Chattanooga Police Incident Data

Recommendations for Improved Privacy Protections

Sonya Chen, Aditya Mengani, Ashley Moss, Ginny Perkey

I. Introduction

In this report, we analyze the privacy of Chattanooga's anonymized Police Incident Data, which is one of many datasets that you have made available to the public through the Chattanooga Open Data Portal. You state that the goal of your data portal is to "give citizens access to community data for solving problems, informing themselves and others, and better interacting with the community around them." We admire and appreciate this goal, and your website makes it easy for members of the public to access the provided information. Users can search the data catalog to find datasets, studies, and visualizations of interest, and they can filter and visualize the data using easy-to-use tools on the website. Data portals like yours increase transparency within a

community, and they make research and innovation accessible to a broader range of people, not just government employees or academics who have institutional subscriptions to databases.

The police incident data in particular can help community members understand how Chattanooga is being policed. Transparency about policing is a distinguishing feature of democratic societies like ours. Without transparency, law enforcement operations may serve illegitimate purposes such as monitoring political rivals. Earlier in 2021, Russian police conducted arbitrary arrests at political gatherings to dampen support for Vladimir Putin's political rival, Alexei Navalny (HRW, 2021). Transparent data on police activity provides a check against governmental overreach. There are also everyday uses for this type of information. Crime incident data can help inform residents' personal purchasing, transportation, and safety decisions. Residents may refer to neighborhood crime statistics as part of the homebuying process. A woman planning to travel alone at night could optimize her route using an understanding of city crime patterns. When reserving parking space for a trip downtown, a citizen might refer to theft data to assess the relative safety of different parking lots.

Part of what makes the dataset so useful to researchers and to everyday citizens is that it contains a wealth of useful details about the policing incidents it tracks. Each row in the dataset represents a specific interaction between members of the public and the police, and includes the following information:

- Date and time
- Address, city, state, zip code, latitude, longitude, and neighborhood
- Jurisdiction, police district, zone, council district
- Incident code, incident type
- Case number, case status, case status description

We were surprised by how comprehensive the information was, especially regarding the location data. Although we value how the level of detail increases transparency, we are concerned that

people involved in these incidents may be identifiable, despite the absence of names. Your office and the Chattanooga Police Department (CPD) have taken steps to protect sensitive personal data. To protect vulnerable classes of people like juveniles, CPD leaves out certain types of incidents in accordance with Tennessee state law. The website states that the records are generalized to block level and randomly offset to protect victims. You also caution users that "all crime data posted is preliminary and may or may not have been reviewed and approved by the Chattanooga Police Department's (CPD) quality control process." The text further advises that "users should not make decisions as it relates to safety solely based on the data provided on this website, but should seek independent verification directly through CPD's Crime Analyst Unit." Despite these privacy protections, making this comprehensive dataset so easily accessible could cause harm to the individuals and communities who were involved in these police incidents. In the following section, we list some of the concerns that we have about this dataset, and we demonstrate some of the issues that could arise as a result of the information that is being shared.

II. Potential Issues

The public at large may be unaware of the CPD's practice of publishing incident data, which could cause a public relations problem for the police department, especially given current controversies around policing in America. In the context of police contact, citizens likely expect their data to support internal CPD purposes such as investigatory research (Nissenbaum, 2011). Law enforcement's stated purpose is to "protect and serve" the public, so citizens may feel exposed and surprised to see incidents posted as soon as the same day in such an easily searchable format. In the pre-digital world, we would never expect to see a police car at our neighbor's house and find out what happened there the next day from the comfort of our desk. Sharing searchable, unverified data online so quickly, without an opportunity for those involved in the incidents to consent to the data sharing, may damage public trust in the department.

Additionally, because the data is preliminary and unverified, the citizen's right to due process is called into question. The CPD discloses the specific nature of an incident before that information passes through a thorough verification process. The Fourteenth Amendment specifies that citizens should not be deprived of "life, liberty, or property, without due process of law" (Cornell, 2021). Under this doctrine, citizens have the right to a fair review before being harmed in any way by the government (Mulligan, 2016). Based on the possibility of re-identification as outlined below in case studies, releasing this data may impede the constitutional right to due process.

The internet disseminates information more widely than ever before, which threatens individual privacy (Solove, 2002). The location and exact time of an incident may be coupled with prior knowledge or other sources to re-identify individuals. It is very easy to look up a street address online and find out who lives there. If that address is associated with a crime in this dataset, a user may assume that the person who resides at that address committed the crime. Furthermore, because the data is perturbed and truncated to block level, users may attempt to identify someone from the dataset from an address that has been truncated, without realizing that the associated address has been made inaccurate in the name of privacy.

As we demonstrate below, cross-referencing other websites lets us construct a portrait of a citizen, whether accurate or not. Anyone with sufficient financial or personal incentive can use the data to make decisions that have real social impacts, such as employment decisions or insurance rate adjustment. This lends the data to secondary purposes beyond what the city intended (Solove, 2002). Our concerns about the location information included in this dataset, as well as the potential dangers in making this data easily accessible online, led us to investigate how different use cases of this dataset might result in harm to those represented in the dataset.

Case Study I - Employment

We began our investigation by considering how an individual might be impacted by inclusion in the dataset. Employers go through various methods of vetting potential candidates for jobs before hiring them. Formal background checks may be a part of this process, but they can be costly, and they are regulated by both the US Equal Employment Opportunity Commission (EEOC) and the Fair Credit Reporting Act (FCRA). Employers who are unaware of these constraints or unable to fund a proper background check may resort to a do-it-yourself background check solution.

Let's consider a case in which a prospective employer draws inaccurate conclusions based on the incident data and decides not to hire the candidate. Robert Builder, hiring manager of a local construction business, has interviewed a Chattanooogan named James Jackson for a job at his company. He wants to know more about his background before making a job offer, so he uses the phone number on James's resume to search for information on USPhoneBook.com. From there, Robert recovers James's age and previous addresses. Next, Robert searches police incident data on James's current address, 1800 Wilcox Blvd. Here, Robert makes an error, not realizing that CPD has replaced the last two digits of the address with '00'. To his untrained eyes it looks like a real address, 1800 Wilcox Boulevard, and the findings are alarming: the police recorded a Forcible Rape incident there in July 2021.

Robert begins to suspect that James has a criminal background, so he searches municipal court records for prior offenses using James' age and previous addresses from the US Phone Book. Yes, Robert uncovers charges for larceny and hit-and-run from 1993. Rape, larceny, and hit and run? Robert denies James employment, and James is punished for the action of a neighbor that falls outside his control.

While this scenario is hypothetical, it is plausible, and situations like these should be factored into the decision to release this dataset with the level of information it currently includes. Although privacy protections are in place within the dataset, the communication and display format makes the precision level of each individual record unclear to an end user. While lessening the probability of actual re-identification, the government creates new potential for misinterpretation and inaccurate matching.

Case Study II - Non-violent incident

We also attempted to identify people associated with non-violent events, such as field interviews. In the database, we looked at the police incident information for a field interview that took place at 3300 Curtis St, Chattanooga, TN on July 13, 2021. On the Chattanooga's assessor website, we found the name of the house owner, as well as detailed information about the owner's other properties, including their addresses, the year they were purchased, and the sale price. We searched the address on Zillow and found detailed historical property tax information. We then went to the TruePeopleSearch website, entered the house owner's name, and several people with the same name popped up. From the public information of the purchase year of the property on the government assessor's website, we were able to identify which of the people owned the house. We also found information about the owner's phone numbers, age, month and year of birth.

Imagine if your address was included in this dataset as a result of mistake or the way that blocks were aggregated. At first, it may not seem like a big problem, especially considering that the dataset has a disclaimer that the incident record website is preliminary and may not be accurate. However, there are reasons for concern. In your community, a family member of a crime victim might use this dataset to find out who the police have interviewed regarding that crime. They

could be angry, and use your address to come find and confront you, incorrectly thinking you were involved in the crime because your address was mistakenly in the dataset.

Another potential harm is that corporations might use that "dirty," preliminary and inaccurate data, which has your name or address, to train models that are used to predict crime. According to the article "Police across the US are training crime-predicting AIs on falsified data" from MIT Technology Review (Hao, 2019), tech companies such as Palantir partnered with the city to train "deploy predictive policing systems. The system used historical data, including arrest records and electronic police reports, to forecast crime and help shape public safety strategies, according to company and city government materials". If your address was inaccurately associated with a crime, this system might predict that you would be involved in crimes in the future.

A dataset like this could also be used to create software for background checks. If an employer used this software to check your address and name, the model may suggest that you have a higher percentage of risk because the model is trained with a perturbed dataset that has your name or address. Even if you were given the opportunity to explain how this might have happened, a "risk-averse" employer might still prefer other "safer" candidates. Technology is quickly advancing, and companies are developing new software everyday that relies on machine learning models that must be trained on data of varying quality. Although these situations are hypothetical, there is potential for harm as technology advances faster than privacy regulations.

Case Study III - Neighborhoods and Communities

After establishing how easy it would be to identify individuals using this anonymized dataset, we were interested in how this dataset might impact groups of people. Research indicates that there are racial disparities in policing in the United States, particularly when it comes to traffic stops and search decisions (Pierson). This analysis on Policing and Racial Equity, which was created by

Chattanooga's Office of Performance Management and Open Data, indicates that these racial disparities in policing also are an issue in Chattanooga. According to the report, "There is a significant disproportionate amount of economic violation only type traffic stops for the BIPOC community." These disparities in policing are likely to result in non-white Chattanoogans being overrepresented in this police incident dataset.

Chattanooga has several historically Black neighborhoods, many of which are upwards of 90% Black. If Chattanooga is disproportionately policing its Black communities, these neighborhoods would likely be overrepresented in the dataset. We took a closer look at the historically Black, downtown Chattanooga neighborhood called Avondale. According to the 2017 American Community Survey, Avondale has about 2500 residents, 94% of whom are Black. The city of Chattanooga has about 180,000 residents, only 31% of whom are Black. Avondale residents are only about 1.4% of the population of Chattanooga, yet we found that Avondale makes up 9.3% of the police incidents in the dataset that are associated with a neighborhood. This suggests that members of historically Black neighborhoods like Avondale are more likely to be represented in this dataset than members of majority white neighborhoods. These community members will be disproportionately impacted by any potential harms that are caused by the public availability of this dataset.

In addition to the higher risk of harm to individual Avondale residents, this dataset can also impact investment in the community. Search Avondale online and you will find that it is frequently listed as one of Chattanooga's worst neighborhoods, usually because of the crime rate. The public availability of datasets like these can make it difficult for a neighborhood like Avondale to shake this type of perception. A reputation as a dangerous neighborhood can have very negative impacts for communities. Businesses often use datasets like this to determine where they want to open up new stores or otherwise invest in communities, and, according to this study, "private investment,

as represented by building permits, decreases on blocks where crime increases." Neighborhoods like Avondale have long been plagued with underinvestment, going back to Chattanooga's historic redlining in the 1930s, which discouraged banks from making loans to people interested in buying property in majority Black neighborhoods. Avondale was rated a C in this system, and open datasets like this, which exclude many white collar crimes, such as tax evasion, that are more prevalent in higher income, majority white neighborhoods, risk exacerbating the harms of historic discrimination like redlining.

Despite the fact that the police incident data is preliminary and unverified, the Chattadata website publicizes this police incident dataset as a way to learn about where crime happens in the city. It states, "Want to know where crime is occurring in the city? Explore, analyze, and learn from the Police Incident Dataset." A casual user of this site could easily use the convenient visualization tools to make a crime map that would make Avondale and other historically Black neighborhoods look like dangerous parts of the city, without considering the impacts of over-policing or the fact that this dataset includes only incident reports, not confirmed crimes.

III. Recommendations

As part of our analysis, we looked at how Chattanooga's public police incident dataset compared with those of Madison, WI and Palo Alto, CA (see Appendix for additional details). Madison's dataset provides a variety of information including, but not limited to, suspect name, victim name, address, and detailed physical descriptions of some suspects. The information provided in the dataset is determined by the police officer for each incident, and not all incidents include the same level of detail. Compared to Madison, Palo Alto displays very limited information related to the crime incident. Their dataset includes only latitude and longitude, the block-level address, and a description of the crime.

In comparison to these cities, your Chattanooga dataset protects citizen privacy better than Madison's, but it includes more identifiable information than Palo Alto, CA. All three datasets provide geolocation fields, which poses a risk for subject identification. However, privacy and fairness concerns must be balanced with the usefulness and accuracy of the data for valid use cases. Considering these tradeoffs in the following sections, we provide several options for improving the privacy protections in your dataset that strike a balance between transparency and privacy.

Editing or Removing the Address Field

We recommend that you consider removing the address field in the dataset, as it is the most probable attribute to re-identify individuals when combined with other publicly available information. Even though this approach can easily reduce the scope of privacy breaches, it might limit the usefulness of this public dataset for various purposes by general public and research institutions. If you determine that removing the address is not the best option, an alternative is to mask the last two digits of the street number by using a special character like 'X' instead of a digit like '0'. Making the masking process more obvious would reduce the possibility of misidentifying individuals using the data set. This solution can be easily implemented without consuming many financial resources.

Data Aggregation to Census Block

A less convenient but more thorough solution is to aggregate data to a census block level. A census block is the smallest geographic unit used by the US census bureau for tabulation of census data. Each block typically covers around 500-1000 households in a physical location. Census blocks can be grouped into larger sections called census tracts, which contain about 39 census blocks on average and can be bounded by roads, highways, or geographical boundaries. This can be

helpful in identifying datasets at different levels of granularity for better analytics. Census tracts also do not map directly onto neighborhoods. This may be helpful for businesses in communities like Avondale, since users of the dataset will not be able to see if a specific block within an overpoliced neighborhood has experienced more police incidents.

When the data is aggregated at a census block level, it not only safeguards the anonymity of the households in that census block, but it also satisfies the transparency requirements of the publicly available datasets, as most of the publicly available datasets used in research follow some kind of roll-up or aggregation of the data (across census block or group/city/county/state/country etc) as part of predictive analytics and modeling frameworks. The risks or tradeoff with this approach can arise by not being able to pinpoint or drill down the incident to a particular location, for validation or auditing the quality of the original data, which might be required in some cases for research purposes.

As a workaround, we propose that a neutral agency, such as your department or another office within the Chattanooga city government, should control the aggregated datasets and original unaggregated datasets as a way to avert or limit the Police department's ability to manipulate the data. When a research team needs access to raw data, they should be able to access them by reaching out to this department, who can evaluate the use case and determine if it is appropriate to share the data with the researchers.

Differential Privacy

A differential privacy approach to the police incident dataset would preserve privacy by restricting the amount of inference that can be obtained regarding a single individual by making the smallest possible single arbitrary substitution in the dataset. This is similar to cryptography of the data, and would restrict the amount of publicly decipherable data available there by enhancing

the privacy of the dataset. This is a robust alternative to the aggregation approach for census blocks mentioned earlier, and it strikes a similar balance between transparency and privacy.

Although this approach has many benefits, it has risks as well. Differential privacy will be harder to implement than our other recommendations. According to this study published in the Harvard Data Science Review, differential privacy can pose risks that commonly occur with data pertrubation efforts by randomizing the data, censoring and adding noise to the original data. This approach can create biases while establishing relationships, as a random measurement error and selection bias does. It can also add a layer of non-sampling error that can propagate into the final analysis. Moreover, a lot of requirements surrounding usage and privacy of the data need to be known before actually implementing differential privacy. This could be a challenging task. Finally it can end up using more complex statistical methods to account for non sampling errors in the data, limiting the scope and complexity of the research questions.

The study concludes that , due to these risks, when applying differential privacy, social science researchers should not just be held accountable for privacy but also for enforcing statistical rigor (Oberski and Kreuter, 2020). Therefore, if you decide to implement differential privacy for the police incident dataset, you should provide detailed guidelines and enforce statistical rigor in your datasets. You will also need to create a mitigation plan to resolve and address any statistical errors. Your administrators and researchers should adhere to the principle of beneficence while applying differential privacy, so as to adhere to the welfare of the participants in the datasets and corresponding studies. Finally, because differential privacy systems tend to demand more technical background from users, it will be important for you to consider how to develop documentation and training to make the datasets accessible to less tech-savvy users.

Future Regulations for Public Datasets

In addition to the above proposals, we also highly recommend that you and the Chattanooga Police Department work closely with other cities and states across the nation to propose and implement a generalized framework for sharing public police incident datasets. This framework should standardize the data sharing practices across all cities and aim to improve the quality of the data provided to the public without compromising privacy concerns. You might begin by collaborating with police departments from other cities and come up with a pilot dataset.

With this dataset, you can implement practices that attempt to satisfy the tradeoffs between privacy vs accuracy and privacy vs transparency concerns of these datasets. If the datasets follow data perturbation techniques, they should adhere to strict guidelines and enforce privacy and statistical rigor in the datasets as discussed earlier in the differential privacy section to protect misinterpretation of information. These steps, when implemented, can help Chattanooga Police Department to play a pivotal role in creating these futuristic public police incident datasets which can immensely improve their credibility and transparency efforts in the general public and research community.

More General Recommendations

In addition to the above recommendations, we propose that the Chattanooga Police Department address the privacy concerns and accuracy of the data raised by the general public in a more transparent manner by providing public comment sessions, where the general public get an opportunity to share their grievances about their records available as part of the datasets. This would be helpful for community members and groups who may be experiencing harms due to this dataset.

IV. Conclusion

Although we identified privacy concerns within your dataset, we want to commend you on your mission and how well your website is meeting the goals of transparency and data access within Chattanooga. The police incident database provides valuable information for the citizens of Chattanooga, and researchers can easily use the different datasets to learn more about a wide range of topics related to policing in Chattanooga. Because of the issues that we identified in our analysis, we recommend that you prioritize implementing one of the approaches to anonymizing address and location information that we discuss above. Although these changes decrease the usability of the dataset, they will go a long way to protecting the members of your community whose life experiences are reflected here.

References

Chattanooga Municipal Record Search. Chattanooga City Court. (July 2021). https://www.municipalrecordsearch.com/chattanoogatn/Cases

City of Chattanooga Office of Performance Management and Open Data. 2017 5 Year Race Ethnicity. Retrieved from https://www.chattadata.org/dataset/2017-5-Year-Race-Ethnicity/3jq9-a5da

City of Chattanooga Office of Performance Management and Open Data. Chattanooga Historic Redlining. Retrieved from https://www.chattadata.org/Economy/Chattanooga-Historic-Red-Lining/biep-kjnx

City of Chattanooga Office of Performance Management and Open Data. Policing and Racial Equity. Retrieved from https://www.chattadata.org/stories/s/26bg-ejs3

City of Madison Map Data. (2017). Police incident reports. Retrieved from https://data-cityofmadison.opendata.arcgis.com/datasets/cityofmadison::police-incident-reports/about

Cornell Legal Information Institute. U.S. Constitution, 14th Amendment. (July 2021) https://www.law.cornell.edu/constitution/amendmentxiv

Fair Information Practice Principles (FIPPs). International Association of Privacy Professionals. (July 2021). https://iapp.org/resources/article/fair-information-practices/

Güss, C. D., Tuason, M. T., & Devine, A. (2020). Problems With Police Reports as Data Sources: A Researchers' Perspective. Frontiers in psychology, 11, 582428. https://doi.org/10.3389/fpsyg.2020.582428

Hirsch, Tad, et al. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. *Proceedings of DIS (Designing Interactive Systems Conference)*, 2017: 95–99.

Human Rights Watch (HRW). Russia: Arbitrary Detentions at Pro-Navalny Protests. *Human rights Watch*. (April 2021). https://www.hrw.org/news/2021/04/22/russia-arbitrary-detentions-pro-navalny-protests

Karen Hao (February 13, 2019). Police across the US are training crime-predicting AIs on falsified data. *MIT Technology Review*. (July 2021). https://www.technologyreview.com/2019/02/13/137444/predictive-policing-algorithms-ai-crime-dirty-data/

Lacoe, J., Bostic, R. W., & Acolin, A. (2018). Crime and private investment in urban neighborhoods. Journal of Urban Economics, 108, 154-169. doi:https://doi.org/10.1016/j.jue.2018.11.001

Mulligan, Deirdre K., Koopman, Colin and Doty, Nick (2016). Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Philosophical Transactions of The Royal Society A: Mathematical Physical and Engineering Sciences*, 374(2083):20160118 (December 2016). http://doi.org/10.1098/rsta.2016.0118

Nissenbaum, Helen F. (2011). A Contextual Approach to Privacy Online. *Daedalus* 140:4 (Fall 2011), 32-48. https://ssrn.com/abstract=2567042

Oberski, D. L., & Kreuter, F. (2020). Differential Privacy and Social Science: An Urgent Puzzle. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.63a22079

Pierson, E., Simoiu, C., Overgoor, J. et al. A large-scale analysis of racial disparities in police stops across the United States. Nat Hum Behav 4, 736–745 (2020). https://doi.org/10.1038/s41562-020-0858-1

ProximityOne.Census block groups and block group codes. Retrieved from http://proximityone.com/geo_blockgroups.htm

Solove, Daniel J. (2002). Conceptualizing Privacy. *California Law Review* 90.4 (July 2002). https://doi.org/10.15779/Z382H8Q

Appendix: Comparison with other cities

- Madison, WI(Worse)
  - Suspect name
  - Victim name
  - Address
  - Selected by officers
  - Not inclusive of all incidents
  - Body characteristics
  - Case number

- Chattanooga, TN(Better)
  - Latitude/Longitude
  - Address
  - Case number
  - Devoid of protected classes
  - Generalized to block level(some*)
  - Preliminary
  - Jurisdiction

- Palo Alto, CA(Best)
  - Latitude/Longitude
  - Address(compressed to block level)
  - Description