CRIME AND SOCIAL MEDIA:
THE GEOGRAPHY OF CRIME DISCUSSION ON REDDIT

BY

SEAN KUNZ

BS, Binghamton University, 2020

THESIS

Submitted in partial fulfillment of the requirements for
Master of Science in Computer Science
in the Graduate School of
Binghamton University
State University of New York
2022

Accepted in partial fulfillment of the requirements for
Master of Science in Computer Science
in the Graduate School of
Binghamton University
State University of New York
2022

August 9, 2022

Weiyi Meng, Chair
Department of Computer Science, Binghamton University

Jeremy Blackburn, Faculty Advisor
Department of Computer Science, Binghamton University

Kyoung-Don Kang, Member
Department of Computer Science, Binghamton University

Abstract

The relationship between crime and the media is an oft studied subject. Researchers have long analyzed how the ubiquity of crime reporting in traditional media can affect people's view of public safety in their communities and how various factors can impact this trend. However, in the age of the Internet, social media has become increasingly relevant when compared to traditional media. Despite this, the level of research on crime and social media has not rivaled social media's rise.

In this paper, we analyze the submission of crime-related posts to communities representing the 384 Metropolitan Areas of the United States on the social media website Reddit. We find that crime discussion is more common in larger communities and communities that lean more liberal politically. Reported crime rates do not have a strong impact.

Dedicated to my friends, family, and all those who helped me along the way

Contents

List of Tables

List of Figures

List of Abbreviations

API Application Programming Interface

FBI Federal Bureau of Investigation

MSA Metropolitan Statistical Area

REST Representational State Transfer

TD The_Donald

UA Urban Area

UC Urban Cluster

UCR Uniform Crime Reporting

US United States

1 Introduction

Crime has long shaped people's lives. The fear of crime has inspired historical movements [22] and shaped elections [18]. Through mediums like television [37] and print [12], researchers have long analyzed how the prevalence of crime reporting might affect consumer's perceptions of public safety.

While traditional media is frequently the subject of crime-relationship research, it has grown increasingly irrelevant. As traditional media declines [23], social media has risen in its place. Due to this shift in media consumption patterns, an increasing number of people receive news through social media [17], including stories that focus on crime. However, social media faces some challenges not found in the traditional media landscape. Due to the free and open nature of who can post on social media, submissions containing misinformation [24] and manipulation of content [2] have remained a challenge for social media companies to reign in.

As social media begins to fill traditional media's niche, one might wonder how the landscape of crime reporting and discussion might change. In this work, we analyze the discussion of crime on the social media website Reddit. Specifically, we examine how crime discussion takes shape on 384 unique communities in the United States. To measure crime discussion, we have developed a Reddit crawler to fetch submitted posts from the 384 selected communities. We then utilize textual analysis techniques to identify posts that discuss crime while minimizing false positives. We analyze the data collected by the crawler to determine a community's level of crime discussion in comparison to its population, voting patterns, and documented crime rate. Lastly, we dive into the posting histories of users who post to Reddit's New York City community to identify any further trends or patterns.

2 Background and Related Work

2.1 Geography

2.1.1 Scope

Before we begin the gathering of the data, we must first understand the scale at which the data should be collected. First, we are collecting data that pertains to the United States. Much research has focused on the relationship between crime and media specifically within the United States. The United States Census Bureau also provides plentiful contextual and reference data. Second, in a general sense, we want to measure at a city-level scale. The city is the base-level at which a critical mass of population can support an online community; a city provides a geographic center and a name to associate with a community. However, the definition of a city is rather murky. The definition varies on a state-by-state basis due to the nature in which a state delineates county-level subdivisions. All states are broken down into counties, but differ in how counties are partitioned. For example, New York breaks down counties into towns and cities, while North Dakota only subdivides counties into cities, with excluded areas remaining unincorporated. In this way, we can think of city limits and city populations as somewhat arbitrary. There is no consistent, uniform definition, so one should be careful when comparing cities for analysis purposes.

2.1.2 Census Defined Statistical Areas

To best equalize differences in city definitions, we use statistical area delineations provided by the United States Census Bureau to normalize city population and sizes. The first and most precise definition is of Urban Areas (UAs) and Urban Clusters (UCs). UAs are dense clusters of people that house over 50,000 people. UCs are similarly measured dense clusters of people with between 2,500 and 50,000 residents. UAs and UCs are computed by first identifying a dense cluster of census blocks, the smallest census geography. From here, the UA or UC snakes off from the central core to include housing that fits the density requirements in less dense census blocks [4]. This process creates the "true" city limits, which are unaffected by municipal boundaries.

While Urban Areas are an effective method of measuring the true sizes of cities, they are difficult to map and obtain statistics on. Their shapes are generally compact and do not have any sort of uniform governing body. To alleviate these issues, we choose Metropolitan Statistical Areas (MSAs) as our scale. MSAs consist of the county or counties in which an Urban Area lies, alongside any counties in which 25% of the population commutes to a core county [5]. In this way, we gain the population measurement accuracy of UAs with the ease of data access from state-defined municipalities. As of March 2020, there are 384 Metropolitan Statistical Areas [5], each with a corresponding Reddit community.

2.1.3 Geography and Social Media

When analyzing our selected Reddit communities, we need to consider the geospatial aspects of social media. In some sites, specific geographic location is intrinsic to the functionality of the product, whereas in others, anyone can browse a community for somewhere far away. For example, on Reddit there is no barrier preventing someone from, for example, Los Angeles from posting in a New York City community. Scellato et al. [36] measure the geographic nature of social networks through the comparison of geospatial proximity of graph-connected users on four social media sites. They found that sites, for example FourSquare, that had an explicitly geographic nature produced a higher degree of locality. Blackburn et al. find a similar trend on the Steam gaming community [3]. However, sites that focused more on general information exchanges, like Twitter, showed less geographic clustering.

The Reddit communities we analyze hold an interesting space between these two categories; they are both tied to place and detached from it. There is no geographic component stopping a user from posting to a distant community, but the geospatial nature of the community itself can lead to geographic clustering. In this way, we must determine if those posting in a certain community are truly local.

2.2 Traditional Media

Traditional media, including television, newspapers, magazines, and radio, has long been intertwined with a fascination for crime. Live manhunts for suspects, including OJ Simpson [15] and the Boston Bombers [38], have captured and enthralled viewers and readers. However, while high-profile cases garner much of the attention, local and lower-profile cases often inhabit newspaper columns and television segments. It is with these cases that people grow uneasy with the sense of public safety in their community. A flurry of reporting on violent acts can increase worry, stress, and fear that crime is rising, even if the documented crime statistics show that crime has not increased [12].

An increase in fear of crime can have implications beyond elevated negative emotions. Fear of crime has far-reaching social [12], political [15], and institutional [33] impacts. Belief in excessive amounts of crime can lead to loss of community cohesion; Heath and Gilbert [12] found that as television news media consumption increased, distrust both in the local community and the world "out there" increased. Additionally, media creators often sensationalized stories, turning unfortunate incidents into grim spectacles. Those that perceived crime to be worse than others were more likely to perform excessive crime prevention actions; for example, those with an increased fear of crime might not leave their house or engage with the neighborhood.

Liu et al. [18] describe how the prevalence of crime discussion affected the 2016 presidential election in the United States. The authors found that voters who consumed more television in turn consumed more campaign advertisements. About 15% of campaign advertisements covered issues including crime and violence. Frequent crime media consumers assumed crime and violence issues to be worse than they really are [37]. Furthermore, the authors found that excessive crime reporting and media exposure had impacts on voting patterns; those that believed crime to be worse tended to vote for more conservative candidates.

Worry of crime victimization can lead to the creation of institutional policies meant to combat the perceived threat. One notable example is the idea of Broken Windows Theory [43] and a practice that stemmed from its development, known as stop and frisk [19]. Broken Windows Theory states that visible signs of crime and disorder lead to further crimes and disorder [43]. Essentially, the theory states that ignoring minor criminal incidents can lead to more frequent and serious issues. Acting on this theory, the New York Police Department instituted a stop and frisk policy, in which police officers temporarily detained civilians suspected of carrying illegal objects [19]. The policy, while believed to be effective at the time, has since come under scrutiny. Researchers found that only 6% of stops resulted in arrests [19] and that there is no clear proof that stop and frisk contributed to the decline in crime in New York City [42]. Stop and frisk was removed as a policy in 2013 when it was deemed unconstitutional [19].

2.3 Social Media

2.3.1 Social Media and Crime

While the relationship between traditional media and crime has been thoroughly researched, the relationship between social media and crime has not been as rigorously explored. Social media has grown and evolved beyond its humble beginnings. Social media sites are now the main source of news for many Americans [17]. Walker and Masta [41] found that nearly half of Americans consume their news from social media sites like Twitter, Facebook, and Reddit. This trend coincides with the decline of traditional media. Printed media, as well as televised media, have consistently declined since the explosive growth of the Internet [23].

Curiel et al. [8] detail the ecosystem of crime discussion on Twitter. Specifically, they look at how often crime discussion appears in Tweets posted from users in Latin American countries. Using Twitter's geolocation tags, Tweets are subdivided down to the municipality that the original user posted from. From here, the authors compared Twitter crime discussion rates to city-level documented crime rates. They did not find a correlation between high frequencies of crime-related Tweets and high rates of reported crime. The authors conclude that the Tweets do not accurately represent actual rates of crime, but they effectively capture the worry of crime in a community.

The authors introduce an interesting observation: some of the crime-related Tweets are highly sensationalized postings from a group of "involved users" that frequently comment about violence [8]. In this way, social media differs from traditional media. Traditional media acts as a gatekeeper, where only select stories and viewpoints emerge and factchecking generally exists in mainstream publications [33]. Social media, on the other hand, allows for a more diverse set of viewpoints and opinions. However, as moderation policies vary [44] on social media platforms, blatantly false stories can spread across sites if left unchecked. This proliferation of misinformation has become increasingly impactful in the United States. It has affected important events in recent American history, including presidential elections [18] and the response to the Covid-19 pandemic [21, 39]. Without intervention, falsehoods and panics continue to spread across the web.

2.3.2 Reddit

While Facebook [26] and Twitter [39] have many times been the subject of backlash directed towards online misinformation, smaller websites, like Reddit, play a role in the distribution of sensationalized stories and "fake news" as well. Reddit communities, also known as subreddits, allow users to post self-made content or link content from other parts of the web. Subreddits are dedicated to a certain topic, like fishing, political groups, or real-life communities. Subreddits can be created by anyone and are run by volunteer moderators. As moderators are a varied group of individuals, they represent a diverse set of viewpoints not beholden to any set of limits. In this way, misinformation can seep into communities, as there are differing levels of moderation among subreddits. To combat rouge communities and moderators, Reddit has moved to ban certain subreddits that consistently break sitewide rules [31].

Several Reddit communities have been the subject of studies on misinformation. One such subreddit is known as `The_Donald` (TD). TD, which acted as the support community for former President Donald Trump, spread unsubstantiated conspiracy theories, like Pizzagate [25], and engaged in harassment of other communities on the website [9]. TD was eventually quarantined and banned from the site, leading users to create their own Reddit offshoot website [31]. However, TD spread misinformation with little intervention for the nearly five years that it survived on the site [31]. This is largely because Reddit generally avoids intervening in communities, while occasionally banning subreddits with egregious behavior [34]. Several Reddit communities, including `Incels` [32], `FatPeopleHate` [7], and `NoNewNormal` [10] that produce hateful, harassing, or misleading content, have also been banned after significant backlash.

Reddit has also been responsible for sensationalizing crime. In one notable example, Reddit users attempted to solve the identity of the 2013 Boston Marathon Bombers [27]. Reddit users believed that through community effort and brainstorming, the Reddit community could catch the killers quicker than the authorities. After the users performed their investigation, they believed that they had correctly identified the suspect and engaged in harassment of those related to him [27]. Ultimately, the Reddit users had identified the incorrect perpetrator. An embarrassing moment to many Reddit users, the incident nonetheless is important in showing how the fear of crime, and how social media platforms perpetrate this fear, can help radicalize individuals.

3 Analysis

3.1 Data Overview

3.1.1 Reddit Data

The Reddit dataset contains posts from the subreddits of the principal cities of the 384 MSAs of the United States as of March 2020 [5]. From this list of subreddits, we fetch and store all posts that users submit. In total, 161,166 posts were gathered between April 5, 2022 and May 31, 2022. A timeline of this process can be seen in Figure 3.1. The datastore archives posts in two formats. First, posts are written to a system of record to ensure that there is an authoritative data source that can be referenced. Second, pertinent information, like post title and author, is extracted from the posts data stream and placed in a separate datastore.

After storing data, an analysis tool runs to determine if a post is about crime. To do this, a regular expression checks if a post's title or a post's link's URL contains a crime-related keyword. These keywords were derived from actions that constitute violent crime and their synonyms, as well as generic crime-related words. In total, 124 crime-related keywords are used. A full list can be found in Appendix B. If the regular expression returns a hit, a flag is raised in the database denoting that the post is about crime. Overall, 6,792 of the collected posts are estimated

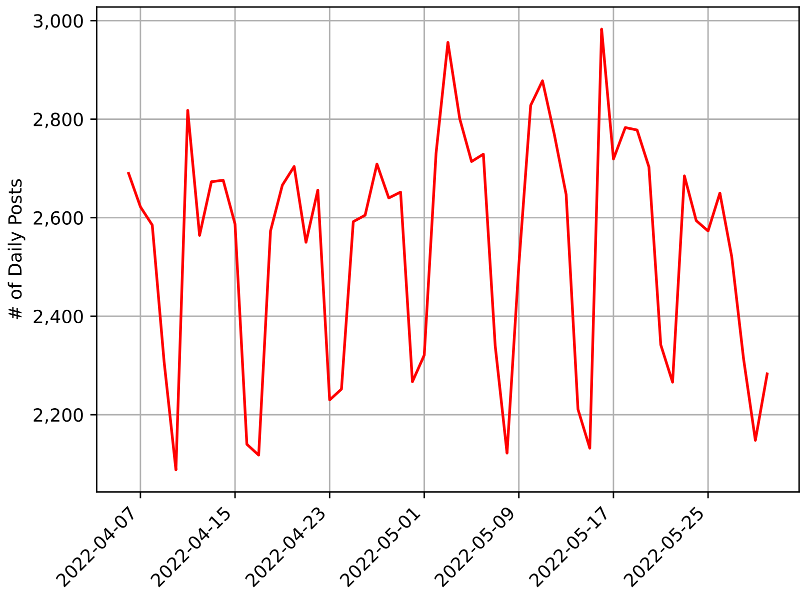

Figure 3.1: Number of daily posts across all subreddits

to discuss crime.

Due to the nature of text analysis, a number of false positives might be considered, while some false negatives go undetected. False negatives can be hard to detect, as they include dog-whistles, local-dependent terms, and ambiguous references [8]. Certain posters may choose to keep their posts ambiguous so as to keep their intentions visible to only a small group [1]. To determine the rate of false positives, we analyze a sample of stored posts. We look for phrases like "I shot a picture" that do not relate to crime, but contain a crime-related keyword. Of a sample of 400 posts, we find that 36 posts do not relate to crime, for an accuracy rate of 91%. Similar analysis done on the social media site Twitter reached a rate of 66% accuracy [8].

3.1.2 Contextual Data

We utilize a variety of data to highlight trends in the Reddit posts. One variable that subreddits are compared to is population size of the MSA. These values range from a minimum of 58,639 (Carson City, NV) to a maximum of 20,140,470 (New York-Newark-Jersey City, NY-NJ-PA). MSAs are categorized into four bins based on their size: large, medium, small, and very small. The population cutoffs are 1,000,000, 500,000, and 200,000. The large population bin contains 56 members, the medium bin contains 54 members, the small bin contains 113 members, and the very small bin contains 161 members.

Another variable that provides context is the political climate of an MSA. For this metric, we consider the results of the 2020 election [28] in terms of the percent of the MSA voting Democrat. In this case, we only document Democrat and Republican votes. These values range from a minimum of 20.2% (Morristown, TN) to a maximum of 80.9% (Santa Cruz-Watsonville, CA). The MSAs are categorized into five bins based on their voting results: Very Conservative, Conservative, Moderate, Liberal, and Very Liberal. The percentage cutoffs for the bins are 37.5%, 47.5%, 52.5%, and 62.5%. The Very Conservative bin contains 112 members, the Conservative bin contains 103 members, the Moderate bin contains 57 members, the Liberal bin contains 73 members, and the Very Liberal bin contains 39 members.

A third variable that is considered is the documented violent crime rate of each MSA. This data is derived from the FBI's Uniform Crime Reporting [13] statistics.[1] Documented violent crime rates range from a minimum of 44.3/100,000 (Bangor, ME) to a maximum of 1,397.2/100,000 (Birmingham-Hoover, AL). We select violent crime rates for our comparison data because they are less ambiguous than property crimes. For example, police departments are less likely to differ

[1]Data from the FBI mostly comes from the year that is most recent, 2019. However, not all municipalities report information, so there are slight estimations in the data [14]

in their definition of a homicide compared to criminal mischief. Additionally, documented crime data is normalized using the min-max normalization method. This method condenses all values in a range between a value of 0 and 1, where 1 is the highest value in the range and 0 is the lowest. In doing so, we can compare crime data to a similarly normalized crime discussion value from the Reddit data.

3.2 Methods

We use one main crawler to gather post data from Reddit. This crawler utilizes multiple different tools to gather disparate pieces of data and place them into two datastores. The crawler checks for new posts once every hour so that posts can be fetched before a user deletes them or a moderator removes them. These communities are generally slow moving, with an average of 111 posts an hour between all 384 subreddits.

To gather the data, the crawler first sends requests to the Reddit REST API [30]. We use the `new` endpoint for each subreddit to check the most recent posts. The post last written to the database for each subreddit is cached in a database table so that we can take advantage of the `after` parameter of the API. This parameter allows us to take posts that have been created since a given post ID. The `after` parameter is especially useful on slower subreddits; without it, we would be fetching posts that might be months old. If the cached value does not exist or is not valid, then the most recent hundred posts are gathered. This is the maximum allowable value by the Reddit API.

After performing error checking to ensure all of our requests went through, each record is sent to two databases. The first is a CockroachDB database that acts as a system of record. Raw posts are stored here in a simple format: `id`, `timestamp`, and `data`, the last of which is the data retrieved from the Reddit API in JSON format. After writing to CockroachDB, the crawler then writes the JSON data from Reddit to ElasticSearch. ElasticSearch provides quick and efficient text analysis that allows for the extraction of certain keywords in post titles and URLs.

The crawler and databases run in Docker containers. Containerization allows the tool, as well as the databases, to run in a clean, isolated environment without worrying about being choked out by other processes. Additionally, the crawler-container is scheduled by Kubernetes to allow for a higher degree of availability and resilience. If something unexpected occurs and the tool goes down, it will simply be restarted by Kubernetes.

3.3 Results

3.3.1 Subreddit Activity

In total, the Reddit crawler gathered 161,166 posts between 384 subreddits. Posts were collected between April 5th, 2022 and May 31, 2022. Table 3.1 details the ten most active subreddits of those monitored. Similarly, Table 3.2 lists the ten most active subreddits when normalized by population.

The most active MSAs by total number of posts share some common, predictable traits. One shared characteristic is that all of the Metropolitan Areas listed have populations above 1,000,000. While expected, some of the most notable MSAs were not included, including New York-Newark-Jersey City, NJ-NJ-PA and Los Angeles-Long Beach-Anaheim, CA, which have populations of 20,140,470 and 13,200,998, respectfully. One possible explanation for these exclusions is the fact

Table 3.1: Most active Metropolitan Areas

Metropolitan Statistical Area	Number of Posts	Population
Austin-Round Rock-Georgetown, TX	4,084	2,283,371
San Diego-Chula Vista-Carlsbad, CA	3,265	3,298,634
Boston-Cambridge-Newton, MA-NH	3,120	4,941,632
Denver-Aurora-Lakewood, CO	2,785	2,963,821
Chicago-Naperville-Elgin, IL-IN-WI	2,673	9,618,502
Sacramento-Roseville-Folsom, CA	2,668	2,397,382
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	2,541	6,245,051
Columbus, OH	2,533	2,138,926
Portland-Vancouver-Hillsboro, OR-WA	2,526	2,512,859
San Francisco-Oakland-Berkeley, CA	2,476	4,749,008

that there is no uniform, consistent moderation policy among subreddits [16]. For example, some subreddits automatically filter out all posts and manually approve them, or they disallow certain types of posts [40]. Subreddits with more laissez-faire moderation policies will have more posts simply because they are not removing a mass of posts that could be considered redundant or irrelevant.

Table 3.2: Most active Metropolitan Areas per population

Metropolitan Statistical Area	Number of Posts	Population
Bellingham, WA	1,381	226,847
Bloomington, IN	799	161,039
Missoula, MT	449	117,922
Bend, OR	660	198,253
Burlington-South Burlington, VT	546	171,415
Corvallis, OR	298	95,184
Eugene-Springfield, OR	1,183	382,971
Charlottesville, VA	638	221,524
Ithaca, NY	299	105,740
Asheville, NC	1,312	469,015

The list of most active MSAs when normalizing by population also share some common traits. These MSAs are generally politically liberal [28] and have large college age populations [6]. Essentially, these areas have large populations of people that fall under Reddit's main demographic [35] and, as a result, have correspondingly active communities.

3.3.2 Crime Discussion

To first understand the relationship between certain socio-economic variables and the rate of crime discussion, we must first look at crime discussion percentages independently. Of the 161,166 posts gathered, 6,792 are determined to discuss crime by matching certain crime-related keywords to post titles and URLs. Table 3.3 lists the Metropolitan Areas with the largest percentage of crime discussion from the total number of posts.

Table 3.3: Metropolitan Areas with most crime discussion

Metropolitan Statistical Area	% Crime Posts	# Crime Posts
Seattle-Tacoma-Bellevue, WA	18.2	344
Anniston-Oxford, AL	15.7	8
Lewiston-Auburn, ME	15.4	14
Bay City, MI	14.0	7
Lewiston, ID-WA	13.0	7
Pine Bluff, AR	12.5	4
Los Angeles-Long Beach-Anaheim, CA	12.0	257
Portland-Vancouver-Hillsboro, OR-WA	11.4	287
Rocky Mount, NC	11.3	7
Decatur, AL	10.9	7
New York-Newark-Jersey City, NY-NJ-PA	10.7	196
Flint, MI	9.7	10
Morristown, TN	9.3	4
Sumter, SC	9.3	5
Vallejo, CA	8.8	9

Furthermore, Figure 3.2 displays a map of all Metropolitan Areas by the percentage of crime posts submitted to their respective subreddits. The darker the shade of green, the more posts discuss crime. The average crime discussion percentage is 3.5%.

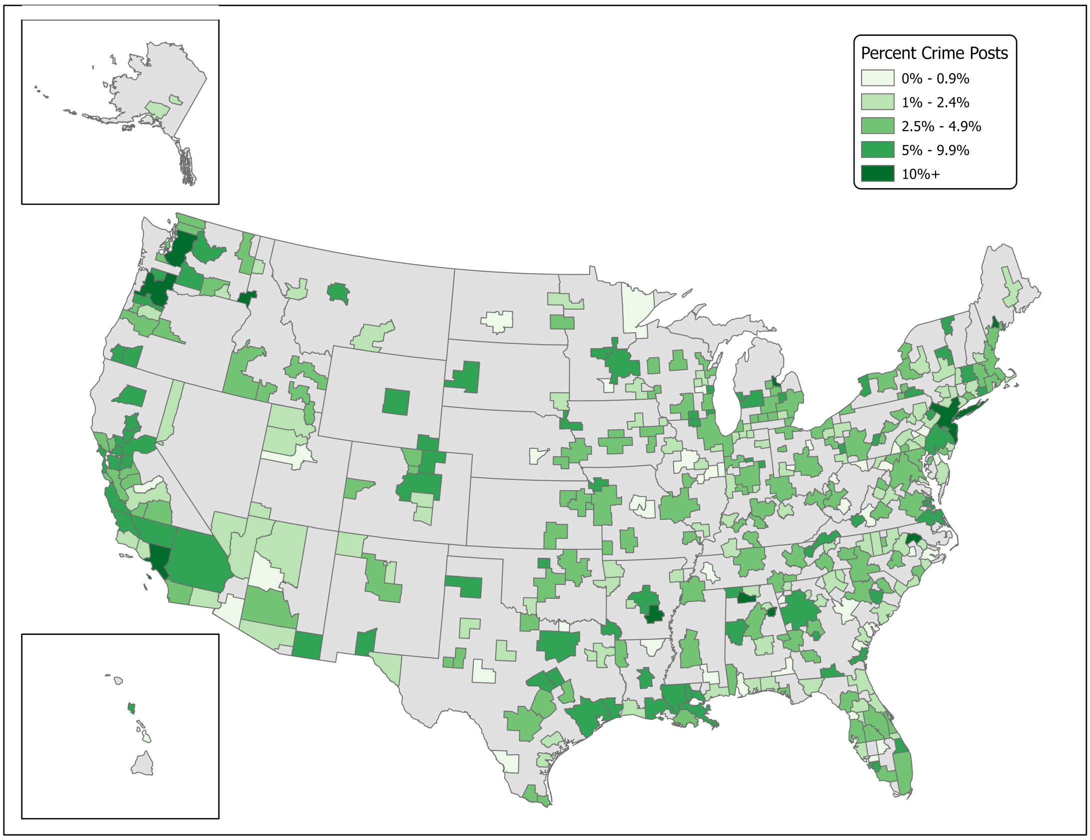

Figure 3.2: Map of crime discussion by MSA

3.4 Contextual Analysis

3.4.1 Population Demographics

The 384 MSAs of the United States vary greatly in terms of their actual population sizes. Once the Urban Area that makes the core of the Statistical Area reaches a population of at least 50,000, a Metropolitan Area forms in the county (or counties) that the Urban Area lies in. As a result of this definition, we break down MSAs into the four categories based on size discussed previously: very small, small, medium, and large. We use these categorizations to help explain trends in data relating to MSA population sizes.

To identify any potential trends in the population data, we take the mean crime discussion value of each bucket category. Figure 3.3 displays the average min-max normalization score for the percentage of crime discussion based on population size. Here, the very small population bucket contains 161 members with an average normalized score of 0.196, the small bucket contains 113 members with an average score of 0.173, the medium bucket contains 54 members with an average score of 0.166, and the large bucket contains 56 members with an average score of 0.260.

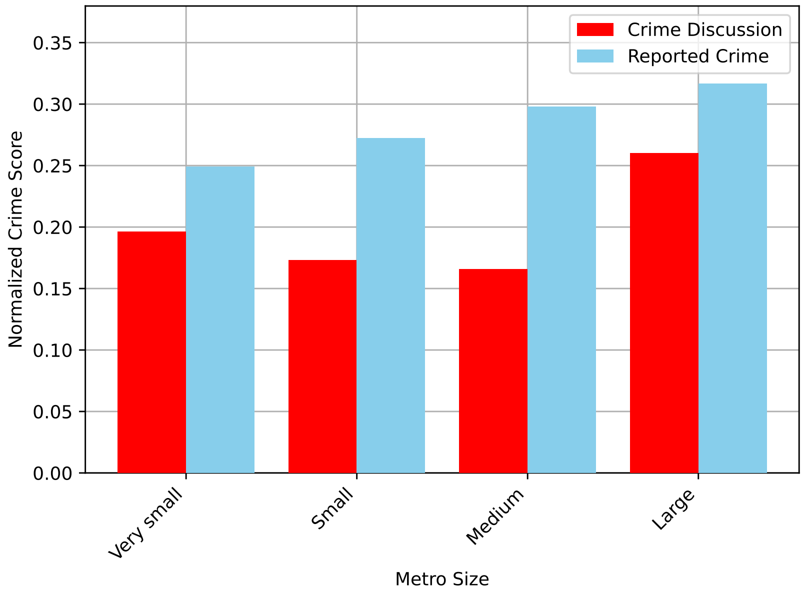

Figure 3.3: Comparison of discussed crime and documented crime by population

From Figure 3.3, we can see that as population increases, documented crime rates gradually increase. This trend is reflected in the crime discussion data. Crime discussion scores are roughly equivalent for the very small, small, and medium buckets. However, the large population bucket has a much higher crime discussion score than that of the other buckets. The large population bucket has a 6% greater score than the medium population bucket when comparing documented crime data, but the large population bucket has a 56% greater score than the medium population bucket when comparing crime discussion data.

Figure 3.4 shows a cumulative distribution function of the data by population bucket. Here, we can see that the members of the large population bucket have higher crime discussion scores. When looking at the very small MSAs, we notice quite a few outliers. We can see that some Metropolitan Areas in this category have among the highest amounts of crime discussion. However, almost 20% of very small MSA subreddits have not discussed crime at all. The small bucket has the next highest percentage of subreddits with zero crime discussion, with about 8% of communities not discussing crime at all.

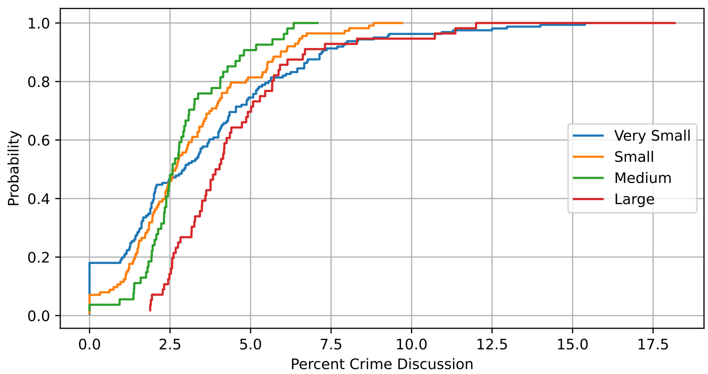

Figure 3.4: Crime discussion by MSA population class

When looking back at the map of crime discussion in Figure 3.2, we can see that the Metropolitan Areas with the most crime discussion are not bound by population size. MSAs both small and large are represented. However, this trend is not reflected among Metropolitan Areas with the least amount of crime discussion. These areas are almost exclusively in the very small and small population buckets, with no large population MSAs being represented. Having at least some degree of crime discussion in all MSAs and significant representation in the highest scoring MSAs leads to large population Metropolitan Areas scoring the greatest average crime discussion among all population classes.

3.4.2 Political Atmosphere

Similarly to population data, political affiliation is one possible variable that can sway crime discussion. To measure this, we look at the results of the 2020 presidential election. We consider the number of votes that the Democratic and Republican parties received as stand-ins for liberal and conservative political views, respectfully. Political affiliation of the 384 MSAs is broken down into the five categories discussed earlier: very conservative, conservative, moderate, liberal, and very liberal. Figure 3.5 shows a map of the political affiliation of all Metropolitan Areas in the United States.

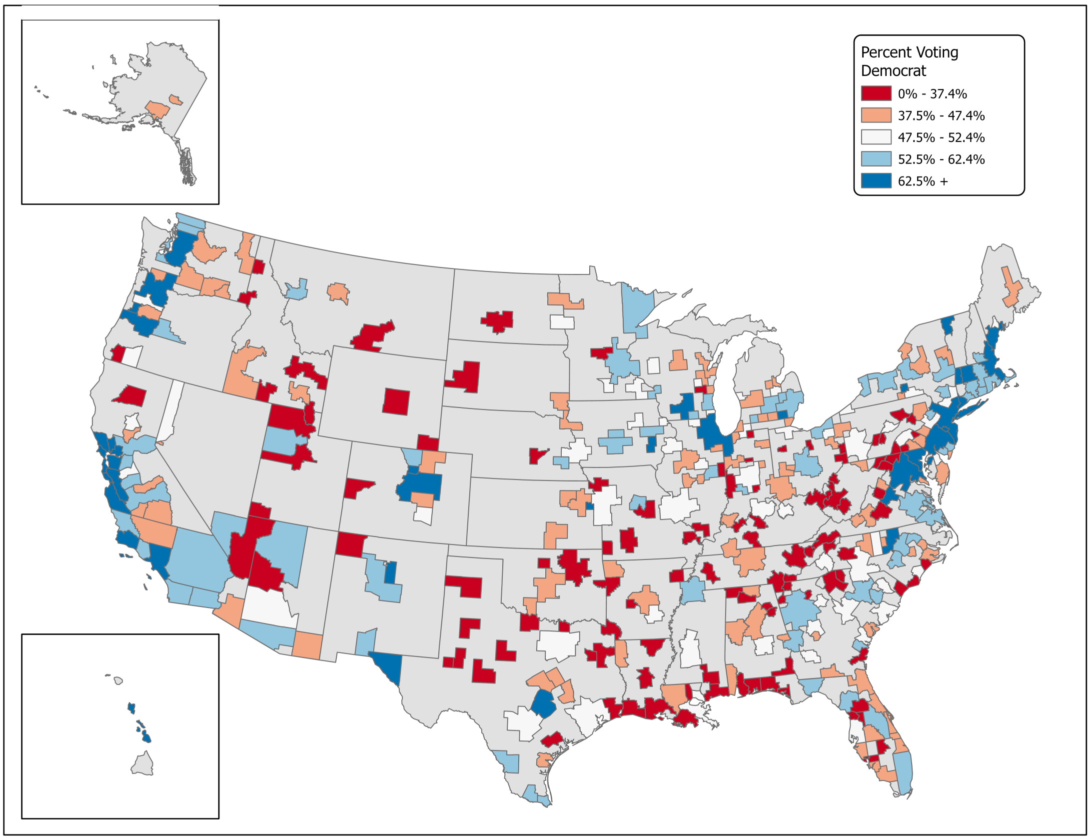

Figure 3.5: Map of 2020 presidential election results by MSA

In identifying trends based on politics, we follow a familiar method; we take the mean normalized crime discussion value and mean normalized documented crime rate of each political class. Figure 3.6 displays the average min-max normalization score for crime discussion and reported crime for each political class. Here, the very conservative bucket contains 112 members with a mean score of 0.170, the conservative bucket contains 103 members with a mean score of 0.179, the moderate bucket contains 57 members with a mean score of 0.207, the liberal bucket contains 73 members with a mean score of 0.201, and the very liberal bucket contains 39 members with a mean score of 0.274.

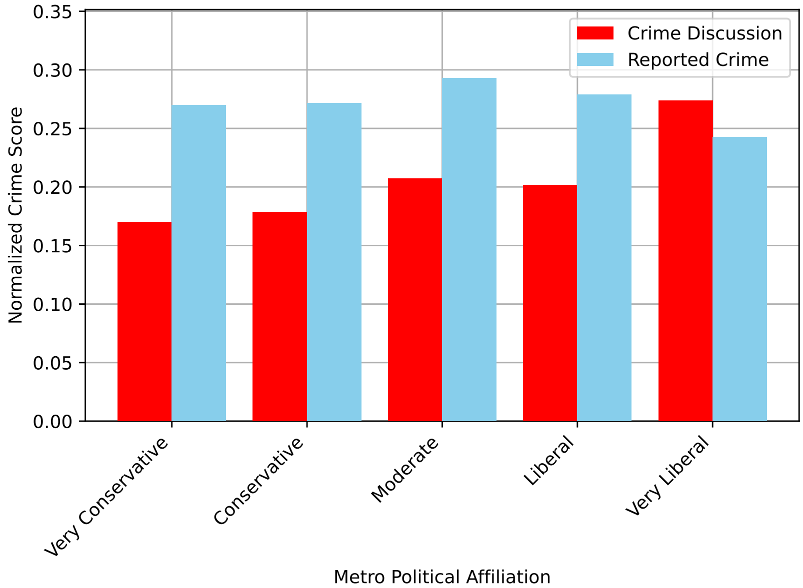

Figure 3.6: Comparison of discussed crime and documented crime by political class

When looking at Figure 3.6, we can see that among crime discussion rates, more liberal cities have the highest average scores. As the share of democratic voters increase, the average crime discussion score increases, with moderate MSAs scoring slightly higher than liberal MSAs. The very liberal class has a 32% higher crime discussion score than the next largest score. Additionally, the two conservative political classes have the lowest mean crime discussion scores. However, these trends are not reflected in the documented crime rate scores. Figure 3.6 shows that very liberal MSAs have a lower mean documented crime rate than more conservative areas, as well as the lowest mean score overall. Very liberal areas are also the only class in which the crime discussion score is higher than the reported crime rate score despite its low average documented crime.

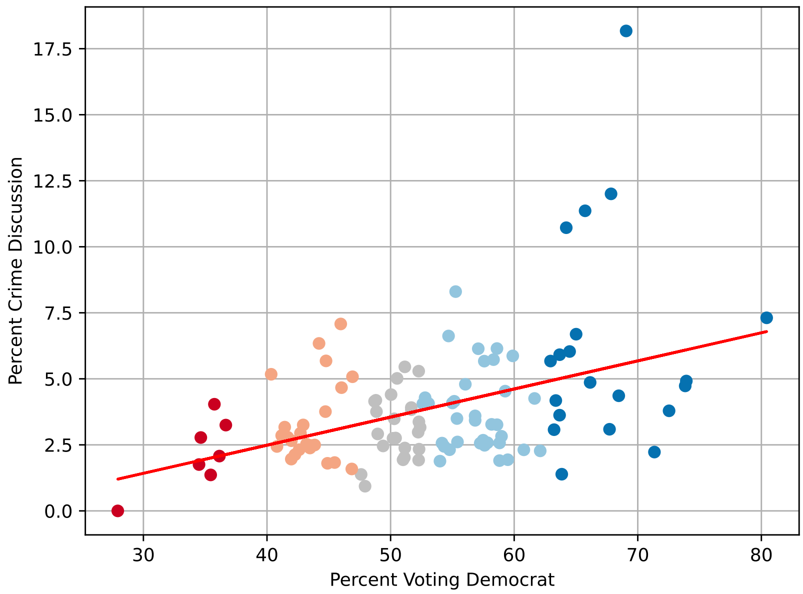

Figure 3.7: Crime discussion by percent voting Democrat

Figures 3.7 and 3.8 show two scatterplots: one which shows the percent crime discussion in relation to politics among medium and large sized MSAs and the other which shows the documented crime rate among the same sample group in relation to politics. The shades of the dots represent their political class. From the trendline in the data, we can see that as areas become more liberal, their crime discussion percentages increase fairly rapidly. The most conservative member of the plot, Provo-Orem, UT, did not include a single post discussing crime, while more liberal areas like Seattle, Los Angeles, Portland, and New York City had values over 10%. In contrast, when looking at the documented violent crime rate among large and mid-size MSAs, more liberal areas generally have lower average crime rates. MSAs that lean conservative, like Birmingham-Hoover, AL and Kansas City, MO-KS have among the highest crime rates, as well as more politically moderate cities, including Memphis, TN-MS-AR.

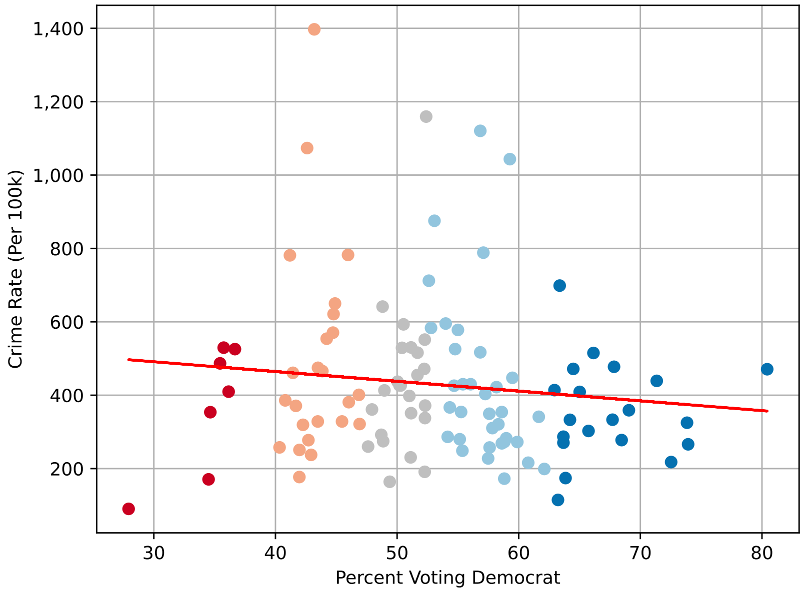

Figure 3.8: Crime discussion by documented crime rate

When considering the data, we can see a strong relationship between how liberal a Metropolitan Area votes to how high its average crime discussion lies. Despite having lower than average documented crime, very liberal Metropolitan Areas have the highest crime discussion rates.

3.4.3 Reported Crime

Another factor to consider alongside crime discussion is the reported crime rate of an MSA. Higher documented crime rates can lead to increased amounts of crime discussion, as crime may affect more members of the community. To measure MSA crime rates, we use the FBI's Uniform Crime Reporting Estimates [13]. The 384 Metropolitan Areas represent a differing level of crime rates, with a high of 1,397.2 from Birmingham-Hoover, AL, a low of 44.3 from Bangor, ME, and an average of 380.9 among all MSAs. Table 3.3 displays the MSAs with the ten highest documented crime rates. In this case, crime rates are defined as the amount of violent crime per 100,000 people. Additionally, the map in Figure 3.9 displays the crime rate for all Metropolitan Areas.

Table 3.4: MSAs with highest crime rates

Metropolitan Statistical Area	Crime Rate	% Crime Discussion
Birmingham-Hoover, AL	1,397.2	2.52
Farmington, NM	1,199.8	1.61
Anchorage, AK	1,194.6	2.36
Kansas City, MO-KS	1,159.4	3.16
Memphis, TN-MS-AR	1,120.5	3.43
Winston-Salem, NC	1,073.6	2.32
Albuquerque, NM	1,043.4	4.53
Danville, IL	935.7	0.00
Pine Bluff, AR	895.4	12.5
Odessa, TX	881.8	4.30

Table 3.4 represents a diverse group of Metropolitan Areas in regards to population, politics, and crime discussion rates. Populations range from very small MSAs to large MSAs. Most of the MSAs lean conservative, but there are some liberal cities included, like Albuquerque, NM. Additionally, a wide array of crime discussion percentages are noted, including Danville, IL at a low of 0% and Pine Bluff, AR at a high of 12.5%.

From Tables 3.3 and 3.4, as well as Figures 3.2 and 3.9, we can see that MSAs with high reported crime rates do not necessarily have high amounts of crime discussion. Examples include Metropolitan Areas like Danville, IL and Farmington, NM. The same goes for MSAs with low crime rates and low amounts of crime discussion. New York-Newark-Jersey City, NY-NJ-PA and Seattle-Tacoma-Bellevue, WA have significantly higher normalized crime discussion rates than normalized documented crime rates.

From the map in Figure 3.9, we can see that many Metropolitan Areas with

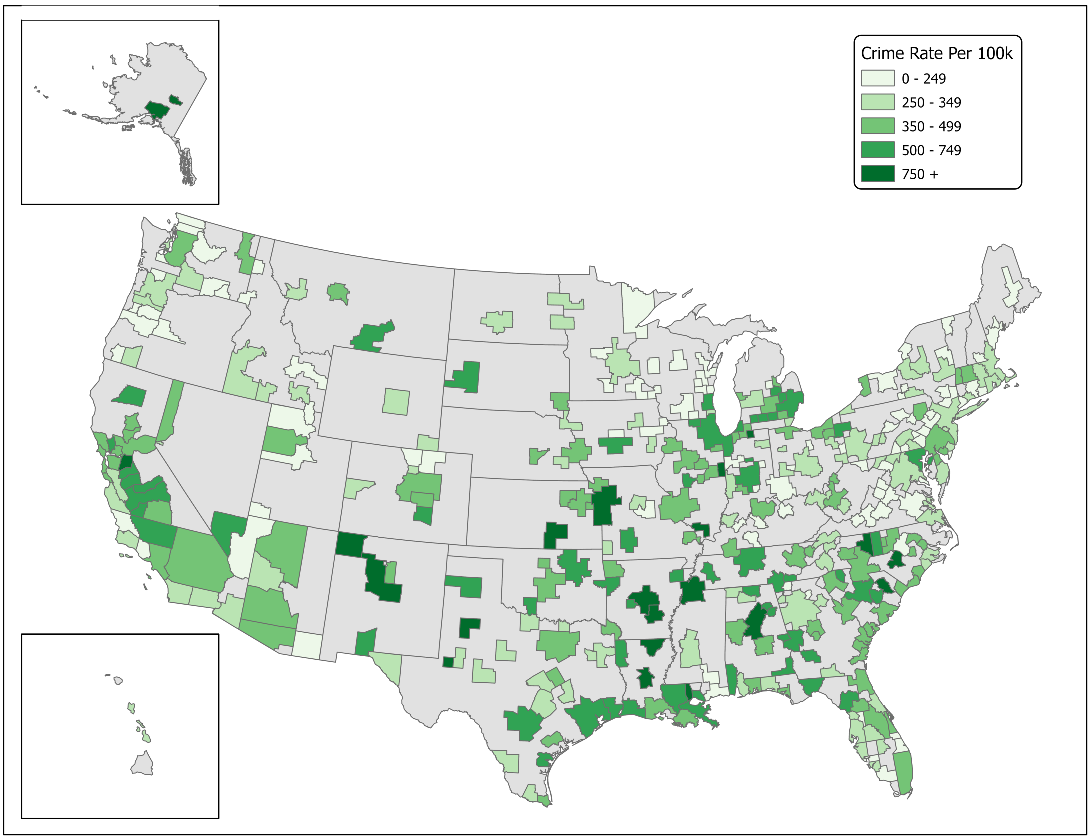

Figure 3.9: Reported Crime Rate by MSA

high documented crime rates lie in the South and Midwest [13]. However, many these MSAs have low amounts of crime discussion. To calculate the discrepancies between crime discussion and reported crime, we calculate the difference between the min-max normalized crime discussion score and the min-maxed documented crime score. Figure 3.10 maps this normalized difference score. Higher positive values indicate that the amount of crime discussion is higher than actual crime, while lower negative values indicate that actual crime is higher than crime discussion. MSAs in which the ratio of crime discussion and documented crime are roughly equal rest around a score of zero.

Table 3.5 displays the 15 highest and lowest scoring MSAs for the crime discussion to crime rate ratio. A full listing can be found in Appendix A. Of the highest scoring MSAs, large and very small Metropolitan Areas are over-represented, while

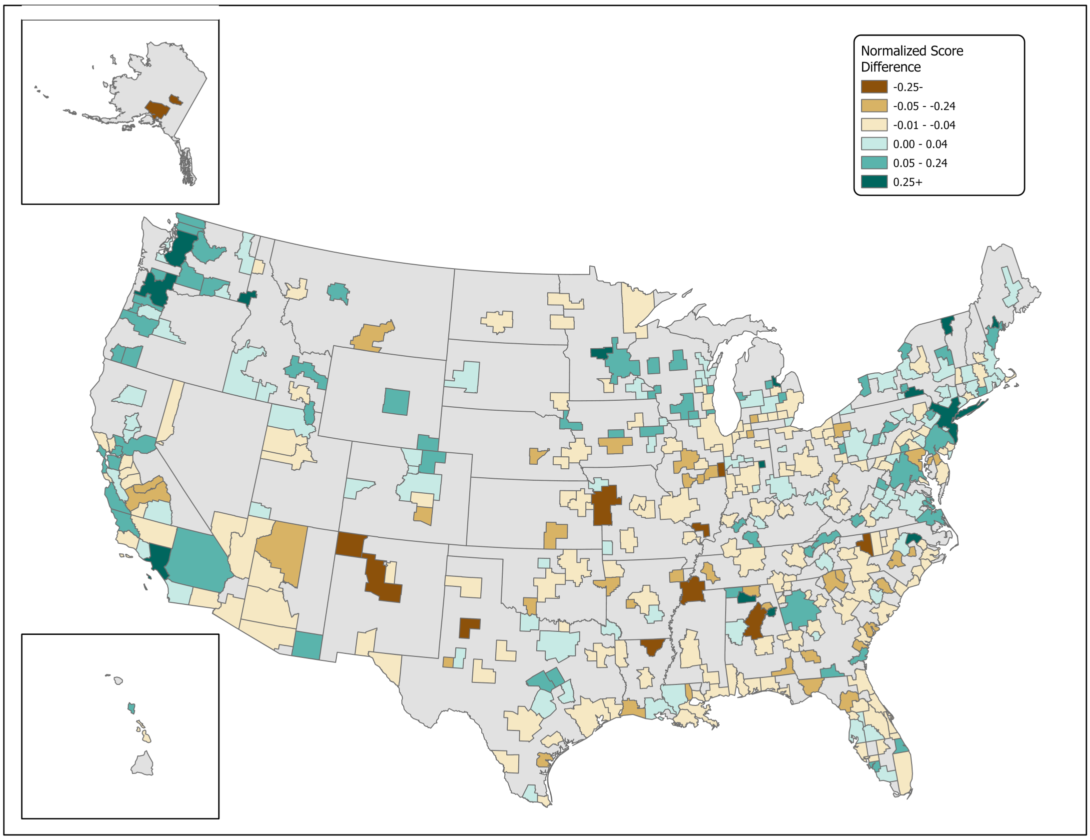

Figure 3.10: Reported Crime Rate by MSA

all population sizes are represented among the lowest scoring Metropolitan Areas. When considering political affiliation, very liberal cities represent 33% of the highest scoring MSAs while only consisting of 10% of the total number of Metropolitan Areas. For the lowest scoring MSAs, conservative and very conservative Metropolitan Areas represent 80% of the lowest scoring members while only making up 56% of the total number of members.

When considering the documented crime data, we can see that high documented crime rates do not necessarily lead to high crime discussion rates, and vice versa. Figure 3.10 and Table 3.5 show that more liberal and larger MSAs in the Northeast and on the West Coast have higher rates of crime discussion than crime rates, while smaller, more conservative MSAs in the South and Midwest have higher crime rates than crime discussion. As a result of these differing scores, no

Table 3.5: MSAs by highest and lowest difference score

Higher Score		Lower Score	
Metropolitan Area	Score	Metropolitan Area	Score
Seattle-Tacoma-Bellevue, WA	0.743	Birmingham-Hoover, AL	-0.861
Lewiston, ME	0.723	Farmington, NM	-0.770
Lewiston, ID-WA	0.614	Anchorage, AK	-0.725
Bay City, MI	0.496	Danville, IL	-0.670
Portland, OR-WA	0.409	Kansas City, MO-KS	-0.656
Decatur, AL	0.381	Winston-Salem, NC	-0.641
New York, NY-NJ-PA	0.352	Memphis, TN-MS-AR	-0.613
Anniston, AL	0.344	Monroe, LA	-0.582
Rocky Mount, NC	0.323	Carbondale, IL	-0.501
Los Angeles, CA	0.319	Albuquerque, NM	-0.497
St. Cloud, MN	0.308	Fairbanks, AK	-0.490
Burlington, VT	0.290	Lubbock, TX	-0.489
Binghamton, NY	0.273	Corpus Christi, TX	-0.448
Muncie, IN	0.262	Jackson, TN	-0.439
Sheboygan, WI	0.238	Dothan, AL	-0.431

relationship between crime discussion and reported crime rates emerges.

3.5 Sentiment Analysis and Overlap

From the data collected by our Reddit crawler, we have seen that crime discussion is widespread among the selected communities. However, while we have analyzed posts about crime, we have not yet analyzed the users that initiate this crime discussion. In doing so, we can potentially identify trends in the data that could highlight motivations behind crime posters or their submission patterns. We can also attempt to identify if crime posters are truly local by seeing if they overlap with other state and local subreddits.

To analyze users' posting histories, we extract their submission and comment data. For this work, we use posters to `r/nyc`, the New York City subreddit, as a sample. We have extracted all 1,197 unique posters from the 1,828 posts to `r/nyc`. 134 of these posters have posted about crime at least once, while 1,063 have not. Seven users have posted about crime five or more times. Additionally, of the 134 crime posters, five have had their account suspended since May 31st, 2022. This accounts for 3.7% of the crime posters. Of the 1,063 non-crime posters, 27 have had their accounts suspended, or 2.5%. After fetching the unique users, we gathered posts and comments from their user history. In total, 11,926 comments and posts were gathered from crime-posting users, while 87,658 comments and posts were gathered from non-crime-posting users.

One method of measuring user interaction patterns is through performing sentiment analysis on post and comment data. In this case, we measure the mean polarity score of users who post about crime and compare it to the mean polarity score of users who don't post about crime. A lower polarity means that the input data is more negative, while higher scores mean that the sentiment is more positive [20]. This can indicate the potential nature of a poster; if their polarity score is low, then they are a consistently negative voice in the community. Enough negative voices can detract from a community and introduce conflict and disarray.

Table 3.6: Sentiment polarity scores for `nyc` users

	Crime posters	Non-crime posters
Within `r/nyc`	-0.07	0.09
Outside of `r/nyc`	-0.07	0.10

Table 3.6 displays the results of the sentiment analysis on users that submitted a post to the New York City subreddit. Comments and posts were divided between those within `r/nyc` and those outside of the community. The sentiment of crime-posters was nearly identical within and outside of the community. These poster were more on the negative side, with neutrality laying at zero. Non-crime posting users were generally more positive, scoring 0.09 within the community and 0.10

outside of the community.

Another method of analyzing user interaction patterns is through identifying other communities that `r/nyc` members post in. This can potentially identify important posting trends. Examples include location identity and political affiliation. To measure this, we use the Jaccard Index [11]. The Jaccard Index takes the following form, where $0 \leq J(A, B) \leq 1$ and A and B represent subreddits:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Table 3.7 displays the subreddits with the highest overlap for both crime and non-crime posters. Crime posters and non-crime posters alike share some common, New York City-related subreddits, like `NYCBike`. However, crime posters have high overlap between two subreddits specifically meant to discuss crime in New York City: `CrimeInNYC` and `NYStateOfMind`.

Table 3.7: Subreddits with highest overlap by poster type

Crime poster		Non-crime poster	
Subreddit	Score	Subreddit	Score
`CrimeInNYC`	0.83	`nycrail`	0.81
`Judaism`	0.42	`AskNYC`	0.64
`AskNYC`	0.29	`RunNYC`	0.48
`NYCbike`	0.25	`astoria`	0.42
`newyorkcity`	0.21	`Brooklyn`	0.43
`NYYankees`	0.16	`NYCBike`	0.42
`rangers`	0.15	`newyorkcity`	0.40
`RealTesla`	0.14	`jerseycity`	0.37
`NYStateOfMind`	0.12	`BostonU`	0.33
`GenX`	0.11	`Baruch`	0.32

From the overlap data, we can see that crime is a frequent topic of discussion even outside the `nyc` subreddit for the identified crime posters. However, the overlap with several New York City-related communities indicates that many of the crime posters have some relation to the city itself. When compared to non-crime posters, crime posters exhibit less of a sense of locality. Nine of the ten communities that non-crime posters have strong overlap with are explicitly related to New York City, while only seven of the ten communities that crime posters overlap with are geospatially defined.

While the two types of posters share some subreddits, their political community overlap is fairly different. To identify political subreddits, we derive the methodology from Rajadesingan et al. [29]. Table 3.8 shows the political subreddits with the largest amount of overlap.

Table 3.8: Political subreddits with highest overlap by poster type

Crime poster		Non-crime poster	
Subreddit	Score	Subreddit	Score
`neoliberal`	0.05	`VoteDEM`	0.16
`LockdownSkepticism`	0.03	`tuesday`	0.10
`Conservative`	0.02	`neoliberal`	0.05

Here, we see that the two groups diverge on political opinions. Those that post about crime trend more towards the right; `LockdownSkepticism` is a community critical of Covid-19 restrictions and `Conservative` is Reddit's largest right-leaning political community. Non-crime posters are active in politically moderate communities like `tuesday` and `neoliberal`, as well as `VoteDem`, which supports the Democratic Party. Crime posting users tend to be more active in political communities, with their top three overlapping political subreddits ranking 19th, 23rd, and 32nd. Non-crime posting users' highest overlapping political communities rank 21st, 47th, and 77th.

From our user-level analysis, we have identified several characteristics of crime posters. First, these users are consistently more negative than non-crime posting users. They post about crime frequently, as indicated by their significant overlap with crime discussion specific communities. However, they are often somewhat local to the communities in which they post, as they have significant overlap to location specific hobby and sport communities. Despite this, crime posters have a smaller degree of overlap compared to non-crime posters. Lastly, those that post about crime are more active in conservative political communities than those who don't post about crime.

4 Conclusion

We have discussed the nature and prevalence of crime discussion on Reddit, a social media website, as it pertains to location-based communities. To obtain the necessary data and information, we have created a highly resilient Reddit crawler that fetches all posts from communities that represent the 384 Metropolitan Statistical Areas of the United States. The crawler gathered 161,166 posts over a two-month time span. Of these 161,166 posts, a total of 6,792 were determined to discuss crime.

Through analysis of user-submitted posts, we have found that communities that represent larger Metropolitan Statistical Areas and more liberal Metropolitan Statistical Areas tend to have higher rates of crime discussion. However, there is no correlation between high documented crime rates and high crime discussion rates. Furthermore, we have found that the users that make submissions about crime are significantly more negative in their posting and commenting style than those that post about non-crime topics. By measuring community overlap, we have found that these users are generally local to the communities they post in, although to a lesser degree than non-crime posters, and more conservative than their community as a whole.

The findings in this work act as a counterargument to popular narratives surrounding crime. While elevated crime discussion rates highlight beliefs of increased amounts of violence in larger, more liberal cities, our research shows that crime discussion is not representative of documented crime rates. Furthermore, we find that users that post about crime to the New York City community lean more conservative, a contrast to the liberal city they live in. In posting about crime, these users might be performing an act of political protest. They are dissatisfied with the community around them and are voicing their discontent, even if it is not grounded in fact. However, by providing a data-driven approach, this paper helps to combat any potential misconceptions or biases regarding crime.

While this paper provides a first look at the crime discussion dynamics on Reddit, there are several paths that continued research could explore. One pathway of future work on the relationship between crime discussion and social media could include a more in-depth analysis of users that post about crime beyond the sample of `nyc` users that we looked at. Another approach could be applying the tool and analysis used on Reddit to other social media platforms, like Facebook or Nextdoor, to determine if this phenomenon is Reddit-specific.

Appendix A

Metropolitan Statistical Area	% Crime Posts	Diff. Score
Abilene, TX Metro Area	1.27	-0.16
Akron, OH Metro Area	2.34	-0.14
Albany, GA Metro Area	3.39	-0.33
Albany-Lebanon, OR Metro Area	1.67	-0.01
Albany-Schenectady-Troy, NY Metro Area	2.48	-0.05
Albuquerque, NM Metro Area	4.53	-0.5
Alexandria, LA Metro Area	8.82	-0.12
Allentown-Bethlehem-Easton, PA-NJ Metro Area	2.46	0.02
Altoona, PA Metro Area	3.45	-0.05
Amarillo, TX Metro Area	6.6	-0.06
Ames, IA Metro Area	4.55	0.11
Anchorage, AK Metro Area	2.36	-0.73
Ann Arbor, MI Metro Area	1.2	-0.19
Anniston-Oxford, AL Metro Area	15.69	0.34
Appleton, WI Metro Area	1.53	-0.03

Asheville, NC Metro Area	3.35	-0.04
Athens-Clarke County, GA Metro Area	1.95	-0.11
Atlanta-Sandy Springs-Alpharetta, GA Metro Area	5.73	0.09
Atlantic City-Hammonton, NJ Metro Area	1.54	-0.1
Auburn-Opelika, AL Metro Area	4.35	-0.13
Augusta-Richmond County, GA-SC Metro Area	0.93	-0.21
Austin-Round Rock-Georgetown, TX Metro Area	3.62	-0.01
Bakersfield, CA Metro Area	5.68	-0.13
Baltimore-Columbia-Towson, MD Metro Area	4.18	-0.27
Bangor, ME Metro Area	1.47	0.05
Barnstable Town, MA Metro Area	1.84	-0.13
Baton Rouge, LA Metro Area	6.34	-0.05
Battle Creek, MI Metro Area	4.88	-0.18
Bay City, MI Metro Area	14.0	0.5
Beaumont-Port Arthur, TX Metro Area	5.94	-0.06
Beckley, WV Metro Area	0.0	-0.24
Bellingham, WA Metro Area	4.06	0.07
Bend, OR Metro Area	2.88	0.04

Billings, MT Metro Area	1.54	-0.3
Binghamton, NY Metro Area	8.68	0.27
Birmingham-Hoover, AL Metro Area	2.53	-0.86
Bismarck, ND Metro Area	0.0	-0.22
Blacksburg-Christiansburg, VA Metro Area	6.67	0.23
Bloomington, IL Metro Area	1.63	-0.16
Bloomington, IN Metro Area	3.38	-0.07
Bloomsburg-Berwick, PA Metro Area	1.89	-0.03
Boise City, ID Metro Area	2.66	-0.03
Boston-Cambridge-Newton, MA-NH Metro Area	4.36	0.04
Boulder, CO Metro Area	4.37	0.05
Bowling Green, KY Metro Area	3.7	0.09
Bremerton-Silverdale-Port Orchard, WA Metro Area	4.11	0.04
Bridgeport-Stamford-Norwalk, CT Metro Area	1.39	-0.05
Brownsville-Harlingen, TX Metro Area	3.19	-0.09
Brunswick, GA Metro Area	7.14	0.13
Buffalo-Cheektowaga, NY Metro Area	8.3	0.2
Burlington, NC Metro Area	4.0	-0.07

Burlington-South Burlington, VT Metro Area	7.88	0.29
California-Lexington Park, MD Metro Area	1.23	-0.08
Canton-Massillon, OH Metro Area	2.67	-0.1
Cape Coral-Fort Myers, FL Metro Area	5.17	0.1
Cape Girardeau, MO-IL Metro Area	2.94	-0.11
Carbondale-Marion, IL Metro Area	1.61	-0.5
Carson City, NV Metro Area	1.25	-0.18
Casper, WY Metro Area	7.14	0.19
Cedar Rapids, IA Metro Area	4.86	0.13
Chambersburg-Waynesboro, PA Metro Area	0.0	-0.12
Champaign-Urbana, IL Metro Area	0.0	-0.35
Charleston, WV Metro Area	3.17	-0.15
Charleston-North Charleston, SC Metro Area	1.94	-0.18
Charlotte-Concord-Gastonia, NC-SC Metro Area	2.75	-0.15
Charlottesville, VA Metro Area	2.35	0.0
Chattanooga, TN-GA Metro Area	3.25	-0.2
Cheyenne, WY Metro Area	7.36	0.19

Chicago-Naperville-Elgin, IL-IN-WI Metro Area	4.86	-0.1
Chico, CA Metro Area	6.52	0.04
Cincinnati, OH-KY-IN Metro Area	3.25	0.01
Clarksville, TN-KY Metro Area	1.34	-0.2
Cleveland, TN Metro Area	3.33	-0.22
Cleveland-Elyria, OH Metro Area	2.56	-0.15
Coeur d'Alene, ID Metro Area	1.55	-0.07
College Station-Bryan, TX Metro Area	2.58	-0.05
Colorado Springs, CO Metro Area	2.5	-0.2
Columbia, MO Metro Area	0.75	-0.15
Columbia, SC Metro Area	4.29	-0.18
Columbus, GA-AL Metro Area	2.58	-0.24
Columbus, IN Metro Area	1.49	-0.02
Columbus, OH Metro Area	2.57	-0.06
Corpus Christi, TX Metro Area	1.44	-0.45
Corvallis, OR Metro Area	4.03	0.11
Crestview-Fort Walton Beach-Destin, FL Metro Area	1.38	-0.15
Cumberland, MD-WV Metro Area	2.08	-0.08
Dallas-Fort Worth-Arlington, TX Metro Area	5.45	0.05
Dalton, GA Metro Area	1.15	-0.08

Danville, IL Metro Area	0.0	-0.67
Daphne-Fairhope-Foley, AL Metro Area	0.0	-0.19
Davenport-Moline-Rock Island, IA-IL Metro Area	2.16	-0.14
Dayton-Kettering, OH Metro Area	1.83	-0.13
Decatur, AL Metro Area	10.94	0.38
Decatur, IL Metro Area	0.0	-0.3
Deltona-Daytona Beach-Ormond Beach, FL Metro Area	2.13	-0.11
Denver-Aurora-Lakewood, CO Metro Area	5.67	0.02
Des Moines-West Des Moines, IA Metro Area	4.06	-0.29
Detroit-Warren-Dearborn, MI Metro Area	3.6	-0.17
Dothan, AL Metro Area	0.0	-0.43
Dover, DE Metro Area	0.0	-0.31
Dubuque, IA Metro Area	1.27	-0.03
Duluth, MN-WI Metro Area	0.62	-0.12
Durham-Chapel Hill, NC Metro Area	2.23	-0.19
East Stroudsburg, PA Metro Area	6.25	0.17
Eau Claire, WI Metro Area	4.55	0.1
El Centro, CA Metro Area	1.92	-0.14

El Paso, TX Metro Area	3.09	-0.07
Elizabethtown-Fort Knox, KY Metro Area	1.92	0.03
Elkhart-Goshen, IN Metro Area	4.92	-0.29
Elmira, NY Metro Area	1.96	-0.05
Enid, OK Metro Area	5.26	0.02
Erie, PA Metro Area	3.36	-0.03
Eugene-Springfield, OR Metro Area	3.97	0.05
Evansville, IN-KY Metro Area	1.12	-0.17
Fairbanks, AK Metro Area	1.42	-0.49
Fargo, ND-MN Metro Area	2.77	-0.06
Farmington, NM Metro Area	1.61	-0.77
Fayetteville, NC Metro Area	4.07	-0.4
Fayetteville-Springdale-Rogers, AR Metro Area	2.44	-0.14
Flagstaff, AZ Metro Area	1.07	-0.27
Flint, MI Metro Area	9.71	0.12
Florence, SC Metro Area	3.51	-0.4
Florence-Muscle Shoals, AL Metro Area	6.45	0.14
Fond du Lac, WI Metro Area	0.0	-0.13
Fort Collins, CO Metro Area	3.0	-0.0
Fort Smith, AR-OK Metro Area	1.52	-0.34
Fort Wayne, IN Metro Area	2.44	-0.08

Fresno, CA Metro Area	1.88	-0.32
Gadsden, AL Metro Area	2.04	-0.27
Gainesville, FL Metro Area	2.57	-0.35
Gainesville, GA Metro Area	1.61	-0.06
Gettysburg, PA Metro Area	2.86	0.06
Glens Falls, NY Metro Area	5.08	0.19
Goldsboro, NC Metro Area	0.0	-0.3
Grand Forks, ND-MN Metro Area	1.32	-0.11
Grand Island, NE Metro Area	0.0	-0.27
Grand Junction, CO Metro Area	4.32	0.03
Grand Rapids-Kentwood, MI Metro Area	5.08	0.05
Grants Pass, OR Metro Area	7.25	0.23
Great Falls, MT Metro Area	7.69	0.17
Greeley, CO Metro Area	6.38	0.2
Green Bay, WI Metro Area	3.62	0.04
Greensboro-High Point, NC Metro Area	2.38	-0.25
Greenville, NC Metro Area	2.06	-0.15
Greenville-Anderson, SC Metro Area	1.36	-0.27
Gulfport-Biloxi, MS Metro Area	1.15	-0.1
Hagerstown-Martinsburg, MD-WV Metro Area	3.41	-0.01
Hammond, LA Metro Area	5.56	-0.3

Hanford-Corcoran, CA Metro Area	4.0	-0.14
Harrisburg-Carlisle, PA Metro Area	1.38	-0.11
Harrisonburg, VA Metro Area	3.72	0.11
Hartford-East Hartford-Middletown, CT Metro Area	2.27	-0.02
Hattiesburg, MS Metro Area	0.0	-0.17
Hickory-Lenoir-Morganton, NC Metro Area	3.57	-0.11
Hilton Head Island-Bluffton, SC Metro Area	0.0	-0.28
Hinesville, GA Metro Area	0.0	-0.3
Homosassa Springs, FL Metro Area	0.0	-0.18
Hot Springs, AR Metro Area	1.98	-0.27
Houma-Thibodaux, LA Metro Area	3.51	-0.07
Houston-The Woodlands-Sugar Land, TX Metro Area	5.02	-0.15
Huntington-Ashland, WV-KY-OH Metro Area	4.11	0.04
Huntsville, AL Metro Area	1.94	-0.29
Idaho Falls, ID Metro Area	4.26	0.09
Indianapolis-Carmel-Anderson, IN Metro Area	4.19	-0.23
Iowa City, IA Metro Area	2.98	-0.03

Ithaca, NY Metro Area	2.68	0.06
Jackson, MI Metro Area	4.0	-0.15
Jackson, MS Metro Area	3.37	-0.06
Jackson, TN Metro Area	0.0	-0.44
Jacksonville, FL Metro Area	2.48	-0.2
Jacksonville, NC Metro Area	1.85	-0.09
Janesville-Beloit, WI Metro Area	5.17	0.12
Jefferson City, MO Metro Area	0.0	-0.16
Johnson City, TN Metro Area	2.13	-0.12
Johnstown, PA Metro Area	7.14	0.2
Jonesboro, AR Metro Area	2.0	-0.29
Joplin, MO Metro Area	4.17	-0.01
Kahului-Wailuku-Lahaina, HI Metro Area	0.98	-0.14
Kalamazoo-Portage, MI Metro Area	2.55	-0.27
Kankakee, IL Metro Area	1.89	-0.17
Kansas City, MO-KS Metro Area	3.16	-0.66
Kennewick-Richland, WA Metro Area	3.94	0.06
Killeen-Temple, TX Metro Area	5.68	0.12
Kingsport-Bristol, TN-VA Metro Area	6.15	0.12
Kingston, NY Metro Area	1.92	-0.0
Knoxville, TN Metro Area	2.78	-0.1
Kokomo, IN Metro Area	6.45	-0.04

La Crosse-Onalaska, WI-MN Metro Area	4.9	0.18
Lafayette, LA Metro Area	6.12	0.0
Lafayette-West Lafayette, IN Metro Area	2.75	-0.01
Lake Charles, LA Metro Area	2.08	-0.28
Lake Havasu City-Kingman, AZ Metro Area	1.96	-0.06
Lakeland-Winter Haven, FL Metro Area	2.94	-0.04
Lancaster, PA Metro Area	1.96	-0.02
Lansing-East Lansing, MI Metro Area	4.79	-0.04
Laredo, TX Metro Area	0.0	-0.23
Las Cruces, NM Metro Area	5.53	-0.09
Las Vegas-Henderson-Paradise, NV Metro Area	2.32	-0.25
Lawrence, KS Metro Area	3.67	-0.08
Lawton, OK Metro Area	1.82	-0.41
Lebanon, PA Metro Area	0.0	-0.12
Lewiston, ID-WA Metro Area	12.96	0.61
Lewiston-Auburn, ME Metro Area	15.38	0.72
Lexington-Fayette, KY Metro Area	2.02	-0.05
Lima, OH Metro Area	4.0	-0.01
Lincoln, NE Metro Area	3.05	-0.08
Little Rock-North Little Rock-Conway, AR Metro Area	7.08	-0.17

Logan, UT-ID Metro Area	3.42	0.1
Longview, TX Metro Area	1.49	-0.15
Longview, WA Metro Area	5.36	0.16
Los Angeles-Long Beach-Anaheim, CA Metro Area	12.0	0.32
Louisville/Jefferson County, KY-IN Metro Area	4.4	-0.07
Lubbock, TX Metro Area	2.02	-0.49
Lynchburg, VA Metro Area	2.7	-0.02
Macon-Bibb County, GA Metro Area	2.88	-0.17
Madera, CA Metro Area	0.0	-0.38
Madison, WI Metro Area	3.79	0.05
Manchester-Nashua, NH Metro Area	1.23	-0.1
Manhattan, KS Metro Area	4.05	-0.09
Mankato, MN Metro Area	0.0	-0.13
Mansfield, OH Metro Area	0.0	-0.16
McAllen-Edinburg-Mission, TX Metro Area	3.26	-0.01
Medford, OR Metro Area	6.74	0.13
Memphis, TN-MS-AR Metro Area	3.43	-0.61
Merced, CA Metro Area	4.29	-0.16
Miami-Fort Lauderdale-Pompano Beach, FL Metro Area	3.28	-0.12

Michigan City-La Porte, IN Metro Area	1.1	-0.23
Midland, MI Metro Area	3.08	0.07
Midland, TX Metro Area	4.3	0.01
Milwaukee-Waukesha, WI Metro Area	4.09	-0.19
Minneapolis-St. Paul-Bloomington, MN-WI Metro Area	5.87	0.13
Missoula, MT Metro Area	2.45	-0.12
Mobile, AL Metro Area	2.79	-0.22
Modesto, CA Metro Area	2.76	-0.23
Monroe, LA Metro Area	0.0	-0.58
Monroe, MI Metro Area	3.28	-0.01
Montgomery, AL Metro Area	2.36	-0.2
Morgantown, WV Metro Area	0.95	-0.09
Morristown, TN Metro Area	9.3	0.23
Mount Vernon-Anacortes, WA Metro Area	3.66	0.09
Muncie, IN Metro Area	7.94	0.26
Muskegon, MI Metro Area	7.21	0.07
Myrtle Beach-Conway-North Myrtle Beach, SC-NC Metro Area	1.74	-0.16
Napa, CA Metro Area	4.72	-0.13
Naples-Marco Island, FL Metro Area	3.72	0.04

Nashville-Davidson–Murfreesboro–Franklin, TN Metro Area	3.76	-0.2
New Bern, NC Metro Area	0.0	-0.2
New Haven-Milford, CT Metro Area	2.57	-0.05
New Orleans-Metairie, LA Metro Area	5.29	-0.1
New York-Newark-Jersey City, NY-NJ-PA Metro Area	10.72	0.35
Niles, MI Metro Area	4.08	-0.17
North Port-Sarasota-Bradenton, FL Metro Area	2.38	-0.1
Norwich-New London, CT Metro Area	6.48	0.22
Ocala, FL Metro Area	2.42	-0.18
Ocean City, NJ Metro Area	1.67	-0.03
Odessa, TX Metro Area	4.3	-0.39
Ogden-Clearfield, UT Metro Area	1.75	-0.03
Oklahoma City, OK Metro Area	3.17	-0.16
Olympia-Lacey-Tumwater, WA Metro Area	3.54	0.02
Omaha-Council Bluffs, NE-IA Metro Area	2.91	-0.14
Orlando-Kissimmee-Sanford, FL Metro Area	2.61	-0.16
Oshkosh-Neenah, WI Metro Area	1.75	-0.04

Owensboro, KY Metro Area	3.17	0.08
Oxnard-Thousand Oaks-Ventura, CA Metro Area	2.31	-0.03
Palm Bay-Melbourne-Titusville, FL Metro Area	2.79	-0.11
Panama City, FL Metro Area	2.78	-0.16
Parkersburg-Vienna, WV Metro Area	0.0	-0.31
Pensacola-Ferry Pass-Brent, FL Metro Area	2.07	-0.18
Peoria, IL Metro Area	0.96	-0.27
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD Metro Area	6.69	0.08
Phoenix-Mesa-Chandler, AZ Metro Area	3.49	-0.11
Pine Bluff, AR Metro Area	12.5	0.05
Pittsburgh, PA Metro Area	3.76	0.01
Pittsfield, MA Metro Area	2.0	-0.16
Pocatello, ID Metro Area	2.53	-0.07
Port St. Lucie, FL Metro Area	5.48	0.13
Portland-South Portland, ME Metro Area	3.07	0.09
Portland-Vancouver-Hillsboro, OR-WA Metro Area	11.36	0.41

Poughkeepsie-Newburgh-Middletown, NY Metro Area	1.92	-0.03
Prescott Valley-Prescott, AZ Metro Area	0.0	-0.2
Providence-Warwick, RI-MA Metro Area	2.82	-0.05
Provo-Orem, UT Metro Area	0.0	-0.06
Pueblo, CO Metro Area	2.26	-0.37
Punta Gorda, FL Metro Area	0.0	-0.14
Racine, WI Metro Area	5.63	0.08
Raleigh-Cary, NC Metro Area	1.9	-0.02
Rapid City, SD Metro Area	6.78	-0.01
Reading, PA Metro Area	0.0	-0.17
Redding, CA Metro Area	8.33	-0.01
Reno, NV Metro Area	2.27	-0.21
Richmond, VA Metro Area	2.68	-0.02
Riverside-San Bernardino-Ontario, CA Metro Area	6.62	0.06
Roanoke, VA Metro Area	0.32	-0.15
Rochester, MN Metro Area	1.79	-0.01
Rochester, NY Metro Area	3.5	0.01
Rockford, IL Metro Area	8.06	-0.01
Rocky Mount, NC Metro Area	11.29	0.32
Rome, GA Metro Area	0.0	-0.24

Sacramento-Roseville-Folsom, CA Metro Area	6.15	0.08
Saginaw, MI Metro Area	4.88	-0.18
Salem, OR Metro Area	5.94	0.13
Salinas, CA Metro Area	5.71	0.08
Salisbury, MD-DE Metro Area	1.72	-0.14
Salt Lake City, UT Metro Area	2.44	-0.13
San Angelo, TX Metro Area	0.0	-0.23
San Antonio-New Braunfels, TX Metro Area	3.83	-0.16
San Diego-Chula Vista-Carlsbad, CA Metro Area	4.26	-0.01
San Francisco-Oakland-Berkeley, CA Metro Area	7.31	0.07
San Jose-Sunnyvale-Santa Clara, CA Metro Area	4.73	0.03
San Luis Obispo-Paso Robles, CA Metro Area	5.51	0.14
Santa Cruz-Watsonville, CA Metro Area	3.07	-0.12
Santa Fe, NM Metro Area	1.16	-0.22
Santa Maria-Santa Barbara, CA Metro Area	1.85	-0.14
Santa Rosa-Petaluma, CA Metro Area	3.03	-0.12

Savannah, GA Metro Area	2.4	-0.14
Scranton–Wilkes-Barre, PA Metro Area	1.59	-0.2
Seattle-Tacoma-Bellevue, WA Metro Area	18.17	0.74
Sebastian-Vero Beach, FL Metro Area	1.89	-0.05
Sebring-Avon Park, FL Metro Area	0.0	-0.19
Sheboygan, WI Metro Area	6.67	0.24
Sherman-Denison, TX Metro Area	0.0	-0.18
Shreveport-Bossier City, LA Metro Area	4.39	-0.16
Sierra Vista-Douglas, AZ Metro Area	5.63	0.15
Sioux City, IA-NE-SD Metro Area	6.1	0.11
Sioux Falls, SD Metro Area	1.23	-0.21
South Bend-Mishawaka, IN-MI Metro Area	2.48	-0.18
Spartanburg, SC Metro Area	1.12	-0.3
Spokane-Spokane Valley, WA Metro Area	4.67	-0.02
Springfield, IL Metro Area	2.73	-0.27
Springfield, MA Metro Area	6.03	-0.01
Springfield, MO Metro Area	3.79	-0.24
Springfield, OH Metro Area	0.0	-0.18
St. Cloud, MN Metro Area	8.0	0.31
St. George, UT Metro Area	2.02	-0.01

St. Joseph, MO-KS Metro Area	6.0	0.04
St. Louis, MO-IL Metro Area	3.91	-0.11
State College, PA Metro Area	4.94	0.18
Staunton, VA Metro Area	1.39	-0.05
Stockton, CA Metro Area	6.14	-0.23
Sumter, SC Metro Area	9.26	-0.0
Syracuse, NY Metro Area	4.15	0.03
Tallahassee, FL Metro Area	1.49	-0.3
Tampa-St. Petersburg-Clearwater, FL Metro Area	4.15	0.02
Terre Haute, IN Metro Area	1.01	-0.14
Texarkana, TX-AR Metro Area	6.67	0.04
The Villages, FL Metro Area	2.04	-0.02
Toledo, OH Metro Area	2.98	-0.17
Topeka, KS Metro Area	3.41	-0.17
Trenton-Princeton, NJ Metro Area	5.36	0.05
Tucson, AZ Metro Area	1.94	-0.21
Tulsa, OK Metro Area	4.04	-0.16
Tuscaloosa, AL Metro Area	5.52	0.02
Twin Falls, ID Metro Area	3.45	-0.02
Tyler, TX Metro Area	1.99	-0.13
Urban Honolulu, HI Metro Area	5.91	0.13
Utica-Rome, NY Metro Area	1.06	-0.14

Valdosta, GA Metro Area	6.76	0.15
Vallejo, CA Metro Area	8.82	0.15
Victoria, TX Metro Area	2.0	-0.22
Vineland-Bridgeton, NJ Metro Area	4.35	-0.06
Virginia Beach-Norfolk-Newport News, VA-NC Metro Area	5.67	0.06
Visalia, CA Metro Area	4.27	-0.03
Waco, TX Metro Area	7.92	0.13
Walla Walla, WA Metro Area	2.47	-0.03
Warner Robins, GA Metro Area	5.08	-0.02
Washington-Arlington-Alexandria, DC-VA-MD-WV Metro Area	4.92	0.08
Waterloo-Cedar Falls, IA Metro Area	1.43	-0.15
Watertown-Fort Drum, NY Metro Area	4.55	0.09
Wausau-Weston, WI Metro Area	4.11	0.07
Weirton-Steubenville, WV-OH Metro Area	5.45	0.19
Wenatchee, WA Metro Area	5.22	0.21
Wheeling, WV-OH Metro Area	1.61	-0.13
Wichita Falls, TX Metro Area	3.85	-0.03
Wichita, KS Metro Area	2.85	-0.4
Williamsport, PA Metro Area	2.99	0.04
Wilmington, NC Metro Area	3.63	-0.07

Winchester, VA-WV Metro Area	0.0	-0.12
Winston-Salem, NC Metro Area	2.32	-0.64
Worcester, MA-CT Metro Area	2.58	-0.08
Yakima, WA Metro Area	5.38	0.1
York-Hanover, PA Metro Area	1.41	-0.1
Youngstown-Warren-Boardman, OH-PA Metro Area	1.8	-0.37
Yuba City, CA Metro Area	5.17	0.02
Yuma, AZ Metro Area	0.86	-0.15

Table 4.1: Metropolitan Areas

Appendix B

Table 4.2: Crime-related keywords

abuse	burglary	larcenies	shoots
abused	burgled	larceny	shot
abuser	carjack	molest	shots
abuses	carjacked	molested	shove
abusing	carjacker	molester	shoved
armed	carjacking	molesting	shoves
arming	carjacks	molests	shoving
arrest	crime	murder	spat
arrested	crimes	murdered	spit
arresting	criminal	murderer	spits
arrests	criminals	murderers	spitted
arson	encampment	murdering	spitting
arsonist	encampments	murders	stab
arsonists	fight	punch	stabbed
arsons	fighting	punched	stabbing
assailant	fights	punches	stabbings
assailants	fought	punching	stabs
assault	harass	rape	standoff
assaulted	harassed	raped	standoffs
assaulter	harasser	rapes	steal
assaulting	harasses	raping	stealing
assaults	harassing	rapist	steals
attack	homicidal	robbed	stole
attacked	homicide	robber	stolen
attacker	homicides	robbery	theft
attacking	kill	robbing	thefts
attacks	killed	robs	thief
burglar	killer	shoot	thieves
burglaries	killing	shooter	thieving
burglarizing	killings	shooting	violence
burglars	kills	shootings	violent

References

[1] Aseel Addawood et al. “Linguistic Cues to Deception: Identifying Political Trolls on Social Media”. In: *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*. Vol. ICWSM. 2019, 11 pages.

[2] Ahmer Arif, Leo G. Stewart, and Kate Starbird. “Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse”. In: *Proceedings of the ACM on Human-Computer Interaction*. New York, NY: ACM, 2018, 27 pages. URL: https://doi.org/10.1145/3274289.

[3] Jeremy Blackburn et al. “Cheating in online games: A social media perspective”. In: *ACM Transactions on Internet Technology* 13 (3 2014), pp. 1–25. URL: http://doi.acm.org/10.1145/0000000.0000000.

[4] U.S. Census Bureau. *2010 Census Urban and Rural Classification and Urban Area Criteria*. URL: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html.

[5] U.S. Census Bureau. *About Metropolitan and Micropolitan Areas*. URL: https://www.census.gov/programs-surveys/metro-micro/about.html.

[6] U.S. Census Bureau. *Age and Sex*. 2020. URL: https://data.census.gov/cedsci/table?q=age&tid=ACSST5Y2020.S0101.

[7] Eshwar Chandrasekharan et al. “You Can’t Stay Here: The Efficacy of Reddit’s 2015 Ban Examined Through Hate Speech”. In: *Proceedings of the ACM on Human-Computer Interaction*. Vol. 1. 2017, 22 pages.

[8] Rafael Prieto Curiel et al. “Crime and its fear in social media”. In: *Palgrave Communications* 6.57 (2020). URL: https://doi.org/10.1057/s41599-020-0430-7.

[9] Srayan Datta and Eytan Adar. “Extracting Inter-Community Conflicts in Reddit”. In: *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM)*. 2019, 12 pages.

[10] Cai Davies et al. “Multi-scale user migration on Reddit”. In: *Workshop on Cyber Social Threats at the 15th International AAAI Conference on Web and Social Media (ICWSM)*. Vol. ICWSM. 2021, 9 pages. URL: https://doi.org/10.36190/2021.13.

[11] Sam Fletcher, Md Zahidul Islam, et al. “Comparing sets of patterns with the Jaccard index”. In: *Australasian Journal of Information Systems* 22 (2018).

[12] Linda Heath and Kevin Gilbert. “Mass Media and Fear of Crime”. In: *American Behavioral Scientist* 39.4 (1996), pp. 379–386. DOI: 10.1177/0002764296039004003.

[13] Federal Bureau of Investigation. *Uniform Crime Reporting.* URL: `https://www.fbi.gov/services/cjis/ucr`.

[14] Federal Bureau of Investigation. *Uniform Crime Reporting Statistics: Their Proper Use.* URL: `https://ucr.fbi.gov/ucr-statistics-their-proper-use`.

[15] Yvonne Jewkes and Travis Linneman. *Media and Crime in the U.S.* Thousand Oaks, CA: SAGE Publications, 2018.

[16] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. "Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit". In: *Proceedings of the ACM on Human-Computer Interaction.* Vol. CSCW. 2019, 27 pages.

[17] Haewoon Kwak et al. "What is Twitter, a Social Network or News Media?" In: *WWW '10: Proceedings of the 19th international conference on World wide web.* New York, NY: ACM, 2010, 10 pages. URL: `https://doi.org/10.1145/1772690.1772751`.

[18] Jiawei Liu et al. "Campaign Advertising and the Cultivation of Crime Worry: Testing Relationships With Two Large Datasets From the 2016 U.S. Election Cycle". In: *The International Journal of Press/Politics* (2021). DOI: `10.1177/19401612211020929`.

[19] Tracey L. Meares. "The Law and Social Science of Stop and Frisk". In: *Annual Review of Law and Social Science* 10 (2014), pp. 335–352. DOI: `10.1146/annurev-lawsocsci-102612-134043`.

[20] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey". In: *Ain Shams Engineering Journal* 5 (4 2014), pp. 1093–1113. URL: `https://doi.org/10.1016/j.asej.2014.04.011`.

[21] Areeb Mian and Shujhat Khan. "Coronavirus: the spread of misinformation". In: *BMC Medicine* 18.89 (2020). URL: `https://doi.org/10.1186/s12916-020-01556-3`.

[22] Peter Mieszkowski and Edwin S. Mills. "The Causes of Metropolitan Suburbanization". In: *Journal of Economic Perspectives* 7.3 (1993), pp. 135–147. DOI: `10.1257/jep.7.3.135`.

[23] Rasmus Nielsen and Richard Sambrook. "What is Happening to Television News". In: *Digital news project, Reuters Institute* (2016). URL: `https://ssrn.com/abstract=2771080`.

[24] Katherine Ognyanova et al. "Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power". In: *The Harvard Kennedy School Misinformation Review* 1.4 (2020). URL: `https://doi.org/10.37016/mr-2020-024`.

[25] Abby Ohlheiser. "Fearing yet another witch hunt, Reddit bans 'Pizzagate'". In: *Washington Post* (2016). URL: `https://wapo.st/2Xvbvae`.

[26] Shruti Phadke and Tanushree Mitra. "Many Faced Hate: A Cross Platform Study of Content Framing and Information Sharing by Online Hate Groups". In: *CHI '20*. 2020, 13 pages.

[27] Liza Potts and Angela Harrison. "Interfaces as Rhetorical Constructions: reddit and 4chan During the Boston Marathon Bombings". In: *SIGDOC '13*. 2013, 9 pages.

[28] Associated Press. *United States Election Results*. URL: `https://elections.ap.org/dailykos/results/2020-11-03/state/US`.

[29] Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. "Political Discussion is Abundant in Non-political Subreddits (and Less Toxic)". In: *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*. Vol. 15. 2021, 12 pages.

[30] *reddit.com: api documentation*. Reddit. San Francisco, California, 2022.

[31] Manoel Horta Ribeiro et al. "Do Platform Migrations Compromise Content Moderation? Evidence from rThe_Donald and rIncels". In: *CSCW*. 2021, 24 pages.

[32] Manoel Horta Ribeiro et al. "The Evolution of the Manosphere Across the Web". In: *Fifteenth International AAAI Conference on Web and Social Media*. Vol. 15. 2021, 12 pages.

[33] Vincent Sacco. "Media Constructions of Crime". In: *The Annals of the American Academy of Political and Social Science* 539.1 (1995), pp. 141–154. DOI: `10.1177/0002716295539001011`.

[34] Haji Mohammad Saleem and Derek Ruths. "The Aftermath of Disbanding an Online Hateful Community". In: *CoRR* abs/1804.07354 (2018). URL: `http://arxiv.org/abs/1804.07354`.

[35] William Sattelberg. *The Demographics of Reddit: Who Uses the Site?* 2021. URL: `https://social.techjunkie.com/demographics-reddit`.

[36] Salvatore Scellato et al. "Distance Matters: Geo-social Metrics for Online Social Networks". In: *3rd Workshop on Online Social Networks (WOSN 2010)*. Boston, MA: USENIX Association, June 2010. URL: `https://www.usenix.org/conference/wosn-2010/distance-matters-geo-social-metrics-online-social-networks`.

[37] James Shanahan and Michael Morgan. *Television and its Viewers*. Cambridge University Press, 1999.

[38] Anne Speckhard. "The Boston Marathon Bombers: the Lethal Cocktail that Turned Troubled Youth to Terrorism". In: *Prespectives on Terrorism* 7.3 (2013), pp. 64–78. URL: `https://www.jstor.org/stable/26296940`.

[39] Fatemeh Tahmasbi et al. ""Go eat a bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19". In: *Proceedings of the Web Conference 2021 (WWW '21), April 19-23, 2021, Ljubljana, Slovenia*. New York, NY, USA: ACM, 2021, 12 pages. DOI: `https://doi.org/10.1145/3442381.3450024`.

[40] r/nyc moderation team. *r/nyc sidebar rules*. URL: `https://old.reddit.com/r/nyc/wiki/sidebarrules`.

[41] Mason Walker and Katerina Eva Masta. *News Consumption Across Social Media in 2021*. URL: `https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/`.

[42] David Weisburd, Elizabeth R. Groff, and Sue-Ming Yang. *The Criminology of Place: Street Segments and Our Understanding of the Crime Problem*. Oxford, UK: Oxford University Press, 2012.

[43] James Q. Wilson and George L. Kelling. "Broken Windows". In: *The Atlantic Monthly* 249.3 (1982), pp. 29–38.

[44] Savvas Zannettou. ""I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter". In: *ICWSM*. 2021, 13 pages.