Assignment #8

MACS 30000, Dr. Evans
Due Monday, Dec. 3 at 11:30am

1. **Identification risk in anonymized data (4 points).** This is exercise #11, parts (a) and (b) from Salganik (2018, Ch. 6) with some clarification. In Salganik (2018, Ch. 6), the author proposes a rule of thumb that all data are potentially identifiable and all data are potentially sensitive. Table 1 provides a list of examples of data that have no obviously personally identifying information but that can still be linked to specific people.

   (a) Pick two of the examples in Table 1 and describe in one or two paragraphs how the re-identification attack in both cases has a similar structure.

   (b) In one or two paragraphs, describe how the data could reveal sensitive information about the people in the dataset for each of your two examples in part (a).

Table 1: Examples of Social Data that Do Not have any Obvious Personnaly Identifying Information but can Still be Lined to Specific People, Salganik (2018, Table 6.5)

Data	Reference
Health insurance records	Sweeney (2002)
Credit card transaction data	Montjoye et al. (2015)
Netflix movie rating data	Narayanan and Shmatikov (2008)
Phone call metadata	Mayer et al. (2016)
Search log data	Barbaro and Zeller (August 9, 2006)
Demographic, administrative, and social data about students	Zimmer (2010)

2. **Describing ethical thinking (3 points).** This is exercise #8 from Salganik (2018, Ch. 6) with some clarification. Researchers often struggle to describe their ethical thinking to each other and to the general public. After it was discovered that the Tastes, Ties, and Time study was re-identified, Jason Kauffman, the leader of the research team, made a few public comments about the ethics of the project. Read Zimmer (2010) and then rewrite Kauffman's comments using the principles and ethical frameworks that are described in Salganik (2018, Ch. 6). Specifically, Zimmer (2010) rewrite the following passages from Kauffman (Sep. 30, 2008b) and Kauffman (Sep. 30, 2008c).

   > "Upon the public announcement of this initial discovery, and general criticism of the research teams attempts to protect the privacy of the

> subjects, Jason Kaufman, the principle investigator of the T3 research project, was quick to react, noting that, perhaps in justification for the amount of details released in the dataset, 'Were sociologists, not technologists, so a lot of this is new to us and 'Sociologists generally want to know as much as possible about research subjects."' [Zimmer (2010) citing Kauffman (Sep. 30, 2008b)]
>
> "[Kauffman] then attempts to diffuse some of the implicit privacy concerns with the following comment:
> 'What might hackers want to do with this information, assuming they could crack the data and 'see' these peoples Facebook info? Couldnt they do this just as easily via Facebook itself? Our dataset contains almost no information that isnt on Facebook. (Privacy filters obviously arent much of an obstacle to those who want to get around them.)"' [Zimmer (2010) citing Kauffman (Sep. 30, 2008b)]
>
> "We have not accessed any information not otherwise available on Facebook. We have not interviewed anyone, nor asked them for any information, nor made information about them public (unless, as you all point out, someone goes to the extreme effort of cracking our dataset, which we hope it will be hard to do)." [Kauffman (Sep. 30, 2008c)]

3. **Ethics of Encore (3 points).** Read the Encore web censorship study by Burnett and Feamster (2015) and the reply and critique of that study by Narayanan and Zevenbergen (2015).

   (a) Write a one-half-to-one-page summary of Narayanan's and Zevenbergen's assessment of the Burnett and Feamster (2015) Encore study. Make reference to the consequentialist framework and to the principle of beneficence.

   (b) In one or two paragraphs, write your assessment of the ethical quality of the Burnett and Feamster (2015) Encore study.

References

Barbaro, Michael and Tom Jr. Zeller, "A Face Is Exposed for AOL Searcher No. 4417749," *New York Times*, August 9, 2006.

Burnett, Sam and Nick Feamster, "Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests," 2015.

Kauffman, Jason, "I am the Principle Investigator...," Blog Comment, MichaelZimmer.org, http://www.michaelzimmer.org/2008/09/30/on-the-anonymity-of-the-facebook-dataset/, Sep. 30, 2008b.

_ , "We did not consult...," Blog Comment, MichaelZimmer.org, http://www.michaelzimmer.org/2008/09/30/on-the-anonymity-of-the-facebook-dataset/, Sep. 30, 2008c.

Mayer, Jonathan, Patrick Mutchler, and John C. Mitchell, "Evaluating the Privacy Properties of Telephone Metadata," *Proceedings of the National Academy of Sciences of the USA*, 2016, *113* (20), 5536–5541.

Montjoye, Yves-Alexandre de, Laura Radaelli, Vivek Kumar Singh, and Alex Sandy Pentland, "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata," *Science*, 2015, *347* (6221), 536–539.

Narayanan, Arvind and Bendert Zevenbergen, "No Encore for Encore? Ethical QUestions for Web-based Censorship Measurement," *Technology Science*, December 15 2015.

_ **and Vitaly Shmatikov**, "Robust De-Anonymization of Large Sparse Datasets," 2008.

Salganik, Matthew J., *Bit by Bit: Social Research in the Digital Age*, Princeton University Press, 2018.

Sweeney, Latanya, "K-Anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty Fuziness and Knowledge-Based Systems*, 2002, *10* (5), 557–570.

Zimmer, Michael, "But the Data is Already Public: On the Ethics of Research in Facebook," *Ethics and Information Technology*, 2010, *12* (4), 313–325.