De-identification: A Potential Solution to Protecting Personally Identifiable Information in Marketing Research?

By Howard Fienberg, PLC

De-identification, those oblique techniques that seemingly render personally identifiable information (PII) harmless, could be the linchpin to any overarching federal data privacy law that would avoid ruining the research industry. It is certainly a big reason why research survived the passage of the HIPAA Privacy Rule.

However, the devil will be in the details. As MRA observed and discussed with numerous experts at a recent all-day privacy conference, the privacy community is divided and confused on some fundamental principles, not to mention the gradations necessary to coalesce around in any useful de-identification standard in data privacy law.

Certain terms raise more combativeness from privacy experts than others, and researchers should take note since many in the profession throw these terms around in blithe fashion (especially in their privacy policies and promises made to respondents). Anonymity? No one seems to believe in the concept anymore. Confidentiality? It may only perfectly apply to data that's been shredded beyond recognition.

The Future of Privacy Forum's "Personal Information" conference on December 5 featured a wide variety of speakers and attendees discussing and debating the de-identification, re-identification, and the definition of personal information.

Peter Swire, Professor of Law at Ohio State University and the person who designed the Privacy Rule when he worked in the Clinton Administration, opened the conference with a look at federal statistics law. He cited endless years of federal experience with the decennial Census, resulting in highly useful de-identified data. The promise of confidentiality – and the legal rules to try to back that up – has been central to the Bureau's ability to conduct the decennial and to use and share resulting data. Such promises were elaborated and codified for federal statistical agencies in the Confidential Information Protection & Statistical Efficiency Act of 2002 (CIPSEA). The basic rule is that if you collect data for statistical purposes, you can only use it for statistical purposes and you can't re-identify it. Of course, that culture of practice has existed for years outside of government, in survey and opinion research of all sorts, and in the Code of Ethics for MRA members.

The HIPAA Privacy Rule developed a safe harbor for de-identified data, as well as provisions for data use agreements stipulating that the data could be shared for research purposes as long as it was not re-identified. The end goal was to allow data to be shared publicly if it had been sufficiently scrubbed; if insufficiently scrubbed, the party receiving the data would be subject to an enforceable contract not to re-identify the data.

However, the Internet changed the game tremendously, since data mining rapidly progressed far beyond the province of specialized researchers. This evolution led to an erosion of practical obscurity: the amorphous "they" may now be able to figure out who "we" are.

The benefits of public data are clearly big, so what are the practical risks of harm from potential re-identification of de-identified data? Professor Swire highlighted three different threats:

1. Insiders (e.g., employees or subcontractors) "peeping" at data they shouldn't see, such as George Clooney's doctor records, the online dating info for their spouse, or data that could be (and then is) used for criminal harm (e.g., identity theft).
2. Outsiders hacking into a system;
3. And, the general public.
4. De-identification can be pretty effective for the first two threats (an employee would be unable to search for or accidently find Clooney's doctor records and a hacker would download a huge cache of seemingly meaningless statistical data), but there was a lot of debate at the conference over how much defense it can provide against the public.

Professor Swire asked, "What if... everything can be re-identified?" That

was certainly the case made by Latanya Sweeney, Director of the Data Privacy Lab at Harvard University, based on a pair of her studies. But as several other researchers pointed out in response, the two biggest re-identification studies were very limited cases and not generalizable.

Date of birth could be considered personally identifiable, since it splits the population into 25,000 cells and can enable re-identification. If you combine such data with a zip code containing only a handful of people in a certain age range, it may be very easy to re-identify. Professor Swire made an analogy to a cop collecting clues. A suspect is male, tall, with red hair. That would not be enough to re-identify, but it would certainly make it easier. It is more a matter of how much legwork, analysis and extra data is available and accurate. That is what weighs against the public being able to re-identify de-identified data.

Conclusion: Re-identification is hard and de-identification can be harmful

Khaled El Eman, a researcher at the University of Ottawa, felt that the data re-identification efforts by Sweeney and company are the exceptions that prove the rule. Most attacks fail miserably. The studies that succeeded are too small, too few, too ambiguous, too heterogeneous and with confidence intervals that are way too large. Eman concluded that, "Re-identification is hard." He suggested that there would need to be 40-50 replicable studies to start to change such a conclusion.

Daniel Barth-Jones, epidemiology professor at Columbia University, warned that excessive de-identification of data can yield huge statistical errors and inaccurate research results. The greater the level of de-identification, the less statistically useful the data becomes. Blanket de-identification would grind statistical research and number-crunching to a halt. Ultimately, there is no point in de-identification to a level where there are significantly easier and cheaper ways of getting the data. Professor Barth-Jones ended with a warning about trade-offs, that the real harm is not the ephemeral threat to privacy but the real threat of "not catching the next HIV epidemic".

Look for further discussion of this issue and some of the interesting debates from this conference on the MRA website, and feel free to contact me about it at howard.fienberg@marketingresearch.org

---

Howard Fienberg, PLC is MRA's director of government affairs.

[i]Swire, Peter. "Peeping," 24 Berkeley Tech. L.J. 1164 (2009). http://www.peterswire.net/Peeping.pdf