Dear Citywide Analytics team of Boston,

We are writing to express our concern with one of the open datasets hosted on Analyze Boston, Boston's Open Data Hub. The Employee Earnings Report publishes payroll data for government employees each year. We found, through our ethical and legal analysis below, that the dataset can be exploited, and the individuals' personal privacy is violated in a number of ways. Below is our analysis, specifically with regards to the police department of Boston, and we hope that you will take steps to understand how these negative implications may affect these men and women's lives.

Introduction of the Dataset and Sleuthing

The open-sourced data that we analyzed is referred to as the "Employee Earnings Report" which is published by the City of Boston's (Department of Innovation and Technology) payroll data of employees. This dataset contains payroll data for years 2011 to 2021 containing employee names, job details, and earnings information including base salary, overtime, and total compensation for employees of the city. For the purpose of this paper, we focused primarily on "Police Department" data to analyze the impact and level of privacy concern that is readily available.

The purpose of allowing access to similar information, among other reasons was to increase operational efficiencies, reduce cost, improve service along with the increase in transparency to improve accountability. While the concept of open data started in the 1950s, when "congressional deliberations that culminated a decade later in the Freedom of Information Act (FOIA), open government was seen (and the term first used) as a pathway to more government accountability".[1] On January 14, 2019, the Foundations for Evidence-Based Policymaking Act ("Evidence Act"), which includes the OPEN Government Data Act, was signed into law.

Our dataset is viewable on the website and can be downloaded (in .csv) format for further analysis. Within the data set for each individual (employee) is represented by their Name, Department Name and Title. However, the addition of the zip code which is the residence of the given individual/officer is also provided in the dataset. The compensation data is broken down into regular, retro, other, overtime, injured, detail and Quinn education incentive that aggregates to the total gross pay is provided. The values for 2022, which was updated on February 23, 2022 (updated annually) and includes information about 22,546 employees (and 5 rows with no data). The data is classified under 230 departments with "Boston Police Department" having the largest number of entries, 3,094 or 13.7% within the dataset.

According to Project Open Data: Open Data Policy states, in order to strengthen measures to ensure that privacy and confidentiality are fully protected and that data are properly secured, the act articulate the notions that agencies should note that the mosaic effect demands a risk-based analysis, often utilizing statistical methods whose parameters can change over time, depending on the nature of the information, the availability of other information, and the technology in place that could facilitate the process of identification." [2] Specifically for employees who own a home using the name and leveraging the Massachusetts land/property records across the State which is available in (https://www.masslandrecords.com/) can easily identify the address. The database or land and property records are split at the county level, but the zip code seems to reduce the choice set to specific individuals. While a name by itself will provide all property records in terms of ownership, given the salary zip code the level of accurate identification is significantly higher. A simple search was able to yield the address with a high degree of certainty.

This again seems to cause concern that would not be in agreement with the Circular A-130 -- Management of Federal Information Resources that states: "The individual's right to privacy must be protected in Federal Government information activities involving personal information."[3].

Privacy Analysis

The Massachusetts Public Records Law (Public Records Law) and its Regulations provide that each person has a right of access to public information. While there are exemptions to the Public Records Law, the exemption for "personnel records" is very limited because a major purpose of the Public Records Law is to enable taxpayers to monitor government activities, expenditures and employees. Personnel information that is useful in making individual employment decisions could be exempted from disclosure. Therefore, ordinary evaluations, performance assessments, and disciplinary determinations are exempt personnel records under the statute. However, the employee's name, address, and base and overtime pay are not considered as "useful in making employment decisions", and therefore are not exempt under the "personnel records" prong.

Although it is understandable that the payroll related information should be revealed to the public since it relates to how the government has used the taxpayers' money, releasing payroll along with other personal identifiable information such as name and job title raises serious privacy concerns. In this section, we will look at the privacy concerns through the lens of the commonly used privacy analytics frameworks.

Solove's Taxonomy

Information Collection:

The information collection could be a continuous process. It initially started at the moment when a police officer accepted the job offer, and the data collection continues as the police officer progresses in his or her career journey. When the police officer works overtime, gets injured, changes position, or gets salary adjustments, the data will be captured on an ongoing basis.

Information Processing:

The information should be processed primarily by the Finance and HR department of the Boston government. However, payroll processing could be outsourced to third party service providers as well, therefore, it is also likely that the raw payroll data could have been accessed and handled by external parties. Data has been aggregated and reported on an annual basis.

Information Dissemination:

The Public Record Law enables the taxpayers to monitor the government activities and employees, which has led to a full disclosure of a police officer's payroll information. We can assume that the police officers who work for the Boston government must have given consent for the government to release their information to the public when they accepted the job offer. Every year, Boston releases an employee earning report to the public, which includes full name, department, job title, base salary, overtime pay and injury status. The dataset is posted on the government's official website and can be downloaded by anyone for further analysis.

Nissenbaum's Contextual Integrity

In modeling privacy, we should first understand the context in which information is being shared. The government discloses the personnel information in order to enable the public to monitor the government's effort in realizing transparency and accountability. In this context, a data subject could be any police officer who worked for the Boston government in 2021. The police officers send their information to the Boston government either proactively when applied for, accepted the job offer or reported injury or overtime, or passively, when received payroll or new position during their tenure at the Boston government. The Boston government is the recipient and the holder of the information, and finally becomes the sender of the information when releasing the police officer earning information to the public. While we can see why the individual level information needs to be disclosed in the context of government transparency, the level of details raises serious concerns about the police officers' privacy and the potential harm that could derive from it.

Mulligan et al.'s analytic,

Balancing privacy protection while maintaining the openness in an open dataset seems to be extremely challenging because privacy and openness are fundamentally contrast concepts. Privacy is supposed to protect personal information from being disclosed and eventually used by someone in an inappropriate or unlawful manner. However, it seems that there is barely any measure to protect the privacy of the individual police officers, the data subjects. Privacy is not justified because their personal identifiable information has been disclosed to the public and more personal information such as past job experience, home address or home ownership history could be further identified with the information that has been disclosed. There is really no more sensitive data than personal identifiable information such as name and address. Especially for police officers who are on the frontline to fight against crimes, putting their full name and home address out there raises serious concerns about their personal safety and could give someone who is up for no good a chance to find out where a police officer and his or her family are. Besides physical safety, the personal identifiable information could be used by someone to conduct a wide range of crimes such as opening a credit card with a certain police officer's name, stealing a police officer's tax refund or even pretending to be a police officer when being arrested or to conduct other fraud against public security.

As the police officers' employer, the Boston Government ought to implement measures to protect their employees, the police officers' privacy. However, as a government which has adopted an open data policy, the Boston government has no mechanism to protect the police officers' privacy. This could have become a social practice where a full disclosure of detailed personal information is expected in the context of government transparency. However, the fact that each of the earning reports since 2011 is still available for the public to access raises questions about whether it complies with the data minimization principle and what their data retention policy is.

Ethical and Legal Implications

Although the publication of employee salaries in the government fosters an environment of transparency and accountability, there can also be many ethical and legal problems. We will first discuss these ethical issues at hand and subsequently how to address them through both technical means and more laws and regulations.

A major ethical issue that we identified with this dataset has to do with two specific data fields: Name (first and last name) and Postal (geographic zip code). As demonstrated earlier in this paper, we were able to identify a police officer on multiple websites. The first and last name is a key identifier for a multitude of things—things that can be as harmless as searching one's alma mater to things as serious as being able to search on the National Sex Offender Registry or Public Criminal Records. The National Sex Offender public website is searchable via a single key, Full Name. There is also the option to search by Full Name and Location. This registry is undoubtedly created for the safety of the public, but is there a breach of ethics here? Does this registry violate human rights laws? Even if an individual commits a heinous crime, that does not necessarily decisively determine all future actions made in the future. The Boston Public Records website has the ability to publicly search a Name and City for Boston arrest, court, and public records. In a similar form, this is an ethical dilemma where the formerly incarcerated individual does not want this information accessible by the public. This information can affect their livelihood—it can create a barrier to proper employment, limit child custody rights, revoke driving privileges, and negatively affect renting and leasing opportunities.

The second data field, the individual's Zip Code, presents an additional ethical dilemma. Members of the police department are especially negatively affected because of their role in society. Police officers are the ones who must deal with individuals that break the law. These same individuals, potentially vindictive, will have access to these public records. It would not take much effort at all to search a name and a specific zip code, eventually finding public homeowner

information, tax information, etc. This criminal will ultimately pinpoint exactly where that police officer lives. The targeted officer is now at risk of his privacy being violated, or worse, his safety will be jeopardized. This is just one example of many of how police officers can be disproportionately negatively affected by this publicly available data.

Before we discuss ways to address these types of ethical dilemmas, it's important to understand the current governing law, in this case, the Massachusetts Public Records Law that is the backbone of this public data. The Public Records Law broadly defines "public records" to include "all books, papers, maps, photographs, recorded tapes, financial statements, statistical tabulations, or other documentary materials or data, regardless of physical form or characteristics, made or received by any officer or employee" of any Massachusetts governmental entity. It's worth nothing that this public records law is effective and relevant to requests made to the Records Access Officer or RAO. Essentially, anyone may request the data from the RAO, and they will help to complete your request in a timely manner. This creates a moderate barrier to entry since the state must store a record for each request.

Unfortunately, the open payroll dataset hosted on Analyze Boston does not have any type of request format similar to the RAO interaction mentioned earlier. This is of course, intentional, as this new open data platform is a part of the city of Boston's Open Data to Open Knowledge project. However, we urge you to consider creating some barrier of entry, similar to the process that the Massachusetts Public Records Law adheres to.

Although the city of Boston does indeed have an Open and Protected Data Policy, the policy focuses on working definitions for Open Data, along with information on how it is to be published, reviewed, and licensed—and not so much about how to protect the data subjects, specifically. Although there "may also transform protected data sets so that a relevant and appropriate view of that data can be published as open data", this vagueness does not properly protect this vulnerable population of police officers. Therefore, it is our recommendation that Open Data should ideally be governed by a much stricter and specific set of standards.

The process to make this a reality is lengthy, so it would behoove you to do this piecemeal. Following a similar workflow as the traditional request for public records via a RAO, you can, at a minimum, create some sort of submission form before granting access to this Employee Earnings Report. Notifying potential requestors that their request will be documented may deter some individuals from using the data in a malicious way. This barrier to entry is easy to implement technically and can go a long way until you can develop proper guidelines with Open Data sharing.

In order to prevent your police department members (and all other government members for that matter) from being exploited, consider the following two actions: (1) minimizing your data and/or (2) sample or perturb the data. For data minimization, look to the context of the dataset, for example. This dataset is supposed to contain payroll data. The Zip Code data, a field that can be exploited in tandem with First and Last Name, may not be a field that is actually required in this context. In this case, our dataset would have fewer columns, and therefore fewer positions of harm in the data dissemination workflow. Second, it could be beneficial to mask the individuals in the dataset. Ask yourself to probe questions such as "why do we need the exact salary of this particular person?" The role and general distribution of the population should be enough for most requests, so long as they are not malicious requests.

All in all, there are many ethical and legal implications of this open dataset that should be considered. It is best to take actions piecemeal and you may find that low cost solutions (such as creating a request form) to be a moderate barrier to entry for any pernicious behavior. While creating a stricter Code of Conduct, it is our suggestion that you be more specific. For example, explicitly stating "do not include the Zip Code in the dataset" will make things more actionable,

taking small steps towards protecting the privacy rights of all individuals in the city of Boston's open datasets.

Conclusion

We hope you have found our analysis insightful and actionable. Please feel free to reach out to us with any questions or concerns.

Thank you,
UC Berkeley's School of Information W231 Group: Ferdous, Kelly, & Sissie

References

1. Americans' Views on Open Government Data, BY John B. Horrigan and Lee Rainie https://www.pewresearch.org/internet/2015/04/21/introduction-21/#fn-13338-3
2. Project Open Data: Open Data Policy — Managing Information as an Asset https://project-open-data.cio.gov/policy-memo/#fn:27
3. Memorandum for the Heads of Executive Departments and Agencies, CIRCULAR NO. A-130 Revised https://georgewbush-whitehouse.archives.gov/omb/circulars/a130/a130trans4.html