Doxing and Hate Crimes: Identifying Risk at the Intersection

Project Overview

People motivated enough by hate to commit a hate crime have more access than ever before to a potential victim's home, place of work, and other personal information. This could lead to an increased risk of hate crimes where there is both a growing trend of hate crimes and an increased ability to dox someone. Understanding where this intersection lies is critical to identifying if and where these issues exist. Conversely, identifying best practices can also provide guidance for the municipalities who are missing the mark in terms of victim protection.

While there are many possible stakeholders, in this project, we are focusing on elected officials in municipalities as primary stakeholders, with police departments and the Federal Bureau of Investigation (FBI) as secondary stakeholders. We assume elected officials will be the most responsive to the optics of either low or heightened risk. Police departments may not have the same priorities, as they often have many crimes to investigate with limited resources. Also, while investigating hate crimes is a priority for the FBI[1], it is a large federal bureaucracy that, by definition, is likely to be more slow-moving than a municipal government.

Therefore, this product does the following:

- Identifies municipalities in which we are most and least likely to see the greatest risk of hate crimes, combined with a relative ease of doxing
- Identifies best practices and recommendations for municipalities struggling with this issue
- Presents an interactive data explorer for stakeholders to explore questions and trends that reflect their own interests and priorities

Background

Hate Crimes

According to the FBI, a hate crime is any traditional offense that is in part or in whole driven by the offender's bias against a person's membership in a protected group. These protected groups include race, religion, disability, sexual orientation, ethnicity, gender, and gender identity. While hate and/or bias is not a crime, acting on this in a way that harms others is a crime[2].

Hate crimes go beyond harming someone's person or property in a traditional sense. These crimes violate the victims' civil rights, in addition to the traditional harm. In the United States, every person, regardless of group, deserves equal treatment from both the government and members of their community. It is this prioritization of upholding civil rights that motivates the FBI's focus on and prosecution of hate crimes as such[3].

Doxing

Doxing (or doxxing) is a form of online abuse by which one party releases personally identifiable information (PII) and/or sensitive information. This is done for many reasons, all with the intention of harm, but started in the online gaming world[4]. According to the FBI[5], hacking victims and members of the law enforcement community are at an increased risk of being doxed.

While the FBI report[6] associated doxing risk with prosecution of hacker groups such as Anonymous, we are assuming that because the ability to dox requires less sophistication (depending on data availability), the risk is more widespread now. In fact, cities are faced with a need to balance data transparency with protecting individual privacy[7].

Levels of Personal Information Identifiability

Below are various levels that can be used to identify people[8]:

- Direct Identifier: Name, Address, social security number
- Indirect Identifier: Date of Birth, Zip Code, License Plate, Medical Record Number, IP Address, Geolocation
- Data Linking to Multiple Individuals: Movie Preferences, Retail Preferences
- Data Not Linking to Any Individual: Aggregated Census Data, Survey Results
- Data Unrelated to Individuals: Weather

De-Identification and Re-Identification of Personally Identifiable Information

Anonymization, or de-identification, is a common response to privacy concerns in our digital economy. The way this is done is by removing PII from a dataset. However, because of re-identification, anonymization is not a guarantee of privacy.[9] In fact, Hintze and El Emam find that pseudo-anonymized data is much closer to non-anonymized data than anonymized data.[10]

Re-identification is where PII that had been anonymized is accurately matched with the original owner or subject. This is often done by combining two or more datasets containing different information about the same or overlapping groups of people. Often, this risk occurs when de-identified data is sold to third parties, which then re-identify the particular individuals.[11][12][13][14]

One example of this is seen in Sweeney's 2002 paper, in which she was able to correctly identify 87% of the U.S. population with just zip code, birthdate, and sex.[15] Another example is work by Acqusiti and Gross, in which they were able to predict social security numbers with

birthdate and geographic location.[16] Other examples include Kondor, et al., who were able to identify people based on mobility and spatial data. While their study only had a 16.8% success rate after a week, it jumped to 55% after four weeks. With higher frequency data collection, they expected higher success rates in even shorter periods of time.[17]

There are four general types of de-identification[18]:

1) Removing Data
    a) According to Health Insurance Portability and Accountability Act (HIPAA), only the first 3 digits of a zip code can be reported
2) Replacing Data with Codes or Pseudonyms
    a) Using unique identifiers instead of names or social security numbers is not enough
    b) Pseudonyms only work if they cannot be reversed
3) Adding Statistical Noise
4) Aggregation

El Emam presents a de-identification protocol for open data[19]:

1) Classify variables according to direct, indirect, and non-identifiers
2) Remove or replace direct identifiers with a pseudonym
3) Use a k-anonymity method to de-identify the indirect identifiers
4) Conduct a motivated intruder test
5) Update the anonymization with findings from the test
6) Repeat as necessary

As Narayanan, et. al. suggest, the true risk of re-identification is not just unknown, but unlikely to be truly unknowable.[20] However, assuming re-identification is always possible (albeit difficult or inconvenient), we can measure the relative ease (or lack thereof) with which re-identification is possible.

Products/Algorithms/Model

To assess the combined risk of doxing and hate crimes, we took a mixed-methods approach. Specifically, we compiled a list of all U.S. municipalities with populations above 100,000, per the 2010, 2000, and 1990 census reports[21]. Then, we assessed law enforcement data for each of these to determine the individual risk of doxing. We also used hate crime data from the FBI Data Explorer[22] to create time series models to forecast the trend of hate crimes at the national, state, and, for the municipalities in the dataset, local levels. From there, we assessed the combined risk of both.

For a more detailed overview, please see the Methodology: Doxing Risk Assessment and the Methodology: Time Series Model pages on this site.

Endnotes

1. Federal Bureau of Investigation. (2019). Hate Crimes [Folder]. Retrieved February 10, 2019, from https://www.fbi.gov/investigate/civil-rights/hate-crimes
2. ibid.
3. Federal Bureau of Investigation. (2019). Hate Crimes [Folder]. Retrieved February 10, 2019, from https://www.fbi.gov/investigate/civil-rights/hate-crimes
4. Snyder, P., Doerfler, P., Kanich, C., & McCoy, D. (2017). Fifteen minutes of unwanted fame: detecting and characterizing doxing. In *Proceedings of the 2017 Internet Measurement Conference on - IMC '17* (pp. 432–444). London, United Kingdom: ACM Press. https://doi.org/10.1145/3131365.3131385
5. Federal Bureau of Investigation. (2011, December 18). (U//FOUO) FBI Threat to Law Enforcement From "Doxing" | Public Intelligence [FBI Bulletin]. Retrieved February 3, 2019, from https://publicintelligence.net/ufouo-fbi-threat-to-law-enforcement-from-doxing/
6. ibid.
7. Valenta, Blake. (2017, October 6). How to Open Data While Protecting Privacy. *Government Technology*. Retrieved from http://www.govtech.com/data/How-to-Open-Data-While-Protecting-Privacy.html
8. Lubarsky, Boris. (2017). Re-Identification of "Anonymized" Data. *Georgetown Law Technology Review*. Retrieved from https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/
9. Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of "personally identifiable information." *Communications of the ACM*, *53*(6), 24. https://doi.org/10.1145/1743546.1743558
10. Hintze, Mike, & El Emam, Khaled. (2017). Comparing the Benefits of Pseudonymization and Anonymization Under the GDPR. In *Privacy Analytics White Paper*. International Association of Privacy Professionals. Retrieved from https://iapp.org/media/pdf/resource_center/PA_WP2-Anonymous-pseudonymous-comparison.pdf
11. Porter, C. C. (2008). De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information. *Shidler Journal of Law, Commerce & Technology*, *5*, 1.
12. Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of "personally identifiable information." *Communications of the ACM*, *53*(6), 24. https://doi.org/10.1145/1743546.1743558
13. Lubarsky, Boris. (2017). Re-Identification of "Anonymized" Data. *Georgetown Law Technology Review*. Retrieved from https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/
14. Center, E. P. I. (2019). EPIC - Re-identification. Retrieved February 2, 2019, from https://epic.org/privacy/reidentification/
15. Sweeney, L. (2002). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570. https://doi.org/10.1142/S0218488502001648
16. Acquisti, A., & Gross, R. (2009). Predicting Social Security numbers from public data. *Proceedings of the National Academy of Sciences*, *106*(27), 10975–10980. https://doi.org/10.1073/pnas.0904891106
17. Kondor, D., Hashemian, B., Montjoye, Y. de, & Ratti, C. (2018). Towards matching user mobility traces in large-scale datasets. *IEEE Transactions on Big Data*, 1–1. https://doi.org/10.1109/TBDATA.2018.2871693
18. Lubarsky, Boris. (2017). Re-Identification of "Anonymized" Data. *Georgetown Law Technology Review*. Retrieved from https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/
19. El Emam, Khaled. (2016). A de-identification protocol for open data. In *Privacy Tech*. International Association of Privacy Professionals. Retrieved from https://iapp.org/news/a/a-de-identification-protocol-for-open-data/
20. Narayanan, A., Huey, J., & Felten, E. W. (2016). A Precautionary Approach to Big Data Privacy. In S. Gutwirth, R. Leenes, & P. De Hert (Eds.), *Data Protection on the Move* (Vol. 24, pp. 357–385). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-7376-8_13
21. Bureau, U. C. (n.d.). Decennial Census Datasets. Retrieved April 13, 2019, from https://www.census.gov/programs-surveys/decennial-census/data/datasets.html
22. CDE :: Home. (n.d.). Retrieved April 13, 2019, from https://crime-data-explorer.fr.cloud.gov/