1.

(a).

In this question, I'd like to pick Sweeney's[1] and Zimmer's[2] data for further description. From their paper, we could clearly see that their re-identification processes have high similarity. They both used an anonymized dataset, which contains sensitive information. Then, they tried to link this dataset with existing information to interpret it. The existing information could be another dataset with personal identification or simply the codebook of the anonymized dataset. Even though a single Zip code can't reveal anything, the combination of several information like Zip code, birthday, race, sex and so on in a same row is clear enough to point out a specific person with existing information. Thus, in this way, the author did their re-identification attack.

(b).

To be specific, in Sweeney's paper, she used the health insurance records which includes "patient's ZIP code, birth date, gender, and ethnicity", but without patients name. Nevertheless, she purchased another dataset about voter's information. This dataset not only has voter's name, but it also includes "address, ZIP code, birth date, and gender of each voter." Thus, she could link these two dataset by "ZIP code, birth date, and gender"[3] and finally get a rather comprehensive information about a person. She showed the power of this linkage by even point out a very famous person called "William Weld, a governor of Massachusetts"[4].

In Zimmer's paper, he mainly talked about the ''Tastes, Ties, and Time'' project. The data of this project includes "demographic, relational, and cultural information on each subject"[5], to be specific, like "each subjects' gender, race, ethnicity, hometown state,

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

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

[3] Sweeney's paper, page 2 of that paper, Example 1.Re-identification by linking

[4] Sweeney's paper, page 3 of that paper, Example 1.Re-identification by linking

[5] Zimmer's paper, P314, The ''Tastes, Ties, and Time'' project

and major".[6] Just like what I have said in question (a), even though a single information like gender points to nobody, the combination of all of this information is clear enough to make identification. With the help of the codebook, we could even know that there are some unique students from specific states like "Delaware, Louisiana, Mississippi, Montana, and Wyoming" or from specific countries like "Albanian, Hungarian, Iranian, Malaysian, Nepali, Philippi no, and Romanian", which makes it even easier for us to identify a person's identity. Therefore, "the privacy protection made by the T3 project was far more than sufficient by purely remove the name and identification number of the student or a delay of release"[7].

2.

(1).

In his comments of first part of 2008b, Kauffman tried to show that his team strictly followed "the principle of beneficence[8] and the consequentialism framework[9]." By saying that "Sociologists generally want to know as much as possible about research subjects"[10], he tried to convey that they just want to "maximize possible benefits"[11] of publishing this comprehensive dataset. However, I don't think "sociologist is not technologist" is a good excuse of failing to comply with another important part of beneficence that the data "should not harm" the subjects. Even though we could find that, from his words, they had tried their best to protect the privacy of those subjects, which followed the rule of consequentialism framework ("balancing risk and benefits"), it's not quite enough to hide the privacy information sufficiently.

In the second part of Kauffman's 2008b comments, he tried to show that his team also

---

[6] Zimmer's paper, P319, The insufficiency of privacy protections in the T3 project

[7] Zimmer's paper, P315, Partial re-identification and withdrawal of dataset

[8] Salganik, Matthew J., Bit by Bit: Social Research in the Digital Age, Princeton University Press, 2018, Chapter 6.4.2

[9] Bit by Bit, Chapter 6.5

[10] 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.

[11] Bit by Bit, Chapter 6.4.2

followed “the principle of justice”[12]. By showing that “they didn’t add any information besides the Facebook, which is already public on the internet”[13], Kauffman actually argued that their dataset didn’t increase the risk of privacy disclosure for those subjects, which is unfair for them.

As for the 2008c, Kauffman tried to show that his team followed “the principle of respect”[14]. By saying that “they only used the data on the Facebook page and never reach out those subjects”[15], he argued that the information they used was published by those subjects themselves, thus, didn’t against their wills to use additional privacy information and definitely respect those subjects.

3.

(a).

In their paper, Narayanan and Zevenbergen recognized the improvement and innovation of “measuring Internet filtering and censorship worldwide”[16] by the Encore system. However, authors also used “the Belmont Report and the Menlo Report”[17] to analyze the ethnical problems faced by this system.

To start with, to make a comprehensive judgement of Encore, authors must clarify its stakeholders. However, this task is rather impracticable that the Encore system involved “personal users, governments, third parties, and so on and their user IPs cannot be used directly.”[18] What’s more, the question that “whether the subjects of Encore is human or not”[19] is inconclusive. Even though the Encore system only collected the IP address,

---

[12] Bit by Bit, Chapter 6.4.3

[13] 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.

[14] Bit by Bit, Chapter 6.4.1

[15] Kauffman, Jason, “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.

[16] Narayanan, Arvind and Bendert Zevenbergen, “No Encore for Encore? Ethical Questions for Web-based Censorship Measurement," Technology Science, December 15 2015. Page 2, Abstract

[17] Narayanan, Arvind and Bendert Zevenbergen’s paper, Page 8, Methods

[18] Narayanan, Arvind and Bendert Zevenbergen’s paper, Page 9, Analysis

[19] Narayanan, Arvind and Bendert Zevenbergen’s paper, Page 10, Is Encore human-subjects research?

which is purely an index for machine, we could still use it to make a link to the user of that machine who finally was involved in this project. Thus, we cannot guarantee that the Encore system only focuses on technical index and have no interact on the user of Internet.

Then, authors "used the principle of beneficence to identify the benefits and risks of Encore system"[20]. For benefits, it is obvious that "it helped to illuminate censorship—both its motivations and the technologies behind it". Nevertheless, the censorship itself is a rather controversial topic. As for the harms, Burnett and Feamster said that "normal web browsing exposes users to the same risks that Encore does, saying "the prevalence of malware and third-party trackers itself lends credibility to the argument that a user cannot reasonably control the traffic that their devices send"[21]. However, Narayanan and Zevenbergen view this as "ethical race to the bottom that credentialed researchers and respected academic organizations should not participate in and facilitate".

For the use of consequentialist framework, considering that "it highlights the balancing risk and benefits"[22], we could find the counterpart that the Encore system "tried to mitigate the harm by limited the set of URLs that the script induced users to measure". However, authors also stated that there are still many things need to be done to lower the harm and risk.

(b)

To start with, it is quite obvious that the team of Encore violated the principle of respect. They never "asked for the subject's consent" when they are collecting their data. What's more, the team didn't follow the principle of justice. Subjects of Encore system were exposed to higher risk that they could even violate the law of their country if the worst case happens. However, the team tried to obey the principle of beneficence that they

[20] Narayanan, Arvind and Bendert Zevenbergen's paper, Page 11, Identifying Potential Benefits and Harms
[21] Narayanan, Arvind and Bendert Zevenbergen's paper, Page 13, Harm: does Encore present more than minimal risk?
[22] Bit by Bit, Chapter 6.5

indeed tried to “mitigate the harm by limiting the set of URLs. In a word, we should admit that the Encore system is indeed helpful for the study of censorship, but the ethical problems it faced also deserve more attention.

Reference

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.

Salganik, Matthew J., Bit by Bit: Social Research in the Digital Age, Princeton University Press, 2018, Chapter 6.

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.

Kauffman, Jason, “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.

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.