Legal and Ethical Analysis of Re-Identification

W231 Wednesday 4:00 PM

Date: April 2021

Lucas Lam, Dan Ortiz, Mohan Sadashiva, Darian Worley

Executive Summary

Data scraped from the City of Chicago’s open data program was de-anonymized and aggregated with multiple first and third party data sources to profile one of its members of the business community. Chicago's open data is at serious risk of web scraping and misuse by bad actors who can harm the city's citizens. The city needs to take immediate action to mitigate this risk to protect its citizens from privacy invasions and potentially greater harms.

Since the advent of the Open Government Directive, city and government publications of public/open data has exploded. Today, public records like land ownership and employee salaries are easily available to scrape and aggregate resulting in intrusions into citizens everyday lives. While there are legitimate reasons why data should be open and public, like transparency in land ownership, it is unwise for the government to publish open data sets without assessing the right level of information. Governments need to assess the potential harms, including joining with third party data sets, releasing a data set could have on its citizens regardless of intentions.

The city of Chicago is one of the largest U.S. cities to embrace the Open Government Directive, providing endless amounts of city data to anyone on the internet. Although some data sets require special access, most do not even require an account to gain access to. Enter the Problematic Landlord (PLL) data set.

The Chicago Building Scofflaw Ordinance (Section 2-92-416 of the Municipal Code of Chicago) prevents landlords that refuse/refrain to remedy building code violations from receiving city contracts, including subsidized housing. Violations of the code includes failure to provide adequate heat, hot water, and working smoke/carbon monoxide detector. Landlords who have been found liable in two or more Administrative Hearing cases within a 24 month period and have three or more serious building code violations are added to the PLL. Information such as the legal ownership entity, address of the property in violation, and the property's geographic location are included in the dataset.

There are serious consequences for landlords who are on the PLL. These include the inability to obtain a business license, obtain building permits unrelated to the cited violations, and in some cases forfeiture of the property. Due to the steep consequences for being placed on this list, it is not unreasonable to believe an individual is associated with the entity that owned the property is negligent. Leveraging other open government data sets, it's easy to put a name to this entity,

To further illustrate this concern, the team took a business entity from the website and constructed a profile around the owner. PII has been removed as to not cause any further harm to the individual.

Olaf is a 65 year old male who lives in the Grand Rapids Michigan area. His primary job is a dentist and studied at both the University of Michigan and Michigan State. He owns his own business, has two locations, and his wife is a nurse who works with him. In addition, Olaf has two children and the family is Catholic. Olaf is one of 5 brothers and sisters and his father was a miner and his mother was a nurse. Olaf currently owns three LLC not including his dentistry practice and his residence is valued over $600K and is located in a secluded part of town.

The key to constructing this profile was joining the data from the states business registry to the PLL data set as part of the information in the registry iis the owner's name and address (note: this does not have to be a person, this could be another company). This information gave us the ability to find the subject on LinkedIn, which provided his profession, age, education, primary business, and confirmed his general location. Then using Google we were able to find additional information about him and his family.

This aggregation took a little more than two hours as the individual used a network of legal entities and trust to hold his assets and had a limited self-promoted internet footprint. If our subject was a little more intentional on how he used different legal entities, and where he incorporated them, discovering his identity would have become increasingly difficult. This raises concerns of pay-for-privacy in government data that protects the powerful and wealthy while subjugating everyone else.

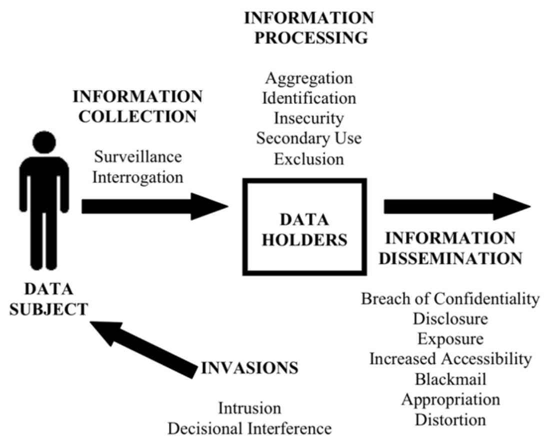
While this aggregation is stalker level creepy and is easy enough today's kindergartener can do it, it is not illegal. This is so important we'll restate it. This is not illegal. The PLL is a government data source released to the public in the interest of transparency. The intent is to inform the residents of Chicago of problematic landlords. In addition, the information in the data set are business and property ownership records, which is public information. However, the ethics of "cyber-stalking" have something to say.

The key question here is "What is the value of having this dataset public?". If the goal of the dataset is to inform citizens about problematic landlords in Chicago, there are means to limit the possibilities of mis-use of this kind of data. When looking at the impact to citizens' privacy through the three frameworks, Solove's Taxonomy, Nissenbaum's Contextual Integrity, and Mulligan/Koopam Multidimensional Privacy Analytics, we can assess the potential harm of releasing the data.

Solove's Taxonomy (see Figure 1) assesses privacy from four different perspectives: 1) information collection, 2) information processing, 3) information dissemination, and 4) invasion.

Figure 1: Solove's Taxonomy

Figure 1



Applying Solove's Taxonomy to the PLL dataset we note the following:

- *Information Collection:* The information was collected legally. However the landlord was likely not aware at the time of information collection of the potential exposure of this information to the public and its ramifications.
- *Information Processing:* PLL processes this information and does not specifically include owner name and owner address.
- *Information Dissemination:* The city of Chicago is publishing data that can be easily joined with other first party data sets to reveal personally identifiable information about the landlords.
- *Invasion:* A combination of the information made public along with other data sources constitutes an invasion of the privacy of the individual.

In summary, although the information was collected legally, the PLL data can be combined with other data and potentially lead to harm to the individual. This information dissemination could lead to a breach of confidentiality and ultimately an invasion of privacy.

Next we consider Nissenbaum's contextual integrity framework. Nissenbaum's contextual integrity framework attempts to model privacy in terms of the expected flow of personal information modeled with the construct of context-relative norms. The key parameters of informational norms are: 1) Actors, 2) attributes, and 3) transmission principles. Applying Nissebaum's framework, we focus on the transmission principles. Considering the context of the information posted on the website as a matter of record, posting the PLL data doesn't seem to do any harm when considered in isolation. However, again we see that when we use the data outside of its intended context, we see that the data can cause harm to the subjects (the landlords) in the dataset.

Finally, we consider the Mulligan/Koopman multidimensional privacy analytic (see figure 2). The Mulligan/Koopman multidimensional privacy analytic is a very detailed approach to applying a privacy construct. This analytic evaluates privacy using the following 5 privacy dimensions: 1) dimensions of theory, 2) dimensions of protection, 3) dimensions of harm, 4) dimensions of provision, and 5) dimensions of scope. Applying the Mulligan/Kopman multidimensional privacy analytic to the PLL data (see Table 1), we note that there is no indication of how long the data will be kept. This is alarming since a PLL may be on this list but could have paid the necessary fines to resolve the outstanding issue. This would be misleading and could cause harm to an individual on this list.

Table 1: Application of the Mulligan/Koopman Multidimensional Privacy Analytic to the PLL dataset

Privacy Dimension		
Level 1 Dimension	**Level 2 Dimension**	**Application to PLL data**
Dimension of Theory	Object	control over personal information
Dimension of Protection	Target	Personal information
	Subject	Business Owners
Dimension of Harm	Offender	The government and other individuals using the data for unintended purposes
Dimension of Provision	Mechanism	Social norms are intended to protect and individual's privacy
	Provider	Government
Dimension of Scope	Temporal Scale	There is no indication regarding how long data will be kept

Recommendations to the City

This assessment has proven that bad actors can access and misuse the information the city provides to the public. Since the city is disseminating this data, the city has an ethical obligation to reduce the harm this data can inflict on its subjects. The team formally recommends the following actions be taken to reduce the potential for misuse of Chicago's public data.

1. **Licence/Terms of Use:** The City of Chicago should adopt a licences and/or terms of use for the use of its open data. This would provide legal recourse for the city and its subjects under contract law if the data is misused.
2. **Control Access to Open Data:** The City of Chicago should require users to create a free account to access the data in the website. This will increase the difficulty for web scrapers to grab the open data.
3. **Log and monitor access:** Log what account access what data, and the pertinent unique identifiers for a defined time period. This can be used in the event of suspicious activity.
4. **Define data retention for subjects in the data set:** What is the relevant time period this data will help the citizens of Chicago. If it no longer has utility it should be removed from the PLL set. The underlying data should be kept according to the data retention policy in the appropriate system. This will result in the data being used for intention.
5. **Validate all data in the data set is required to meet the need:** The City of Chicago should analyze it's open data to ensure that just enough data is being

distributed in its data set to meet the needs of its intentions. Any data that is not needed should be removed from the data set to reduce the risk of linking to other data sets resulting in greater harm to its citizens.

6. **Clear Notice and Data Correction Policies:** The City of Chicago should be notified of addition of their data to an open data set and be given the opportunity to correct the data before it is published. Incorrect data has the potential to become codified in other systems that get its data from the open data set, perpetuating the false information.

The City of Chicago needs to implement tighter controls around its open data to prevent harm to its citizens. This assessment has demonstrated how easy it is to associate data provided by the city to other data sources to build profiles around citizens within a few hours. The PLL is one data set, out of hundreds, disseminated by the city which compounds the opportunity for bad actors to misuse the data. In order to serve the public interest and protect its citizens, the City of Chicago needs to revise its current controls around its open data program.

Work Cited

C. (n.d.). *City of Chicago Open Data*. Retrieved April 2, 2021, from https://www.chicago.gov/city/en/depts/bldgs/supp_info/building-code-scofflaw-lst.html.

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*, (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). https://ssrn.com/abstract=2567042

Open government directive. (2014, July 10). Retrieved April 02, 2021, from https://www.fcc.gov/general/open-government-directive

Solove, Daniel J. (2006). A Taxonomy of Privacy. *University of Pennsylvania Law Review,* 154:3 (January 2006). https://ssrn.com/abstract=667622