Published on *The National Law Review* https://natlawreview.com

The Data De-Identification Spectrum

Article By:

Peter J. Guffin

This post is part two in a series examining privacy and transparency issues in the context of public access to digital court records, building on my essay "Digital Court Records Access, Social Justice and Judicial Balancing: What Judge Coffin Can Teach Us."

Given the significant risk of harm to individuals stemming from data re-identification, it is imperative that the SJC account for data identifiability in determining which information in court records will be made accessible to the public through its soon-to-be-launched digital system. Data identifiability is a factor wholly separate and distinct from the sensitivity of data. It exists in varying degrees along the de-identification spectrum.

In its 2012 report, "Protecting Consumer Privacy in an Era of Rapid Change, the FTC concluded that the process of removing personally identifiable information (PII) from data is not a silver bullet, acknowledging the broad consensus that "the traditional distinction between PII and non-PII has blurred and that it is appropriate to more comprehensively examine data to determine the data's privacy implications."

To that end, the Future of Privacy Forum (FPF) has published "A Visual Guide to Practical Data De-Identification" (FPF Guide) which delineates various categories of data on the de-identification spectrum based on the interplay of three main variables: direct identifiers, indirect identifiers, and the types of safeguards and controls that an organization places on the way data are obtained, used, or disseminated.

Without getting into the finer nuances, direct identifiers are data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain. These include names, social security numbers, or basic contact information. Keep in mind though that not all direct identifiers are equal in terms of relevance. For example, an email address may be just as quick to identify an individual as a social security number, but the latter is still considered more sensitive given its persistence, resistance to change, and common use as a key to additional sets of PII.

In contrast, indirect identifiers are data that help connect pieces of information that can be used to identify a person. Some common indirect identifiers include date of birth, age, gender, ZIP code, and other demographic information. Identifying an individual based on any one of these data elements ordinarily is not possible. By combining one of these data elements with additional indirect identifiers,

however, an identity can emerge. The typical ways in which organizations deal with indirect identifiers to minimize the risk of re-identification include: suppressing or removal; generalizing values as sets or ranges; swapping data between individual records; and perturbing or adding noise to the data set.

In the two Maine cases referenced in my previous post, both of which involved court orders granting discovery motions to compel production of nonparty patient records, Justices Murray and Walker appropriately dealt with direct and indirect identifiers by, among other things, suppressing the identifiers (*e.g.*, date of birth, age, sex, race) entirely or generalizing values (released data included ***only*** the year of the surgery).

Looking at the FPF Guide, at one end of the spectrum are data that contain direct identifiers thus making the data "Explicitly Personal." Data fall into this category when no attempt has been made to obscure either direct or indirect identifiers.

At the other end of the spectrum are data that are "Anonymous" or "Aggregated Anonymous." Anonymous data feature mathematical and technical guarantees that are sufficient on their own to distort the data so as to prevent re-identification. With aggregated anonymous data, the data are so highly aggregated that additional safeguards or controls are no longer relevant (*e.g.,* published census statistics).

In between on the spectrum are a number of other categories, including "Pseudonymous Data" and De-Identified Data," as well as subcategories of each.

Although the terms de-identified data and anonymous data are often (incorrectly) used interchangeably, it is important to recognize that de-identified data is in a higher risk category and is not equivalent to anonymous data or aggregated anonymous data.

As stated by leading privacy scholars, "[d]e-[i]dentified data is the focal point for much controversy within the technical de-identification community, based in large part around different views as to whether certain information in a database is likely to be useful – now or at some point in the future—as an indirect identifier." "Experts . . . frequently question the merits of de-identification, joined by data scientists skeptical that risk-based de-identification is sustainable or sufficient."

While conceding that "it may be impossible to predict which items of data will *in the future* become linkable indirect identifiers," these same privacy scholars also astutely observe, however, that "[a]t the same time, data points that can be used to pierce through de-identification often carry important social benefits in areas like healthcare, education and science, rendering their suppression undesirable." They argue in favor of taking a practical, risk-based approach to de-identification.

As it happens, virtually all privacy regulatory regimes in the U.S., including the California Consumer Privacy Act (CCPA), which became effective this year and is widely viewed as one of the most privacy protective consumer laws in the nation, essentially take a risk-based approach to de-identification. The CCPA, for example, defines "personal information" as "information that identifies, relates to, describes, is capable of being associated with, or could **reasonably** be linked, directly or indirectly, with a particular consumer or household."

The FPF Guide represents a risk-based approach to de-identification. In addition to factoring into the risk calculus the existence of direct and indirect identifiers, the FPF classifies data based on consideration of the safeguards and controls that an organization places on the way data are obtained, used or disseminated. The latter safeguards include internal administrative controls (*e.g.*,

security controls, access limits, and data deletion practices) as well as external contractual controls (*e.g*., contractual controls that restrict use and disclosure, remedies, and audit rights to ensure compliance).

As I mentioned in my previous post, releasing de-identified data under the data use agreements model is a common safeguard used by many organizations (including the courts) to reduce the risk of re-identification. Putting aside the issues of feasibility and enforcement, the SJC might consider making public access to court records conditional on the recipient's agreement to abide by certain use and disclosure restrictions. It would be an alternative to the "release and forget model" where data are published publicly or made available on the internet without any strings attached, granting complete impunity to users of the information.

In crafting rules regarding public access, one of the key questions that must be answered is how far to push information down the identifiability spectrum to minimize the risk of re-identification. There are only a limited number of variables with which to work. The SJC can remove or manipulate direct and known indirect identifiers from case records in a manner that reasonably breaks the linkage between the information and an individual. It also can put in place reasonable safeguards and controls to complement and buttress its technical de-identification measures.

The trade-off is clear: the less strict the safeguards and controls, the more attention should be paid to dealing with direct and indirect identifiers to achieve a sufficiently low risk of re-identification.

In determining which data are made accessible to the public, I urge the SJC to take a practical, risk-based approach to de-identification, aligning itself with most, if not all, of the privacy regulatory regimes in the U.S. This means considering the state of the data itself (*i.e.,* the degree of its identifiability), in addition to such other factors as the data's sensitivity, accessibility and permanence.

As of this date, in terms of redaction, the SJC has provided us with no clue as to where it thinks data from case records should be on the data identifiability spectrum. Nor can or do we know at this time whether there will be sufficient compliance with the rules by litigants to get us there. Until we have answers to these questions, it is anyone's guess as to whether the digital court records access rules being crafted by the SJC will be adequate to protect Maine citizens' privacy.

National Law Review, Volumess X, Number 149

Source URL:https://natlawreview.com/article/data-de-identification-spectrum