Summary of Discussion for
How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems
by Bradley Malin and Latanya Sweeney
Journal of Biomedical Informatics 37 (2004) 179–192
http://dx.doi.org.libproxy.lib.unc.edu/10.1016/j.jbi.2004.04.005

Presented by Noel Fiser
INLS279: Bioinformatics Research Review
February 15, 2006

The discussion presented in this paper relates to the protection of personal information in a distributed network (e.g., the Internet). Specifically, the discussion focuses on the protection of personal identification data in conjunction with personal genomic data and the need to *strongly de-couple* these two data sets.

Unfortunately, both of these data sets exist in quasi-public domains: limited personal information released in hospital admission records, and genomic information released for research purposes. In order to raise awareness of how this information can be re-connected, the authors present a computational model of how to re-identify genomic data based on patients visiting more than one location. They selected on a small subset of the data from public records of the state of Illinois (1990-1997), focusing on eight rare diseases.

This re-identification algorithm, which they label REIDIT, comes in two flavors: complete and incomplete, based on the completeness of the population data. The “complete” algorithm (REIDIT-C) assumes a complete data set and simply curses through each genomic record and tries to associate it with any personal information in the system. Based on the fact that the information is complete, this correlation becomes trivial: it is just a matter of matching up the genomic data when and where it is released to the personal data as it is recorded at the multiple hospitals the patient visited.

The situation where all sets of data are released and accessible is, however, very unlikely, and thus a level of complexity must be accounted for when there are gaps in the data. In this case the “incomplete” algorithm (REIDIT-I) is applicable. This algorithm is very similar to REIDIT-C; it just includes an extra pass over the data in order to try to compartmentalize these unmatchable data points.

In their experiment, Malin and Sweeney were able to use these algorithms to successfully re-identify the various patient groups at different rates, ranging from 32.9% of the cystic fibrosis patients up to 100% of the Refsum’s disease patients. These results were particularly striking in indicating a direct relationship between re-identification and the number of patients visiting a hospital. Not surprisingly but still made starkly clear, the more patients that visit a hospital, the less likely those

patients can be re-identified with these algorithms. Or conversely, the fewer patients that visit a hospital, the easier it is to correlate those patients' genomic and personal data.

In order to better disassociate these two data groups, the authors look at two different solutions, the latter of which they support. deMoor proposes a central repository solution maintained by a trusted third party (AKA an "honest broker" arrangement). They point out that the flaw of deMoor's solution is that one can still "trail" patients because it maintains the number of locations visited instead of the exact names of these locations. As a result we can correlate the number of visits in one data set with that number in another and if these are unique, then you have made a successful re-identification. The solution proposed by deCODE Genetics is embraced by the authors because it only associates personal patient information with the *first* location (hospital) that they visit, thus disabling this entire set of algorithms since they rely on the correlation between two data matrices with multiple patients and multiple locations.

While the class appreciated the completeness of the argument made in the article, we did have a few unanswered questions and concerns about the paper. First, we questioned just how much of this data was actually being released into the public domain. While it was surprising to see such data from the State of Illinois (and how frighteningly accessible it is), this seems to be a rare situation where such genomic and personal data on the same population are made so wide open.

We also noted that the true scope of such research is limited to a very small subset of even those populations for which you can get such data. These experiments were conducted only on known rare diseases, largely as a proof of concept. But as the study indicates, while the concepts are applicable to any genetic sequence of the population that you have access to, as the subgroup on which you are focusing gets bigger and as the number of patients visiting similar locations also gets bigger, re-identification becomes less effective. Thus, while these rare diseases are exposed, the great majority of the population will not be equally identifiable without looking for longer and more specific genetic markers. Still, we fully recognized that any such possibility of re-identification does indicate a flaw in the data storage and release process that should be actively addressed. We discussed the proposed solutions involving a third-party, "honest broker" system and agreed with the authors that this method is indeed the best way to protect personal and genomic data as they become more and more ubiquitous in the modern medical world.