At A Glance: De-Identification, Anonymization, and Pseudonymization

February 28, 2016

De-identification of data refers to the process used to prevent personal identifiers from being connected with information. The FTC indicated in its 2012 report Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers that the FTC’s privacy framework only applies to data that is “reasonably linkable” to a consumer. The report explains that “data is not ‘reasonably linkable’ to the extent that a company: (1) takes reasonable measures to ensure that the data is de-identified; (2) publicly commits not to try to re-identify the data; and (3) contractually prohibits downstream recipients from trying to re-identify the data.” With respect to the first prong of the test, the FTC clarified that this “means that a company must achieve a reasonable level of justified confidence that the data cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer, computer, or other device.” Thus, the FTC recognizes that while it may not be possible to remove the disclosure risk completely, de-identification is considered successful when there is a reasonable basis to believe that the remaining information in a particular record cannot be used to identify an individual.

De-identification is not a single technique, but rather a collection of approaches, tools, and algorithms that can be applied to different kinds of data with differing levels of effectiveness. In 2010, the National Institute of Standards and Technology (NIST) published the Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) that provides a set of instructions and de-identification techniques for federal agencies, which can also be used by non-governmental organizations on a voluntary basis. The guide defines “de-identified information” as “records that have had enough PII removed or obscured, also referred to as masked or obfuscated, such that the remaining information does not identify an individual and there is no reasonable basis to believe that the information can be used to identify an individual.”

NIST has identified the following five techniques that can be used to de-identify records of
Suppression: The personal identifiers can be suppressed, removed, or replaced with completely random values.

Averaging: The personal identifiers of a selected field of data can be replaced with the average value for the entire group of data.

Generalization: The personal identifiers can be reported as being within a given range or as a member of a set (i.e., names can be replaced with “PERSON NAME”).

Perturbation: The personal identifiers can be exchanged with other information within a defined level of variation (i.e., DOB may be randomly adjusted -5 or +5 years).

Swapping: The personal identifiers can be replaced between records (i.e., swapping the ZIP codes of two unrelated records).

The following provides snapshot information concerning de-identification, anonymization, and pseudonymization.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>18</strong></td>
<td><strong>&lt;.25%</strong></td>
<td><strong>4</strong></td>
</tr>
<tr>
<td>The number of specific types of data that must be removed from a health record to qualify under the HIPAA “Safe Harbor” De-Identification Method.</td>
<td>The re-identification risk found by two studies of health records that had been de-identified using field suppression methods.</td>
<td>The number of randomly chosen observations of an individual that could be used to uniquely identify 95% of “mobility traces” (a record of locations and times that a person or vehicle visited over a year).</td>
</tr>
</tbody>
</table>

Key Definition: “Anonymization” of data refers to a subcategory of de-identification whereby data can never be re-identified. This differs from de-identified data, which is data that may be linked to individuals using a code, algorithm, or pseudonym.

Key Definition: “Pseudonymization” of data refers to a procedure by which personal identifiers in a set of information are replaced with artificial identifiers, or pseudonyms.
Key Definition: **Aggregation** of data refers to the process by which information is compiled and expressed in summary form.


[2] Id. at iv.

[3] Id. at 21.


**RELATED PRACTICES**

<table>
<thead>
<tr>
<th>Antitrust and Competition</th>
<th>Data Privacy &amp; Cyber Security Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td></td>
</tr>
</tbody>
</table>