Assignment 8

Ellen Hsieh

1. Identification risk in anonymized data

(a)

The re-identification attack in both the health insurance records case (Sweeney 2002) and the case of revealing social data about students (Zimmer 2010) has a similar structure. First, the dataset released by the researchers still contains sensitive identification attributes like ZIP code and birth date, even though the researchers already removed certain critical characteristics that might be easy to identify the person such as name and ID. Second, it is too easy to obtain other datasets that might be able to link to the main dataset to reveal the identification of the participants. For instance, in the health insurance record case, Sweeney (2002) only spent twenty dollars to get the voter registration list for Cambridge Massachusetts. Also, in Harvard College case, the codebook is even downloadable without submitting any application. Therefore, based on the two characteristics mentioned above, it shows that these two attacks share a similar structure and demonstrates that re-identification is not as hard as the researchers think.

(b)

For the first case, the health insurance record studied by Sweeney (2002), the medical record certainly contains a lot of private information about the individuals. For instance, the ethnicity might imply the socioeconomic status; the diagnosis can indicate the health status of one person and what things could harm his or her health according to the disease; the total charge might imply the income of the individual since the rich people are able to spend more on sustaining their health, on the other hand, the poor might not be able to spend that much. As for the case of revealing the demographic, administrative, and social data about students, the dataset includes many sensitive attributes such as hometown state, race, ethnicity, and college major, which are really unique and not too hard to be personally identified using external sources like codebook or reference comments. What's more, from those external sources, people can learn those specific individuals' political views or sexual preference, which are significantly private.

2. Describing ethical thinking

‘‘We’re 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’’

We’re sociologists, the people who dedicate their life to better understand the behavior of human being and the society in order to better off the development of this world. Therefore, using those online data is still new to us, we are not technologists who are familiar with how to manipulate the given data and the perfectly encode every private information. What we want to do with the data is to know as much as possible about research subjects in order to better off the society.

What might hackers want to do with this information, assuming they could crack the data and ‘see’ these people’s Facebook info? Couldn’t they do this just as easily via Facebook itself?

If hackers want to obtain this private information and do something bad with it, then why not just go directly to Facebook page and collect them? We already got approval and permission from Facebook and IRB to collect those data and do our research, and we believe that this research can help us to understand more about human behavior, thereby advancing this society.

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).

To complete our research, we only need the information that are available on Facebook. There is nothing else more to collect directly from the students. Therefore, it makes no sense for us to be blamed for using this information since it is publicly posted on Facebook. Also, our goal for the study is to observe the behavior of those students and how the social network changes over time.

Even though Kaufman insisted that his research group already properly protected the privacy while releasing the data, they still violated the four principles mentioned by Salganik (2018) while facing ethical uncertainty. To begin with, the principle “Respect for Persons” points out the importance of treating people as autonomous (Salganik 2018, p.295). In Kaufman’s

research, he did not ask for any consents from the students being observed. That is to say, the participants in this study were not aware of this, and they might not be willing to be in part of it.

The second principle is “Beneficence”, which is about not to harm any participants and maximizing the possible benefits and minimizing the possible harms. (Salganik 2018, p.296) However, in order to achieve Kaufman’s goal, he ignored the possible harm to his research objects by releasing the dataset that is easy to link to other external sources, which enables the people to learn the participants’ private and sensitive information. Making the sensitive information public can cause troubles for those students or make them feel uncomfortable.

Next is the principle of “Justice”, which is about ensuring the distribution of risks and benefits is fair. (Salganik 2018, p.298) Apparently, in this study, only the student bore the burden of the research. Kaufman should think of some way to compensate the participants in order to better strike the balance between the burdens and benefits of his research.

Lastly, “Respect for Law and Public Interest” which encourages the researchers to include law in their consideration. (Salganik 2018, p.299) Even though Kaufman’s research already got approval from IRB and his goal and methods were pretty clear, there are still more for him to consider. One way to do is to hire some lawyers to help him reviewing his work and to make sure that his action will not violate any law that protect the participants or for the good of the society.

It is obvious that Kaufman’s work emphasized more on consequentialism than deontology. Even though he insisted that he already tried his best to protect those research subjects, apparently, he underestimated the power of the technology and the data. In order to achieve his goal to observe the change of the social network, it is clever to use the social media like Facebook to analyze how people interact. Nevertheless, he ignored the possible ethical problem of using the online data and releasing the data to other researchers.

3. Ethics of Encore

(a)

In Narayanan’s and Zevenbergen’s paper (2015), it talks about the Encore study by Burnett and Feamster (2015). Encore is a study that focuses on the web censorship across different countries. Its purpose is to study the behavior of censorship systems. However, there are some ethnic concerns for this research.

First, the researchers are usually required to know who the stakeholders are, but in this case, in order to maximize the scalability, it seems infeasible to follow this rule. Second, is Encore a human-subjects research? The authors of Encore consider themselves as conducting a research on technical system. Nevertheless, the Internet is usually understood as a sociotechnical system rather than a simply technical system (Narayanan and Zevenbergen 2015, p.11). In addition, whether IP addresses are regarded as PII is still an ongoing debate. Even though some proponents of viewing the computer security research as human-subjects research propose that such researches need human being to participate to obtain the useful results, Narayanan and Zevenbergen points out if the research replaces human with robots, the research could be better off since the biases in the measurement would be minimized (Narayanan and Zevenbergen 2015, p.13).

Another important guiding principle of research ethnics is "Beneficence", which is to minimize the risks and maximize the benefits. However, apparently, it is hard for Encore to fully adhere to this principle due to its global scale. For instance, it is difficult for Encore to define the harms for each Internet user. Also, the censored contents differ from country to country due to the different motivation behind the censorship. Even though Encore considers itself only presenting the minimal risk, it might still increase its users' risk by exposing the users to surveillance agencies (Narayanan and Zevenbergen 2015, p.16).

Although seeking informed consent from users might mitigate the harms, doing so might decrease the capability of the research. Instead, Encore contains a statement and an opt-out option for the users at the bottom of the website. Undeniably, Encore is a fascinating study that provide us more insight in the behavior of web censorship. However, when it comes to the issue of research ethnics, it is still controversial that whether the way how Encore collects its data is justified.

(b)

Even though the authors of Encore study claim that it is a study focuses on the web censorship, undeniably, there are still some people involved so that the study can be completed. Therefore, Encore shouldn't be regarded as a non human-subjects research. To consider the study with the four basic principle for research ethnics, it seems that Encore violates most of them.

Firstly, Encore does not properly inform the users, even though it contains a statement at the bottom of the website. It might be not so easy for users to notice even if they scroll down the website. Second, the risks and the benefits are not justified since Encore does harm its users to some degree, which depends on the country or region the users are. For example, in some conservative or communist countries, the consequence of being caught by browsing censored website could be extremely serious such as considering it as a crime.

What's more, the burdens and benefits are not equally distributed in the research. Only the people that are randomly picked by Encore need to bear the potential risks but the benefit of Encore study are to everyone in the society. In order to reach its goal, to maximize the scalability, the study Encore neglects some critical ethnic issues, it puts more emphasis on consequentialism. Nevertheless, the authors of Encore should also value more of the deontology for their research.

References

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

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

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.