Computational Social Science Use Case
Summary Description

Computational social science is rapidly changing the study of humans, human behavior and human institutions. Initially the most visible aspect of this change was the advancement of statistical and data analytic methods. It is now clear that, taken as a whole, the evidence base of social science is shifting [1,2,3].

The characteristics of "big data" and "big analysis" used in this next wave of social science research is vast, and often integrates data from multiple sources. Millions of political blogs offer a window into public opinion, at the same time as the response rates on large traditional public opinion surveys are plummeting, endangering the latter's validity. Social scientists are rapidly developing methods capable of tapping these new data sources. Cutting-edge social science research now incorporates the analysis of social network data, harvested textual data, professional and amateur video content, and fine-grained geospatial data including the traces of movements in space and time. Cell phones and other portable devices are beginning to offer the opportunity to collect continuously sampled, geospatially and temporally identified records of people's lives; virtual worlds like Second Life give us the ability to record every aspect of subjects interactions within them for later analysis; and massively open online courses are enabling interaction with learning objects and other learners to be instrumented in previously unimagined detail.

The range of methodologies in use is equally broad. While descriptive statistics, contingency tables and regression models remain common in social science publications, the most influential research, and research engaging with new forms of evidence, and published in leading social science methods journals [4] increasingly employs Markov Chain Monte Carlo and other Bayesian computation methods; supervised and unsupervised heuristic clustering and other data mining methods; statistical methods for topic, stylographic, and sentiment analysis of text; and dynamic visualization.

Traditional disclosure limitation was based on ad-hoc deidentification of data – typically through suppression of extreme values and generalization of measurements. However, new forms of evidence also pose new difficulties for confidentiality. The general problem is that these new forms of information tend to have more detail, richer structure, and quite different distributional properties from "traditional data". This makes it difficult to predict and ameliorate the risks to confidentiality associated with release of the data.

The scope of the consent of the individuals being studied using new data sources may be unclear. One problem is that researchers may seek to use information collected by a commercial service according to terms of service and privacy policies that may or may not disclose third-party research uses. Often, future uses of the data are not apparent at the time of collection, when notice and consent may be given, and data may be collected over a long period of time under evolving terms of service. The examples below provide further discussion on a number of the types of problems encountered with new data sources.

References:

[1] Altman, M., Rogerson, K., & U, D. (2008). Open Research Questions on Information and Technology in Global and Domestic Politics – Beyond "E-." PS Political Science and Politics, 41(4), 1-8.
[2] Lazer, David, et al. "Life in the network: the coming age of computational social science." Science (New York, NY) 323.5915 (2009): 721.
[3] King, Gary. 2011. Ensuring the Data Rich Future of the Social Sciences. Science 331, no. 11 February: 719-721.
[4] See for example, *Political Analysis; Sociological Methods & Research; Geographical Analysis; Econometrica*

Examples

1. The "Netflix Problem" illustrates the difficulty of anonymizing data in which the measures have unusual distributional properties. Movie viewing has a long tail – while most people have seen movies that are very popular, each of us also watches a large number of movies that are not popular. Moreover, individuals tend to rate these movies differently. As a result, a short list of unpopular movies, or an even shorter ranking of them is unique to an individual.

   Narayanan and Shmatikov were able to use these distributional properties (high-dimensionality and sparsity) to identify individual users in the Netflix database. This database was distributed in nominally anonymized form as a research challenge. Although the database did not contain names, Narayanan and Shmatikov had, for some users, been able to obtain a very small subset of their rankings of movies from another source, using this fragment of rankings, they were able to positively identify the entire ranking history of these individuals in the Netflix database.

   See: Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." Security and Privacy, 2008. SP 2008. IEEE Symposium on. IEEE, 2008.

2. The "GPS Problem" illustrates the difficulty of rich locational data. When data is collected continuously over a long period, it is difficult to effectively mask with current statistical techniques which add noise / aggregation to individual observations. Although an individual observation of a position is easy to perturb (such that it is indistinguishable from other observation) and unlikely to positively identify an individual if discovered, a full trace of an individual's movements through time is hard to disguise and more likely to be positively identifying. For example, if we aggregated all positional information to zip code (greatly decreasing the utility of the data) we could not identify any individuals based on a single observation. But if we knew for each subject which zipcode they were in every hour or even every day, we could positively identify many people if supplied with a small portion of their business or recreational travel

histories with enough precision as external information.

See: D.L. Zimmerman, C. Pavlik , 2008. "Quantifying the Effects of Mask Metadata, Disclosure and Multiple Releases on the Confidentiality of Geographically Masked Health Data", *Geographical Analysis* 40: 52-76

3. The "Blog problem" illustrates how rich data sources may carry unexpected identifying information. Pseudonymous blog postings, and other form of commentary on the web are increasingly being analyzed using content and sentiment analysis. But although the author's name may be a pseudonym, or completely redacted, the author's identity can sometimes be recovered through textual analysis of writing style (stylometry).

   See: Faresi, Ahmed Al, Ahmed Alazzawe, and Anis Alazzawe. "PRIVACY LEAKAGE IN HEALTH SOCIAL NETWORKS." Computational Intelligence (2013).

4. The "Facebook Problem" illustrates the difficulty of anonymizing data that is highly structured. In data collected from Facebook, information resides not only in the characteristics of Facebook pages and messages, but also in the network structure induced. For example, information about an individual such as sexual orientation, can be predicted accurately using only information about their friends; thus, having information about the network of relationships and characteristics of some users can reveal (probabilistically) information about other users. Furthermore, the complexity of the systems privacy controls can make it challenging to determine what information subjects have intentionally made public. Moreover, even if all of the characteristic information is stripped, it is possible to identify a small group of users (such as a person and a subset of their friends) by examining the structure of the Facebook graph. Or if only a few nodes are controlled by an adversary and are included in the subset of data, this may also be used to identify the user.

   See: Jernigan, Carter, and Behram FT Mistree. "Gaydar: Facebook friendships expose sexual orientation." *First Monday* 14.10 (2009); L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore Art Thou R3579X?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography," in 16th Conference on World Wide Web (WWW'07), 2007.; Zimmer, Michael. ""But the data is already public": On the ethics of research in Facebook." Ethics and Information Technology 12.4 (2010): 313-325.

5. The "Hubway Problem" illustrates the difficulty of anonymizing data that can be linked to external data sources. In such cases, information in the anonymized data can be used as a key for matching it with external, nominal data, thus re-identifying the data subjects. For example, Boston's bike sharing initiative, "Hubway", released an anonymized data set about its member's rides. Sweeney et al. were able to cross-reference the ride data with tweets, Foursquare check-ins and voter registration data, to re-identify some of Hubway members. The project created a web site, showing members pick-up and drop-off locations, and a Hubway-related Twitter feed.

See: http://www.aboutmyride.org/more.html