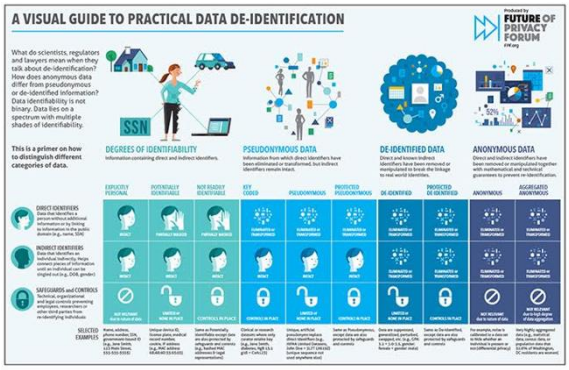
Working Group Meeting Notes

Open to the following FPF working groups: K-12 Privacy, Higher Ed Privacy, K-12 Privacy Leaders, Ad Tech and Location, and Smart Cities

Wednesday, October 11, 2017 1pm-2pm
Topic: De-Identification 101
90 attendees

Kelsey Finch (FPF) on FPF's Perspective of De-Identification

- Looking across sectors, there is a disconnect between the promise of de-identification and the reality of it.
- To be effective, de-identification requires a lot of hard work (both as a technical matter and an organizational policy matter).
- FPF has tried to make sense of the confusion of overlapping legal definitions of technical standards around de-identification by looking at data on a spectrum of identifiability rather than as binary (i.e., personal or not personal, de-identifiable or PII).
    - See our visual guide to practical data de-identification.



- The visual above describes how you can take a record full of PII and by starting to apply different kinds of technical or operational legal control to the direct and indirect identifiers, you can move yourself further down the identifiability spectrum.
- FPF also looks at the risk of re-identification as a spectrum, which depends on the amount of protection that you have in place, what utility you need to preserve the data (de-identifying data inherently removes some utility), and the likelihood that somebody is actually going to try to re-identify that data.
- While you'll never be able to guarantee zero risk of re-identification, there are many tools and concepts that will help you situate and mitigate the risks of re-identification appropriately given your particular circumstances.

Mike Hintz (Hintz Law PLLC) on the general law and policy perspective on de-identification.

Key Points:

- De-identification should be thought of as an important risk mitigation strategy and a key part of legal compliance.
- It is important to strike the right balance when considering de-identification options: Adopt the strongest method of de-identification that's consistent with your data needs.
- You should also document what you are doing. This is particularly important under some emerging legal standards. (Preferred by FTC, Required by EU laws, and it's just a good idea if you can demonstrate and show what you've done to manage risk.)

What is de-identification and to what extent does it protect privacy?

- De-identification is a way to reduce the risk of processing personal data. It's not a silver bullet. There have been a lot of stories about times where de-identification fell short or somebody thought data had been de-identified effectively and it was re-identified. De-identification is one tool in the toolbox. It's not the only thing you should be doing when thinking about a privacy and security program. It can, however, be a very effective risk reduction/risk mitigation tool in combination with other measures that you may take to protect privacy and security of data.

What factors must be considered when choosing a de-identification method?

- There are a wide variety and range of de-identification methods. When you're looking at de-identification options, it's about striking the right balance: finding the strongest methods that you can use to reduce the risk of re-identification that is still compatible with the needs that you have for the data.
- In terms of legal compliance, de-identification can be an effective tool. Some laws call out anonymization or pseudonymization or other types of de-identification as one way you can meet certain legal obligations around protecting data. Most privacy laws are scoped to certain definitions of personal data or PII. These laws describe what makes it personal data as opposed to non-personal data (e.g., "directly tied to an individual"). In some cases, de-identification is a partial way of meeting obligations. In other cases, you can use a de-identification method that goes to the very far ends of the spectrum where you could truly call it anonymous in many cases and that takes you outside the scope of those privacy laws altogether.

Exploring the corollary between de-identification and the concept of personal data:

- Make sure you're being precise about the terms you use and what they actually mean. Anonymous, pseudonymous, de-identified, and aggregated are often used interchangeably. For example, the term "anonymous" is sometimes thrown around. People say, "oh don't worry this data because it is anonymous," but have later found that it was de-identified in a way that it was able to be reversed. Also, using the word "anonymous" for types of de-identification that aren't as strong can also cause problems. Even the FTC has started saying that using the term "anonymous" word for lesser forms of de-identification may actually be a deceptive practice. Don't use the "A" word unless you really mean it.

What kinds of data are considered personal data?

- If there's information that makes it clear who the data relates to, that's clearly personal data.
- If you can easily match a unique identifier in one record to another record that has a name or an address, that makes it personal data.
- If the data allows you to reach out and touch somebody, connect to them directly, communicate with them directly, that's going to clearly be within the scope of personal data. (→ contactable data)
- And, increasingly, personal data includes data sets where you don't necessarily know who the person is in the real world (e.g., online behavioral profiles that are used for advertising.) For example, a cookie ID doesn't tell you who a person is, but it allows you to treat that unique individual associated with that cookie differently by showing a different ad or providing a different personalized experience. That ability to single somebody out and treat them differently will often lead to a conclusion that that data is considered personal data.
- If you can connect two datasets – one that's identified or identifiable and then another that (on its own) might not be, then by association that second data set becomes part of the personal data as well (→ guilt by association concept).
- Determinations of whether data is personal data can be relative and contextual. The IP address is a class example. On one hand, a website that's logging an IP address about a

person's visit might claim this is not personal information because the website owner does not know who this person is. On the other hand, to the ISP that issued that IP address, that data clearly would be personal data because they have an association with a billing account or some other personal information with that IP address.

- Some laws talk about the concept of personal data in very specific terms. It will be like almost a laundry list: name, address, phone number, credit card number, etc. The laundry list inevitably leaves something out.
- Others laws have more general definitions like information (e.g., "allows you to identify an individual"). The trend is towards a more general definition.
- Sometimes, laws have a catch all phrase at the end like "and any other information that allows you to connect or identify somebody."
- The FTC concepts/definitions of personal data has evolved over time. From 2002 to 2012, you can see that they broadened the concept significantly. They dropped the term "individually identifiable information" to just simply call it "covered information." They wanted to get away from people thinking about those terms in very narrow ways. They added things like IP addresses, photos, videos, and physical location. All those things that have become thought of as identifiable in different ways.
    - In 2013, the FTC also expanded the definition of personal information under the COPPA rule in 2013 to include IP addresses, cookie IDs, processor or device serial numbers or other unique identifiers, photographs, videos, or audio files that contain a child's voice or image, and geolocation.
- State breach notification laws define personal data very narrowly. Only certain types of personal information will trigger those and they're fairly narrowly scoped to financial information, social security numbers, government issued IDs, financial account numbers, and the like. Over time, some of them have been expanded to include other types of information insurance or health/medical data, but they are still quite narrowly focused.
- The new definition of personal data under the GDPR is quite broad. I've got sort of a shortcut/decision tree to help you decide whether something is personal data under European law. If you have some data, ask whether it is personal data? The answer is yes. It's almost always yes.

What types of data are most identifiable?

- Certain data types are almost inherently identifiable and it's just good to keep this in mind.
- Obvious examples include names, email address, biometrics (anything that is unique to your body genetic information, unique iris scans, fingerprints, even voice prints).
- If you if you're collecting the **content** of communications where you don't have control over that content, there might be something in that content that identifies someone. For example, if you are collecting voice snippets to create a voice recognition algorithm, this sounds very innocuous because you wouldn't be able to tell who someone is without doing some kind of sophisticated analysis on it. If, however, a voice snippet that you collect says, "Hey, this is Mike. Here's my phone number," obviously that content makes it makes it identifiable. If you are managing content, it's something to keep in mind. It's very hard to anonymize that kind of data.
- If you collect enough **location information** over time it becomes very unique to a person. For example, there's only one person on the planet who travels from my home to my workplace Monday through Friday. You are able to single out somebody at least and potentially identify them based on the collection of location information.
- If you just get enough information, any one piece of which would not be identifiable to a human, that collection over time becomes quite unique, and it's possible to single out somebody (→ contact tapestry effect). For example, years ago, AOL publicly released a lot of search data (terms that people had been entering into their search engine) so that

researchers could do interesting work with it. They replaced the account IDs and IP addresses with a randomly generated unique ID for each account. But the search queries – if there were enough of them connected over time - might be able to narrow it down and, in some cases, identify someone. (For example, you might learn where they live because they're doing a lot of local searches or you might be able to identify them because of the names of the people they are searching, which might be associated with them in some way.)

- Finally, personal data about others might actually reveal data about you. One study, for example, found that if a certain number of people in your social network happened to be openly gay, it's quite easy to predict whether or not you are.

Things that you can do in addition to re-identification to ensure or provide protections against the risk of re-identification:

- The FTC, in talking about how they're thinking about concepts of de-identification of personal data, set out a framework for what data can be reasonably linked to a specific consumer, computer, or device. If the entity takes reasonable measures to ensure that the data is de-identified, publicly commits to not trying to re-identify the data, contractually prohibits downstream recipients from trying to re-identify the data – these provide additional protections against the risk of re-identification.

Under the HIPPA privacy rule, there are two alternative approved methods for de-identification:

- One is called the **Safe Harbor method**, which is sort of the analogous to that laundry list approach to defining personal data. If you remove certain defined data types from a data set, that would be considered de-identified under the HIPPA rule.
- The second method is called the **expert determination method** where you bring in an expert, and they look at your de-identification method. If they attest that it is valid, the de-identification that also can satisfy the privacy rule.

Daniel Barth-Jones on the HIPPA Re-identification Arena

Key Points:

- De-identification will fail some of the time. That doesn't mean that it is completely useless
- Good public policy requires reliable scientific evidence.
- Re-identification risks can be controlled usually without losing much of the statistical accuracy of the data, but this requires a firm grounding and an extensive body of statistical disclosure control and limitation literature.

Misconception #1 about de-identification: it doesn't work and has no efficacy.

- Paul Ohm kind of led this debate with his statements and his broken promises of privacy paper, indicating that it was easy, cheap, and powerful [to re-identify people].
- That was an accurate statement pre-HIPPA.
    - Urban myth: You can re-identify 87% of the U.S. population with three data elements (5-digit code, full birth date, and gender).
    - In fact, this has been studied repeatedly and over the last three decades of U.S. Census data. The number, if you were able to pull off an entirely accurate census, ***was*** probably about 63%.
    - When the study was attempted again in 2013 by the same person who had originally cited the 87%, only 28% of the individuals were able to be re-identified. Note: 28% is not an acceptable risk.
    - Post-HIPPA implementation: The estimate from that same expert is that roughly about 4 people out of 10,000 could be re-identified under Safe Harbor.

- In reality, under HIPPA's de-identification requirements, re-identification turns out to be quite expensive and time consuming. It requires some substantive computer and mathematical skills. That's rarely successful. And the person who attempts the re-identification is usually uncertain as to whether she has actually succeeded.

Misconception #2 about de-identification: de-identification works perfectly or permanently.

- Perfect is de-identification is impossible. It's not able to free data from all subsequent privacy concerns, and it's never permanently de-identified. There's no 100% guarantee that the data would remain de-identified, regardless of what you do with it after that de-identification process.

Essential re-identification concepts.

- **Record linkage**:
  - With the use of linkage keys or quasi-identifiers, we may not individually identify someone, but in combination it can re-identify by somebody.
  - Quasi-identifiers are things that don't individually identify a person – things like gender, age, ethnic group, marital status, and geography – that may be linked in other sources to direct identifiers like their name or their address.
- **Sample unique vs. population unique**:
  - When we have only one person shares particular set of characteristics in your sample, we call them sample unique.
  - When only one person with a particular set of characteristics exists in the entire population we would call them population unique.
  - The reason why it's useful to differentiate between cases is that all of our data in our sample (whether it's health care or education data) is not necessarily unique, but those that are unique are at potential risk of re-identification by linking it up against people who are unique in the larger population.
  - If the records are not unique in the sample, they can't be unique in the population. This means they are not at a definitive risk of being re-identified.
  - If the records are not in the sample at all, that's not part of what you're required to pay attention to because you are the person releasing the data, and you need to attend only to the risks for the data that you are releasing.
  - Records that are unique from the sample but would link to multiple individuals in a larger population: While they have some probability of being re-identified, it can't be re-identified with certainty. So, we have to attend to the idea of what is that probability of re-identification.
  - Then of course only the records that are unique in the sample and the population are at risk of being re-identified with exact record linkage, which can quite easily determine the proportion of sampling makes by analyzing our own data and doing cross classification of all the characteristics. For many characteristics, the likelihood of being unique in a larger population can be estimated from statistical models, often accomplished with U.S. Census data.
- **Straightforward methods for trying to reduce re-identification risks**:
  - Most of the methods that involve distorting the data like adding random noise are non-trivial to implement. They are expensive and complicated to put into place. T
  - Reducing the resolution of the keys that are present and increasing the population reporting sizes are two simplest methods. They are the most practical when we're protecting data that's coming in that continuous data stream.
  - **Reducing the key resolution** reduces both the proportion of samples we need from the data and the probability of that the individual is unique in the larger population. We can do this either by reducing the number quasi-identifiers - completely eliminating them - or reducing the number of categories or values of quasi-

identifiers.

- We can also **increase the population sizes** of the geographic reporting units. This is a pretty easily implemented solution for reducing disclosure risks; particularly, if we use our requirements for minimum population sizes for geographic reporting. That's HIPPA safe harbor did. It implemented that you could only report 3-digit zip codes that contained populations with more than 20,000 individuals. However, it also requires, if you're trying this very small risk approach under the expert determination, that you conduct statistical analyses to determine what these risks are for the particular set of keys that are present in your data. Using larger population sizes for geographic reporting areas is an important method for controlling disclosure risks because increasing the population size decreases the population of the individual, or the probability that the individual will be unique within the reporting area. Ideally we want to choose a method that allows reporting on all or most of the population, but the geographic resolution would be scaled to the underlying population such.

What other risks, other than re-identification, must be considered when choosing de-identification methods?

- Unfortunately, the de-identification leads to information loss, which may limit the usefulness of the resulting health information. Some popular de-identification methods can degrade the accuracy of de-identified data for multiple various statistical analyses or data mining methods. This is well understood by statisticians, but it's not as well recognized and integrated into public policy. So, we are not in the situation, due to mathematical constraints, where we can have our cake and eat it too. We can't have perfect information and perfect protection.
- De-identified data is an invaluable public good. It really provides an essential tool for society and supporting scientific innovation within education and improvement in our education system.
- Poorly conducted de-identification can unfortunately lead us into bad science. And bad decisions about whatever policy or action we're considering taking based on that data. I would suggest that perhaps that's the greater harm - even above and beyond the privacy harms - because now we're harming everybody with our incorrect knowledge of the world.
- We need to balance our risks of disclosure and statistical accuracy. We can receive very good increases in disclosure protection for relatively small losses of information most of the time.

How can we decrease re-identification risk without losing statistical accuracy?

- Re-identification risks can be controlled usually without losing much of the statistical accuracy of the data, but this requires a firm grounding and an extensive body of statistical disclosure control and limitation literature.
- The HIPPA expert determination method allows us to substitute a very small risk instead of eliminating all of the required identifiers.
- Unfortunately, the safe harbor approach eliminates full dates for detailed geography that are often quite important for analyses. It's also challenging to implement in complex datasets, particularly datasets that are changing over time where we have to preserve referential integrity. And, for example, process encrypting data does not result in de-identification. HIPPA safe harbor requires its removal, so encrypting it does not get you there. (It can be used under the expert determination methods.) And safe harbor is really not suitable for cases where we have to produce test data or development data or demonstration data because it requires the removal of these data elements and software development.
- The important thing to understand about the expert determination method is that the risks needs to be very small (it doesn't specify what it needs to be zero), and that there's some

possibility that this data could be linked back.

What do you make of the phrase – "there is no such thing as anonymous data?"

- This is an essentially meaningless and perhaps even misleading statement. If we don't give consideration to a specific re-identification context and the data details that are present such as a the particular type of data that present, their coding schemas, whether it's a geographic or spatial data or the network properties. And we also need to consider the de-identification methods that were applied in the experimental design re-identification attack demonstrations.

How do we move beyond anecdotes to rigorous scientific evidence approaches for dealing with re-identification risks?

- As proposed by FPF's Jules Polonetsky, we need to simultaneously require both technical de-identification methods and supplement those with legal and administrative controls.
- On a societal level, we really need a multi-sectoral legislative prohibition against data re-identification. We also need to have exemptions for re-identification research.
- Finally, we need some centers for excellence for a combined graduate training and statistical disclosure and privacy computer science.