Personnel, Use of Force, and Complaints in the Chicago Police Department (CPD)
Datasheet

Reproduced from: [1]

I. Motivation

I For what purpose was the dataset created?

The original raw data files were sought by J. Kalven, a journalist in the City of Chicago, as part of his investigation into police abuse. After the original FOIA requests and legal case, the non-profit Invisible Institute (https://invisible.institute) began to collaborate with Kalven and the University of Chicago's Mandel Legal Aid Clinic to follow up on earlier FOIA requests and to file new ones. The data disclosed in response to these earlier and now ongoing FOIA requests were made available online as part of the Citizens Police Data Project.

I Who created the dataset, and on behalf of which entity?

The Chicago Police Department (CPD), Civilian Office of Police Accountability (COPA), and the City of Chicago produced the raw data files in response to FOIA requests. The raw data were curated and released publicly by the Invisible Institute and its collaborators. The cleaned and linked data were produced as part of research by the authors of this document.

I Who funded the creation of the dataset?

The acquisition of the original raw data was funded by the Invisible Institute.

II. Composition

II What do the instances that comprise the dataset represent?

There are multiple types of instance in this data.

- Officer: information about an individual police officer
- Unit assignment: a single unit assignment for an officer
- Complaint: a complaint filed against a police officer, either internally or by a civilian
- Tactical Response Report: a form that an officer is required to fill out after their response requires use of force
- Award request: a request to grant an award to an officer
- Salary: a record of an officer's salary, pay grade, and position across multiple years

II How many instances are there in total (of each type)?

There are roughly 35,000 unique officers in the cleaned roster appearing in roughly 130,000 profiles throughout the data, 730,000 award request records, 194,000 salary records, 108,000 unit assignment records, 109,000 complaints, and 10,500 tactical response reports.

II Does the dataset contain all possible instances or is it a sample of instances from a larger set?

This data contains information regarding all sworn officers in the Chicago Police Department / City of Chicago databases for the stated date ranges (which differ for each source of raw data).

II What data does each instance consist of?

- Officer: officer unique ID, race, gender, age, appointment date, resignation date, badge number(s), position title(s)
- Unit assignment: officer unique ID, start date, end date, unit number
- Complaint: complaint ID, involved officer IDs, allegation, result of the investigation, resulting sanction (where available)
- Tactical Response Report: report ID, event location, date, and time, environmental conditions, who was notified, weapons discharged, weapon information, subject demographic information
- Award Request: awardee unique ID, requester, request date, award reference number, award type, request tracking number, incident dates, ceremony date
- Salary: officer unique ID, salary, position title, pay grade, year

II Is there a label or target associated with each instance?

Not explicitly. However, labels could be constructed from the data that exists. For example, one could aggregate complaints to produce an integer "number of complaints" for each officer in the data, and use that as the response variable in a prediction task.

II Is any information missing from individual instances?

In the original raw data files, missing data (of all fields) is quite common (see Appendix D in [1]). In the cleaned and linked data files, we are able to aggregate multiple profiles of a single officer appearing throughout the data to "fill in the gaps," although this process is not perfect and there are still missing entries.

II Are relationships between individual instances made explicit?

In the raw data, no. In the cleaned data, we provide a unique officer identification that enables linking the activities and records regarding individual officers across datasets. There is no relational data (i.e., network edges) explicitly contained in the data. However, it is possible to use the data to construct a network, e.g., by linking officers co-listed on complaints.

II Are there recommended data splits?

No, although the officer database is likely to be incomplete prior to roughly 1980.

II Are there any errors, sources of noise, or redundancies in the dataset?

There are redundancies in the raw data, but these are removed by our cleaning and linking procedure. Errors, inconsistencies, and missing data are also present in the raw data; our cleaning and linking resolves much of these issues. However, per Section 4 in [1], the officer database is likely to be incomplete prior to roughly 1980 (as officers were added to the database only gradually over time).

II Is the dataset self-contained, or does it rely on external resources?

The dataset is self-contained: the raw data itself is stored in the `raw/` folder of the repository (with links to the external source files for reference), and the cleaned/linked data is produced by the source code in the repository.

II Does the dataset contain data that might be considered confidential?

No; all of this data was publicly released as part of FOIA requests. Confidential data (e.g., relating to under cover officers) was withheld by the Chicago Police Department.

II Does the dataset contain data that, if viewed directly, might be offensive, insulting, or threatening?

No.

II Does the dataset relate to people?

Yes; it contains records relating to police officers in the Chicago Police Department.

II Does the dataset identify any subpopulations?

Yes; officer records include race, gender, age, appointment date, unit history, badge numbers, position title, salary, awards, complaints, and tactical response reports. Subpopulations of officers can be constructed using these fields.

II Is it possible to identify individuals?

Yes; detailed information is available that could be used to identify individual officers.

II Does the dataset contain data that might be considered sensitive in any way?

The data contains a coarse categorization of racial origins of officers.

III. Collection Process

III How was the data associated with each instance acquired?

The raw data were obtained via FOIA requests to the City of Chicago and Chicago Police Department.

III What mechanisms or procedures were used to collect the data?

The raw data were obtained via FOIA requests to the City of Chicago and Chicago Police Department.

III If the data are a sample from a larger set, what was the sampling strategy?

Not applicable.

III Who was involved in the data collection process and how were they compensated?

Journalists in collaboration with the Invisible Institute were responsible for filing the FOIA requests, and officials within the Chicago Police Department and City of Chicago were responsible for providing data in response to those requests. It is not known explicitly whether or how either party was compensated.

III Over what timeframe was the data collected?

The earliest releases per FOIA request occurred in 2016, and continue to occur as more FOIA requests are filed. The raw data itself pertain to records from the CPD dating back to the mid 20th century. The roster data covers the period up to 2018. The awards data pertains to records from 1967 to 2019. The salary data pertains to the years 2002 to 2017. The unit history data covers records up to 2016. The complaints data pertains to records from 1967 to 2016. The tactical response report data pertains to records from 2004 to 2017.

III Were any ethical review processes conducted?

It is unknown whether the CPD conducted any ethical review processes prior to the release of the raw data. No ethical review process was conducted prior to the activities involved in the present repository, i.e., cleaning the publicly available data.

III Does the dataset relate to people?

Yes; it contains detailed records regarding the activities of police officers in the City of Chicago.

III Did you collect the data from the individuals directly, or obtain it via third parties?

The raw data was acquired from public links provided by the Invisible Institute (https://invisible.institute). The Invisible Institute acquired the data through FOIA requests made to the CPD and the City of Chicago.

III Were the individuals notified about the data collection?

It is unknown whether the individual officers were notified by the CPD when the raw data was released.

III Did the individuals in question consent to the collection and use of their data?

Not explicitly. The Chicago Police Department was compelled by law to produce these records per FOIA requests.

III If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

Not applicable.

III Has analysis of the potential impact of the dataset and its use on data subjects been conducted?

Not known.

IV. Preprocessing and cleaning

IV Was any preprocessing of the data done?

Yes; the main section of this documentation provides details the cleaning and linking of the raw data resulting from FOIA requests made to the City of Chicago.

IV Was the "raw" data saved in addition to the cleaned data?

Yes; the raw data is available in the `raw/` folder in the repository.

IV Is the software used to clean the data available?

Yes; the source for cleaning and linking is provided in the `src/` folder in the repository.

V. Uses

V Has the dataset been used for any tasks already?

Not the newly cleaned and linked version. The raw data itself has been used previously; see e Section 5 in [1] for details.

V Is there a repository that links to any or all papers that use the dataset?

Not that the authors of this work are aware of.

V What (other) tasks could the dataset be used for?

This data set has a rich variety of possible uses; for example, network analysis (and in particular, analysis of dynamic events occurring on networks) and predictive regression/classification. See Section 5 in [1] for more details.

V Is there anything about the composition of the dataset or the way it was collected and cleaned that might impact future uses?

Yes; the data are less reliable in earlier years (e.g., pre-1980). See Section 4 in [1] for more details.

V Are there tasks for which the dataset should not be used?

This data should not be used to single out, study, or identify individual officers.

VI. Distribution

VI Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?

Yes, the data is publicly available.

VI How will the dataset be distributed?

It is available on GitHub at https://github.com/chicago-police-violence/data. Release versions will be marked using the "release" feature on GitHub.

VI When will the dataset be distributed?

It is currently publicly accessible.

VI Will the dataset be distributed under a copyright, other IP license, or terms of use?

Yes; the source code is released under the MIT license, and the data output by the cleaning code is released under the Creative Commons 4.0 BY-NC-SA license.

VI Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

VI Do any export controls or other regulatory restrictions apply to the data?

No.

VII. Maintenance

VII Who is supporting/hosting/maintaining the dataset?

The repository will be hosted on GitHub. As of August 2021, the repository owners are Thibaut Horel, Trevor Campbell, and Lorenzo Masoero, but ownership may change over time.

VII How can the data owner/curator be contacted?

Issue threads on GitHub are the primary channel of contact for the repository maintainers.

VII Is there an erratum?

Not as of yet. For each major release version, notes will be included and hosted in the repository that will detail cleaning/linking errors that have been fixed.

VII Will the dataset be updated?

The original raw source data from FOIA requests will not be modified. More raw data files may be added over time corresponding to new FOIA requests. The data cleaning and linking code will be edited over time to fix errors; release versions will be clearly marked on GitHub. There is no set schedule for updates.

VII If the dataset relates to people, are there applicable limits on the retention of data associated with the instances?

No; this data was released per FOIA requests and is in the public domain.

VII Will older versions of the dataset continue to be supported/hosted/maintained?

Yes; a full version-controlled history of the project exists on GitHub.

VII If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Yes; the repository for the dataset is hosted on GitHub, where pull requests are a usual channel for external contribution.

References

[1] Thibaut Horel, Lorenzo Masoero, Raj Agrawal, Daria Roithmayr, and Trevor Campbell. The CPD Data Set: Personnel, Use of Force, and Complaints in the Chicago Police Department, 2021.