Simson L. Garfinkel
National Institute for Standards and Technology
Open Police Data Initiatives and the Impact on Victims of Intimate Partner Violence
Eisenhower Executive Office Building
April 21, 2016

I know that you all are here today to discuss ways that data related to sexual assault and domestic violence could be released and made available to the public. I don't know anyone who wants to publish victim names and home addresses on the Internet. On the other hand, many advocates and researchers do feel that publishing some kind of microdata online can help address this important social problem—provided that the data do not jeopardize the privacy or safety of victims.

De-identification, which is also called anonymization in Europe, is an approach for stripping the personal information from a dataset. Ideally, de-identification should make it possible for researchers to do their thing with the data without victims suffering any negative consequences from the data release.

Unfortunately, there are many stories of de-identified data being published on the Internet and then being **re-identified** by a group of graduate students or data hackers. These stories have terrified some privacy officials and made others throw up their hands. After all, what's the use of removing names and identifiers if some intrepid data scientist can do a fancy correlation and re-identify the victims? Why should a data provider spend the time and money de-identifying, just to be embarrassed—or worse—face legal consequences? Can any data set really be de-identified? How do you know?

The good news is that there are many ways to remove identifiers from data sets so that the identities of victims are protected. Most of the stories that you've heard of re-identification were performed on data sets that had been improperly de-identified with **ad hoc techniques** that had not been **tested** or **validated**.

The bad news is that removing this information *always* decreases the data quality.

Let me turn your attention for a moment from police data to healthcare data.

Today there are several de-identification standards for Protected Health Information — PHI — under the US Health Insurance Portability and Accountability Act. The HIPAA Privacy Rule has a "Safe Harbor" provision, which specifies a list of 18 kinds of identifiers that are deemed to be linked to a person's identity. Remove all names, geographic subdivisions smaller than a state, complete ZIP codes, dates, phone numbers, email addresses, social security numbers, account numbers—really, all identifying numbers—biometrics and photographs, and medical records are deemed to be legally de-identified. You can publish them on the internet without the patient's permission. But these de-identified records don't work well for some kinds of

research, so HIPAA also recognizes a kind of limited de-identification that allows data to be shared without a patient's permission, but only for specific purposes, and only with a restrictive data-use agreement. These so-called **limited data sets** cannot be put on the internet.

Most discussion of de-identification focus on ***field suppression***—taking a magic marker and blacking out the sensitive columns prior to publication. Field suppression is easy to understand, but it can damage **data quality**, because many of the fields that are suppressed contain the very attributes that we care about. This is especially true when working with law enforcement data, where the attributes that are the subject of concern—attributes like **race**, **age**, **number of children,** and **neighborhood**—are also highly identifying.

Statistical agencies in the US and aboard have been aware of this issue for decades. Over that time, they have come up with a toolkit for manipulating data in a way that preserves some statistical properties while protecting the identity of data subjects.

Generalization and aggregation are two approaches that are used for attributes that can be identifying, but also have analytical value. In generalization, data values may be rounded or reported in buckets. For example, instead of reporting that a 32-year-old woman was harassed, you might report the woman was in her 30s. Aggregation groups together multiple records. For example, instead of reporting that one woman with two children was harassed and another woman with four children was assaulted, you might report that two women with a total of six children were either harassed or assaulted. Other statistical techniques include **field swapping**—actually swapping attributes between several records—and **adding noise**, or fuzzing the numbers.

The purpose of these approaches is to make it difficult, if not impossible, to re-identify the data by matching the released records up with another dataset—for example, a high school telephone directory. These approaches break the one-to-one correspondence between potentially identifying attributes in the micro-data and the external world.

Of course, breaking that correspondence means there may longer be a one-to-one correspondence between people on the ground and records in the data set. Some or all of the data may be **synthetic** after the de-identification is performed. That's okay if the data are being released to help researchers identify general trends and correlations. But the transformations can also degrade some of the specific uses that are envisioned for police data, such as identifying specific patterns of excessive force, discrimination, or non-enforcement. What you might think is a pattern of non-enforcement might actually be an artifact of privacy protection. Journalists and community activists who are unable to find victims because attributes have been swapped may question the veracity of the entire process. Many statisticians are trained in the benefits and limitations of working with synthetic data. It may be useful to publish synthetic data, and restrict the real data to bona fide researchers working in restricted environments.

Modern de-identification techniques give us knobs to control the risk of re-identification, but they don't tell us where those knobs should be set. Is it okay if the data from a domestic abuse

victim can be identified with 1-in-7 chance of being correct, or should the odds be 1-in-40? How about if 95 people out of 100 will have their privacy absolutely protected, but five people will have their identities revealed? These are policy questions that can't be solved by mathematics.

Instead of publishing microdata, another option is publish a query interface that lets researchers explore trends and correlations, but prohibits exporting individual records. These systems can do a much better job protecting privacy, but they less useful to researchers.

To figure out the appropriate setting of the privacy and data quality knobs, organizations need to adopt clear, repeatable procedures for evaluating the risk contained within the data. A Data Review Board, also called a Data Release Board, is one approach that organizations can use to bring together expertise from within an organization—and even from outside the organization—to perform the analysis and make the hard calls. A DRB can weigh the specific public good that will come from the release of a dataset against the risk that the release may pose to the data subjects. DRBs can rely on national standards, but modify them for local needs. They can also evaluate the risk of possible future data releases by other organizations.

Today there is a belief among open data proponents that publishing government data sets is an end in itself—that data somehow promotes transparency and accountability. Of course, it only does that if someone takes the time to analyze the data. Poorly published, data that contain highly sensitive information have the potential to be highly endangering while contributing little to the public policy goals of open data. By focusing up front on the potential for benefit and harm, it's possible to reverse that calculus, and minimize harm while maximizing the public good.