Assignment #8
MACS 30000, Dr. Evans
Due Monday, Dec. 3 at 11:30am
Haowen Shang

1. Identification risk in anonymized data

(a) I picked the "Search log data" (Barbaro and Zeller, August 9, 2006) and "Demographic, administrative and social data about students" (Zimmer, 2010), both of which have a similar structure of re-identification attacks. In these two cases, although the personally identifying information, such as names and identification numbers, were removed from the dataset, the data was still not indeed anonymous. This dataset can expose some sensitive information when linking with other datasets which contain some personally identifying information and share some unique information with this dataset such as age, gender and race. By this linkage, specific individual in this dataset can be re-identified.

(b) AOL released a person's search history over a three-month period. In order to "anonymize" this searcher, the AOL company removed the searcher's name and assigned a unique number, 4417749 to her. However, this searcher was still re-identified by her search history which contains information of "landscapers in Lilburn, Ga", "homes sold in shadow lake subdivision gwinnett county georgia" and the last name Arnold (Barbaro and Zeller, August 9, 2006). Although the information just contains the searcher's address, recent activity and last name, it can be linked with other datasets such as house selling records and demographic data. It's not hard to identify that this searcher was "Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga". (Barbaro and Zeller, August 9, 2006)

For the "Demographic, administrative and social data about students", although students' names and identification numbers were removed, the Facebook profile dataset still has information of students' gender, race, ethnicity, home state, political views and college major (Zimmer, 2010, p.315). It also contains some information of the students' university. For example, the university was private, co-educational, located in northeast American, and whose class of 2009 initially had 1640 students (Zimmer, 2010, p.316). By the above information of the university and the unique majors of the students, the source of the data was identified to be Harvard college. After knowing the student's university, it's not hard to find students' identity and access their privacy information. For example, because some nationalities are only represented by one student from this university (Zimmer, 2010, p.316), it's easy to identify some students using their nationality information.

References:

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

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

2. Describing ethical thinking

Jason Kaufman, the principle investigator of the "Tastes, Ties, and Time" research project, made a few public comments to describe their ethical thinking.

Firstly, he said that he and his colleagues provided such a large dataset because they are sociologists and just want to know "as much as possible about research subjects" [Zimmer (2010) citing Kauffman (Sep. 30, 2008b)]. This comment showed the ethical framework of "Consequentialism" and the ethical principle of "Beneficence" (Salganik, 2018, p.296, p.302). He thought that this project was valuable and its benefits were obvious, because this project contained sufficient data and so many detailed information of students, which could help researchers to understand the nature and dynamic of social networks (Zimmer, 2010, p.314). He also believed his team had already mitigated the risks, because they did many primary steps to protect students' privacy, such as deleting or encoding the identifying information (Zimmer, 2010, p.317). However, he didn't make an appropriate balance of the benefits and risks, because while the beneficial outcomes are obvious, some potential risks are underestimated. By releasing such a large amount of data, this project also released so many students' privacy information.

Secondly, he mentioned that if the hackers want to get the students' information released by the data, they could also easily get them from Facebook itself. This comment also showed the ethical framework of consequentialism (Salganik, 2018, p.302). He believed that using the data they released had the same consequence with using the information directly from Facebook, because the data they gathered was already public available (Zimmer, 2010, p.321). Thus, he believed that releasing this dataset didn't increase the risks of hurting student's privacy. However, this project used an "in-network research assistants" to collect data, who were in the network with students and could collected data which only visible by people at Harvard (Zimmer, 2010, p.320). Thus, some data was not truly public available and was much more difficult to get in a public square (Zimmer, 2010, p.321). By releasing the students' data from a specific university, he violated the principle of "Justice", because students from Harvard college "bore the burdens of the research and society as a whole benefited" (Salganik, 2018, p.299).

Thirdly, he said that they got all information from Facebook and they didn't contact students to interview or ask for additional information [Kauffman (Sep. 30, 2008c)]. He thought that this project followed ethical principal of "Respect for Persons" (Salganik, 2018, p.295), because it just used public available information and didn't interview or ask anyone about sensitive information. He also believed the project followed the framework of deontology (Salganik, 2018, p.302), because this project respected students' privacy during the process of collecting data. However, student's Facebook profile was intended to be used in "social networking among friends and colleagues", but not to be used as "fodder for academic research" (Zimmer, 2010, p.322). They didn't get any consent from students to use the data into research or other secondary uses. Also, they didn't follow the ''terms and conditions for use'' of Facebook when collecting data. Thus, without any consent, they didn't follow the principle of "Respect for Persons" and "Respect for Law and Public Interest" (Salganik, 2018, p.295, p.299).

References:

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

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

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.

3. Ethics of Encore

(a) Encore study is a project to measure censorship by making people’s browser download and execute a piece of code without their awareness, which raised some ethical discussion (Narayanan and Zevenbergen, 2015, p.1-2). Narayanan and Zevenbergen (2015) followed Menlo Report and made an assessment of this Encore study. From the view of “Consequentialism” framework and the “Beneficence” principle, they think Encore study “generates significant positive benefits with some potential harms that can be mitigated” (Narayanan and Zevenbergen, 2015, p.1)

Firstly, the benefits for this study is obvious. In the field of computer science, measuring censorship from diverse vantage points is complex and hard. Encore creates an easy way to measure censorship through a global scale and illuminates the technologies behind censorship (Narayanan and Zevenbergen, 2015, p.6, p.11). Also, it helps researchers in other fields to understand some questions about internet censorship such as the motivation of the censorship, and it provides “a geographically fine-grained view of measurement” (Narayanan and Zevenbergen, 2015, p.6). Understanding technologies and motivations can help researchers to create much more effective tools of censorship circumvention (Narayanan and Zevenbergen, 2015, p.11).

However, the risks of this study are complicated and difficult to measure. Individual stakeholder analysis is impossible for Encore study because of its scalability and involvement of millions of computers (Narayanan and Zevenbergen, 2015, p.9). Although whether Encore is a human-subject research is still under debate, the internet users can experience harms and repercussions because of this study, especially for those who “live in a regime without due process” (Narayanan and Zevenbergen, 2015, p.7, p.10). The researchers who conducted this project also argue that compared with normal web browsing, users face the same risk in Encore study, so they believe users just face the “minimal risk” (Narayanan and Zevenbergen, 2015, p.7, p.13). Comparing the obvious benefits with the minimal risks, the researchers think they followed the principle of “Beneficence”. However, Encore and much third-party tracking today indeed flout the users' expectations of protecting privacy, and researchers are not supposed to "participate in and facilitate an ethical race to the bottom" (Narayanan and Zevenbergen, 2015, p.13). Also, the Encore researchers think that it's difficult to measure the harm of Encore study because the diverse types of censored website and different reasons the website was censored, but they believe that "the more widespread measurements like Encore become, the less risky they are for users" (Narayanan and Zevenbergen, 2015, p.13-14)

Finally, the authors discussed some methods to mitigate the harm. Obtaining informed consent would mitigate the harm but it's impractical and would impair "the novel measurement architecture" and "increase risk to users by removing plausible deniability" (Narayanan and Zevenbergen, 2015, p.14). Increasing transparency by giving notice to inform website visitors and explain the risks and benefits of the Encore research is also a way to mitigate the harm (Narayanan and Zevenbergen, 2015, p.15). Although a global study of internet law is impossible, the researchers should "accept responsibility for their actions and the consequences, and have the necessary mitigation strategies in place" (Narayanan and Zevenbergen, 2015, p.16).

(b) I want to make an assessment of Encore study from the view of four ethical principles. Firstly, Encore study doesn't follow the principle of "Respect for persons", because researchers didn't get consent from participants --participants were involved in this project without awareness. Although it may not be considered as a human subject research, it indeed put users in risks when they live in a regime that forbids them to visit some censored website, and I think most users are unlikely to consent to Encore's measurements when they are informed. Thus, they didn't respect the privacy and autonomy of participants. Also, it violates the principle of "Justice", because they put people who live in the country with repressive governments face much more risks but benefit other people in a worldwide scale. Whether the project follows the principle of "Respect for Law and Public Interest" is uncertain, because the Encore project was conducted in a global scale, and a global study of Internet law is impractical. From the view of "Beneficence", I agree with the assessment by Narayanan and Zevenbergen (2015). Balancing the benefits and risks for this project is complicated because the benefits of this project are significant, while the potential risks are difficult to measure.

References:

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

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