Berkeley Police Department
2100 Martin Luther King Jr Way,
Berkeley, CA 94704
https://www.cityofberkeley.info/police/

To whom it may concern,

As requested, our Data Science team has analyzed the publicly available arrest log dataset from the Berkeley Police Department for legal or ethical issues and concerns pertaining to privacy. For our analysis, we utilized three separate privacy frameworks (Solove's Taxonomy, Nissenbaum's Contextual Integrity, and Mulligan/Koopmans' Privacy Analytic) to further frame our perspective. Additionally, we performed a reidentification exercise to explore if individuals from the current log chosen at random could be further identified in other public domains as this may further increase privacy risks and issues.

After examining the data in the Berkeley PD arrest dataset through a legal and ethical lens, our team feels strongly that the data that is currently being shared has the potential to cause a subject harm and for a subject's information to be used beyond what it was collected for (secondary use). It's for these reasons that we have recommended a long term solution of revisiting the California Public Act and updating the need to share PII arrest information publicly, and an immediate solution for the Berkeley PD that involves withholding the PII arrest data from publication or to anonymize the data. It is our goal as a Data Science team that the Berkeley Police Department consider some or all of these recommendations to create and/or foster the importance of privacy for all residents of Berkeley, CA.

This full report is attached. Should you have any questions or concerns, we would be happy to discuss further.

Regards,

The W231 Data Science Team:

Jackie Nichols
Autumn Rains
George Rodriguez

Legal/Ethical Analysis of Reidentification in Open Datasets:

Berkeley Police Department Log - Arrests

Jackie Nichols, Autumn Rains, George Rodriguez

Nov 19, 2021

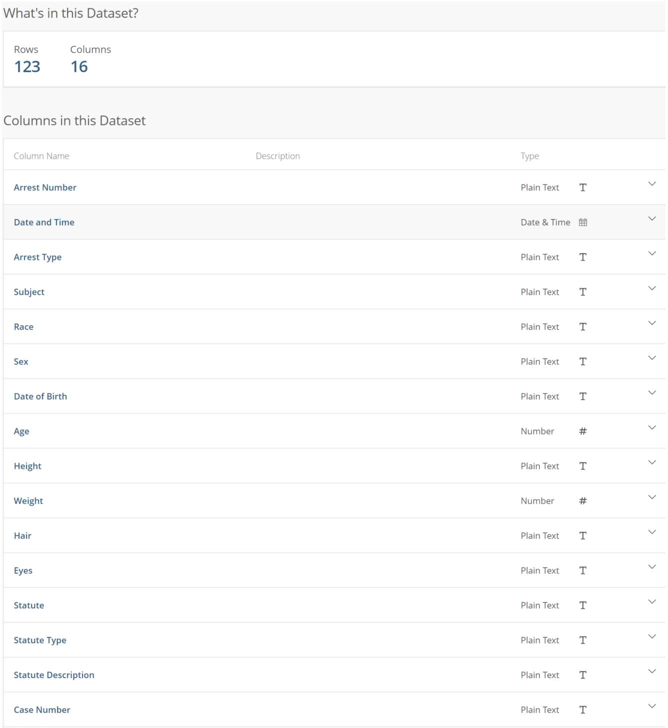
Introduction

Police Departments in the United States create varying levels of reports when investigating crimes. For example, logs are produced upon arresting an individual who may be associated with a crime. In California, the California Public Records Act requires that police reports including arrest logs are available to the public. These arrest logs contain personally identifiable information (PII) like date of birth, gender, occupation, and physical descriptions of those arrested . Given the availability of this information to the public combined with powerful digital search engines that enable reidentification, there are legal and ethical issues and risks regarding privacy for these individuals that should be considered.

Our team will use three privacy frameworks to describe these issues and risks on an open arrest log from the Berkeley Police Department (PD). We will also use the information of six unique individuals found within this arrest log to 'sleuth' across other public domains to determine if identification is possible. Based upon our analysis and sleuthing, we will propose recommendations to the Berkeley PD to improve and protect the privacy for arrested individuals.

Dataset

The dataset "Berkeley PD Log - Arrests" was obtained from the City of Berkeley Open Data portal (Berkeley Police Department, 2021). The owner of the dataset is the Berkeley PD in which they collect and publish the data from their arrest logs, a requirement due to the California Public Records Act. From the one page narrative, the dataset contains adult arrests by the Berkeley PD over the past 30 days - both felonies and misdemeanors - and is updated each business day . Additionally, the arrests in this dataset include persons booked into a jail facility and those cited and released with a future court date. There are currently 16 fields within the dataset and plan to add 4 more in the future. One of the most interesting discoveries of this dataset is that 9 of the 16 fields contain personal identifiable information, everything from name and date of birth to eye and hair color as outlined in Figure 1. Some statistics of the demographics within the dataset include: 88% of the individuals are males, 61% of the individuals are from minority groups, with possession of controlled substances and burglary being the most common statute description for both groups.



What's in this Dataset?

Rows 123

Columns 16

Columns in this Dataset

Column Name	Description	Type	
Arrest Number		Plain Text	T
Date and Time		Date & Time	📅
Arrest Type		Plain Text	T
Subject		Plain Text	T
Race		Plain Text	T
Sex		Plain Text	T
Date of Birth		Plain Text	T
Age		Number	#
Height		Plain Text	T
Weight		Number	#
Hair		Plain Text	T
Eyes		Plain Text	T
Statute		Plain Text	T
Statute Type		Plain Text	T
Statute Description		Plain Text	T
Case Number		Plain Text	T

Figure 1: Columns included in Berkeley PD Arrest Dataset

Assumptions

Given the nature of the dataset, our team felt it important to detail our assumptions about the objectives of publicicing police arrest records as described by the California Public Records Act. Beyond the legal requirements, we assume the Berkeley PD has two primary goals for readily publishing this information. First, we assume it is shared with the public as a means to keep them informed upon criminal and police activity as public servants to the residents of Berkeley, Ca. Second, we also assume this information is shared to the broader public to increase transparency on police activity to aid in research about typical criminal activity or demographics of those arrested.

Risks of Public Records

Once something is out on the internet, it can be nearly impossible to delete it completely or to take it back; because of this, there are multiple risks that could arise with public records being published on the internet. One of the main risks that comes from having information found in this dataset out on the internet is fraud or identity theft. Identity theft can occur by using just a name and date of birth. Another possible risk is secondary use of information. Data from electronic public records files will be used for secondary purposes that stray far from the original public policy purposes for which they were first created, that being government accountability (Privacy Rights Clearinghouse 2002).

Privacy Frameworks

While we see the need for crime reports to be shared with the public, it’s not clear if it’s necessary to provide an arrest dataset with individuals name and date of birth. If the goal of the dataset is to inform citizens about police activities in the Berkeley area, there are means to limit the possibilities of misuse of data. In addition, we must remember that these are arrests and not convictions and while that is an important distinction, the general public may not make that distinction when viewing the dataset. We will leverage three privacy frameworks, Solove's Taxonomy, Nissenbaum’s Contextual Integrity, and Mulligan/Koopam Multidimensional Privacy Analytics, to help assess the potential harm of releasing the PII data found in the arrests dataset.

Solove’s Privacy Taxonomy

Applying Solove’s Taxonomy and the four different perspectives of privacy (Figure 2) to the Berkeley Arrests dataset we note the following (Solove, 2006):

- Information Collection: The information was collected legally as a result of an arrest that was made.
- Information Processing:
    - Aggregation and Identification: The dataset does identify the subjects by their name and date of birth which when aggregated can prove fruitful for gathering additional information about a person. There does not appear to be any attempt to anonymize the data per California law.
    - Secondary Use: It's not clear if there is any secondary use that the Berkeley PD is using this data for beyond arrest statistics but it's clear that the data could certainly be used by others as a secondary use.
    - Insecurity: There does appear to be a level of negligence that could lead to potential future harm of an individual by providing a subject's name and date of birth.
- Information Dissemination:
    - Exposure: There is data in the dataset that when exposed can cause great harm, embarrassment, and humiliation. In the dataset updated November 8 2021, there is a statute description of "contact minor with intent sex" that could prove to be extremely harmful to a wrongly accused individual. While this is an arrest dataset, the harm from being accused of such an act can be long lasting. In addition, race and gender are shared which could lead to fueling public perception about a certain race and/or gender taking part in criminal behavior. For example the dataset we sampled showed 88% of the individuals are males, 61% of the individuals are from minority groups. From this one may believe that males that fall in the minority groups are more likely to commit a crime. This goes beyond causing harm to the individual subject but the possibility of causing harm to a more generalized group of people.

    - Increased accessibility: The Berkeley PD is publishing data that can be easily joined with other first party data sets to reveal additional personally identifiable information about the subjects. The subjects were likely not aware at the time of information collection of the potential exposure and ramifications of publishing this information to the public.
    - Appropriation and Distortion: The data as it is could be used as a secondary use for black mail purposes or could easily be distorted.
- Invasion: If the subject is innocent of the crime, the public arrest information with other data sources constitutes an invasion of the privacy of the individual and may also be used as a secondary use.

Crime data is something the public demands and can be used to inform the public on where to live, work etc. and does certainly have a place in society. Although the arrest information was collected legally, the Berkeley PD arrests data that shares a subject's name and date of birth which can be combined with other data and potentially lead to harm to the individual even if they are innocent of the crime. This information dissemination could lead to a breach of confidentiality and ultimately an invasion of privacy and long-term effects to the subject. The Berkeley PD could accomplish the same results of sharing arrest data by omitting subjects names and date of birth. At a minimum this will make it more challenging for people to combine data to determine who was arrested. A secondary update we suggest is to update the dataset with the outcome of the arrest: were the subjects found innocent or guilty of the charge.

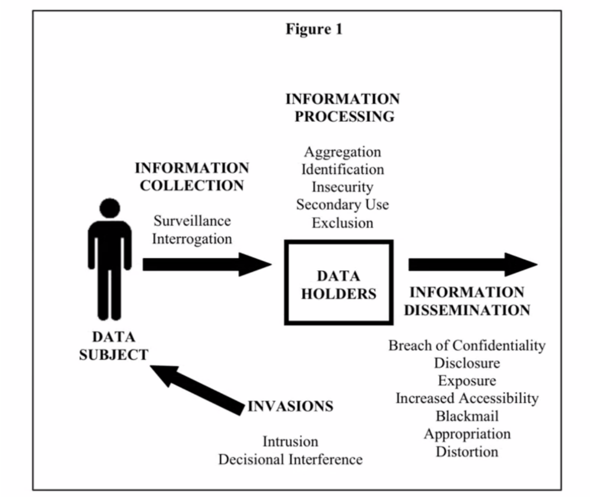


Figure 2: Solove's Taxonomy

Nissenbaum's Contextual Integrity

Helen Nissenbaum states that "contextual integrity holds when context-relative informational norms are respected; it is violated when they are breached" (Nissenbaum, 2011). There are five parameters when considering context-relative informational norms: subject, sender, recipient, information type, and transmission principle. In the context that Berkeley PD is required to release all of the PII information by law, the personal identifiable information of those arrested is being done so appropriately. While the agency is sharing the information appropriately, there could be reasonable concern for the information included in the dataset. The fact that an individual was arrested is not proof that they engaged in criminal conduct. If an individual were to be wrongfully arrested, their information is now on the internet and will most likely forever exist on the internet. With so many identifiable attributes in the dataset like name, date of birth, sex, and age, this could potentially lead to risks such as fraud/identity theft or a negative impact to the individual's reputation.

Mulligan/Koopman's Privacy Analytic

The Mulligan/Koopman Analytic is a tool that allows multiple interested parties to view privacy from varying perspectives using multiple dimensions. The table below uses this analytic from the perspective of the individuals to explore privacy given their information is publicly available. Given the permanence of the arrest log and public availability, the potential for immediate harm now or the future regarding financial or professional negative impact, it is crucial to consider these dimensions when reevaluating privacy risks. (Mulligan, et. al, 2016)

Table 1: Privacy Analytic Framework

Dimension	Description
Theory: Object	The object of privacy seeks to answer "what is privacy for?" For this dataset, privacy would afford the individual arrested dignity as well as the ability to control the record and dissemination of their personal information.
Theory: Contract Concept	Contract concept is a way to determine within a specific scenario what is not private. Based on the structure of the database, it is quite clear that personal information is not private. In fact, the California Public Records Act requires the publication of the data fields found within the Berkeley PD arrest log.
Protection: Subject	Subject, or whose privacy is at stake, is the privacy of the individual arrested. However, because this information can be used to search on digital platforms like LinkedIn or Facebook, the privacy of those connected to the individual arrested may also be at stake. Certainly, immediate family and friends may encounter greater harm to their privacy based upon their connection to the individual.

Harm: Action	Action is a dimension that details how privacy has been violated. Referencing Solove’s Taxonomy, the dissemination of personal information after collection online is the action that violates the privacy of the individuals and their immediate professional and/or personal connections.
Provision: Mechanism	Mechanism details how privacy is provided. From the perspective of the individuals and those connected to them personally or professionally, there is no privacy provided. The data published online is fully transparent and permanent.
Scope: Temporal Scale	Temporal scale is the amount of time in which privacy should be installed. In the instance of this dataset, it is a rolling log of arrests for 30 days. However, this log belongs to a larger dataset. Therefore, as long as this larger dataset is accessible to the public, privacy should be applied indefinitely.

Impact of PII

Six individuals were selected at random to see if the personal identifiable information from the dataset could make them easily identifiable on the internet. Two of the six individuals - Laveka Brown and Mohein Allawe Hassan - had local news articles published of their arrests, which could negatively impact them in the future. Another individual - Ezra Donatelli Blackwood - had a facebook profile where more personal information could have been viewed about their family and friends. Having someone’s PII along with other personal information through social media accounts, social engineering could occur which can lead to identity theft. Finally, we found the LinkedIn profile for Darryl Adam Dewitt, who was arrested in October for possession of controlled substances and burglary. This individual is an internet business owner and was a downtown Berkeley ambassador for more than four years, aiming to “produce a clean and attractive Downtown experience.” People who were to do a google search can find his LinkedIn profile and arrest information at the same time, which would not be good for his professional career. Government agencies publishing public records with so much detailed PII can not only lead to possible identity theft through social engineering, but can also possibly have a negative impact on someone's professional career or personal life.

Recommendations

We recognize police departments such as the Berkeley PD follow the guidelines outlined in the California Public Act regarding arrest information that should be shared publicly. We feel strongly that the California Public Act be revisited regarding information shared about arrests. The court of public opinion is very strong and has a lasting effect even if an individual is found not guilty of a charge and is sometimes referred to as mob justice (Schneier, 2013). It's because of this that we feel the California Public Act should be reviewed to determine the harm caused by sharing a subject's name and date of birth along with arrest information. Guidelines should be developed around sharing arrest data for police departments to follow that demonstrates how to perturb data instead of sharing the entire dataset. This solution would provide consistency across all police departments in California and would provide a more robust and fair solution than each police department performing their own scrubbing of data while protecting an innocent person from being convicted before trial.

While we understand that delinking and scrubbing PII may not be enough to prevent private data from leaking, we believe that an attempt to protect the privacy of accused individuals should be taken. If we consider what Article 29 suggests, anonymization techniques should prevent the following (Article 29, 2014):

1. Singling out: Can you single out an individual's data from the dataset? (age, birth date, name etc.)
2. Linkability: Can you link records across datasets pertaining to an individual? (Think of the Netflix and IMDB example)

3. Inference: Can we infer information concerning an individual? (Can determine attribute X for John Smith, e.g., smoking status or health diagnosis)

We can pull from what the GDPR does to anonymize data to provide some recommendations for this dataset. For example, an anonymization process could be applied that excludes/masks/hides data that can be used to single out or link across datasets (Cohen, 2020). In general it doesn't matter if the data is PII or a random GUID as both are treated equally but in this case we recommend anonymizing name and date of birth at a minimum.

Changing the California Public Act is not a given, and if it is changed will take time. In the meantime, there are things that the Berkeley PD can do to protect individuals. While the California Public Act states that the subjects full name and date of birth be included in the information collected during an arrest, it also states that information can be withheld if *"disclosure of a particular item of information would endanger the safety of a person involved in an investigation or would endanger the successful completion of the investigation or a related investigation"* (Police Records, 2021). From this it appears that the information that each police department shares is somewhat up to their own discretion.

As mentioned previously, the arrest log could cause harm to the subject especially if the subject is exonerated from the crime. In addition to the recommendation of anonymizing data, we recommend adding a column that indicates conviction status and that these columns are updated once upon case resolution. An alternative would be to provide a link in the user interface that allows a user to click to see the status of the case and to also color code the rows (yellow for pending, red for convicted, green for not guilty). While it is difficult to change the public's perception, these measures attempt to provide additional information regarding a case and reinforce the notion of "innocent until proven guilty" (Presumption of innocence, 2021).

In addition, there does not seem to be a way for an individual to notice-and-comment on the PII that has been released. At a minimum, if no other changes are made, we would recommend the Berkeley PD allow for a notice-and-comment process for individuals who have been exonerated. This could be done via an online process where an individual submits their request along with proof of exoneration. Upon approval the record would be updated and removed from the public website.

Conclusion

After examining the data in the Berkeley PD arrest dataset through a legal and ethical lens, our team feels strongly that the data that is currently being shared has the potential to cause a subject harm and for a subject's information to be used beyond what it was collected for (secondary use). It’s for these reasons that we have recommended a long term solution of revisiting the California Public Act and updating the need to share PII arrest information publicly, and several immediate solutions for the Berkeley PD that involves withholding and/or anonymizing the PII arrest data from publication and in storage.

References

ARTICLE 29 (2014) Retrieved from https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf on November 8, 2021

Berkeley Police Department (2021, November 18). *Berkeley PD Log - Arrests* . City of Berkeley Open Data. Retrieved from https://data.cityofberkeley.info/Public-Safety/Berkeley-PD-Log-Arrests/xi7q-nji6 on November 18, 2021

Berkeley Police Department. (n.d.). *GUIDELINES FOR RELEASE OF REPORTS AND INFORMATION THEREIN*. Retrieved from https://www.cityofberkeley.info/uploadedFiles/Police/Level_3_-_General/GO%20R-23_attach_09Jan12.pdf from November 18, 2021

California Public Records Act - Section 6254. Codes Display Text. (n.d.). Retrieved from https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=7.&amp;chapter=3.5.&amp;lawCode=GOV&amp;title=1.&amp;article=1 on November 1, 2021

Cohen, et. al (2020) Towards Formalizing the GDPR's Notion of Singling Out Retrieved from https://arxiv.org/pdf/1904.06009.pdf on November 9, 2021

Mulligan, et. al (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. Retrieved from http://doi.org/10.1098/rsta.2016.0118 on September 20, 2021

Nissenbaum, Helen F. (2011). A Contextual Approach to Privacy Online. Daedalus 140:4 Retrieved from https://ssrn.com/abstract=2567042 on September 20, 2021

Police records. Berkeley Advanced Media Institute. (2020, December 4). Retrieved November 19, 2021, from https://multimedia.journalism.berkeley.edu/tutorials/police-records/.

Presumption of innocence (2021) Retrieved from https://en.wikipedia.org/wiki/Presumption_of_innocence on November 10, 2021

Public Records on the Internet: The Privacy Dilemma. Privacy Rights Clearinghouse. (2002). Retrieved from https://privacyrights.org/resources/public-records-internet-privacy-dilemma on November 18, 2021.

Schneier, B (2013) The Court of Public Opinion Is About Mob Justice and Reputation as Revenge Retreieved from https://www.wired.com/2013/02/court-of-public-opinion/ on November 9, 2021

Solove, D. J. (2006). A Taxonomy of Privacy. University of Pennsylvania Law Review. Retrieved from https://ssrn.com/abstract=667622 on September 22, 2021