Reading Response 2/18

Kritika Chugh
SUID: 882046659

kchugh@syr.edu

Privacy and sparse data – How much can we know about an individual from the data available to a keen eye.

Sometimes we visit hospitals and give our information, mode of payment and previous medical records and have no idea how is that information used. Specifically, in public hospitals they make this data public and, in most cases, freely available on the government websites. In one of the papers, I saw how much it is easy to match a patient's name with the publicly available data even when the original data has no names in the first place. What is ethically wrong here is that most people are unaware of these data sets and in other cases such as Netflix rating data they might not even care.

But data community should take care and educate as these data produces surprising revelations about a target and in most cases when harm is done to a target, they might not even be able to link that harm (say hate crime or snooping) back to the shared data. Some people might even say these data sets made public are for academic purposes but in many cases having these data made public can have problems such as employers doing a background check on the health and demographics of an individual (a kind of a bias), finding credit worthiness, data mining companies making personal products and minting profits and even more personal such as friends or family snooping on each other. Because let's face it re-identification is not all that difficult amidst all these data freely available or can be purchases as low as dollar 75 for a year plan.

Furthermore, combining automated process with the human identification is like double trouble as the later is more probabilistic in approach. Example Netflix now only shares ratings and movie names. There is no other auxiliary information available to see. But given how little auxiliary information is required to de- anonymize the individual one should be still careful as partial de-anonymization can be dangerous.

The solution does not call for putting a stop at sharing data, but to be smart about how and what we share because in most cases the top buyers are the private companies and non-researchers. From the Netflix dataset we saw how much we can know about the movie history of an individual and this information can reveal the political opinions, biased inferences about someone. The right thing to do is to give power to the individual on how much they want their data to be revealed not the company itself.