Identifying Risk in Anonymized Data

Among the various research projects discussed in Table 1, the social data described in Zimmer (2010) and Barbaro and Zeller (2006) are cases where seemingly harmless data can have very harmful consequences. In both the cases, the *intention* was not to cause any re-identification or for the data to be easily identifiable, yet it was fairly trivial to crack the database.

The data in Zimmer (2010) involved evolution of Facebook data over a span of 4 years, and in Barbaro and Zeller (2006) it was AOL Search Log data. In both cases there was no need for sophisticated engineering to identify the dataset - just by using sensitive data *already* present in the databases it was possible to narrow down the set of people who are represented by a data point and then start guessing with a high level of accuracy who the person is. In the AOL search data for example, it was possible to narrow down the location and name of an individual by examining the search patterns. In the Facebook data it was possible to identify which University was being talked about by examining the class size and the courses offered. A little bit of careful guesswork and with the process of elimination it is scary how easy it is to find individuals in the dataset.

As for the kind of information which can possibly be revealed, in both cases it is very personal information which can be used in multiple ways to harass the person identified. Being able to access anybody's search history means you have access to the mental state of that human being - this information can be used to threaten the individual, or gain leverage over the person in social interactions. What is meant to be personal information is, without consent, made public and thought to be free game - this means it is possible to even look for people who are in the same city or region as you by filtering the searches in the logs.

The Facebook data described in Zimmer (2010) is just as harmful if not more; by allowing the University name to be easily identifiable, and by also retaining data on the nationality of the individual, it becomes trivial to identify certain individuals. Once someone is identified, it is possible to identify a digital stamp based on the social media circle, the tastes and activities performed on Facebook. Once again, this data being at the mercy of the public opens up various avenues of harassment.

By arming those who are determined with a trove of data about the online habits of individuals, both research projects and data set releases had conducted huge ethical faux passes.