Before the

FEDERAL COMMUNICATIONS COMMISSION

Washington, D.C. 20554

In the Matter of	)	
	)	
A National Broadband Plan for our Future	)	GN Docket Nos. 09-47, 09-51, 09-137
Comments – NBP Public Notice #29	)	
	)	
	)	

COMMENTS OF THE CENTER FOR DEMOCRACY & TECHNOLOGY: APPENDIX B

Applying Privacy by Design Principles to the Data Minimization FIP Principle and to Sensitive Data

I. Using Privacy by Design to implement the Data Minimization FIP principle

Adherence to the Data Minimization FIP principle requires careful attention to the principles of Privacy by Design. The data minimization principle requires that attention is paid to end-to-end lifecycle protection of data; data practices must take into account which data will be collected, how long it will be retained, and in what form it will be stored. Below, we describe in greater detail the considerations that companies and government agencies that are handling individual-level data should remain aware of as they seek to implement the Data Minimization FIP principle in a manner consistent with Privacy by Design.

A. Understanding data types

The Data Minimization principle requires that entities only collect data "that is directly relevant and necessary to accomplish the specified purpose(s) and only retain [that data] for as long as is necessary to fulfill the specified purpose(s)."[1] A determination of which data fulfills these criteria should hinge, in part, on the degree of data identifiability necessary to fulfill the specified purpose. If knowing that an individual user is a 20-year-old male is sufficient to the purpose at hand, then the data should be rendered pseudonymous, the user's name replaced with a unique ID. But if the goal is simply knowing the geographic breakdown of a site's visitors (what percent come from New York? What percent come from Kansas?), then no individual-level data needs to be maintained – aggregate statistics about users' location are sufficient.

---

[1] *See* U.S. Department of Homeland Security, *Privacy Policy Guidance Memorandum, The Fair Information Practice Principles: Framework for Privacy Policy at the Department of Homeland Security* (Dec. 2008), *available at* http://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf ("DHS FIPs").

Distinctions between data types have long been over-simplified, with vast quantities of data split into two supposedly distinguishable bins: one that holds “personally identifiable information” (“PII”) and one that holds “non-personally identifiable information” (“non-PII”). PII has traditionally consisted of direct identifiers such as name or Social Security Number. Non-PII constitutes almost anything else. But the rules and the assumptions on which the distinction between PII and non-PII is constructed need to be reevaluated. Research has consistently shown that information can appear “de-identified” (and “anonymous”) when alone, but when combined with other data, can help construct an identifying image of an individual. There have been a number of high profile examples of such re-identification. More then a decade ago, then-MIT graduate student Latanya Sweeney used Massachusetts residents' ZIP code, birth date, and gender - all found in public voter rolls - to identify individuals whose "anonymized" hospital records had been publicly released;[2] in 2006, AOL’s infamous release of “de-identified” search terms led to the identification of individuals based on their search history. Meanwhile, researchers have shown that supposedly “anonymized” data released by Netflix was anything but anonymous.[3]

Recognition is finally growing in some quarters that information can identify an individual even absent the individual’s name or Social Security number. In its 2008 staff report on online behavioral advertising, the FTC included within the scope of its behavioral advertising principles “any data collected for online behavioral advertising that reasonably could be associated with a particular consumer or computer or other device,” regardless of whether the data is “personally identifiable” in the traditional sense.[4]

CDT believes this phrasing represents a significant change in the discourse. More broadly, collected data should be evaluated on a spectrum that ranges from identifiable data to pseudonymous data to aggregated data. Principles that guide data collection, protection, and use practices should appropriately reflect the pseudonymity of the data.

In 2009, CDT worked with companies and other advocacy organizations in our Internet Privacy Working Group (IPWG) to establish a workable and specific vocabulary to describe how data is stored and used online. Below, we present a set of definitions to measure data identifiablity that is based on IPWG’s work.[5] We are confident that definitions like these can help move the discourse in a direction that is better aligned with reality and research. These definitions measure data identifiability from the perspective

---

[2] *See* Sweeney, Latanya *Recommendations to Identify and Combat Privacy Problems in the Commonwealth Before the H. Select Comm. on Information Security*, *Statement of Latanya Sweeney, Associate Professor, Carnegie Mellon University* (Oct. 2005) *available at* http://privacy.cs.cmu.edu/dataprivacy/talks/Flick-05-10.html

[3] *See e.g.* Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, IEEE Symposium on Security and Privacy (2008), *available at* http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf; Michael Barbaro & Tom Zeller, Jr., *A Face is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES, Aug. 9, 2006, *available at* http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=1&scp=1&sq=a%20face%20is%20exposed%20for%20AOL%20searcher&st=cse (AOL incident highlights the difficulties in making data truly anonymous).

[4] *See* Federal Trade Commission Staff Report, *Self-Regulatory Principles For Online Behavioral Advertising: Behavioral Advertising Tracking, Targeting & Technology* (Feb. 2009), *available at* http://www.ftc.gov/opa/2009/02/behavad.shtm.

[5] *See* Center for Democracy & Technology, *Threshold Analysis for Online Advertising Practices* 16 (Jan. 2009), *available at* http://www.cdt.org/privacy/20090128threshold.pdf.

of the entity collecting and using data for online advertising (as opposed to an outside observer or statistician, for example). How easy or hard it may be for such an entity to use data to identify an individual depends on the other data sources available to the entity, the capabilities of the entity, and the time, effort, and cost required to identify individuals. Note that all inferably identifiable data is pseudonymous, but all pseudonymous data is not necessary inferably identifiable.

- Aggregate data – Data about multiple individuals that cannot reasonably be used to directly or inferably identify any single individual.
- Directly identifiable data – Data that directly and overtly identifies an individual, such as name, address, email address, phone number, government identifier, or financial identifier.
- Inferably identifiable data – Data from which an individual's identity can be reasonably inferred, including combinations of data elements or data sets that would not, on their own, identify an individual. All inferably identifiable data is pseudonymous.
- Pseudonymous data – Data associated with a unique identifier that does not directly identify an individual.

1. "Aggregate" data

Information about an individual that has been aggregated with information about others is often difficult, if not impossible, to re-associate with that individual. But if this aggregate data, known as tabular data, is subdivided into a sufficient number of categories and contains information about a small enough number of people, then information about specific individuals could reasonably be parsed out of the data set. A number of publications have detailed best practices for ensuring that individual information within tabular data is sufficiently protected, and they should be used as guidance for those who compile tabular data. In their 2008 statement entitled "Data Access and Personal Privacy: Appropriate Methods of Disclosure Control, the American Statistical Association recommends a subset of these guidelines."[6] Of these, a 2005 paper by the Federal Committee on Statistical Methodology (OMB) stands out. This paper lays out appropriate tests for determining if individuals whose information is stored in tabular data are at high risk of being identified and describes methods for managing and merging data such that data quality is preserved and privacy is better protected.[7]

---

[6] *See* American Statistical Association, *Data Access and Personal Privacy: Appropriate Methods for Disclosure: A Statement by the American Statistical Association* (Dec. 6, 2008). *available at* http://www.amstat.org/news/statementondataaccess.cfm.

[7] *See* Federal Committee on Statistical Methodology, *Statistical Policy Working Paper 22 – Report on Statistical Disclosure Limitation Methodology* (2005), *available at* http://www.fcsm.gov/working-papers/spwp22.html

B. Principles to guide data minimization

The collection of large amounts of individual-level data and the accumulation of this data in directly identifiable or pseudonymous form long after it is no longer useful presents some of the greatest privacy risks for consumers.

As mentioned above, in 2009, CDT submitted comments to the OMB on their proposed revision to the federal policy on Web tracking technologies. In these comments, we outlined a set of Data Minimization principles that should guide federal agencies that collect individual-level data for measurement purposes.[8] We have updated these principles to make them applicable to corporate entities that collect individual-level data for measurement purposes as well as for those who collect this data for other purposes.

Entities collecting individual-level data and their commercial partners should take the following steps in connection with limiting data retention:

- **Purpose correlation.** *Only data directly relevant and necessary to accomplish a specified purpose should be collected and individual-level data should only be retained for as long as is necessary to fulfill a specified purpose.*

- **Immediate deletion.** *Elements of individual-level data logs that are not relevant to the analyses proposed in the manner specified by the Purpose Specification FIP principle should be deleted as soon as possible after the data is collected. If, for example, the entity has promised to pseudonymize data, IP addresses should be deleted (and possibly replaced with their corresponding geographic or ISP information) soon after collection, if not immediately.*

- **Disclosure**. *Data retention time frames should be published in privacy policies.*

- **Technical enforcement.** *Expiration time frames for cookies and other technologies that store data on users' computers should be set to match, not exceed, the data retention time frames adopted by the entities who place the cookies.*

- **Partner contracts.** *If an entity contracts with a commercial partner who will collect measurement data, place cookies, or otherwise collect or store individual-level information, the data retention time frames that apply to individual-level data collected by the partner should be explicitly stated in the contract.*

---

[8] *See* Center for Democracy & Technology, *Comments Regarding the Office of Management and Budget's Proposed Revision of the Policy on Web Tracking Technologies for Federal Web Sites* (August 2009), *available at* http://www.cdt.org/privacy/20090810_omb_cookies.pdf.

Using these principles to guide implementation of the Data Minimization FIP principle will promote consumer privacy in a manner that does not interfere with innovation.[9]

II. Applying Privacy by Design to sensitive data

The services made possible by broadband Internet create and collect a wide range of data and much of it can be extraordinarily sensitive in nature. For example, people tell search engines things they would not tell their friends, spouses, or therapists. They store information about their diet, exercise, medicines, and illnesses in Personal Health Records, information they might not tell their bosses or insurance companies. Meanwhile, banking, mortgage, and tax information is increasingly shared with cloud-based applications, while smart phones serve as accurate tracking devices – crumbs of zeroes and ones mark exactly which stores, offices, and residences each customer visits, data that third-party applications are all too eager to mine.[10]

CDT believes that the collection and use of sensitive data by companies operating in the online space necessitates an extra level of protection; it is imperative that companies that plan to collect or use sensitive data pay particularly close attention to the Privacy by Design principles, and use these principles to guide their implementation of FIPs. This is especially important with respect to location information, a relatively new and rapidly expanding data type that promises great benefits, but also incurs substantial privacy risks, for consumers.

As Congress and federal agencies work to establish rules for managing information collected online and deliberate about what extra protections certain information deserves, they will need a precise definition for the term "sensitive data." We believe that the term "sensitive data" should be defined, at a minimum, to include:

---

[9] The data retention elements outlined in this set of principles would represent an improvement over the FTC's current data retention principle, at least in the online behavioral advertising space. With respect to online behavioral advertising, the FTC only recommends that "companies should retain data only as long as is necessary to fulfill a legitimate business or law enforcement need." This is inadequate. The FTC's version of the principle does not guard against unanticipated uses because data retention is not tied to the purpose for which the data was collected in the first place. *See* Federal Trade Commission Staff Report, *Self-Regulatory Principles For Online Behavioral Advertising: Behavioral Advertising Tracking, Targeting & Technology* (Feb. 2009) at 47, *available at* http://www.ftc.gov/opa/2009/02/behavad.shtm.

Indeed, it is not the FTC but a company that recently pushed the market toward a higher standard for data retention. Yahoo! recently made changes to its data retention policy so that the company now removes directly identifiable data and some inferably identifiable user log data after three months. Yahoo's decision was based on its determination that the purpose for which the data was initially collected would not be served by data more than three months old. Three months might be a suitable retention limit for some data; the research showing that the data most relevant for marketing purposes is a mere 24 hours suggests that shorter periods may be appropriate for marketing data. Since Yahoo's announcement, Microsoft has also announced that it will reduce the amount of time that it stores IP addresses. *See e.g.,* Walaika Haskins, *Yahoo Pledges to Forget You Sooner*, TECHNEWSWORLD (Dec. 17, 2008), *available at* http://www.technewsworld.com/story/65545.html?wlc=1255717445; Kevin J. Obrien, *Microsoft Puts a Time Limit on Bing Data,* New York Times (January 19, 2010), *available at http://www.nytimes.com/2010/01/20/technology/companies/20search.htm.*

[10] Often, these applications use location data for the benefit of consumers. *See, e.g.,* Google Latitutde *available at* www.google.com/latitude; Loopt, *available at* www.loopt.com.

- Information about past, present, or potential future health or medical conditions or treatments, including genetic, genomic, and family medical history information of an individual;
- Financial information about an individual;
- Information about an individual's sexual behavior or sexual orientation;
- Social Security Numbers or any other government-issued identifiers;
- Insurance plan numbers;
- Information indicating the precise geographic location of an individual when he or she accesses the Internet.[11]

A. Location Information and the Dawn of the Location Enabled Web

The ubiquity of increasingly high-powered mobile devices has already spawned the Internet's first generation of location-based services and applications. As the accuracy of location data improves and the expense of calculating and obtaining it declines, location may well come to pervade the online experience. While the increasing availability of location information paves the way for exciting new applications and services, the increasingly easy availability of location information raises significant privacy concerns that have not yet been adequately addressed.

1. Definitions to guide discussions of location Information

The IPWG Threshold Analysis also establishes workable and specific definitions for the types of location information that are commonly created and collected online. The location definitions are based on terminology used by technical standards bodies that focus on location information and privacy, most notably the Internet Engineering Task Force (IETF) Geographic Location/Privacy Working Group.

> Civic location data – Data that describes the geographic location of an individual in terms of a postal address or civic landmark. Examples of such data are room number, street number, street name, city, ZIP+4, ZIP, county, state, and country. The precision of this data can be reduced by removing elements (for example, the precision of the combination of city, state and ZIP can be reduced by only using state).
>
> Geodetic location data – Data that describes the geographic location of an individual in a particular coordinate system (for example, a latitude-longitude pair). The precision of this data can be reduced by specifying a geographic area of particular spectrums rather than a point (for example, a circle with a 300 meter radius centered at 40° North, 105° West). However, the limits of such a precision

[11] Both in the following section and in CDT's 2009 Threshold Analysis for Online Advertising Practices, we subdivide location data into five categories: civic location data, geodetic location data, mobile location data, fixed location data, and nomadic location data. We believe that all mobile location data, fixed location data, and nomadic location data is sensitive in nature. Geodetic location data and civic location data can be sensitive depending on the precision o the data. Center for Democracy & Technology, *Threshold Analysis for Online Advertising Practices* 16 (Jan. 2009), *available at* http://www.cdt.org/privacy/20090128threshold.pdf.

specification can be circumvented by repeatedly sampling an individual's geodetic location.

Mobile location data – Civic or geodetic location data that identifies the whereabouts of an individual or his or her device in real or near-real time.

Fixed location data – Civic or geodetic location data that describes a fixed location associated with an individual. Examples include a home or office location.

Nomadic location data – Civic or geodetic location data that identifies the whereabouts of an individual using a device that may be moved occasionally from its fixed location. For example, if an individual occasionally uses his or her laptop at an Internet cafe, the location of the laptop would be considered nomadic.

2. Special considerations for sensitive location information

Because individuals often carry their mobile devices with them, location data may be collected everywhere and at any time, often without user interaction, and it may potentially describe both what a person is doing and where he or she is doing it. Location information can also be highly identifiable: for many people, there is one location where they spend their daytime hours (at work) and one location where they spend their nighttime hours (at home). After a day or two of collecting just those two data points about a person, it becomes fairly obvious whom those data points describe.

The year 2009 saw the dawn of the location-enabled Web, as all of the major browser vendors began integrated location awareness into their browsers. For example, with the release of the iPhone 3.0 software, the latest version of the Safari web browser running on the iPhone is now location-enabled. This means that any Web site can ask Safari for the user's location, and Safari can provide it by using the location positioning technologies built into the phone (including GPS, among others). Apple has implemented a simple interface (based on a draft of a W3C standard) that Web sites can use to request location. Firefox, Opera, and Chrome are now all providing similar functionality.

The browsers provide strong baselines for consent to location sharing. On the iPhone, each Web site that wants to use location has to first obtain the user's permission not once, but twice. Those permissions are reset every 24 hours. This is a good example of "Privacy as the Default," one of Cavoukian's seven foundational principles for Privacy by Design.[12]

But in terms of providing more granular control and transparency, the browsers are lacking. Given the privacy interests at stake and the relative lack of protection in the law, we would expect location controls to be better than other kinds of technological controls on the Web, to offer users more choices about what happens to their data and to be especially transparent about when location data is being passed around. For example,

---

[12] Anne Cavoukian, *Privacy by Design: The 7 Foundational Principles* (August, 2009)*, available at* http://www.ipc.on.ca/images/Resources/7foundationalprinciples.pdf.

much like the “lock” icons that indicate a secure connection, an icon could be displayed on the browser whenever Safari is transmitting location data. A similar regime for location data could encourage good practices from application providers.