GENETICS, SOFTWARE

ANONYMIZING GENETIC DATA

13.12.2007

I have currently a paper under submission at the EJHG that covers ethical issues of genetic testing. One of the key messages is that genetic data are not anonymous if having simply stripped of names.
A story in a completely different field confirms my fears. According to a NYT article

> Last October, Netflix, the online movie rental service, announced that it would award $1 million to the first person or team who can devise a system that is 10 percent more accurate than the companyâ€™s current system for recommending movies that customers would like.

but things turned worse by an article of Narayanan und Shmatikov available at arxiv.org

> We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge. We apply our de-anonymization methodology to the Netflix Prize dataset ...

Basically, they use a scoring function to assign a numerical score to each record in the master dataset based on how well it matches the adversaryâ€™s auxiliary information. The matching criterion is the algorithm that the adversary applies to the set of scores assigned by the scoring function to determine if there is a match. The ultimate record selection is done by a â€œbest-guessâ€ based on the record or a probability distribution, if needed. As far as I can see, this is a rather clever approach, even better than what I am describing in my article.
Think of an individual (with "wet earwax" and even a few other strange characteristics) and I will be able to match his/her online genetic dataset. Yea, yea.