February 28, 2014

Chairman Tom Wheeler
Commissioner Mignon Clyburn
Commissioner Michael O'Rielly
Commissioner Ajit Pai
Commissioner Jessica Rosenworcel
Federal Communications Commission
445 Twelfth Street SW
Washington, DC 20054

Ilya Mironov
Researcher
Microsoft Research
1065 La Avenida
Mountain View, CA 94043

Re: WC Docket No. 13-306, *Public Knowledge Petition for Declaratory Ruling*

Dear Chairman Wheeler and Commissioners:

I am a Researcher at Microsoft Research, where my work focuses on privacy-preserving data analysis, cryptography and applied security. I hold a Ph.D. in Computer Science from Stanford University. I have authored more than 35 papers in peer-reviewed publications, and served on more than 15 program committees[1].

The views expressed in this document are my own and in no way reflect the views of Microsoft Corporation or Microsoft Research. Although I am a Researcher at Microsoft Research, I am also a scholar with an independent interest in this proceeding.

Often, de-identified data sold by companies can be re-identified. Those with an interest in re-identifying records have access to a vast universe of information that enables them to re-identify particular individuals or entire datasets. Current anonymization techniques do not adequately protect privacy in de-identified datasets in light of the existence and availability of such information.[2]

[1] My CV is available at http://research.microsoft.com/people/mironov/cv.pdf.
[2] In this letter I address only the technical possibility of re-identification, without regard to whether re-identification under particular circumstances would violate any law or breach contractual terms. Applications of de-identification as a means of disclosure control within organizations are outside the scope of this letter.

Current Anonymization Techniques Do Not Adequately Protect Privacy

I am an expert in privacy-enhancing technology, and in my opinion it is not possible to de-identify datasets that contain personal information in a way that protects privacy of all individuals while retaining the practical value of the data. Over the past two decades, researchers have proposed several different anonymization techniques, but each new technique has ultimately proven inadequate to protect privacy.[3][4]

In 1997, Latanya Sweeney proposed an anonymization technique called *k*-anonymity, in which a dataset is considered sufficiently de-identified if no combination of "quasi-identifying" attributes (such as ZIP code and age) occurs fewer than *k* times, and *k* is calculated to ensure that an individual in the dataset cannot be re-identified using just those quasi-identifiers.[5] But datasets anonymized in this way still reveal sensitive information about individuals.[6] Even if specific individuals cannot be re-identified in the dataset, the fact that an individual appears in a dataset at all can be revealing.[7]

To address this vulnerability with *k*-anonymized databases, researchers proposed *L*-diversity, which also proved inadequate. *L*-diversity requires that each group of quasi-identifying attributes, such as each area code and billing zip code combination, be associated with at least *L* different values of each other attribute that may contain sensitive information.[8] But datasets anonymized using *L*-diversity can still compromise individuals' privacy, because the sensitive values within a group may each be distinct while still revealing information about

---

[3] That is, it is not possible to de-identify datasets if the data must remain at least marginally useful. The publisher could fully anonymize its data by erasing it or replacing it with random noise, but then it would be both useless and valueless. *See* Dwork, C.: An Ad Omnia Approach to Defining and Achieving Private Data Analysis. PinKDD 2007, LNCS 4890, pp. 1-13 (2008), pg. 2 (discussing the requirement that data remain useful).

[4] There is rich literature on methods for releasing *aggregate information* while protecting privacy of individual records, allowing rigorous and quantifiable trade-offs between utility and privacy. These techniques exist outside the anonymization frameworks discussed in this letter. *See* Dwork, C.: A firm foundation for private data analysis. Communications of the ACM, vol. 51(1), 2011, pp 86-95.

[5] Sweeney proposed *k*-anonymity following her successful attempt to use publicly available voter registration records to re-identify personal medical records of the Massachusetts governor in an "anonymized" insurance dataset that retained patients' birthdate, sex, and ZIP codes. Samarati, P., Sweeney, L.: Protecting Privacy when Disclosing Information: *k*-Anonymity and its Enforcement Through Generalization and Specialization. Technical Report SRI-CSL-98-04, SRI Intl. (1998). As an example of *k*-anonymity, consider a dataset of three people with the area code 218 and four with the area code 219. The dataset is not 5-anonymous because some area codes have fewer than five members, but if the 218 and 219 area codes are instead generalized to 21*, it becomes 5-anonymous.

[6] Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: *l*-Diversity: Privacy Beyond *k*-Anonymity. In: Proceedings of the 22nd International Conference on Data Engineering (ICDE 2006), p. 24 (2006).

[7] *Id.* at 2 (describing homogeneity attacks and background knowledge attacks).

[8] *Id.* at 6-8 (introducing *L*-diversity). The intuition is that this raises the amount of background knowledge required to re-identify someone. The authors also proposed two alternate ways of measuring whether sensitive values are sufficiently well-represented for *L*-diversity. *Id.*

that group if they are similar to each other or skewed relative to the general population,[9] and because if multiple versions of the same dataset are released over time they can be combined to re-identify individuals.[10]

Researchers developed *t*-closeness and *m*-variance to address *L*-diversity's weaknesses,[11] but these new measures don't guarantee privacy either. Indeed, no approach to anonymization has been able to protect against the possibility that multiple de-identified datasets can be combined both to discover enough information to successfully complete a linkage attack, and to learn more details about some individuals in the data. For an example of how two hypothetical de-identified datasets of call detail records from different phone carriers could be combined, see Appendix A.

Outside Information Can Be Used to Re-Identify Individuals and Datasets

With the right type and sample of outside information in hand, a party interested in re-identifying de-identified information can perform what is called a "linkage attack" to re-identify a single specific individual or many individuals in an entire dataset. In a linkage attack, an attacker uses common attributes to link records in the de-identified dataset to records in a dataset that has not been de-identified. For example, if a de-identified dataset tells me that customer 12345 is a man born 11/10/1969 in ZIP code 59032 and that he likes green peas, and another dataset tells me that Big Bird is the only man born 11/10/1969 in ZIP code 59032, then by linking the two datasets together I can re-identify customer 12345 and learn that Big Bird likes green peas. To see an example of what a linkage attack might look like using information about an individual customer and a hypothetical set of de-identified call detail records, see Appendix B.

Linkage attacks may rely on any unique attribute or combination of attributes to provide the key that links two datasets together. For example, in 2006 computer scientists

---

[9] Li, N., Li, T., Venkatasubramanian, S.: *t*-Closeness: Privacy Beyond *k*-anonymity and *l*-diversity. ICDE 2007, pp. 106-115 (2007) (describing skewness attacks and similarity attacks).
[10] Xiao, X., Tao, Y.: *M*-invariance: Towards privacy preserving re-publication of dynamic datasets. SIGMOD 2007, pp. 689–700 (2007). For example, someone who knew that a target individual had placed calls in June but not July could get closer to re-identifying that person by comparing the two months' datasets.
[11] Li *et al.*, *supra* note 9. (*t*-closeness builds on *L*-diversity by requiring that the distribution of each sensitive attribute within each class be close enough to the distribution of that attribute in the dataset as a whole); Xiao, X., Tao, Y.: *M*-invariance: Towards privacy preserving re-publication of dynamic datasets. In: SIGMOD 2007, pp. 689–700 (2007).

demonstrated that Netflix customers' movie ratings are unique and can be used to re-identify individual customers in a de-identified dataset of movie ratings.[12]

Another type of linkage attack relies on the relationships between de-identified individuals. Research shows that even after names have been removed, individuals within a social network can be re-identified using knowledge of a small set of friends within the network.[13] In 2009, researchers linked together anonymized datasets from Twitter and Flickr using only social network structure, and were able to correctly identify a third of the Twitter users who had Flickr accounts with only a 12% error rate.[14] And in 2011, they demonstrated that a dataset describing relationships between de-identified Flickr users could be re-identified by performing a new crawl of Flickr at a later date, constructing a similar dataset with usernames intact, and then comparing the two.[15]

A related re-identification attack uses statistical analysis to descramble information that has been scrambled or replaced with a random value. For example, in a search log in which queries have been scrambled by replacing words with random tokens, the process can be reversed based on statistical properties that are preserved under this transformation[16]

With enough outside information, datasets can be re-identified even after identifiers and quasi-identifiers have been replaced with random values and artificial errors have been

---

[12] *See* Narayanan, A., Shmatikov, V. Robust De-anonymization of Large Sparse Datasets. In: 2008 IEEE Symposium on Security and Privacy, pp. 111-125 (2008), pg. 1 (discussing the fact that data about specific individuals are "sparse," meaning that the average record about an individual has no records with "similar" attributes).

[13] This works because friendship patterns are distinctive enough that most people in a social network can join with a small number of their friends and identify the part of the anonymized graph that contains them. Once that part of the graph is identified, they can map their outside knowledge onto the graph to identify individuals to which they are connected and learn otherwise private information about them. Backstrom, L., Dwork, C., Kleinberg, J. Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. IW3C2 2007.

[14] Narayanan, A., Shmatikov, V. De-anonymizing social networks. In: Proceedings of the 2009 30th IEEE Symposium on Security and Privacy, ser. SP '09, pp. 173–187 (2009). This worked even though users had only a small number of friends that were common to both networks. As the authors explain, "We show that a third of the users who are verifiable members of both Flickr and Twitter can be recognized in the completely anonymous Twitter graph with only 12% error rate, even though the overlap in the relationships for these members is less than 15%!" *Id.* at 2.

[15] Narayanan, A., Shi, E., Rubinstein, B. Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge. IJCNN, pp. 1825-1834 (2011).

That said, linking outside information to a dataset does require some amount of commonality between the two. A party seeking to de-anonymize a dataset would naturally do so by using another dataset with related details, as in the Netflix and IMDb example, not an unrelated dataset or one that provided no additional information.

[16] This is called "priv-analyzing." The researchers used this technique to identify searches even after they were obscured by replacing words with random tokens. Kumar, R., Novak, J., Pang, B., Tomkins, A. On Anonymizing Query Logs via Token-based Hashing. IW3C2 2007.

introduced. Human preferences and interactions are so distinctive that it is easy to re-identify individuals in many contexts, even with limited information.[17]

Purchasers of De-Identified Consumer Phone Records Have Access to Additional and Related Information About the Same and Other Phone Customers

For those with an interest in re-identifying records obtained from a phone carrier, there are many sources where valuable outside information can be obtained. Types and sources of outside information may include:

- Additional records from the same carrier for a different time period
- Additional records for the same time period from a different carrier
- Consumer information available for purchase from other sources
- The purchaser's own data and knowledge about phone users
- Information about the records obtained through conversations with the seller
- Information about the records obtained through the purchaser's own purchase request (e.g., a purchaser who requests records of users ages 18-26 will know that individuals in the dataset are 18-26)
- Publicly available information, including information on the Internet (e.g., social networking sites)

De-identified call records are not purchased in a vacuum; those who purchase them have motivations for obtaining the data and reasons for thinking the data will be of value.

An Anonymization Technique Cannot Be Evaluated Unless It Is Disclosed

As I explained above, I do not believe it is possible for a dataset containing personal information to be de-identified in a way that protects privacy of all individuals while retaining the practical value of the data. I am immersed in this field, and I am not aware of any de-identification technique that is sufficient to protect the privacy of individuals.

[17] Importantly, linkage attacks can reveal information about individuals even where the party performing the attack does not have enough information to identify particular people with certainty. There is a common (and false) counterargument that linkage attacks are ineffective if they cannot prove with certainty that the two linked records belong to the same person. This argument is false because even if a match isn't guaranteed to be correct (for example, if more than one man was born on 11/10/1969 in zip code 59032), we can use statistical techniques to establish the probability that the two linked records belong to the same person.

As long as this probability is better than a random guess, the linkage attack has revealed information. And if the probability of a correct match to an individual is high enough, or if the fact revealed is sensitive enough, that information can harm the individual.

I could be wrong. But if someone has developed a new anonymization technique that achieves robust protection against re-identification, it could only be fully trusted if made public for evaluation. There is a saying in security engineering that "security through obscurity is no security at all."[18] In other words, if a security approach has not been made available to facilitate the discovery (and ultimately the repair) of its vulnerabilities, then the vulnerabilities still exist and have simply not been discovered. The same can be said for privacy-enhancing technologies. An anonymization technique that is kept secret should be presumed to have vulnerabilities.

Conclusion

At this time there is no de-identification technique that has not been shown to have vulnerabilities. De-identified datasets can often be re-identified. With the right type and sample of outside information, those with an interest in re-identifying de-identified datasets can do so. And there is a wide variety of valuable outside information available to assist them. If phone carriers have nevertheless developed a new technique that is more effective than past innovations, it should be disclosed for peer review.

Respectfully submitted,

/s/

Ilya Mironov

Researcher
Microsoft Research
1065 La Avenida
Mountain View, CA 94043

[18] As the Department of Defense put it, "[i]n general, 'Security by Obscurity' is widely denigrated." U.S. Dep't of Defense, Open Source Software FAQ (discussing the fact that hiding source code from third parties "is obviously *not* a security advantage"), *available at* http://dodcio.defense.gov/OpenSourceSoftwareFAQ.aspx.

Appendix A

To understand how de-identified records from different carriers could be combined to augment the total amount of de-identified information available for re-identification, imagine that party R, who is interested in re-identification, has obtained the following hypothetical de-identified records from two different phone carriers:[19]

Carrier A | Customer 2003612345

Date	Time	Direction	Destination	Duration
2013/05/01	**16:56:18**	**OUTBOUND**	**1000354321**	**00:00:12**
2013/05/01	**16:56:48**	**INBOUND**	**1000354321**	**00:07:07**
2013/05/01	18:32:39	INBOUND	9403032323	00:16:20
2013/05/01	20:01:16	INBOUND	9021040404	00:02:24

Carrier B | Customer 5058675309

Date	Time	Direction	Destination	Duration
2013/05/01	12:32:40	OUTBOUND	8112450582	00:24:42
2013/05/01	**16:56:20**	**INBOUND**	**4424042981**	**00:00:10**
2013/05/01	**16:56:47**	**OUTBOUND**	**4424042981**	**00:07:08**
2013/05/01	16:58:22	OUTBOUND	9692875209	00:03:56

Based on the times and durations of the bolded calls, R can infer that Carrier A's customer 2003612345 is denoted as customer 4424042981 in Carrier B's records, and that Carrier B's customer 5058675309 is denoted as customer 2003612345 in Carrier A's records.

[19] In these hypothetical examples, the carrier has de-identified the records by replacing each customer's phone number with an anonymous identifier, but has left intact the starting time of each customer voice or text event, the direction of each event, and the duration of each event.

Appendix B

To understand what a linkage attack might look like using information about an individual customer and a hypothetical set of de-identified call detail records, imagine that party R, who is interested in re-identification, has the following information about John Doe:

- He lives in Brooklyn, New York
- 10/08/13–10/20/13 traveled to Arizona

Imagine that R also has the following de-identified phone records:[20]

Customer 1121534567

Date	Time	Direction	Destination	Cell Site Lat. & Long.	Duration
2013/10/03	20:15:21	OUTBOUND	8500323456	40.670145, -73.972324	00:09:32
2013/10/08	14:18:02	OUTBOUND	8500323456	40.647224, -73.785422	00:02:49
2013/10/08	20:32:32	INBOUND	5700698765	33.436321, -112.02234	00:00:07
2013/10/09	16:15:24	OUTBOUND	1121565432	33.467420, -112.026657	00:12:07
2013/10/20	09:16:59	OUTBOUND	1121565432	33.436321, -112.02234	00:02:24
2013/10/20	14:54:14	OUTBOUND	8500323456	40.647224, -73.785422	00:01:30

From these records, R can determine that the customer:

- Has a billing address in Brooklyn, New York
- 10/03/13 at 8:15 PM called a customer in Arizona
- 10/08/13 at 2:18 PM called the same customer in Arizona from JFK Airport
- 10/08/13 8:32 PM was at Phoenix Sky Harbor (PHX) Airport in Arizona
- 10/09/13 4:15 PM called a customer in Brooklyn from a location in Phoenix
- 10/20/13 9:16 AM called the same customer in Brooklyn from PHX Airport
- 10/20/13 2:54 PM called the Arizona customer from JFK Airport

[20] In this hypothetical example, the carrier has de-identified the records by replacing each customer's phone number with an anonymous identifier consisting of the 5-digit billing ZIP code and a unique integer, but has left intact the starting time of each customer voice or text event, the direction of each event, the duration of each event, and the locations of the starting and ending cell towers associated with each event.

This is comparable to the approach taken in Isaacman *et al.* Identifying Important Places in People's Lives from Cellular Network Data. In: Pervasive 2011, LNCS 6696, pp. 133-151 (2011), which used anonymized call detail records that replaced each phone number with a 5-digit billing zip code and a unique integer; the records also contained the starting time, duration, and starting and ending cell tower locations for each call, and whether the phone was associated with a business or individual. *Id.* at 135.

Using the information R already had about Doe in combination with the information in the de-identified records, R can likely determine that customer 1121534567 is Doe.

In the event that more than one customer with a Brooklyn billing address had the same travel itinerary as Doe, R can single out Doe from among the possibilities with a bit more information. For example:

- If R knows the location of Doe's workplace, she can map the cell site locations where each candidate's cell phone travels on weekdays when in town to discover which one is consistently where Doe's workplace is located
- If R knows what city Doe's mother lives in, she can search each candidate's records for evening and weekend calls to a customer whose billing address is in that city
- If R knows what time any particular call(s) that Doe made or received took place, she can search each candidate's records for a call that took place at that time