

**Before the
Federal Communications Commission
Washington, D.C. 20554**

In the Matter of

In the Matter of the Petition of Public Knowledge <i>et</i>)	
<i>al.</i> for Declaratory Ruling Stating that the Sale of)	WC Docket No. 13-306
Non-Aggregate Call Records by)	
Telecommunications Providers without Customers		
Consent Violates Section 222 of the		
Communications Act		

Comments of The Information Technology and Innovation Foundation

January 17, 2014

Doug Brake
Daniel Castro
Information and Technology & Innovation Foundation
1101 K Street NW Suite 610
Washington, DC 20005

Contents

I.	Summary.....	1
II.	Carriers Should Be Allowed to Innovate with De-Identified Data.....	1
	A. Re-identification concerns are over-stated	2
	B. Carriers already take reasonable steps to anonymize data	4
	C. Data innovation holds great promise and is in the public interest	6

I. Summary

The Information Technology and Innovation Foundation¹ (ITIF) respectfully submits these comments in the above captioned proceeding.² ITIF does not present a stance on the specific facts alleged in the Public Knowledge *et al.* petition (petition) or the broader policy question of whether non-aggregate anonymized data should be marketable without user consent, instead we narrowly argue that anonymized data can, in many circumstances, be used without fear of re-identification.

II. Carriers Should Be Allowed to Innovate with De-Identified Data

The risk of re-identification of individuals from anonymized data, while not non-existent, is significantly lower than indicated by the petition. Furthermore, well-known anonymization techniques likely used by carriers further protect against re-identification. The economic and social value of innovations in data analytics should encourage the Commission to promote data collection and sharing where appropriate safeguards exist. The Commission should hesitate before changing its treatment of anonymized data.

¹ ITIF is a non-partisan research and educational institute – a think tank – whose mission is to formulate and promote public policies to advance technological innovation and productivity internationally, in Washington, and in the states. Recognizing the vital role of technology in ensuring prosperity, ITIF focuses on innovation, productivity, and digital economy issues.

² In the Matter of the Petition of Public Knowledge et al. for Declaratory Ruling Stating that the Sale of Non-Aggregate Call Records by Telecommunications Providers without Customers Consent Violates Section 222 of the Communications Act, WC Docket No. 13-306, (rel. December 18, 2013) (.

A. Re-identification concerns are over-stated

The petition claims that information in anonymized data sets can frequently be “re-identified” and matched to a specific individual. The risk of re-identification presented in the petition is over-stated and not an accurate representation of the ongoing debate in privacy circles. Several scholars have argued that we should not over-react to the risks of re-identification,³ and, furthermore, common sense tells us that we can get a great deal of utility out of anonymized data with reasonable privacy safeguards.

Consider, for example, the risk of re-identification with age, location, and gender information. The petition cites Latanya Sweeney’s famous study showing that 87% of the U.S. population could be uniquely identified by gender, ZIP code, and date of birth.⁴ Researchers at Palo Alto Research Center have replicated this study using 2010 census data, finding that only 63% of the population is uniquely identifiable given those data categories.⁵ More importantly, the risk of unique identification drops off sharply when given slightly more abstract data. For instance, if the data is limited to gender, ZIP code, and the month and year of birth (instead of the full birthday), the percentage of those uniquely identifiable drops to 4.2%.⁶ Similarly, if one replaces the ZIP code with the county in which a man or woman with a particular birthday lives, only 0.2% of the population is unique.⁷

³ See, e.g. Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1815, 2011; Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARVARD J. OF L. & T. 1, 2011; Felix T. Wu, *Defining Privacy and Utility in Data Sets* 84 U. OF COLO. L. REV. 1118

⁴ Latanya Sweeney, *Uniqueness of Simple Demographics in the U.S. Population*, Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.

⁵ Phillippe Golle, *Revisiting the Uniqueness of Simple Demographics in the US Population*, Palo Alto Research Center, available at <http://crypto.stanford.edu/~pgolle/papers/census.pdf>.

⁶ *Id.* at 2.

⁷ *Id.*

This simple example illustrates that there is a balance between the utility of data and the privacy risk. The more granular data is, the more useful it may be to researchers, but the greater the risk of re-identification. Some types of data can have remarkably long tails – for example, the oldest living person can easily be picked out of the world’s population given only a dataset of birth year. There will always be a risk of targeted re-identification against statistical outliers; the question is whether this risk is so great that the Commission should automatically assume all anonymized data is individually identifiable. ITIF believes that the risk is too small to justify such a radical change.

The petition also points to a pair of researchers that were able to uniquely identify movie-watchers using a Netflix database of movie ratings.⁸ The petition neglects to point out that the researchers were only able to identify two out of 480,189 Netflix users with confidence.⁹ These two were re-identified by comparing the user ratings and dates watched with similar information culled from public Internet Movie Database (IMDb) profiles. Here again, it is the statistical outliers most at risk of re-identification: the likelihood of de-anonymization goes up significantly when users had rated a large number of unpopular movies.¹⁰ While those of us with unusual taste in movies certainly deserve no less privacy than everyone else, the point is that we should not assume all non-aggregate data is automatically “individually identifiable.” The Netflix researchers were able to de-anonymize movie-watchers using indirect identifiers only with the

⁸ Petition at 7 citing Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, U of Tex. At Austin, *available at* http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf (Robust De-anonymization of Large Sparse Datasets).

⁹ Robust De-anonymization of Large Sparse Datasets at 13. The researchers used auxiliary data of only 80 IMDb users, so it may be difficult to draw useful conclusions from this fact alone.

¹⁰ While ITIF refrains from commenting on broader issues like forum appropriateness and harm, it is worth mentioning that Netflix settled a class-action lawsuit regarding this data.

help of “auxiliary data” from IMDb. To put this in context, the privacy harm we are concerned about is the nosy boss that finds out an employee’s rating of several Netflix movies. With enough information, the boss could attempt a targeted de-anonymization attack. ITIF believes such a risk does not justify radical changes in how this data is currently regulated.

In the end, re-identification risks likely do not justify a wholesale reclassification of anonymized data. As explained by Jane Yakowitz, a legal scholar critical of the types of arguments presented in the petition, the risk of privacy harm from re-identification is significantly lower than many risks we take without concern, such as throwing out our trash.¹¹ Felix Wu, a professor at Benjamin N. Cardozo School of Law, claims that there is not much support for the “strongly pessimistic view” that no useful data can be anonymous.¹² He explains that “[a] closer look at the computer science ... reveals that several aspects of that literature have been either misinterpreted, or at least overread, by legal scholars.” This appears to be what has happened here.

B. Carriers already take reasonable steps to anonymize data

The risk of statistically anomalous individuals falling prey to targeted de-identification attacks can be significantly reduced using common sanitization techniques. It is common practice for data holders to employ techniques such as generalization, replacing individual attributes with a broader category, and suppression, removing statistically anomalous data altogether. With these and other techniques, those controlling the data can ensure that no

¹¹ See Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARVARD J. OF L. & T. 1, 40, 2011.

¹² Felix T. Wu, *Defining Privacy and Utility in Data Sets* 84 U. OF COLO. L. REV. 1118, 1124.

combination of indirect identifiers (like the movie ratings in the Netflix case) would identify a group less than a certain threshold.

The petition quotes the Netflix study for the general conclusion that “removing identifying information is not sufficient for anonymity.”¹³ This statement is non-controversial, but also not sufficient to show an alarming risk of re-identification. The question is how we control indirect identifiers in data, especially where easily obtainable auxiliary data matches up with those identifiers. It is common knowledge that steps must be taken to anonymize data that go further than simply removing identifiers like names or phone numbers. Just as some researchers have found new techniques to re-identify data, others are working hard on countermeasures to prevent re-identification. In addition, carriers who make data available to third-parties may require data use agreements that prevent re-identification.

Carriers are already taking these steps to make their data as anonymous as possible while preserving utility. AT&T’s privacy policy states that they “remove data fields (such as name, address and telephone number) that can reasonably be used to identify you” and that they also “use a variety of statistical techniques and operational controls to anonymize data.”¹⁴ It is very likely that other carriers take similar steps to protect privacy, reducing the risk of re-identification.

¹³ Petition at 7.

¹⁴ AT&T Privacy Policy FAQ, Aggregate and Anonymous Data, *available at* <http://www.att.com/gen/privacy-policy?pid=13692#menu>.

C. Data innovation holds great promise and is in the public interest

Advances in computing technology are unlocking opportunities to collect and analyze data in ways never before possible. Data collected through mobile phones may have many important uses for consumers. For example, recent studies have shown large-scale mobile phone data can help us to better understand traffic patterns and design our road networks to minimize congestion.¹⁵ We are only just beginning to understand the promises these changes hold – regulators should be cautious and allow these developments room to grow.

Respectfully submitted,

/s/ Doug Brake

Doug Brake
Daniel Castro
Information and Technology & Innovation Foundation
1101 K Street NW Suite 610
Washington, DC 20005

¹⁵ See Pu Wang, *et al. Understanding Road Usage Patterns in Urban Areas*, Scientific Reports, Dec. 20, 2012, available at <http://www.nature.com/srep/2012/121220/srep01001/full/srep01001.html>.