wake of the January 6th US Capitol insurrection exemplifies the potential for agile, collaborative evidence gathering during a crisis situation. This paper studies the r/DataHoarders subcommunity of Reddit and the collective and spontaneous archiving project that users initiated. Users were drawn to the thread out of a desire to contribute to law enforcement efforts, enact punitive justice upon the rioters, engage in public discourse, and preserve information for posterity. They did this by gathering and preserving social media evidence that may have otherwise been lost. I discovered that this constituted a crowdsourced archive of potential legal and historical significance. This paper also overviews the ethical implications of related practices such as doxing and open-source intelligence gathering (OSINT).

ACKNOWLEDGMENTS

Edward Chapman

I am grateful for the guidance and support of my advisor, Dr. Michael Zimmer, the community and kinship of my fellow graduate students in the Social and Ethical Computing Lab at Marquette, and especially for the loving encouragement of my darling partner, Elise McArdle.

TABLE OF CONTENTS

INTRODUCTION

On January 6th 2021, thousands of right-wing demonstrators stormed the US Capitol in an attempt to disrupt the certification of the 2020 presidential election results. Demonstrators were upset over the projected loss of incumbent candidate Donald Trump and had rallied themselves in online forums and communities. The demonstrators were able to infiltrate the security perimeter of the building and access the congressional halls where the certification process was being held. Multiple fatalities, both of police and demonstrators, occurred during the event. Ultimately, the results of the presidential election were certified, removing Trump from office. The sudden and violent nature of the demonstration shocked viewers in the US and abroad. It has since been labeled an insurrection and possible attempted coup. Criminal investigations have targeted over 600 participants in the riot and continue to this day.

Beyond the physical violence and mayhem at the Capitol, the January 6th insurrection was notable for the prevalence of self-recorded photos and videos made by the rioters during the demonstration. An explosion of digital content was posted to social media platforms by the rioters, documenting their identities and participation in the illegal demonstration. As the influx of highly polarizing content made its way across social media platforms, other users rallied to archive and analyze the material for purposes of future accountability. Projects such as *Faces of the Riot* and *homegrownterrorists* attempted to match names with faces from primary material collected from the insurrection. Unbeknownst to many of the rioters, the incriminating and personally identifiable nature of their social media postings left them vulnerable to

personal and professional humiliation. Self-appointed online vigilantes sought to hasten this process and aid in law enforcement's effort to bring accountability to the rioters.

Among these projects, the Reddit community of r/DataHoarders initiated a crowdsourced archive project in which volunteer participants collected and archived social media evidence from the Capitol insurrection. Entitled "MEGATHREAD: Archiving the Capitol Hill Riots," the discussion thread attracted over one thousand participants and produced a substantial archive of incriminating and historically significant material. The public nature of the Reddit platform preserved a record of the Redditor's interactions and deliberations during the crowdsourced archiving project, making it a rich source of data regarding digital collaboration and the ethics of online investigations.

Open-source intelligence (OSINT) and the practice of doxing, targeted deanonymization, are increasingly prevalent concepts in discourse surrounding online data ethics. Both involve the linkage of disparate, publicly accessible information to create a dossier on a topic or subject. Upon discovering the r/DataHoarders archive project, I noticed significant overlap with concepts of OSINT, doxing, and crowdsourced archival work. OSINT and doxing can be used for accountability but run the risk for harm, harassment, and arguably unethical public shaming. I was curious if the use of these tactics in this thread constituted a witch hunt or frenzy, as observed in prior Reddit investigations, such as the hunt for the Boston Marathon bombers in r/FindBostonBombers. Additionally, I sought to discover to what extent the Reddit thread provided a beneficial repository of information more aligned with the democratizing potential of collaborative online investigations and OSINT. My research also focused on

unearthing the stated motivations of Redditors in the thread in order to analyze the different factions, political ideals, and levels of participation at play.

Using the Reddit API and Python scripting, I collected all comments posted to the r/DataHoarders archive thread, including comment text, upvote score, timestamps, and usernames. I imported the dataset into qualitative coding software and performed open qualitative coding to determine the types of activities Redditors engaged with in the archive thread. I merged and split code labels as needed throughout the coding process, arriving at twelve codes for distinct user activities, such as evidence submission and archive torrent seeding. The resulting code data provided insight into the different roles that users performed during the collaborative archiving event. An examination of existing literature on crowdsourced investigations, doxing, and OSINT provided a framework to understand the activities and ethical risks present in the r/DataHoarders archive project.

My findings indicate that while ethical transgressions and potential for misuse were found in the r/DataHoarders archive project, the bulk of activity was focused on practical matters of digital archiving, namely the identification, acquisition, and preservation of evidence related to the January 6th Capitol insurrection. Calls for doxing and other punitive measures were present in the archive thread, but these initiatives were most often proposed by outsiders to thread who commented once and did not return to the archive project. Rather than a crowdsourced witch hunt, this project was more in line with a community archiving project conducted by amateur archivists. The ethical risks and implications of this framing are still notable, as archiving bears responsibility for the preservation of evidence in the shaping of collective memory, especially concerning traumatic events such as human rights violations. The practice of archiving contemporary

social movements is perhaps the best place to begin when considering how to mitigate the ethical harms of crowdsourced archiving projects such as the r/DataHoarders thread in the future.

BACKGROUND

The physical breach of the US Capitol was highly significant from a national security perspective. My research positions the events of January 6th as a primarily digital phenomenon resulting in a unique digital ecosystem of politically inflammable and criminally consequential content. Tracing the lifespan of these digital artifacts reveals tensions and resonances among contemporary issues such as digital surveillance, national security, content moderation, privacy ethics, and internet vigilantism.

2.1 Digital Ecosystem

Images and videos of the Capitol insurrection depict a startling amount of violence and property destruction. Participants, variously armed with formal and improvised weaponry, can be seen assaulting Capitol police officers and smashing windows. But for all of the egregious conduct of the Capitol rioters, the unifying activity of the day was digital content creation. IPhones, more than mace or molotovs, were the weapon of choice and can be seen in nearly every hand in footage of the riots. Whether streaming, tweeting, or taking selfies to remember the moment, the participants of the Capitol insurrection generated an immense stream of real-time digital content.

This content connected the participants at the Capitol to the online movements and communities that had brought them there: QAnon conspiracy theories, COVID-deniers, MAGA Twitter and white nationalist movements and subcultures. The Capitol insurrection was not a spontaneous event and in fact had been planned for weeks on platforms such as Parlor and Twitter [1]. In this way, the January 6th Capitol insurrection was not a physical climax to an organic, unified social movement, but rather a temporary physical convergence of multiple online communities and subcultures. Their occurrence

at the peak of the 2021 COVID-19 pandemic further centers the Capitol Insurrection as a digital phenomenon as it happened at a time when many formerly physical activities had shifted to digital formats. Thus, it is important to understand the creation, lifespan, and eventual archiving or removal of this content in order to properly contextualize the Capitol Insurrection and its significance and consequences.

2.1.1 Content Creation

Content created at the Capitol varied from keepsake selfie to monetized livestream. Mainstream newscasters vied with fringe political streamers for real-time coverage of the fray. Family photos not so different from a regular Capitol tour mixed with braggadocious broadcasts proclaiming the deeds of the day to followers back home. Masks were few and far between, and faces, names, and hometowns were shared freely on videos that would later be presented as criminal evidence. An analysis of the archives which later collected January 6th content can give insight into where digital content was originally published during the insurrection. Reddit's "Archiving the Capitol Hill Riot" project produced a 1TB archive of videos and images. Discounting unsorted and miscellaneous content, figure 2.1 shows the amount and volume of January 6th content collected from each source: YouTube, Parler, and Twitter stand out as the biggest contributors of archived content. Sources such as Facebook and Instagram may be underrepresented in the archive due to the difficulty of accessing and scraping user content from their platforms.

2.1.2 Content Moderation and Removal

The political potency and criminal liability of Capitol insurrection content ensured that photos and videos created that day would not just fade into the background. A

combination of content moderation and user action put an expiration date on much of the evidence from January 6th. The images and videos outraged and inspired viewers in equal measure. The political potency and criminal liability of this content put pressure on social media platforms and challenged their claims of transparency and readiness in content moderation.

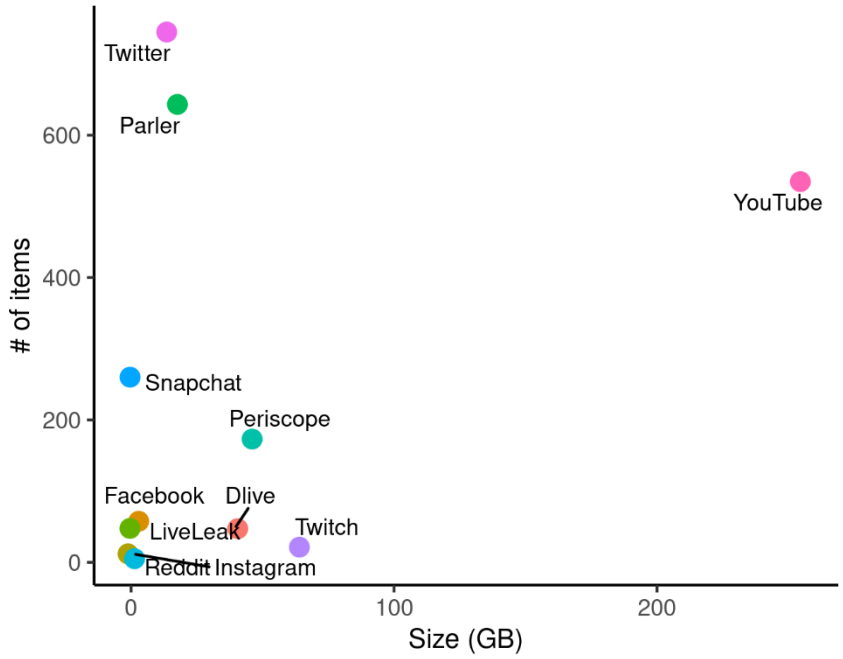

Figure 2.1: **Content sources archived by r/DataHoarders**

2.1.2 YouTube, Facebook, and Instagram

YouTube livestreams are a lucrative source of viewers and income through tipping mechanisms like the “super chat” and “super sticker” features. Viewers can also be directed to third party platforms such as PayPal and Patreon. On January 6th, far-right streams from the Capitol could be found alongside mainstream news feeds. YouTube’s policy forbids monetization of videos encouraging or inciting violence. YouTube made a statement on Twitter the following day warning that election misinformation would result

in a “strike” against the publishing accounts [2]. Some streams and videos encouraging violence at the Capitol were removed by the platform. On others, YouTube added a warning label stating “the electoral college has confirmed Joe Biden as president-elect” [3]. Facebook and Instagram did not limit searches or provide warning labels for “#stopthesteal” until crowds had dispersed from the Capitol [3]. However, their search features do not sort content for newness, making it difficult to find live or real-time content on the day of the riot.

2.1.2 Removal by Users

Content removal occurred most frequently at the discretion of the content creators themselves, who may have realized the incriminating nature of their posts. In my analysis of the Reddit insurrection archive, volunteer archivists were frequently thwarted by content which had been removed by its original poster. Archivists were often motivated to prioritize the most incriminating and high-profile content available due to the risk of removal.

2.2 Content Afterlife

The political and criminal nature of content created at the Capitol on January 6th made it valuable and sought after. The photos and video of the riot would find many uses beyond what their creators had intended. As users and platforms rendered Capitol insurrection content inaccessible, operations were underway to collect and preserve the incendiary media. These archival projects are noteworthy for their spontaneity and effectiveness. The diversity of groups and individuals interested in the material and their divergent goals and methods raised ethical questions on the use of facial imagery, leaked private data, and social movement documents.

2.2.1 Archive Projects

The first projects to appear were urgent attempts to collect and preserve material related to the Capitol insurrection. Three main projects worked in parallel to achieve this goal. The Netherlands-based independent investigatory organization Bellingcat called for photo and video submissions via Google Forms and Google Sheets [4][5]. They instructed volunteers to prioritize livestreams and to ignore low-resolution media and screenshots of videos. Bellingcat ran a similar crowdsourced data-collection project following the 2017 white supremacist rally in Charlottesville, Virginia [6]. Intelligence X, a search engine and data archive company based in the Czech Republic, also called for video and photo evidence of the Capitol Insurrection [7]. They provided credentials to a secure public FTP server allowing tech-savvy users to transfer media files to the project. The archives were shared in an easily-browsable format [8].

2.2.1 r/DataHoarders Archive

In addition to the Intelligence X and Bellingcat projects, users on Reddit formed a thread in the r/DataHoarders community to collect and archive footage of the riots. Entitled “MEGATHREAD: Archiving the Capitol Hill Riots,” the thread grew to include over 2,500 comments and produced a 1 terabyte media archive after four weeks of activity [9][10]. The r/DataHoarders thread formed spontaneously and was conducted in a decentralized, non-hierarchical manner, unlike the Bellingcat and Intelligence X archive projects. The r/DataHoarders insurrection project provided no unifying reason for data collection nor intended usage of the archive. Users arrived with their own motivations and guided the project from diverse and contradictory perspectives. User comment and voting data for the thread provides a detailed and valuable record of the collaborative

crowdsourcing efforts of the r/DataHoarders archive. While Bellingcat and Intelligence X also relied on crowdsourced submissions, no such record of participation is publicly accessible. All three archives eventually subsumed large parts of the others, producing similarly sized collections. Thus, the Reddit megathread provides the most detailed, accessible record of crowdsourced archiving following the January 6th insurrection. It will serve as the main target of analysis for this research.

2.2.2 Doxing

As crowdsourced data collection projects secured footage of the riots, other projects and individuals sought to make use of the evidence for more partisan means. The preponderance of unmasked faces in images of the riots left the individuals depicted vulnerable to an online revenge tactic known as doxing. Doxing describes the public dissemination of an individual's personal information as a means to deanonymize, intimidate, or humiliate them [11]. The term originated in 1990's hacker culture, where "dropping dox" on a rival could reveal their offline identity and exile them from the community [12]. Today, finding someone's social media accounts or legal name can be enough to bring about an avalanche of online harassment and offline consequences, depending on the target. For their polarizing political views and brazen criminality, participants in the Capitol insurrection stood to lose much from being identified. The sheer quantity of media produced at the Capitol that day left rioters particularly vulnerable. Individuals and collaborative projects knowingly seized on this potential, weaponizing the images against the rioters. The Instagram account, *homegrownterrorists* and the *Faces of the Riot* website stand out as well-organized and influential projects which sought to trace and identify participants at the Capitol insurrection. Elsewhere,

individuals on Twitter and other social media platforms conducted doxing operations of their own, gathering information and identifying suspected participants in their free time. These efforts were more disparate and spontaneous than the projects detailed below.

Figure 2.3: **Calls for doxing** [16]

2.2.2 Faces of the Riot

Faces of the Riot is a website designed to assist in the identification of participants in the Capitol insurrection [13]. The website's creators applied facial recognition technology to scraped videos from the Capitol insurrection to extract and deduplicate every face found in the dataset [14]. The result is something of a school yearbook for right-wing insurrectionists. The viewer is greeted by an extensive grid of zoomed in, cropped faces. Clicking on a face brings up the video it was sourced from, as well as any other videos the face was identified in. The website is an invaluable resource for anyone seeking to track an individual's actions throughout the insurrection, as it links

an individual to multiple video sources. The creators insist that the website should only be used to provide tips to the FBI. The website begins with the disclaimer: “DO NOT ATTEMPT YOUR OWN INVESTIGATION INTO ANYONE SHOWN ON THIS WEBSITE. REPORT THEM TO THE FBI USING THIS LINK INSTEAD” [13]. It is unclear what tools or methodology was used to identify and deduplicate the faces. The creators claim to have spent about five hours removing children and “non-rioters” from the website [14].

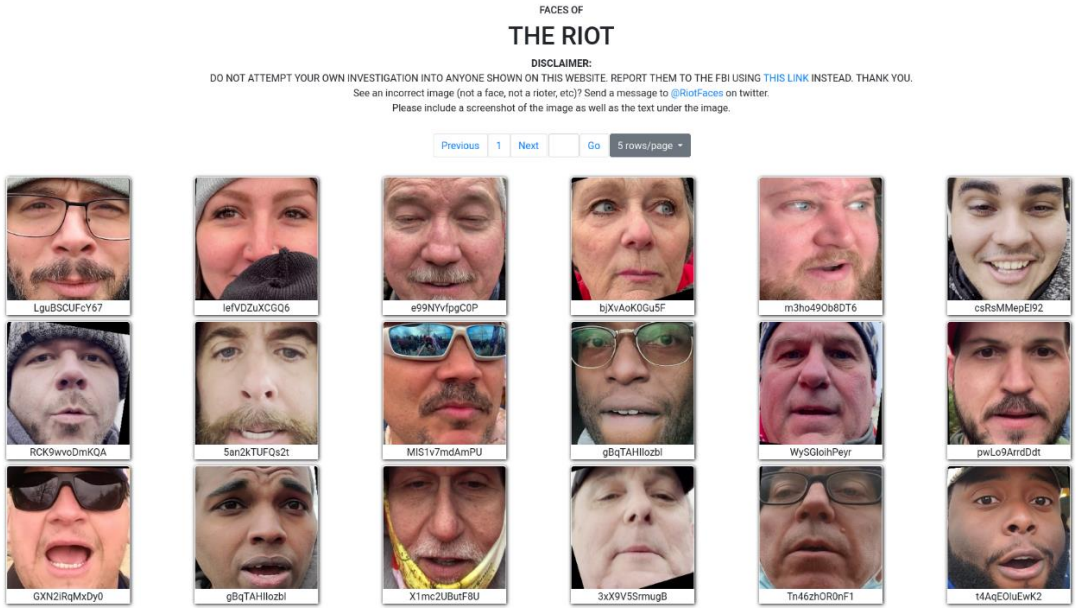

Figure 2.2: **Faces of the Riot** [13]

2.2.2 homegrownterrorists

Another project, *homegrownterrorists* on Instagram, actively seeks to identify and disseminate personal information on suspected participants in the Capitol insurrection. The social media account publishes content related to the identification of Capitol insurrectionists and updates on their legal proceedings to an audience of 366,000

followers [15]. Their bio reads: “It’s like Gossip Girl, but for democracy,” positioning themselves as a mix between celebrity news and activism. And indeed, many of the capitol rioters have taken on a villainous celebrity-like status as their names and faces have circulated since the riot.

2.2.2 Results of Doxing

As a consequence of social media doxing, identified individuals lost their anonymity and faced repercussions in their personal and professional lives. These screenshots were found on Twitter, where users considered the firing of Capitol rioters to be a success.

Figure 2.4: **Example of doxing found on Twitter** [17]

2.2.3 Bellingcat

Bellingcat, one of the organizations behind the crowdsourced archiving of Capitol insurrection media, also used evidence from their January 6th archive to reconstruct the life of Ashli Babbit leading up to the moment she was killed by Capitol security at the insurrection [19]. Investigators at Bellingcat compiled publicly available information about Babbit such as her employment history, social media posting, and comments made by friends and family in the news. They also scoured photos and videos of the Capitol

insurrection to identify moments and locations where she could be spotted. Together, these strands of information create a picture of both her ideological journey to the January 6th riot as well as her actual footsteps leading up to her death. This is significant because it shows how publicly available information can be synthesized to produce a contextual narrative, sometimes with startling accuracy.

2.2.4 Law Enforcement

Shortly after the January 6th insurrection, the Washington D.C. Metropolitan Police Department released a PDF document calling for evidence and identities of several suspected rioters [20]. The document included images and descriptions of suspects, as well as the times and locations of their involvement in the riot. The document was titled "Persons of Interest in Unrest-Related Offenses" and featured 26 individuals charged with unlawful entry. This list of suspects would soon grow to hundreds of individuals as multiple criminal investigations into the January 6th riot intensified. As of October, 2021, Capitol insurrectionists represented 630 federal defendants, including 100 guilty pleas [21].

LITERATURE REVIEW

3.1 r/FindBostonBombers

The 2013 Boston Marathon Bombing and associated subreddit, r/FindBostonBombers provide a historical comparison to the r/DataHoarders Capitol insurrection archive. The event set precedent as the first time a crowdsourced investigation ran parallel to a police investigation for a US terrorist attack [22]. Reddit's ultimately unsuccessful search for the Boston Marathon bombers was notable for the analytic capabilities it mustered from crowdsourced volunteers. It also served as a warning to future endeavors for the misidentification of suspected bombers.

r/FindBostonBombers was created independently of law enforcement in order to assist the FBI with crowdsourced information collection and analysis [23] [22]. Researchers have framed their efforts as an example of crowdsourced surveillance and digilantism (from cyber-vigilantism) [24] [22]. The subreddit served as a hub for news and analysis regarding the bombings. Redditors conducted a "virtual crime scene investigation" by reconstructing the events of the bombing and manhunt in parallel to law enforcement [22]. Users linked photos and videos of the bombing spatially and temporally using EXIF metadata and geolocation [24]. Images of crowds were scoured to identify suspects who matched police descriptions. Individuals looking away from the race or carrying large bags were traced across photos and videos [24].

Analysis and accusations born from r/FindBostonBombers found their way into mainstream new sources. User-created photos were shared by the New York Post and credited to law enforcement [23]. Before the subreddit was shut down by Reddit administrators, multiple individuals were outed as suspects. None of the individuals

named had any relation to the real perpetrators, the Tsarnaev brothers. In fact, none of the photos analyzed in r/FindBostonBombers even contained the Tsarnaev brothers [24].

3.2 Doxing and Crowdsourced Investigations

Doxing describes the public dissemination of an individual's personal information as a means to deanonymize, intimidate, or humiliate them [11]. The term originated in 1990's hacker culture, where "dropping dox" on a rival could reveal their offline identity and exile them from the community [12]. The risk of doxing is ever present in crowdsourced investigations such as the r/DataHoarders insurrection archive, where information collection can be used to enact a witch hunt. In this section, I will contrast the harms and benefits of doxing according to two sources. In "Doxing or Deliberative Democracy: The case of r/Serial," Buozis explores the democratizing and demonizing potential of crowdsourced investigations on Reddit [25]. Douglas' "Doxing: a conceptual analysis" provides a philosophical overview and definition of different forms of doxing [11].

3.2.1 For Doxing

Deliberative digital democracy is a form of online engagement in which rational public discourse produces democratic knowledge [25]. Buozis views crowdsourced evidence gathering such as the r/Serial subreddit as having the potential to enact participatory civic engagement and discoursed-based knowledge production [25]. To him, Reddit is an ideal site for deliberative digital democracy because it melds "the power of crowdsourcing" with "the possibilities for testing hypothetical arguments in public" [25].

This form of crowdsourced online investigation is beneficial because it represents an effort to better societal institutions through participatory civic engagement. Deliberate digital democracy arises out of citizen-led research, discussion, debate, and fact-checking. Through this process, citizens envision ways in which current institutions of governance can be improved. Buozis points out that this form of crowdsourcing avoids problems that traditional media may face in establishing rational discourse. Traditional journalistic media is bounded by the "spacial and temporal demands of the market," something that participatory communities such as Reddit have the potential to sidestep [25]. Additionally, the narratives of traditional media exist within a limited lifetime of the news cycle, and mechanisms for audience feedback and participation are non-existent or ineffective. In contrast, digital communities such as Reddit provide a space where audience engagement can exist outside of these boundaries. Doxing has a complicated relationship with deliberative digital democracy. While it may exist in this realm, Buozis later argues that it also has the potential to thwart democracy, as I discuss later.

Douglas describes doxing as "documentary evidence of identity knowledge" [11]. His paper, "Doxing: a conceptual analysis", positions doxing as a complex side effect of public discourse surrounding identity knowledge. Although Douglas points out many cons of doxing, he does define some of its values as well. Doxing has the potential to establish accountability for wrongdoing or deception, as well as to combat hate speech and other forms of toxic communications [11]. Douglas views doxing as a form of whistleblowing, provided that the proper precautions and discretion have been taken into account. In all cases, doxing must be justified by a "compelling public interest" [11].

Doxing cannot be indiscriminate, and the information released must only meet the minimum amount required to establish wrongdoing.

3.2.2 Against Doxing

Although Douglas does concede these conditional pros, he also provides ample evidence of the ethical problems and risks attached to doxing as a practice. Vindictive or indiscriminate doxing can be "a means of intimidation and incitement to cause harm" [11]. The careful balance between justified whistleblowing and public humiliation is easily lost. Increasing a subject's physical accessibility through the release of their home address or workplace puts them at a greater risk of physical harm. Humiliating and objectifying a subject through doxing can exclude them from participating in "social, political, and public activity" [11] contrary to the aims of deliberative digital democracy. Thus, in the context of discourse-based knowledge production, doxing can silence and alienate "minority or dissenting views" [11]. Doxing has the potential to enact the "private enforcement of public laws and standards," and may lead to vigilantism [11]. At its best, doxing may be a tool for accountability for those who cause harm through anonymity. In other cases, it is simply a means to intimidate those with unpopular opinions.

Buozis mentions doxing as well, although it plays a smaller role in his analysis of crowdsourced investigations. Buozis views the cons of doxing through the lens of deliberative digital democracy and its failure to adequately foster a fair discourse. The social "frenzy" involved in cases of crowdsourced doxing create a form of witch hunt, wherein "messages of control" can actually "blind participants in their frenetic search for information" [25]. In this way, the benefit of crowdsourced democracy is stifled because

the utility of proffered information is compromised. Buozis argues that the actions of doxers “reveal a certain arrogance and impatience for information that contribute little to the production of real democratic discourse” (emphasis mine) [25]. Although the practice of doxing may arise out of a desire to contribute to democratic discourse, it can actually be counterproductive to its creation. One extreme example of inaccurate evidence presentation and social frenzy colliding in a harmful way can be observed in Reddit’s hunt for the Boston Marathon bombers. Based on “pure speculation,” users misidentified a number of suspects, including Sunil Tripathi, a missing student who was later found to have commit suicide prior to the bombing [22]. The sudden influx of threats and harassment to Tripathi’s family only magnified their pain. In this event, what may have begun with good intentions on the parts of Redditors quickly escalated to a situation wherein doxing perpetuated harm.

3.2.3 Doxing Conclusion

Neither paper outright condemns nor endorses doxing. Buozis blames the process rather than the concept, stating that doxing can contribute to inaccurate fact gathering by limiting the transparency of the democratic discourse. Doxing may be a logical side effect of the nature of crowdsourced investigations, especially ones that address politically transgressive topics. However, Buozis expresses most concern when the frenzied nature of doxing limits the beneficial potential for crowdsourced investigations to create democratic discourse. Douglas views doxing as a potential mechanism for accountability, especially for those who use anonymity to cause harm. He also acknowledges the ethical challenges presented by the complex nature of doxing, such as the risks of revealing too

much information about a target, the use of doxing for humiliation, and the danger of a collective vigilante mindset that silences dissenting views.

3.3 Open Source Intelligence (OSINT)

Open source intelligence (OSINT) is the iterative process of gathering, analyzing, and synthesizing intelligence about a target using open data sources [26][27]. Open data sources include news media, social media, public government data, and commercial data which are publicly accessible but may be difficult to locate. OSINT originated as a form of military intelligence in the Second World War [28]. It has since found usage in law enforcement, information security and investigative journalism [27]. In this section, I will review three sources on OSINT from three different disciplines. In "Intelligence in the Internet Age," Glassman and Kang view OSINT as a way to understand internet-mediated human intelligence [28]. Stottlemyre's "HUMINT, OSINT, or Something New?" positions OSINT in relation to other forms of military intelligence [26]. Finally, "The Not Yet Exploited Goldmine of OSINT" by Pastor-Galindo et. al provides an overview of current OSINT practices from the perspective of cybersecurity [27].

Pastor-Galindo et. al describe OSINT as a three-part, iterative workflow consisting of data collection, data analysis, and knowledge extraction [27]. In data collection, the OSINT investigator uses open data sources to expand a dataset around a particular target. Starting with a single piece of information– perhaps a username, email, or IP address, the investigator attempts to identify other data points that refer to the same target. Each piece of information increases the possibility of finding additional items. In the analysis phase, the investigator organizes and connects data points from the collection phase to build a holistic profile of the target. Investigators may rely on natural language

processing, geospatial mapping, and social network analysis to plot the target's attributes, relations, and activities. In knowledge extraction, the investigator applies machine learning techniques to the dataset to detect patterns, correlations, and predict future behavior. Knowledge extraction infers "abstract, complex and juicy issues about the target that are not explicitly published on the Internet" [27].

In "Intelligence in the Internet Age: The emergence and evolution of Open Source Intelligence (OSINT)," Glassman and Kang use a human- and community-centered approach to understanding OSINT. They identify the benefits of OSINT in offering "ways to create cooperative, open, problem solving communities" [28]. Through the use of OSINT, "it is possible and productive to look for and make connections that are not immediately apparent or are even (initially) counter-intuitive" [28]. The community to which Glassman and Kang refers to is not limited to national security organizations or other pre-established loci of investigation, but also ad-hoc, participant-run online networks. They view OSINT as an extension of the open-source software movement which espouses the value of nonhierarchical information collaboration.

In "HUMINT, OSINT, or Something New? Defining Crowdsourced Intelligence," Stottlemyre approaches the question of OSINT from a military intelligence background. He differentiates between OSINT and what he terms "crowdsourced intelligence," but also finds commonality between the two. He defines crowdsourced intelligence as "collecting information that was originally gathered by humans" [26]. He elaborates on this as the collection of information, acquired in a second-hand way, for purpose of aiding national security decision-makers [26]. This differs from HUMINT, the acquisition of information through direct interview or other firsthand methodology, and

OSINT, the secondhand collection of intelligence data that is not acquired by investigators for the explicit purpose of aiding national security. Stottlemyre's perspective rests firmly within the discipline of military intelligence. It does not support my research because the Redditors who acted in the r/DataHoarders insurrection archive did not meet the criteria for "crowdsourced intelligence" as defined by Stottlemyre. Stottlemyre states that crowdsourcing intelligence must be enacted on "behalf of a national security organization" and involve questions issued by national security directly to a group of potential sources [26]. The Redditors do not meet this criteria because they were not in direct communication with the FBI or other law enforcement organizations. Although they may have viewed themselves as procuring "intelligence" as it is defined in national security, they were given no verbal or written direction from authorities regarding their research.

METHODS

In this section I will describe my choice and acquisition of the r/DataHoarders Capitol insurrection archive dataset as well as the use of the qualitative coding as a methodological approach.

4.1 Dataset

My dataset is a Reddit comment thread entitled "MEGATHREAD: Archiving the Capitol Hill Riots" posted to the r/DataHoarders community on January 6th, 2021 at 3:13pm CT. The thread calls for volunteers to accumulate and archive photo and video evidence of the January 6th Capitol insurrection. The thread was active for 44 days before it was locked by community moderators on February 18th due to excessive arguing and the completion of its stated goal. The dataset contains 2,788 comments posted by 1,311 different Redditors. The activities of the thread culminated in the production of a 1TB archive of Capitol Insurrection evidence hosted by MEGA. The r/DataHoarders archive thread was one of three concurrent crowdsourced archiving projects that responded to the January 6th insurrection. However, it is the only project that left a public record of participation and thus was my choice for analysis.

4.1.1 Data Collection

Reddit provides a public API for programmatic access to community, user, and comment data. The PRAW Python package enables users to interact with the Reddit API using Python. I wrote a Python script using PRAW to retrieve comment text and metadata from the r/DataHoarders megathread. I saved the contents of the thread to a CSV file. Metadata fields included comment author, comment timestamp, comment text

in HTML and plain text, permalink URL, identifiers for comment's position in the thread tree, and upvote and downvote counts for the comment.

4.1.2 Data Ethics

Reddit content is highly accessible. Discounting a minority of private subreddits, most user discussion can be viewed without a Reddit account. A user's comment history can be tracked across multiple subreddits. Reddit accounts have minimal sign-up requirements– users need only supply a unique username and password to create an account and post comments. It is commonplace for a user to possess multiple Reddit accounts and switch between them fluidly. While a Reddit account may not map to a user's offline identity to the same degree as a Facebook account, I still choose to treat unique user information cautiously as an ethical consideration. My research does not publish Redditor's usernames or inspect their account activity outside of the r/DataHoarders archive thread. I do share unedited user comments, which can be searched for and linked to the posting accounts. I consider this a concession to the predominately public nature of the platform.

4.1.3 Data Limitations

Reddit's userbase skews young and male. Although Reddit accounts do not specify demographic information, site administrators estimate that 56% of users are male and 58% are between the ages of 18 and 34 [29]. The lack of account demographic data makes it difficult to compare my dataset to the demographic profile of the platform as a whole. However, it can be assumed that the demographic profiles are similar because the megathread was briefly featured at the top of r/all, the most visible and accessible content on the platform.

4.2 Qualitative Coding

To analyze and understand the material in the r/DataHoarders insurrection archive project, I applied qualitative coding to the comment text in the thread. Coding is an approach to qualitative data analysis used to organize and extract meaning from textual data. Brief segments of the text are assigned descriptive labels known as codes [30]. The list of codes used for a project is developed throughout the coding process. Code labels may be derived from phrases found in the text (in-vivo) or from knowledge of existing research in the field [31]. For this project, I used qualitative coding to categorize different forms of participation in the r/DataHoarders insurrection archive.

I began by drawing a random sample of 100 comments from the dataset. I used this sample to develop an initial set of codes. My codes focused on the types of activity conducted by Redditors in the thread: *evidence* for sharing URLs to photos and videos of the riots; *strategy* for posing questions or giving direction regarding the general archiving process; *analysis* for discussion, speculation, and reconstruction of events of the January 6th insurrection; *arguing* for hostile or antagonistic political and personal discussion. As I worked through the full dataset, I added codes for salient topics in my research such as *doxing* for encouraging the exposure of rioters' personally identifiable information and *fbi* for calls to submit evidence to law enforcement agencies. I also removed or merged codes that were needlessly specific, such as distinctions between requests for technical help and responses to these requests, both of which were merged into *strategy*. I did not code every comment. I skipped comments which were deleted, comments directed at or posted by bots, very brief comments, and comments whose contents were contextually irrelevant to the aims of the thread.

RESULTS

5.1 Descriptive Results

In this section I will share the results of my qualitative coding and describe the themes and patterns present in the dataset. Submitting URL links to photo and video evidence of the Capitol insurrection was by far the dominant activity in the r/DataHoarders archive project. While many of the submitted URLs were duplicates of each other, submitting evidence was still the most obvious and accessible way to participate in the thread and contribute productively. Strategy comments were the next most common form of thread activity. These comments directed Redditors to potential sources to investigate, gave technical advice, and defined the scope of the archive. Beyond evidence and strategy posts, other activities occurred at comparable rates. Tangential or counterproductive contributions, such as arguing, doxing, and analysis of insurrection media did not outweigh on-topic activities focused on the collection and archiving of photos and videos.

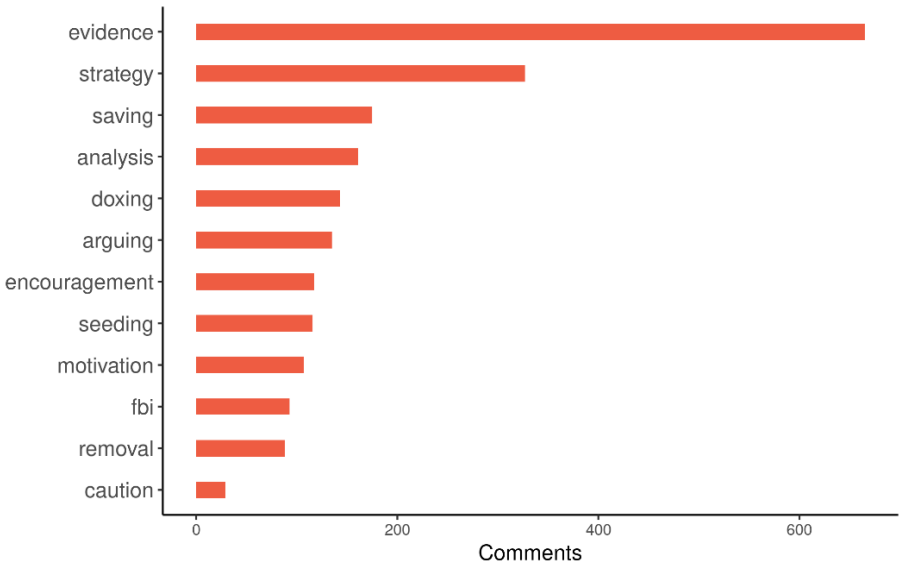

Figure 5.1: **Thread activities**

5.1.1 Time Periods

The megathread was posted at 3:13pm CT on Wednesday, January 6th, 2021. The first few hours were the most active period, peaking at 232 comments per hour between 4 and 5pm. Commenting slowed down after the first afternoon, but the thread remained highly active (100+ comments per day) through Sunday, January 10th. Redditors continued to engage with the thread for the next week, posting an average of 30 comments a day through Saturday, January 16th. At this point, commenting slowed to a trickle of 1 or 2 comments per day until commenting was disabled by admins on February 18th.

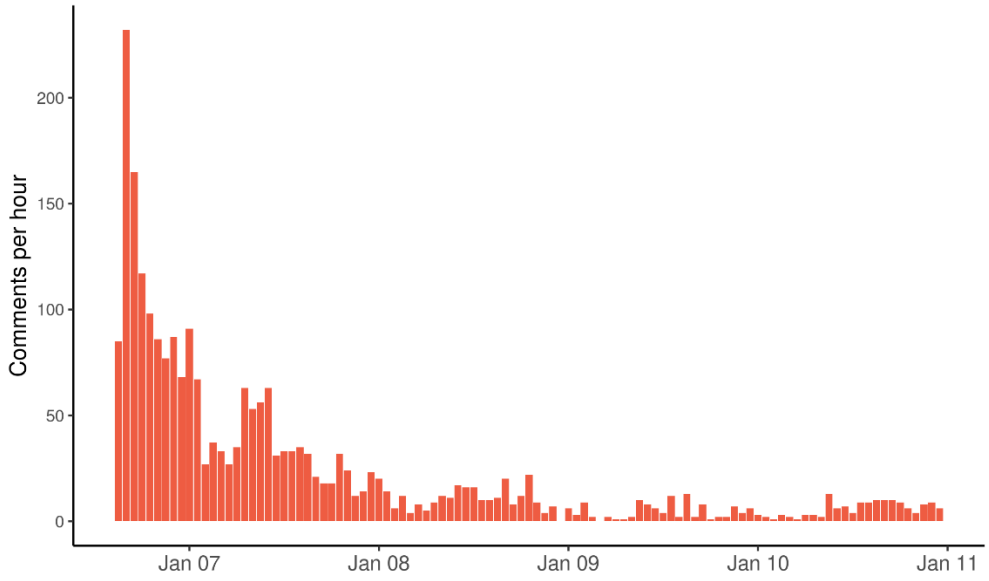

Figure 5.2: **Comments per hour during peak period**

5.1.2 User Info

A total of 1,311 Redditors commented 2,683 times on the megathread during its lifetime. While most commentors (68.6% of accounts) posted only once, repeat commenters accounted for the majority (66.5%) of the thread's total comments. Redditors in the thread can be grouped into three categories based on the number of times

they commented: low for single comment accounts, medium for accounts posting two to five times, and high for accounts posting six or more times in the thread. Table 5.1 shows the number of accounts in each category and the total number of comments each group posted.

Comments per user	*Category*	*Total accounts*	*Total comments*
1	low	899	899
2-5	medium	352	936
6+	high	60	798

Table 5.1: **User engagement levels**

Each user engagement category accounted for roughly one third of total coded comments. Figure 5.3 shows the number of comments each user engagement category contributed to various activities in the thread. While some activities such as seeding were evenly shared among users, most activities display stratified contributions across user engagement levels.

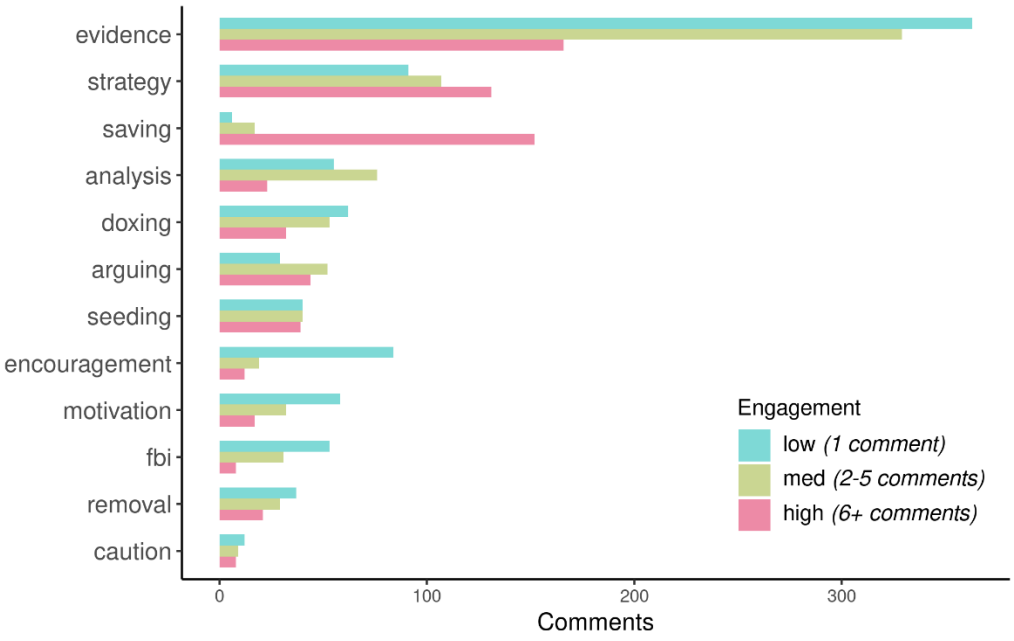

Figure 5.3: **Thread activities by user engagement level**

5.2 URL Analysis Submitting

URL links to photos and videos of the Capitol insurrection was the most common form of participation in the megathread. Redditors collected 1,354 URLs in total. 907 of these came from social media platforms such as YouTube, Facebook, and Twitter. The rest were links to news broadcasters, Google Drive accounts, and miscellaneous web sites. Of the 907 social media submissions, 713 were unique and 194 were duplicates of a previous submission.

By default, the Reddit interface displays the comments with the highest voting score. Especially in long and active threads such as the r/DataHoarders insurrection archive project, low-scoring comments are filtered out of view and have little chance of attracting interaction from viewers. This platform affordance may be counterproductive in the context of crowdsourced data collection, as data submissions may be buried before they can be accounted for. The following chart shows the sequential postings of duplicated URL submissions and the number of votes they received. The submission with the highest number of votes is highlighted in red. It can be observed that a URL submission is often not recognized by other Redditors until it has been posted multiple times. This illustrates a weakness in the Reddit interface when collecting crowdsourced data. Another possibility is that certain pieces of evidence gained visibility over time as they were shared across social media. These examples would be more likely to be repeatedly submitted and might attract more attention from voting Redditors upon subsequent submission.

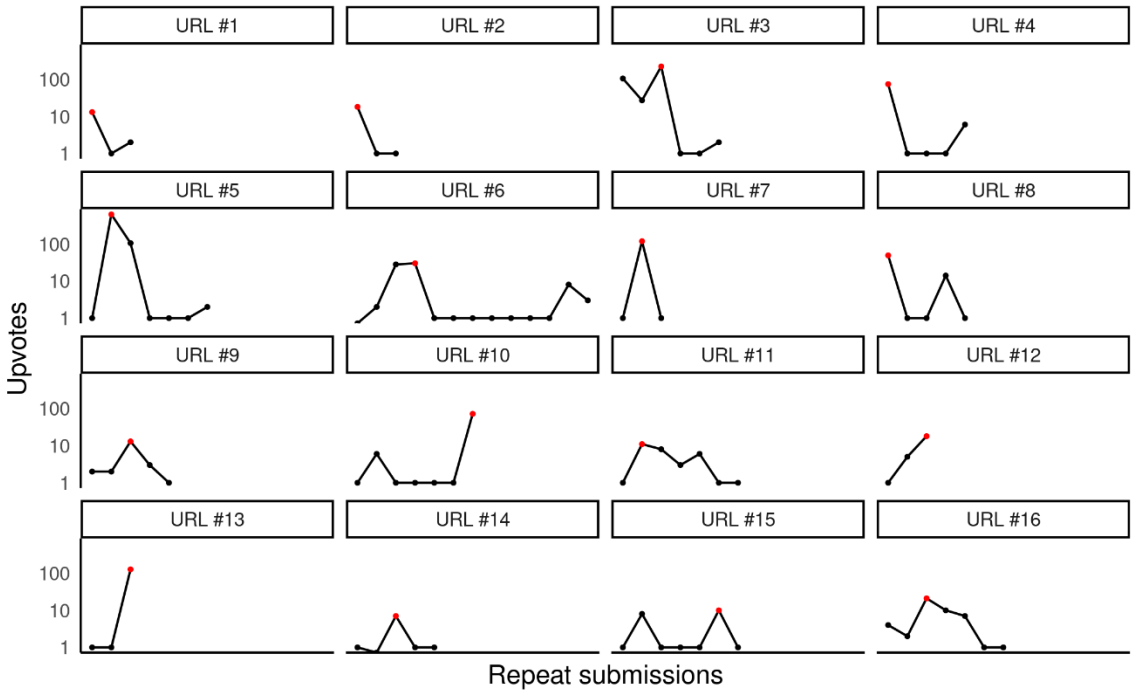

Figure 5.4: **Scores of duplicate evidence submissions**

ANALYSIS

6.1 Intro

In this section I will examine user motivations that I quantified earlier in the results section in order to apply salient themes to existing research on crowdsourced investigations and their intersections with participatory online democracy. Because this project was autonomously initiated and volunteer-run, participants arrived with a number of divergent philosophies. I will begin by highlighting users who were motivated by anger and a desire to punish the Capitol rioters. Next, I will focus on Redditors who acted out of a perceived moral and civic duty to archive the riots. Finally, I will inspect other apolitical motivations at work in the r/DataHoarders archive project.

6.2 Angry Redditors

Outrage was among the most common motivations stated by Redditors in the r/DataHoarders archive thread. As you can see in the quotes below, Redditors took personal offense to the actions of the rioters. Many expressed a sense of duty to strengthen the case against the rioters, either legally or through personal humiliation. They sought to gather material in order to preempt denial or defense of the rioter's actions. Redditors displayed open hostility towards the rioters and used demeaning language to illustrate their frustration. Words like "these idiots" and "these fucks" show the emotional charge in the thread. I can speculate that these users were drawn to the thread as an outlet to express their frustrations with the social and political landscape.

> *I can't wait to save this shit to throw back at people when all these fucks try to deny it. (gicv5ol)*

> *Archive everything. Never let these white supremacists walk away free. (giddz5w)*

Other Redditors framed their actions as explicit assistance to future criminal investigations. They conceptualized law enforcement agencies as allies in their efforts. These users found comfort in the thought that their combined efforts would aid the state in enacting punitive justice on the rioters.

> *Please download and archive. Please also preserve text. A lot of which will be incriminating evidence and proof of motives. (gievwit)*

> *Sure they have cameras in the Capitol, but these will have better angles and will show that these idiots committed crimes knowingly and intentionally and posted it on social media. (giek1wb)*

Inevitably, feelings of outrage led Redditors beyond archiving to more personal punishment.

> *It isn't just the videos... I hate to say this but we need names.... people need to lose their jobs... they tried to overthrow the government... (gie3ksy)*

These quotes espouse the consequences that some Redditors envisioned for the rioters, namely public humiliation and job loss in additional to legal repercussions.

6.2.1 Doxing

As I reviewed earlier in the literature section, doxing is one of the most serious and harmful outcomes of crowdsourced investigations such as the r/DataHoarders archive project. While calls for consequences, including vindictive punishment, were common throughout the thread, I did not observe significant engagement with the actual practice of doxing. In my results, I showed that comments which encouraged doxing were well outweighed by archive-focused activities such as the identification, downloading, and preservation of evidence. When categorizing users in the r/DataHoarders archive thread according to the number of comments they made, I found that most calls for doxing were made by users who contributed only a single comment. Highly-engaged Redditors who

made six or more comments in the thread tended to focus on archival activities. At no point did I observe the sort of dossier-building that characterized the r/FindBostonBombers subreddit. While the archived materials produced by r/DataHoarders certainly found use outside of Reddit, the Capitol riots archive project lacked the Reddit-to-mainstream media pipeline that made r/FindBostonBombers so dangerous.

The r/DataHoarders archive thread shows the fine line between collective outrage and vindictive targeting as exemplified by doxing. It is important to distinguish that the anger and outrage expressed in the quotes above did not themselves constitute doxing. As I reviewed in the literature section, doxing is a meticulous and involved process that requires careful connection of disparate digital threads. Plainly, doxing was not a viable tactic for enacting accountability towards the Capitol rioters due to the sheer number of targets they presented. In contrast to the r/FindBostonBombers subreddit, in which a crowd of participants searched for one or two individuals, the January 6th archive was a crowd-on crowd engagement. The risk of doxing and other vigilante activities in this case may have been mitigated by the symmetrical composition of the Redditors and rioters, both being volunteer and largely non-hierarchical efforts.

6.2.2 FBI

As with doxing, my results show that calls to submit evidence to the FBI were a significant, though not dominant element in the r/DataHoarders archive project. Calls to collaborate with the FBI typically took the form of a link to the FBI's evidence submission portal with a suggestion that users in the thread share the work they've been doing. Users ran into trouble with the FBI's website, which only permitted 4 uploads at a

time with a maximum size limit of 1024MB. These limits were clearly at odds with the scope and breadth of the r/DataHoarders archive.

> *I'm uploading to FBI now, for good measure. They only take four files per submission limited to 1024MB. Taking a while... (gif6op3)*
>
> *I think their server is over worked because the submit button isn't responding (gifx8fb)*
>
> *FBI website would actually break every time I tried to upload (giixt5q)*

As you can see in the above comments, users in the thread found that sharing evidence with the FBI was a more tedious process than actually collecting and archiving the material. Ironically, this was largely due to their own collective fixation with the necessity of sharing the information with the FBI in the first place. Interestingly, many users also proclaimed a belief that the FBI was actively observing and benefiting from the thread, rendering submission of materials pointless as they were presumably "already all over it".

> *Dude, the FBI has been all over this thread for hours, 100% We are basically doing their job. (gif2rhq)*
>
> *I mean I think the FBI knows how to use BitTorrent, is very likely watching this thread, and the Magnet link is right there in the post. (gikpoax)*

We can understand these Redditors as taking part in an ad-hoc, crowdsourced investigation running parallel to law enforcement. This can be defined as digilantism [22]. It can be argued that the first instance of digilantism in response to a terrorist event in the US occurred following the 2013 Boston Marathon bombing [22]. The r/FindBostonBombers subreddit was a hub for self-directed information gathering and dissemination by a highly motivated crowd of amateur sleuths. In addition to outrage and sympathy posts, users scrutinized photo and video evidence of the bombing, attempting to glean new information or leads that may have been missed by law enforcement.

Although Redditors did not identify either of the bombers, multiple individuals were incorrectly outed as the perpetrators. The failure of the r/FindBostonBombers subreddit to restrain their harmful motivations is memorialized by the phrase "We did it, Reddit," which is used throughout the Reddit platform, including in the insurrection archive thread as a cheeky reminder of the shame and regret caused by previous investigative oversteps.

6.3 Self-Righteous / Morally "Good" Redditors

While many Redditors in the r/DataHoarders archive project felt inspired by anger and punishment, others were motivated by a sense or moral or civic duty. These users attempted to engage with the democratizing potential of crowdsourced investigations posited by Buozis and Glassman and Kang. Users felt empowered by the archive project and felt that it gave them a chance to push back against societal ills. Rather than simply seeking punitive justice through law enforcement or personal vengeance, the quotes below show a desire to step up as informed, engaged citizens. These users felt an obligation to contribute their time and resources toward the betterment of society, more so than to the downfall of individual rioters. Redditors saw a chance for their online community to counter anti-democratic forces and bolster societal institutions with the hope of a more just future. To them, the insurrection archive was more than an opportunity to punch back, but a chance to "do the right thing." Their comments express optimism toward the potential of the Reddit community to harness crowdsourced action in a meaningful way.

> *You guys are like the Ham Radio operators! Gearing up in times of crisis. (gicvb37)*

> *Thank you for your hard work protecting the country from terrorism (gif7x40) To me, this is a volunteer preservation effort - a social duty which should be separate from any monetary gain. (gin035z)*

Glad I can always count on you guys to fight the real fights (gicujki)

Other Redditors felt a sense of moral superiority in their actions. They seemed emotionally driven in a way that paralleled the angry Redditors but without the spite and personal offense. Like the angry Redditors, these users were polarized by the actions of the rioters. Similar to the "social duty" Redditors, they felt a sense of obligation to act. By aligning themselves with religion and patriotism, they claimed that the archive was not just an effective participatory response, it was morally just and even necessary on a higher calling.

People in this thread doing the actions of true patriots (gidpnox)

You have done God's work, and I applaud you. I never even thought how important it would be to save and share this ... thank you. (gifrcwv)

Feels like I'm doing the lords work. (gifuhcg)

6.4 Cautionary Statements

Not all Redditors were so quick to take action or assign moral value to their investigations. I discovered in my findings that the shadow of the r/FindBostonBombers subreddit loomed large in the memory of the Redditors in r/DataHoarders, and informed suggestions of care from more discerning users. While precautionary comments were a minority of the thread activity, multiple users specifically brought up the harms that resulted from past crowdsourced witch hunts on Reddit.

Also, I believe a guy committed suicide after being falsely identified as the internet cat killer. Reddit does not have a good track record of these sort of things. (gijqgdf)

Dude, the FBI has been all over this thread for hours, 100% We are basically doing their job. And if it's not to catch those guys, it's to make sure that Reddit doesn't go rogue [sic] as detective again lol (gif2rhq)

As these quotes show, past excursions into vigilantism are a source of shame and regret among the Reddit community. Even years later, instances such as the r/FindBostonBombers subreddit are still referenced as an example of what can go wrong when well-meaning internet sleuths lose sight of the human costs of their investigations.

6.5 Apolitical Redditors

The remaining segment of Redditors claimed motivations that were less emotional or political than the quotes shared above. These users possessed the skills and resources needed for the archive thread and saw an opportunity to contribute. They were motivated by a drive to archive (or "hoard") data for historicity, future analysis, or simply out of an interest and ability to engage in digital archival practices. In any case, these users can be distinguished from previous categories by their relative stated indifference toward the punitive and democratizing potentials of the archive.

6.5.1 Data Hoarding

The eponymous hoarders of the r/DataHoarders subreddit often weighed in on how the insurrection archive intersected with their beliefs and skill sets. Users expressed a drive to archive data not for being valuable or incriminating, but because it was at risk of removal.

> *You guys are doing great work. In the digital age nothing should be able to be completely removed so thanks for archiving this before the idiots take it down and deny it (giflouc)*

The r/DataHoarders subreddit is largely an apolitical community focused on tools and resources for acquiring, handling, and storing digital content. Some users felt alienated by the stark political rhetoric of the insurrection archive and attempted to distance their subreddit from the influx of new commenters. Still, these quotes show a

tolerance for the multitude of motivations that brought the 1,311 Redditors together for the archive project.

> *Yesterday's events were truly insane on many levels, but we are not a political sub. Some may DH for political reasons, and some for sociological reasons. Whatever your reason, we ask everyone to keep it about datahoarding/archiving in here. (gifmaje)*
>
> *Some may archive for political reasons, others like myself archive, for posterity [sic]. (gigdzr9)*

The breakdown of thread activities in figure 5.1 shows that the top three activities, evidence, strategy, and saving account for over half (54%) of the total comments in the archive thread. These activities, along with seeding form the core tasks of digital archiving: locating, downloading, storing, and preserving evidence. Archival skills such as these are commonly discussed on the r/DataHoarders subreddit and represent a collective interest that binds the community. In figure 5.3, the same chart is broken down by user engagement level, or the number of comments a user made in the archive thread. We can see that the most engaged users who commented six or more times allocated the majority of their efforts to these top archival tasks. The saving activity, representing the actual downloading or scraping of submitted evidence, was entirely dominated by high-frequency users. This leads me to believe that the core values and leadership of the r/DataHoarders subreddit retained a high degree of influence and control over the insurrection archive thread, even as the thread's popularity and reach extended far beyond the niche community where it originated.

6.5.2 Historicity

A unique characteristic of this dataset is that participants were self-aware of the historicity and uniqueness of this endeavor and devoted time and space to discussing why it was important.

> *I'm backing this up to document the day fascism has erupted from the bowels of this "democracy". (giek8ro)*

The Reddit "RemindMe" bot is an automated account which notifies users about a requested thread at a specified time in the future. By commenting "RemindMe! 3 days," the user will receive a notification in three days with a link to their original comment. It is typically used as a short-term reading list for returning to interesting content at a more convenient time. Multiple Redditors in the r/DataHoarders insurrection archive thread called on the RemindMe bot to remind them far into the future, up to 4 years. Perhaps these users wanted to see if the thread and the events of January 6th would stand out as a historic moment in retrospect.

> *RemindMe! 1 year (gif22ms)*
>
> *RemindMe! 4 years (gif24b7)*
>
> *Thanks we need this piece of history archived so we can look back on this moment over and over again. (gif6l2h)*
>
> *I truly believe its important we preserve all the footage of this event as possible. This was truly a terrible and historic event. (gifpj10)*
>
> *Descriptions of today's events by future historians are only as good as what we provide them to study. We must continue to archive. (gigdzr9)*
>
> *Thank you admins (u/macx333) for leaving this up. This is data of historical significance and, IMO, the highest calling for hoarders like myself (gigjkgg)*
>
> *Downloading now, and plan to seed for a while. A month? A year? I don't know, my link is slow but we need to keep it alive for history. This feels ... important. (gilwauj)*

These quotes demonstrate an awareness of the historical value not just of the insurrection archive project, but of archiving as a whole. Through language such as "highest calling for hoarders like myself" and "we must continue to archive," these Redditors express a belief in the inherent value of archival work. These quotes speak to a future community outside of the immediate social reception of the events at the Capitol

riot. Data hoarding Redditors expressed interest in the utility of the material to an audience outside of law enforcement or judicial usage. While previously discussed Redditors sought to preserve evidence of the riot to achieve immediate judicial ends, these historically-minded users espoused the inherent value of information preservation, envisioning future uses which cannot be fully understood in the context of the present moment.

6.5.3 Archiving Social Movements

In "Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations," Jules et. al outline the value of archiving social movements through social media as well as the ethical challenges and considerations it presents [32]. Speaking about the Ferguson Black Lives Matter movement, the authors state, "digital content adds a new layer of documentary evidence that is immensely valuable to those interested in documenting, researching and interpreting contemporary events," expressing the necessity of archives to interpretation [32].

They state, "the level of participation in these movements as they play out on social media makes them rich scholarly resources deserving of collection, preservation and study" [32]. The ability of social media and other digital platforms to capture ongoing events, and the evolving, interactive nature of these platforms adds additional value and context. In the case of the r/DataHoarders archive, the heavy use of social media by Capitol rioters provided a deep source of evidence, information, and context behind their collective actions on January 6th. The Redditors engaged in archival work by documenting the scope and depth of the digital content produced by the rioters.

Additional value of archiving, as well as criticisms and cautions, is discussed in "Archives and Human Rights: Questioning Notions of Information and Access" by Caswell and Punzalan [33]. Caswell and Punzalan posit that both libraries and archives contribute positively to discursive formation through providing, respectively, information and evidence of human activity [33]. Archives are defined by their future use value (evidence), whereas libraries are a source of pure informational value. In the context of my paper, we might understand the pure informational value through association with OSINT, and the r/DataHoarders thread as primarily an archival project because many users conceptualized their research as evidence gathering. Yet these two things are not mutually exclusive, and the archive/evidence value and the OSINT/informational value of the r/DataHoarders research are both at play. The authors believe that through documenting potential evidence of human activities or events, human rights questions are called into account. They in fact believe that "archives stress preserving evidence of human rights infringements for accountability" [33]. The Redditors who engaged in the Capitol archive thread sought to see the bigger picture of the American political narrative and envisioned their work as a documentation of a political moment that may some day be associated with a rupture in democratic governance.

Caswell and Punzalan posit that "by stewarding evidence, archival institutions are also key actors in efforts at redress, reparation, and reconciliation in societies undergoing reconstruction and healing in the wake of human rights abuse" [33]. Additionally, "archives also play a complicated and shifting role regarding the shaping of collective memory of traumatic events" [33]. This argues that historicity relies upon shared national and social memory in order to create and interpret narratives about human events.

Archives play an important role in preventing the “forgetting, eliding, and silencing” of selective aspects of these events [33]. However, they also argue that archives can have the exact opposite effect due to their inherent power and role in the shaping of collective memory and historicity. If archivists allow themselves to become aligned with biased state interests, they can themselves become a tool for silencing marginalized voices through removing them from the narrative. In the case of the Capitol riots, many Redditors viewed the insurrection as a potential human rights abuse that risked being elided from history. They envisioned the January 6th event as the moment that, though attempted coup, the naked face of demagogic rule was bared. Through their archival efforts, they sought to preserve the most accurate picture of what they viewed as a brazen lunge for political power. Therefore, their self-conceptions of their archival work often have a reconstructive tone, seeking to preserve a narrative they anticipated would be challenged or erased.

CONCLUSION

My research concludes that the practice of archiving contemporary social movements is perhaps the best place to begin when considering how to understand crowdsourced archiving projects such as the r/DataHoarders insurrection archive. The bulk of participation in the thread was focused on archival tasks and was led by dedicated users who expressed an interest and belief in the value of archival work. Although I also found numerous proposals to dox the rioters and seek punitive justice, also known as digilantism, these elements were outweighed by the volume and stated value of archiving material for its own sake. Redditors recognized the historical importance and ephemeral nature of the Capitol insurrection material and acted accordingly, preserving it for posterity.

The r/DataHoarders dataset was a rich source of insight into the process of crowdsourced archiving due to the public deliberation and discussion found within it. My research showed many examples where users expressed a desire to participate positively and democratically in the archive project. Coding the dataset for themes and user activities allowed me to understand the practices and roles with which the Redditors engaged. I found it valuable to read the self-stated motivations that users shared in the thread. However, a more robust understanding of the archive participants could have been achieved by contacting individual Redditors or evaluating their Reddit activities outside of the archive thread. Due to software limitations, I coded each comment individually without being able to review the comments that preceded them. Some context to my dataset may have been lost in this process.

The memories of January 6th still loom large in national consciousness. The explosion of digital content posted to social media platforms that day exemplifies the intersection of digital self-documentation and political volatility that drew people to threads such as the r/DataHoarders archive project. As of November 2021, a large number of legal cases are in process and being continuously updated with new information on publicly accessible sources such as USA Today. Many of these cases rely on photo and video evidence posted by the defendants and their co-conspirators. Without projects like the r/DataHoarders insurrection archive, these pieces of evidence may have been deleted or otherwise lost. Future research in this area could potentially focus on further developing ethical guidelines for amateur and crowdsourced archival projects. A deeper dive into the community of r/DataHoarders might involve interviews and calls for direct participation in an ethnographic study. Finally, the ethical implications of breaching terms of service in the process of social media archiving warrants further study as well. Regardless of the historical singularity of the January 6th insurrection, the social and technological impetus for crowdsourced archiving guarantees that these questions will persist.

BIBLIOGRAPHY

[1] J. Lytvynenko and M. Hensley-Clancy, "The Rioters Who Took Over The Capitol Have Been Planning Online In The Open For Weeks," *BuzzFeed News*, Jan. 2021.

[2] YouTubeInsider, *YouTubeInsider on Twitter*, Tweet, Jan. 2021.

[3] J. Alexander, "How Facebook, Twitch, and YouTube are handling live streams of the Capitol mob attack," *The Verge*, Jan. 2021.

[4] Bellingcat, *Storming of the Capitol Visual Material Collection*, https://docs.google.com/forms/d/e/1FAIpQLSfgcEYSRfoKmUPmFwcKB0pSJNvyETkAZ8qEGXZB528f0AlEQ/viewform?usp=embed facebook, 2021.

[5] Bellingcat, *Bellingcat on Twitter: "Thanks again to everyone who has submitted visual materials from the Capitol attack! You can submit any photos/videos (if you've put the video on a Google Drive/Dropbox, that helps a lot!) here and we'll get to it. We have about 100 backlogged entries now. https://t.co/iRnP24KvDJ" / Twitter*, Tweet, Jan. 2021.

[6] A. Toler, *Database of August 12 Charlottesville Videos*, Aug. 2017.

[7] intelx, *Archiving Capitol Hill riots' media – Intelligence X*, Archive, Jan. 2021.

[8] Intelligence X, *Capitol Hill Riot Archive*, https://intelx.io/?did=814b39fe-ad98-45a1- 9f44-0346bc9f9b94, Archive, 2021.

[9] A. Lynch, *Trump protest Jan 06 2021*, https://mega.nz/fm, Cloud Storage, 2021.

[10] *r/DataHoarders - MEGATHREAD: Archiving the Capitol Hill Riots*, https://www.reddit.com/r/DataHoarder/comments/krx449/megathread archiving the capitol hill riots/, Social Media, Jan. 2021.

[11] D. M. Douglas, "Doxing: A conceptual analysis," *Ethics and Information Technology*, vol. 18, no. 3, pp. 199–210, Sep. 2016.

[12] M. Honan, "What Is Doxing?" *Wired*, Mar. 2014.

[13] *Faces of the Riot*, https://facesoftheriot.com/.

[14] S. Musil, *Website features faces from Parler's Capitol riot videos*, https://www.cnet.com/news/website-features-faces-from-parlers-capitol-riot-videos/.

[15] @homegrownterrorists, *@homegrownterrorists Instagram profile • 337 photos and videos*.

[16] @aleximenez, *Alejandra on Twitter*, Jan. 2021.

[17] Tre Ward, *Tre Ward on Twitter*, Tweet, Jan. 2021.

[18] Goosehead Insurance, *Goosehead Insurance on Twitter*, Tweet, Jan. 2021.

[19] Bellingcat Investigation Team, *The Journey of Ashli Babbitt*, Jan. 2021.

[20] Metropolitian Police Department, *Persons of Interest in Unrest-related Offenses*, 2021.

[21] Z. Tillman, *100 Capitol Rioters Have Pleaded Guilty. Here's What They Did And What They're Facing*. Oct. 2021.

[22] J. Nhan, L. Huey, and R. Broll, "Digilantism: An Analysis of Crowdsourcing and the Boston Marathon Bombings," *The British Journal of Criminology*, vol. 57, no. 2, pp. 341–361, Mar. 2017.

[23] L. Potts and A. Harrison, "Interfaces as rhetorical constructions: Reddit and 4chan during the Boston Marathon bombings," in *Proceedings of the 31st ACM International Conference on Design of Communication*, ser. SIGDOC '13, New York, NY, USA: Association for Computing Machinery, Sep. 2013, pp. 143–150, ISBN: 978-1- 4503-2131-0.

[24] N. Lally, "Crowdsourced surveillance and networked data," *Security Dialogue*, vol. 48, no. 1, pp. 63–77, Feb. 2017.

[25] M. Buozis, "Doxing or deliberative democracy? Evidence and digital affordances in the Serial subReddit," *Convergence: The International Journal of Research into New Media Technologies*, vol. 25, no. 3, pp. 357–373, Jun. 2019.

[26] S. A. Stottlemyre, "HUMINT, OSINT, or Something New? Defining Crowdsourced Intelligence," *International Journal of Intelligence and CounterIntelligence*, vol. 28, no. 3, pp. 578–589, Jul. 2015.

[27] J. Pastor-Galindo, P. Nespoli, F. G. Marmol, and G. M. Perez, "The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends," *IEEE Access*, vol. 8, pp. 10 282–10 304, 2020.

[28] M. Glassman and M. J. Kang, "Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT)," *Computers in Human Behavior*, vol. 28, no. 2, pp. 673–682, Mar. 2012.