**BEFORE THE PUBLIC UTILITIES COMMISSION OF THE**

**STATE OF CALIFORNIA**

| | |
|---|---|
| Order Instituting Rulemaking to Consider Smart Grid Technologies Pursuant to Federal Legislation and on the Commission's Own Motion to Actively Guide Policy in California's Development of a Smart Grid System | Rulemaking 08-12-009<br>(Filed December 18, 2008)<br>Phase III Energy Data Center |

# M E M O R A N D U M

**To:** Participants of Working Group organized pursuant to Administrative Law Judge's Ruling Setting Schedule To Establish "Data Use Cases," Timelines For Provision Of Data, And Model Non Disclosure Agreements, from Rulemaking Proceeding No. 08-12-009

**From:** Electronic Frontier Foundation and the Samuelson Law, Technology & Public Policy Clinic at the University of California, Berkeley, School of Law

**Date:** April 1, 2013

**Re: Technical Issues with Anonymization & Aggregation of Detailed Energy Usage Data as Methods for Protecting Customer Privacy**

## INTRODUCTION

This memorandum is one of two memoranda offered by the Electronic Frontier Foundation (EFF) and the Samuelson Law, Technology & Public Policy Clinic at the University of California, Berkeley, School of Law to aid in Working Group discussions outlined in Judge Sullivan's February 27, 2013, titled *Administrative Law Judge's Ruling Setting Schedule to Establish "Data Use Cases," Timelines for Provision of Data, and Model Non-Disclosure Agreements*, No. 08-12-009 ("Ruling"). This memorandum addresses the technical issues surrounding aggregation and anonymization of customer data. The other memorandum covers particular privacy rules and laws that apply to the disclosure of energy consumption data.

Thus far, this proceeding has established basic principles and a targeted framework—in the form of the Rules Regarding Privacy and Security Protections for Energy Usage Data

("Privacy Rules"),[1] adopted by the California Public Utilities Commission ("Commission") in D. 11-07-056 ("2011 Decision")[2] and set forth in Attachment D to that Decision—for managing customer data collected by smart meters. This proceeding has already established the serious implications for privacy in the home that come from releasing customer energy consumption data.[3] Accordingly, the Privacy Rules adopted by the Commission govern the release of "covered information:" customer usage data that can identify the customer or be re-identified after some identifying information has been removed. The Privacy Rules are discussed in further detail in our companion memo *Legal Considerations for Smart Grid Energy Data Sharing* regarding applicable law.

In this next phase, the proceeding aims to implement the Privacy Rules and other relevant legal requirements, in part by devising effective, secure protocols for manipulating customer energy data so that it can be shared with third parties without unduly compromising customer privacy. We offer this memorandum to help the Working Group understand the practical realities of known aggregation and anonymization techniques in light of computer science research demonstrating the characteristics of these techniques in protecting customer privacy, including their limitations. We also explain the need to involve technical experts working in the fields of data privacy and re-identification in order to develop protocols that effectively protect customer privacy and provide useful data to researchers.

This phase of the proceeding has thus far focused its attention on protecting privacy through anonymization and aggregation techniques. Unfortunately, a known set of technical problems that come with these techniques can make them highly vulnerable to re-identification of individual households or ratepayers included in the data set. While the terms "anonymization" and "aggregation" have not yet been clearly defined in the proceeding,[4] individual methods that have been discussed—including the "15/15 Guideline," zip code aggregation, and census-tract aggregation—are all vulnerable to these threats.

---

[1] *Rules Regarding Privacy and Security Protections for Energy Usage Data*, in *Attachment D*, Decision Adopting Rules to Protect The Privacy And Security of the Electricity Usage Data of the Customers of Pacific Gas & Electric Company, Southern California Edison Company, And San Diego Gas & Electric Company, Rulemaking 08-12-009 (July 29, 2011) [hereinafter Privacy Rules].

[2] Decision Adopting Rules to Protect The Privacy And Security of the Electricity Usage Data of the Customers of Pacific Gas & Electric Company, Southern California Edison Company, And San Diego Gas & Electric Company, Rulemaking 08-12-009 (July 29, 2011) [hereinafter 2011 Decision].

[3] Decision Adopting Rules To Protect The Privacy And Security Of The Electricity Usage Data Of The Customers Of Pacific Gas And Electric Company, Southern California Edison Company, And San Diego Gas & Electric Company. D. 11-07-056.

[4] *See* Ruling No. 08-12-009 at section titled "Definitions."

The first Working Group is expected to discuss various threshold definitions, including definitions for "aggregate" and "anonymous" data. The Working Group has also been charged with proposing standards for data anonymization and aggregation that "ensure the anonymity of data, protect customer privacy, and prevent the reverse engineering of the aggregated data."

In order to effectively engage with these tasks, Working Group participants first need to consider existing and ongoing research in the computer science community. To help with this task, we have consulted with technical experts in the field, and requested analysis from them. As part of this analysis, we are pleased to attach as Appendix A to this memorandum a paper titled *Privacy Technology Options for Protecting and Processing Utility Readings*, written as background for the Working Groups by computer security and privacy expert George Danezis. Unfortunately, analysis of the existing research demonstrates that existing techniques for anonymization or aggregation of data, taken alone, are insufficient protections for customer privacy. Anonymizing data (removing identifiers) and aggregating data (processing data and releasing only sums or patterns) have proven inadequate for protecting customer privacy because attackers and researchers can manipulate these data sets to re-identify individuals. As the Privacy Rules explicitly limit the release of data that can be re-identified, these proven workarounds must be taken into account when deciding what protocols to put in place for protecting customer privacy.

Accordingly, to devise the appropriate measures for protecting customer privacy without the risk of data re-identification, we believe that it is critical for the Working Groups to consult technical experts to help develop more robust solutions, beyond mere aggregation and anonymization (see, for example, the suggestions under "Robust Privacy Technology Options" in Appendix A). More robust solutions will help to prevent re-identification of "covered information," as required by the Privacy Rules, and to provide researchers with useful data that contributes to valuable energy research.

## DISCUSSION

### A. Disclosure of the Detailed Customer Energy Consumption Data Collected from Smart Meters Creates Serious Risks to Customer Privacy.

Since the late 1980s, scientists have reported the ability to derive detailed behavioral information about a household or other premise from electrical meter readings.[5] For example, Non-intrusive Appliance Load Monitoring (NALM) "use[d] temporally granular energy consumption data to reveal usage patterns for individual appliances in the house."[6] These usage patterns revealed, for example, time away from one's home, cooking and sleeping habits, or the number of inhabitants in a particular household. Not long after its development in 1989, scientists described this technology as capable of remotely identifying patterns based on externally available meter information. In a 1989 paper, NALM creator George Hart simultaneously noted that identifying these patterns created the potential for invasions of private information.[7] By tracking the daily energy usage of a household, it is possible to create a consumption profile and deduce behavior for that household.[8] It exposes not only energy consumption patterns overall, but also intimate behavioral information that most customers would not suspect is being shared, including travel, sleeping, and eating patterns, occupational trends, and even detailed information such as when children are home alone.[9] This type of profiling is attractive for a number of purposes, from behavioral research to marketing. For an example of such consumption profiling used in the retail industry, Target Corporation used data on women's shopping habits to develop a pregnancy detection method so reliable that it often

---

[5] According to one employee of Siemens Energy:

> We, Siemens, have the technology to record [energy consumption] every minute, second, microsecond, more or less live. From that we can infer how many people are in the house, what they do, whether they're upstairs, downstairs, do you have a dog, when do you habitually get up, when did you get up this morning, when do you have a shower: masses of private data.

Quote from Martin Pollock of Siemens Energy in Gerard Wynn, "Privacy Concerns Challenge Smart Grid Rollout" Reuters, June 25, 2010, *available at*: http://uk.reuters.com/article/idUKTRE65O1RQ20100625.

[6] Jennifer M. Urban, *Privacy Issues in Smart Grid Deployment*, at 6-7, in SMART GRID AND PRIVACY (forthcoming 2013).

[7] Hart, George W. (1989), 'Residential Energy Monitoring and Computerized Surveillance via Utility Power Flows', *IEEE Technology and Society Magazine,* 8 (2), 12-16 at 13; F. Sultanem (1991), "Using Appliance Signatures for Monitoring Residential Loads at Meter Panel Level," *IEEE Transactions on Power Delivery,* 6 (4), 1380, 1381, col. 2 (showing load graphs of various appliances and a fluorescent light). The reader can find a lay introduction to NALM technology in Quinn, Elias L. (2009) 'Privacy and the New Energy Infrastructure', *Social Science Research Network,* 09 at 21-25.

[8] D. 11-07-056.

[9] *Id.*; *See also,* Presentation of Chris Vera at January 15 workshop (slides available at ftp://ftp.cpuc.ca.gov/13011516_EgyDataWorkshop).

allowed for targeted advertisements before a woman had even revealed her pregnancy to others.[10] Similar predictive algorithms can be used to extend noticeable trends in energy consumption data, such as using real-time data to determine when an occupant is at home for solicitation by the utility or some third party. To continue with family formation as an example, an occupant's consumption profile might indicate a new baby in the house. This would violate the home occupants' privacy and create risks of leaking personal information that the customer had not even considered exposed in the first place.[11]

Working Groups will need to consider both existing profiling capabilities and those that are likely to arise in the near future. More recent scientific research on techniques for ascertaining information from energy data describes the developing ability to discern what video content is being viewed on a television or computer monitor. Known as "use-mode detection," this method relies on collecting energy data in real time. Lab scientists tested multiple television sets to determine that the content viewed on those devices left uniquely identifying energy signatures, known as electro-magnetic interference (EMI). The same video content would produce the same repeatable EMI traces, even across different television sets. Under laboratory conditions, researchers were able to identify 1200 movies at a 92% accuracy rate by reviewing these trace EMI patterns.[12]

Given the present and developing abilities to use energy data to detect appliance usage, discern regular household habits, and review the in-home consumption of video content or online information, the Working Groups must implement protections that guard such personal information and align with the requirements of the Privacy Rules.

## B. Known Limits to Anonymization and Aggregation as Methods for Preventing Re-identification and Protecting Privacy.

As described further below and in Appendix A, scientists now recognize that aggregating or anonymizing data to sufficiently prevent re-identification of an individual is almost impossible. As such, instead of relying directly on these techniques, instances of re-identification have prompted new efforts among computer science and privacy experts to "balance the risks

[10]Presentation of Ashwin Machanavajjhala at January 15 workshop (slides available at ftp://ftp.cpuc.ca.gov/13011516_EgyDataWorkshop).
[11] Presentation of Lee Tien, EFF at January 15 Workshop (slides available at ftp://ftp.cpuc.ca.gov/13011516_EgyDataWorkshop)
[12] Jawurek, et. al., "SoK: Privacy Technologies for Smart Grids – A Survey of Options" at 5, *available at* http://research.microsoft.com/pubs/178055/paper.pdf.

and value of data sharing in a de-identification regime."[13] Existing and developing re-identification capabilities must inform the Working Group's decisions on the dynamic definitions of aggregated/anonymized data to give privacy-protecting protocols any value.

In this section, we summarize for the Working Group some of the research shared in the workshops and previous proceedings, from consulting with experts, and from scientific literature, showing that these techniques fail to effectively protect customer privacy, and that data that have been anonymized or aggregated remain subject to the Privacy Rules, which cover all information about the customer that is "reasonably re-identifiable." For more detail, please see George Danezis' analysis in Appendix A.

### *1. Anonymization*

Anonymization techniques attempt to protect anonymity of data subjects by removing personal identifiers, such as names and addresses, from the data. Although anonymized data do not, on their own, point to specific individuals, numerous examples demonstrate that re-identification can be achieved by comparing anonymized data with external information that contains corresponding data points. See, for example, Appendix A, which offers the example of cross-referencing a customer's load profiles against external information about that customer's occupancy, allowing someone to re-identify the individuals referenced in the data.[14] It explains that a customer's (sometimes public) travel schedule, mobile phone location records, or even a short period of observation of the customer's house might be enough external information to match the anonymized load profile to a particular utility customer.

As evident in the case studies below, the removal of key identifiers, such as the data subject's name, address and birthdate, is insufficient to protect customer privacy.

#### *a. Examples: Netflix and AOL Research Datasets*

Professors Jennifer Urban and Ashwin Machanavajjhala both noted the Netflix Prize privacy breach at the January workshop. Netflix offered a prize for the contestant who could develop the best algorithm for matching users to films and released anonymized, customer-specific data to get them started. University of Texas-Austin researchers Arvind Narayanan and

[13] Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," 57 UCLA Law Review 1701 (2010); Jane Yakowitz, "Tragedy of the Data Commons" (March 18, 2011). Harvard Journal of Law and Technology, Vol. 25, 2011. Available at SSRN: http://ssrn.com/abstract=1789749.
[14] George Danezis, *Privacy Technology Options for Protecting and Processing Utility Readings*, Mar. 1, 2013, p. 3.

Vitaly Schmatikov, however, combined the data with available information from the Internet Movie Database, allowing them to re-identify users.[15] This brought Netflix under legal process and the scrutiny of the FTC; ultimately, Netflix chose not to pursue further similar competitions.

Professor Machanavajjhala also highlighted a privacy breach experienced by AOL as a further example. In 2006, AOL decided to publish search logs, containing user search queries, to help researchers communities improve searching algorithms. AOL user IDs were replaced by random numbers. No names or other traditional identifying information was included with the search queries. Within two hours, researchers were able to reveal a photograph of a particular user, based on review of the search queries. The fact that the anonymization attempt was broken in only two hours demonstrates how trivial it would be for an attacker to identify specific households within an "anonymized" energy usage data set with a small amount of external information about that customer's energy consumption. Disclosure of supposedly anonymized data for energy research purposes, such as to multiple third parties to assess energy efficiency programs, could create similar problems for the utilities, the Commission, or researchers, highlighting the need to address these risks in developing data protocols.

*b. Example: Massachusetts Government Health Data*

Professor Machanavajjhala additionally noted the Massachusetts government breach involving medical information. In 1997 the Massachusetts government began making "anonymized" health records of state employees available to researchers. Patients' names, addresses, and SSNs were removed from the health records, which otherwise remained intact. The governor assured his citizens that it would be impossible to re-identify individual patient information. Within two days, an MIT graduate student was able to identify the Governor's health records by cross-referencing them against voter registration records. She mailed the Governor's health records to him in an envelope.[16]

Professor Machanavajjhala referred to data points shared with data from external sources—like the voter registration records the researcher used here—as "quasi-identifiers" because they can identify an individual, but require comparison with other data sets in order to

[15] Arvind Narayanan and Vitaly Shmatikov "Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)," Feb. 5, 2008, U. Tex. at Austin, *available at* http://arxiv.org/pdf/cs/0610105v2.pdf.

[16] Erica Klarreich, "Privacy by the Numbers: A New Approach to Safeguarding Data," in *Scientific American*, at 1 December 31, 2012 (*available at* http://www.scientificamerican.com/article.cfm?id=privacy-by-the-numbers-a-new-approach-to-safeguarding-data) (Hereinafter Klarreich)

do so. In the energy world, a number of other data points could qualify as quasi-identifiers, including sets of appliances, devices, or vehicles, patterns of appliance usage, sleep patterns, and potentially a variety of other information. At the January workshop, some presentations included intentions to compare energy data to external sources, such as state-wide and county assessor maps, as well as data on building characteristics.[17] Knowing that researchers seeking anonymized energy use data intend to combine that data with additional information sources highlights the need for Working Group members to take seriously the potential risk to utility customer privacy that could occur via re-identification techniques.

*c. Example: Amazon Purchase History*

In 2011, researchers showed that it is possible to determine an online shopper's personal purchase history simply by studying the displays on Amazon.com's product recommendation feature. The researchers noticed that the aggregate-level statements—"Customers who bought this item also bought A, B and C"—changed over time, based on a shopper's own purchase history. By cross-referencing the product recommendations with customers' public reviews of purchased items, the researchers could successfully infer that a particular customer had bought a particular item on a particular day, even before the customer had posted a review of the item.[18]

Energy data similarly changes over time, allowing for noticeable patterns to appear. Unique energy signatures become personally identifying characteristics when compared to external information with shared data points. In addition, many of the same characteristics, such as name, address, birthdate, etc., are collected by utilities, as were in the Massachusetts government health data breach or by online service providers like Amazon, Netflix, and AOL. Further, many of these characteristics are available to the public on other databases, making it possible to identify an individual through linking other data.

These examples, among others, explain why anonymizing data by removing a few key identifiers unfortunately does little to prevent re-identification. In some cases, it was only a matter of hours before data considered "anonymized" was cross-referenced with external data and re-identified, compromising the data subject's privacy. As such, data that has been "anonymized" is often easily re-identifiable. Accordingly, data that has been processed with

[17] *See* Presentations of Lauren Rank, Mike McCoy, and Paul Matthew from January 15 workshop. (slides available at ftp://ftp.cpuc.ca.gov/13011516_EgyDataWorkshop)
[18] Klarreich at 3.

these types of anonymization techniques, without additional protective steps, would still be considered "covered information" under the Privacy Rules. As a result, it can only be released with consent or otherwise pursuant to the Rules, and without additional steps in place, could expose customers to re-identification risks

### *2. Aggregation*

The use of the term "aggregated data" has not been consistent throughout this proceeding. Based on the scientific literature in this area, we understand aggregated data not to include micro-data—i.e., the underlying, discrete records about individuals from which the aggregation is derived. Unlike attempts to anonymize data, for example by removing certain identifiers from individual records, aggregating data requires processing it such that there are no individual-level records, for example by computing the sum or the average of a group of individual households' energy usage information. For our purposes, "aggregated data" would not include the total annual or average annual energy usage for an individual household, precisely because the data pertains to a specific household.

Despite excluding micro-data, aggregated data can still leak private information. Traditional privacy protections for aggregation, such as the 15/15 Guideline, are sometimes referred to by computer scientists as "naïve aggregation rules" because of the uncomplicated techniques for circumventing their restrictions.

To use an historical example, this one from as far back as World War II, it is now well-known that re-identification of naively aggregated Census Bureau data helped the U.S. military locate and transfer Japanese-Americans to internment camps during World War II. Although naïve aggregation was considered an acceptable privacy policy in the 1940s, today's Census Bureau employs a series of complex data-blurring techniques to promote data integrity but maintain heightened security in response to such re-identification risks.[19]

The 15/15 Guideline is the most prominent "aggregation" model in this proceeding.[20] Although burying an individual's data within a larger data set like this may seem like a reasonable means to protect privacy, the shortcomings of this approach are well documented.

---

[19] Douglas A. Kysar, Book Review, *Kids & Cul-De-Sacs: Census 2000 and the Reproduction of Consumer Culture*, 87 Cornell L. Rev. 853, 873-874 (2002) (footnotes omitted); *Id*. at n. 124.

[20] The 15/15 Guideline is a model that permits a database to generate query results, only if the results represent an aggregate data set consisting of 15 or more individual utility customers and no one utility customer in the set constitutes 15% or more of the total aggregated data.

Specifically, a carefully crafted series of queries can generate aggregate results that, when looked at together, reveal customer-specific information. A brief explanation of how queries can work around the limits imposed by the 15/15 Guideline is given below, followed by an example of the risks of cross-referencing aggregated data with external sources. Please see Appendix A for further discussion of data security issues with the 15/15 Guideline.

### a. *Likely Smart Grid Data Leaks from Naïve Aggregation Rules*

The 15/15 Guideline and similar well-intentioned standards unfortunately exhibit fundamental flaws that render them unable to effectively defend customer privacy. Numerous researchers have addressed how a combination of queries can enable the re-identification of individuals represented in aggregate data, even though neither query on its own infringes the individual's privacy. [21]

To illustrate, imagine a quantitative query system[22] under a standard like the 15/15 Guideline, which ignores requests when the number of results is less than a particular threshold. In such a case, one need only ask two questions that meet that threshold to obtain an answer otherwise forbidden by the rule:[23]

> *The first question:*
>
> How many people in this database exhibit power usage patterns consistent with using a television and video games in the afternoon, but patterns consistent with additional appliances, electric vehicles, and lights in the evening?

---

[21] Salil Vadhan, et. al. Comment on "Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators" HHS-OPHS–2011–0005 at 6.

> [In an] interactive system designed to answer queries about the health care expenses of the Harvard faculty, which allows queries of the form "how many Harvard faculty satisfy X" where X is a search criterion that can involve attributes like age, health care expenses, and department. While "how many" questions may seem relatively safe when computed over a population of 2000+ individuals, they are not. By asking the question "How many Harvard faculty are in the computer science department, were born in the U.S. in 1973, and had a hospital visit during the past year?," it is possible to find out whether one of the authors of these comments (S.V.) had a hospital visit during the past year (according to whether the answer is 0 or 1), which is clearly a privacy violation. A common "solution" to this sort of problem is to only answer queries whose answers are sufficiently large, say at least 10. But then, by asking two questions --- "how many Harvard faculty had hospital visits during the past year?" and "how many Harvard faculty, other than those in the computer science department and those born in the U.S. in 1973, had hospital visits during the past year?" --- and taking the difference of the results, we can obtain an answer to the original, privacy-compromising question.

[22] For example, how many individuals in this data set have characteristic X?

[23] Klarreich at 2.

*The second question:*

> How many people in this database who exhibit power usage patterns consistent with using a television and video games in the afternoon, but patterns consistent with additional appliances, electric vehicles, and lights in the evening, do not live at 100 Main Street?

Although both questions provide aggregated results, the combination of these two questions has effectively "leaked" information about 100 Main Street. The first question essentially asked for the total number of homes where children are likely to be home alone in the afternoon. The second question sought the same information but excluding 100 Main Street. If the answers to these two questions are the same, then one can reasonably infer that there are no latchkey children at 100 Main Street; if the answers differ by 1, then one can reasonably infer that there are. See Appendix A for further detail regarding problems with the 15/15 Guideline.

Unfortunately, it is very difficult for computer programs to detect the query combinations that breach customer privacy in advance.[24] Professor Machanavajjhala pointed out at the January workshop that energy data is dynamic, not static. If aggregated data changes, then individuals can be uniquely identified in ways that computers were not programmed to protect against. For example, if data shows a new house on the block, then an attacker can look at changes in the neighborhood's energy consumption and subtract the new information to attribute change to the new home.

Because this simple, two-query process for overcoming the 15/15 Guideline defeats its protective purpose, data masked in this manner is likely to remain re-identifiable. As such, like data that has been subjected to basic anonymization techniques, data aggregated according to these techniques would still be considered "covered information" under the Privacy Rules, and would expose customers to re-identification risks if released without additional protective protocols in place.

### *b. Attacks Using Pre-existing Information about an Individual*

If an attacker or researcher has background information about an individual represented in an aggregated data set, re-identification becomes even easier. For example, in 2008, a research team, led by Nils Homer, then a graduate student at the University of California at Los Angeles,

[24] Klarreich, at 2.

showed that in many cases, knowing a person's genome can help determine, beyond a reasonable doubt, whether that person had participated in a particular genome-wide test group.

Homer's research team demonstrated the risks of disclosing aggregate information from genome-wide association studies, one of the primary research vehicles for uncovering links between diseases and particular genes. These studies typically involve sequencing the genomes of a test group of 100 to 1,000 patients who have the same disease and then calculating the average frequency in the group of something on the order of 100,000 different mutations. If a mutation appears in the group far more frequently than in the general population, that mutation is flagged as a possible cause or contributor to the disease.[25]

After Homer's paper appeared, the National Institutes of Health reversed a recently instituted policy that had required aggregate data from all NIH-funded genome-wide association studies to be posted publicly.[26] In this example as in others, the comparison of supposedly "safe" data to external, background data led to re-identification.

Energy data is susceptible to the same sorts of attacks on other types of personal data. If an attacker knows the unique combination of appliances that a utility customer has in their kitchen, he can examine aggregate energy usage patterns to determine if the data signature corresponding to that combination of appliances fits the aggregate profile, which would lead to an inference that the customer was or was not included in the data.

Accordingly, with certain background information and data manipulation, data aggregated according to these techniques, as well, can easily be re-identified—especially as researchers, marketers, or others combine datasets—and would still be considered "covered information" under the Privacy Rules.

The Working Groups will need to consider carefully protocols to protect energy usage data in order to find methods that take attacks like those we have described into account. As noted next, we believe specific technical expertise is required in order for the Working Groups to sufficiently consider the issues and develop appropriate approaches.

[25] Klarreich at 2–3.
[26] Klarreich at 3.

### C. Technical Expertise Is Required to Develop More Robust Privacy Solutions Because Anonymization and Aggregation Techniques Alone Fail to Protect Private Customer Data

We hope this background is helpful to the Working Groups. As made clear during our analysis and in the examples above, when devising protocols for the disclosure of customer data, Working Group participants should be aware that neither aggregation nor anonymization can be defined or evaluated in static terms if privacy is to be protected. Re-identification is a dynamic concept. Each time there is an influx of publicly available data, an advance in computer technology, or additional collection of personally identifying characteristics, re-identification strategies will evolve. This means that the techniques required for the "safe" release of smart grid data will likely also change. Any definitions adopted by the Working Groups will need to accommodate this reality. In order to do this, the Working Groups need to consult experts in the fields of computer science, consumer privacy, and data security at each stage of developing data disclosure procedures, in order to understand the unfortunate, but genuine challenges in securely sharing data and to develop feasible solutions that overcome the known shortfalls of anonymization and aggregation.

### D. Summary and Next Steps

In summary, we hope this memorandum has supplied the Working Group with useful background information to move forward in this proceeding, acknowledging that:

- Both scientific research and live, real-world examples show that basic techniques for anonymizing or aggregating data do not by themselves provide sufficient protections to customer privacy.
- Unfortunately, the 15/15 Guideline and similar well-intentioned aggregation standards cannot be relied on to protect customer specific data because of simple workarounds that neither human beings nor computer programs can reliably predict.
- The dynamic nature of energy data and the constantly developing technologies for de-identification and re-identification should each be considered by the Working Groups in developing definitions and proper disclosure procedures.

- ❖ Consultation with technical experts in is necessary at all stages of this proceeding to determine:
    - o What types of data can be released or should not be released under the requirements of the Privacy Rules;
    - o What privacy solutions have been shown from experience to adequately or inadequately protect customers' private information; and
    - o What feasible solutions can the Commission use to impart sufficiently robust protections of customer privacy while still providing useful energy data for valuable research purposes. (See, for example, the suggestions under "Robust Privacy Technology Options" in Appendix A.)

Respectfully submitted this April 1, 2013 at San Francisco, California.

/s/ Jennifer Urban

JENNIFER URBAN, Attorney
Samuelson Law, Technology & Public Policy Clinic
University of California, Berkeley School of Law
396 Simon Hall
Berkeley, CA 94720-7200
(510) 642-7338
Attorney for ELECTRONIC FRONTIER FOUNDATION

/s/ Lee Tien

LEE TIEN, Attorney
Electronic Frontier Foundation
454 Shotwell Street
San Francisco, CA 94110
(415) 436-9333 x102
Attorney for ELECTRONIC FRONTIER FOUNDATION

# Appendix A

# PRIVACY TECHNOLOGY OPTIONS FOR PROTECTING AND PROCESSING UTILITY READINGS

George Danezis
Paris, Friday, 1 March 2013

## SCOPE OF THE DOCUMENT

This document discusses the privacy concerns surrounding the collections and processing of granular readings from next generation utility architectures, such as smart electricity grids. New generation distribution systems rely partially on computerised meters installed in households and businesses that record more information than previous electromechanical meters, and have facilities to transmit them regularly to the energy operators and distributors. A modern smart meter is capable of recording consumption of electricity, as well as production, at a very fine granularity, close to "real time." Most deployments in the US[27] and Europe[28] are presently working toward readings every 15 minutes to 30 minutes respectively (48 or 96 readings per day) uploaded as a single "load profile" about once a day. These are collated with other readings from the same household to build larger load profiles over months or years. This document is concerned with the management and privacy of those detailed readings – other information such as billing details, demographics and subscriber information are broadly similar to information already gathered and benefit from established processes to ensure their security and privacy.

The management of the electricity grid is special, compared to water and gas, in that production and consumption has to be balanced very carefully at all times. Some production requires significant planning to start or stop, and the use of renewables adds uncertainty as to the capacity. These make forecasting and demand response mechanisms important. On the other hand, gas and water provision is also undergoing computerization in its control and distribution, since better recording of consumption could be used to optimize the delivery of those services (like detect leaks). Those attempting to manage privacy issues in smart grids, and the regulatory and technical solutions applied, should therefore foresee that they will create a precedent for the management of other utility data. Furthermore those undertaking privacy impact assessments for managing and processing utility readings should be mindful that combined readings from all utilities may be available at some point, providing a multi-dimensional view into household habits.

[27] Guidelines for Smart Grid Cyber Security: Vol. 2, Privacy and the Smart Grid. National Institute of Standards and Technology. NISTIR 7628., August 2010.
[28] Smart metering implementation programme data access and privacy consultation document. United Kingdom Department of Energy and Climate Change, Consultation Document, April 2012.

Readings and load profiles have direct and indirect uses. They are used directly by the energy industry to monitor and balance production / consumption, forecasting energy needs in the short and long term data, plan for future distribution capacity, and bill customers at a coarse or fine granularity. Where the energy sector is private and competitive, meter readings are also used to settle contracts in the energy market. Billing customers according to the time they consume electricity is particularly promising to provide incentives to reduce consumption at peak time, and is generally called time-of-use tariffs.

Indirect uses are also foreseen for detailed readings for both research and operations: they can be used for monitoring and providing advice on energy efficiency of homes and devices, understand penetration of smart vehicles in different areas, insurance, marketing of renewables, risk management of credit, etc. These are indirect uses since they are not vital for the day to day operation of electricity provision, and may not be performed by the traditional players in the energy industry. In fact, indirect uses are of great interest since they may create new services, or optimize and economically "disrupt" existing ones. Research is a particularly important area that requires data, and by its very exploratory nature, it might require more access than an operational system.

The focus of this document is to provide an overview of technical and other options that support processing of the meter readings to support both direct and indirect uses, and their benefits, while minimizing the exposure of the readings and providing protection of the privacy of households, businesses and government agencies making use of modern grid technologies.

## OVERVIEW OF THREATS

Fine grained meter readings recorded by smart meters from households are widely recognized as privacy sensitive. NIST[29], in the US, recommends they are processed as PII (Private Identifiable Information) and jurisdictions with horizontal data protection regimes (Canada and the EU) consider that load profiles fall under their provisions[30]. Substantively, detailed smart meter reading provide a record of activity from within a household that might otherwise be difficult to infer. This activity might be sensitive for occupants. We outline here a number of possible privacy and security threats resulting from the collection and mining of readings:

- Meter readings at the granularity of 15-30 minutes can be used to infer the occupancy of a home, since aggregate half-hourly consumption goes when one is at

[29] Guidelines for Smart Grid Cyber Security: Vol. 2, Privacy and the Smart Grid. National Institute of Standards and Technology. NISTIR 7628., August 2010.
[30] Opinion 12/2011 on smart metering. Article 29 Decision, April 4 2011.

home. They leak information about when occupants may be away on holiday, at work or not. As a result compromised readings contain information that could be used to target homes for burglary when they are empty. Interestingly, one of the earliest cases of widespread indirect use of meter readings involved inferring occupancy to detect safe houses of German terrorists[31]. This particular practice was later deemed unconstitutional by German courts.

- Similarly, granular readings can be used to estimate the number of inhabitants at a particular time. Third parties also profile inhabitants in relation to their family situation: for example to discover whether a spouse is working or not. Houses shared by multiple unrelated occupants also exhibit a different pattern of electricity consumption than houses inhabited by a single family.
- Detailed smart meter readings contain information about the sleeping patterns of inhabitants, which can be surprisingly intrusive. Sleeping patterns are associated with specific religious groups: comparatively early morning activity in the months of Ramadan is a sign of a practicing Muslim household. Erratic patterns of sleeping are also indicative of poor health: irregular use of electricity at night may be indicative of early stages of prostate cancer. A change in the use of electricity (for frequent washes) as well as night time patterns of use may indicate to a third party a household with a young child.
- Non-intrusive appliance monitoring[32] techniques detect which appliances are in a home, and when they are used, from fine grained readings of a whole household. While the frequency of readings in current smart-metering deployments is too coarse for a direct application of those techniques, it is clear that some information on appliances, such as the presence of an electric vehicle, a fridge, air-conditioning, or an electric oven can be inferred. It is noteworthy that modern smart meters can be configured, even remotely and without the knowledge of the household, to take readings at a finer granularity. More recent studies have demonstrated under laboratory conditions that electricity consumption can even leak information about which TV channel is being watched[33].
- Even more intrusive information can be inferred when combining electricity with other utility readings, for example water and gas readings. Such combined readings can be used to detect different patterns of cooking in a household, since cooking activity exhibits correlated uses of electricity, gas and water. Similarly, the frequency of use of a dishwasher or washing machine can be inferred. Finally, the combined use of large volumes of water along with either gas or electricity can be

[31] B. S. Amador. The federal republic of Germany and left wing terrorism. Master's thesis, Naval Postgraduate School, Monterey, CA, December 2003.

[32] G. W. Hart. Residential energy monitoring and computerized surveillance via utility power flows. IEEE Technology and Society Magazine, June 1989.

[33] M. Enev, S. Gupta, T. Kohno, and S. Patel. Televisions, video privacy, and powerline electromagnetic interference. In Proceedings of the 18th ACM conference on Computer and communications security, pages 537–550. ACM, 2011.

used to infer how often members of the household have showers. Electricity and water provides information about night time patterns of sanitation, and even how often and when inhabitants use the toilet overnight.

Besides the above sample privacy threats, the rationale for storing and processing of meter readings is the extraction of some level of information about a consumer. As such any argument about the value of meter readings at the granularity of a household becomes an argument about potential privacy invasion, as the information originates from, and characterizes, a household. In line with fair information practices[34] this information should only be used with the knowledge and consent of the household, to ensure their best interests are at the heart of any indirect processing.

Besides legal or substantive privacy concerns, smart meter deployments have been jeopardised partly through the poor handling of customer privacy and protection concerns. For example, the smart meter deployment in the Netherlands[35] had to be put on hold due to consumer revolt.

As a result of the above we consider there are serious risks associated with the bulk storage, processing and availability of detailed utility meter readings. First of all, organizations holding such data can be compromised, or lose the data due to mishandling. This is a serious threat to consumers, and the reputation of the entity that that is a victim of a cyber-attack or a mistake. Organizations holding data may also be compelled to reveal the readings they hold, though the legal process of countries they operate in. In some jurisdictions even divorce or private dispute cases can lead to organizations being compelled to reveal information about their customers. Finally, organizations themselves may be tempted to process the readings to gain an unfair advantage in their commercial dealings with customers.

## PARTIAL SOLUTIONS AND CAVEATS

A number of solutions are popular to mitigate the perceived risks of handling and processing detailed meter readings. In particular opt-in/opt-out mechanisms, anonymization, and naïve aggregation rules are popular due to their conceptual ease, and relative low cost of implementation. Despite being valuable parts of a larger strategy, in themselves, these mechanisms cannot guarantee the level of protection one would hope for the privacy of readings and households.

[34] FTC Fair information practices (http://www.ftc.gov/reports/privacy3/fairinfo.shtm)

[35] Cuijpers, Colette and Koops, Bert-Jaap, Smart Metering and Privacy in Europe: Lessons from the Dutch Case (February 15, 2013). In: S. Gutwirth et al. (eds), European Data Protection: Coming of Age, Dordrecht: Springer, pp. 269-293 (2012).

### OPT-IN/OPT-OUT

Both guidelines for processing PII in the US (fair information processing practices) and data protection regimes consider that, where possible, the informed consent of the data subjects should be sought for any otherwise non-necessary processing. The UK regulator DECC[36] has proposed a gradual system of consent to enable processing of increasingly invasive data: the provision of one reading a month per household is absolutely necessary and therefore obligatory; the provision of a reading per day is subject to customer opt-out, but in its absence collection and processing can go ahead; finally any finer grained processing (as for computing time-of-use tariffs) requires an explicit opt-in from the customer.

The requirement to obtain consent for collection and processing is in itself positive, particularly for indirect uses of readings, where a customer may not have reasonably foreseen it. Yet, it does not alleviate all risks: despite consent to collect and process, readings are still sensitive, and could still be lost or compromised. Therefore some technical protection is still necessary to ensure this sensitive information is stored and processed to minimize its exposure to external or internal risks. Furthermore once bulk readings are available in clear it is difficult to audit what they are used for, to ensure that only authorised processing is taking place.

Finally, a key limitation of solely relying on opt-in as a privacy protection is purely economic. In case time-of-use tariffs become the norm, and added value services relying on energy readings are commonplace, households opting out will find themselves marginalized or possibly unable to benefit from the best prices for the goods and services they receive. Therefore they will be faced with a harsh choice of either opting into a system with poor privacy or being charged a premium for opting out. For this reason it is important to consider additional technical privacy protections even for customers opting in advanced services.

### ANONYMIZATION

One option for minimizing the danger to households, from the processing of any private information is to first anonymize it. Anonymization[37] removes any personal identifiers from the data in an attempt to make it difficult to link it back to a specific individual or household. Anonymization is an extremely flexible mechanism: full load profiles over time are available to researchers and any function can be computed on them. Sadly, robust

[36] Smart metering implementation programme data access and privacy consultation document. United Kingdom Department of Energy and Climate Change, Consultation Document, April 2012.

[37] C. Efthymiou and G. Kalogridis. "Smart grid privacy via anonymization of smart metering data." 2010 First IEEE International Conference on Smart Grid Communications, pages 238–243, 2010.

anonymization of load profiles is extremely difficult due to this abundance of data on one side, and the abundance of side information on the other.

Firstly, household energy consumption is rather regular over time. This means that the availability of a short period of non anonymized data can be used to link anonymized load profiles back to the household[38]. Concretely this means that an entity that has a short period of readings from a household, for example a month, can use those readings to pick a longer anonymized load profile related to the same household. To do this, a number of markers would have to be extracted from the raw identified load profile, such as the presence of certain household devices, number of occupants, typical patterns of occupancy related to the schedule of inhabitant's work, school or recurrent appointments. Then the anonymized profiles can be sieved according to the same markers, looking for a match. Different households may be susceptible to this matching to different degrees but some, with very stable unique markers, will be trivially re-identifiable.

Secondly, detailed load profiles are correlated with activities in the home that may be known, public or discoverable by others. Thus markers can be constructed to match other activities linked with specific individuals with anonymized load profiles. Any side-information associated with occupancy can be used[39]: public traffic schedules, a short period of direct physical observation of the home, mobile phone location records or internet access records can be used to construct markers. Thus anyone in the possession of such data sets can create an approximation of a load profile over time, and then attempt to match it with the database of anonymized load profiles. This technique is likely to be much more successful than the previous one, since it does not rely on regularity of habits over time.

For the sake of clarity we present a concrete de-anonymization attack using side-information:

- Consider an on-line web service, like webmail, on which a known target user has an account and checks periodically both from home and outside the home.
- The service logs contain a time series of accesses, and the network address (IP address) of these accesses. The network address leaks whether the user is at home or outside the home, through differentiating between a home internet service provider and a mobile or business internet service provider. Using a different computer at home than at work, can also be leveraged to mount the re-identification attack.

[38] M. Jawurek, M. Johns, and K. Rieck. "Smart metering de-pseudonymization." In ACSAC, pages 227–236, 2011

[39] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin. "Private memoirs of a smart meter." In Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, BuildSys '10, New York, NY, USA, 2010. ACM.

- The service is then provided with a large number of anonymized electricity load profiles, and wishes to re-identify a target user. To achieve this, the service makes the reasonable assumption that a user at home consumes more electricity than a user outside the home.
- For each anonymized trace the services computes this simple statistic: it adds the readings corresponding to times the target user was observed at home, and subtracts the readings when the target user was observed outside the home.
- The anonymous trace corresponding to the target user should achieve a high value of this statistic – ultimately the highest value.

This is the result of the actual trace matching perfectly the observations of occupancy, while other traces being partially independent of it. The more side-information the service has about the user, meaning more accesses to the on-line service, the better the estimation of the statistic and the more confident it can be the de-anonymization attack will be successful. This example illustrates that mounting a de-anonymization attack against an anonymized load profile is computationally cheap, and the side information required only needs to be vaguely related to occupancy – and as such is plentiful and in the hands of many third parties.

De-anonymization techniques may be new in the field of smart-grids, but general techniques are already very mature in related fields of statistical databases privacy or social network privacy. Recently, researchers have demonstrated the inherent dangers of publishing rich anonymized datasets: they managed to de-anonymize a number of users from a dataset of movie preferences published by the Netflix Company using side information from other public sources[40]. In that work they used particular combinations of movie preferences attached to known persons as "markers", and then detected those markers in the anonymized data set to link it to individuals.

Thus, anonymization through the mere removal of obvious identifiers is now recognized as a very weak privacy protection mechanism[41]. It could be used to protect load profiles from mistakes or accidental disclosure, but it is fundamentally a mechanism to keep honest people honest. It cannot protect against a malicious entity that, for example compromised the dataset and is trying to identify specific households.

### NAÏVE AGGREGATION RULES

In terms of flexibility another option, besides anonymization, involves providing an "aggregation service" that computes aggregate statistics on specific data items on request,

[40] Narayanan, Arvind, and Vitaly Shmatikov. "How to break anonymity of the netflix prize dataset." arXiv preprint cs/0610105 (2006).

[41] Ohm, Paul. "Broken promises of privacy: Responding to the surprising failure of anonymization." UCLA Law Review 57 (2010): 1701.

and returns only the aggregate results. The hope is that aggregation obscures information about individual households, alleviating privacy concerns. Rules are put in place to ensure each datum is computed on the basis of many households and rounding or suppression can be used to obscure items that do not conform to the rule. One such example is the so-called "15/15 Guideline" that stipulates that at least 15 households are involved in any aggregate.[42]

Sadly there is an extremely mature[43] and rich[44] literature outlining generic attacks against systems that provide the facility to query datasets and return statistics in a naïve manner, despite complex sanitization rules. It has been shown that special queries (called "Trackers") can be crafted, each conforming to the rules, but jointly leaking private information.

Building a tracker for the 15/15 rule is simple. The rule stipulates that a query can only be performed if it concerns a certain minimum number of households: an analyst can submit a query that concerns a large number of specific households (say 1000); then a second query over the same households plus an additional one (namely 1001 records) is performed. The result of the two queries jointly leaks all information about the record that was included in the second query, despite the fact that the queries are compliant with the 15/15 rule. Furthermore, one can show that it is very expensive to audit for sets of queries that are crafted to leak information about single records: one would have to consider the potential leakage of all subsets of queries – and the number of these subsets is very large indeed.

Thus, while allowing querying of a database of records provides flexibility, it has to be supported with great care to ensure no information about individual households is leaked. Positive guarantees of security and privacy must be proven for any sanitization rule to ensure that tracking queries cannot be crafted to extract information.

## ROBUST PRIVACY TECHNOLOGY OPTIONS

Privacy protection through procedures or technology is an exercise in risk management that has to balance the benefit of processing the data and the potential privacy risk to households. It is important to note that the benefits of indirect processing may in fact not directly benefit households. Therefore regulators must be very cautious to ensure those benefiting from the processing do not choose alone what constitutes an acceptable risk. In many cases, technology can help to minimize risks, while also maximizing benefits, and thus privacy does not have to be a zero-sum game. A privacy-by-design methodology can

[42] Audrey Lee, Marzia Zafar. "Energy Data Center". Briefing paper. September 2012.
[43] Denning, Dorothy E., Peter J. Denning, and Mayer D. Schwartz. "The tracker: A threat to statistical database security." ACM Transactions on Database Systems (TODS) 4.1 (1979): 76-96.
[44] Adam, Nabil R., and John C. Worthmann. "Security-control methods for statistical databases: a comparative study." ACM Computing Surveys (CSUR) 21.4 (1989): 515-556.

be applied to identify the privacy issues throughout the development of a smart-metering system[45], and appropriate privacy technologies can be deployed to support privacy policies[46].

### Sampling Load Profiles, anonymizing & licencing

The first, mostly procedural, option for processing detailed readings is to establish a scheme to provide sampled anonymized load profiles to clearly identified, authorized and overseen researchers for pre-determined uses. In that case anonymization is used to ensure that data leaks do not happen accidentally. A high sampling rate, of say one household in 100-1000 could be used to ensure that any compromise would not leak a very large volume of information, and that any specific target household for which there might be a lot of information is not likely to be in the set of load profiles available for analysis.

Yet, providing anonymized data under a licence or an NDA is not a perfect protection, and some household may have valid reasons to object to taking this risk. It is worthwhile considering explicit opt-in from households for use of load profiles in indirect processing for research through such a scheme. To be fully honest consent should be obtained under the assumption the sharing of the data is not fully anonymized, and possibly financial incentives should be provided to participating households.

On the technical side, getting data under licence should be accompanied with a robust audit of an organizational operations and technical procedures to ensure the security of that data. This should include secure authentication, storage, transport, audit, deletion mechanisms and an ownership structure that ensures the data will be processed according to the licence.

This mechanism is ideally suited for advanced R&D that requires access to full load profiles for exploration. It might also be used to perform computations as part of operations, when complex calculations need to be performed on full load profiles.

### Aggregation & Query privacy

The workhorse of most processing is likely to be access to aggregates and statistics based on a number of load profiles. For example, it is legitimate to monitor the aggregate consumption per region, changes over time, or even extract "average" load profiles for researching tariff structures or to train forecasting models. All those uses require readings only as a means to aggregating them into statistics, and not to make decisions on individual

[45] "Operationalizing Privacy by Design: The Ontario Smart Grid Case Study." Information & Privacy Commissioner, Ontario, Canada. February 2011.
[46] "Smart Meters in Europe: Privacy by Design at its Best." Ann Cavoukian, Ph.D. Information and Privacy Commissioner, Ontario, Canada. April 2012.

households. A number of privacy technologies allow access to those aggregates without making available detailed readings.

To compare to the naïve aggregation rule architectures, architectures that allow secure privacy friendly aggregation rely on a centralized party (or parties) holding the readings, and accepting queries to be performed on the data. Once the query is performed the answer is returned, possibly with some slight modification to ensure that information is not leaked. Queries can be pre-registered and data streams for each query can be produced ahead of time and made available to third parties in real-time.

For simple aggregation, involving sums and weighted sums, a very high degree of privacy can be provided through the use of appropriate encryption technologies[47] [48]. Meter readings can be stored encrypted, thus preventing even the storage service from accessing them in detail. Queries are performed on the encrypted readings, for example to compute encrypted sums over time or space, and returned to the relying services. Special encryption techniques can be used that "unlock" the results of queries to uncover the results, without giving access to any individual readings, with the help of a set of authorities overseeing the privacy policy. This architecture ensures that only the final aggregate result is available to anyone processing the readings. No one has access to raw readings, neither the storage service, nor the authorities nor the party receiving the result. Queries can be overseen by authorities for compliance to any policy, or to ensure they are appropriately rate limited to avoid exposing too much information to the any single entity.

Some aggregation is more complex than simple weighted sums. For example non-linear operations might have to be performed on readings before they are aggregated. In those cases the storage service needs to keep the readings in clear and process them to get the results. As we discussed, it is important to ensure no information can leak from specific or repeated tracker queries. One principled framework for achieving this is to ensure that statistics computed are differentially private[49], namely they are not overly influenced by the existence or absence of any single record, irrespective of the others (to protect against side information attacks).

We describe here two example mechanisms for ensuring an arbitrary statistic is differentially private:

---

[47] Klaus Kursawe, George Danezis, Markulf Kohlweiss: "Privacy-Friendly Aggregation for the Smart-Grid." PETS 2011: 175-191

[48] Marek Jawurek, Florian Kerschbaum: Fault-Tolerant Privacy-Preserving Statistics. Privacy Enhancing Technologies 2012:221-238

[49] Cynthia Dwork: A firm foundation for private data analysis. Commun. ACM 54(1): 86-95 (2011)

- The first differentially private mechanism is called "the *Laplacian* mechanism"[50]. One first computes the sensitivity of the statistic, as the maximum difference the inclusion or exclusion of any single item could make to the result of a query. Then some random noise is added to the result, drawn from a specific noise distribution, to mask any specific item, while providing information about the aggregate.
- The second mechanism is called "the *Subsample and Aggregate* mechanism"[51]. It is based on splitting a data set into smaller sub-sets; computing the statistic on each set; and then aggregating the result with some noise. Despite the fact the results are noisy, the average magnitude of the noise added is constant, therefore not overly influencing or biasing the result of queries on larger datasets.

The architecture of submitting queries to a service and getting back results, instead of processing load profiles locally, might be a departure from the habits of some researchers. In case few load profiles are processed a scheme based on licencing a sample of them may be preferable. Yet, in case large volumes of readings have to be processed, centralized processing in a data centre or private cloud may be the best option irrespective of privacy concerns. In that case the privacy-friendly architecture, that requires submitting queries to a service, aligns perfectly with the remote processing that would have to take place anyways, and is easy to add to existing computational models such as map-reduce[52]. Query based privacy mechanisms are highly scalable, and provide the ability to audit activity, and very flexible processing. There is no impediment to registering queries ahead of time, and receiving results in real time.

Privacy-friendly query systems can be made very privacy friendly. For simple statistics, they ensure that no single entity can ever get access to raw readings while providing real time access to aggregates and statistics. More complex computations require a storage service to store and process data in clear, but differential privacy mechanism ensure that the results cannot be used to infer much about any single household. They are also very efficient and scale to very large datasets.

### User Authorization & Data Export

Ultimately some who would make indirect uses of meter readings may prefer per-household detailed load profiles. In those cases none of the previous privacy technologies are applicable, since they rely on sampling or aggregation. In such cases the reading storage service can still incentivise a privacy friendly use of the data by third parties by managing user authorization of processing.

---

[50] Cynthia Dwork, Frank McSherry, Kobbi Nissim, Adam Smith: Calibrating Noise to Sensitivity in Private Data Analysis. TCC 2006: 265-284

[51] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In STOC, pages 75–84. ACM, 2007.

[52] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.

Conceptually, the storage service can manage the authentication of households to whom the data belongs, as well as services that wish to use the data. The storage service then ensures that permissions to access customer information have been granted by customers for each service. This is not dissimilar to the permission model used by modern mobile platforms (such as *Android* or *Windows Phone*) when an application wishes to access personal data from users. Social network platforms such as *Flickr* or *Facebook*, implement a similar authorization service for third party applications to access user feeds. Google dashboard also provides a model of an interface where a customer can go to manage their authorizations to applications, view and delete the results of computations. Providing such authorization and transparency mechanisms in one central place is highly advised.

Besides providing a well-defined API that allows third party services to access the data, after proper authorization and authentication from customers, the reading storage service can also provide to authenticated users their own household readings to use as they wish. In fact, one of the gravest challenges to privacy – in its information self-determination sense – is that a plethora of services may have access to customer information, when the customer does not. Besides providing access to raw readings, special cryptographic techniques can be used to ensure customer applications can process the data and compute results that can be used with third party services in a privacy friendly manner -- even without leaking the raw readings[53]. These facilities can be used, for example, to produce verifiable time-of-use bills on customer devices, without leaking the raw readings.
Any central store of information has a key role to play when it comes to facilitating and enabling a privacy friendly eco-system of applications. If it does not support core privacy services like private aggregation and queries, rich interfaces for authentication, authorization and data export it might block valuable applications due to privacy concerns, or force privacy invasive practices as the only option.

## DESIGN FOR PRIVACY

The generic privacy protections presented are quite flexible, but specific applications using electricity readings may have features that make them amenable to other mechanisms for protecting privacy. It is therefore important to include in any R&D program a component that looks at the most privacy friendly way to gain value out of data, and provide rich services.

Unlimited and full access to vast amounts of data and all load profiles in R&D is detrimental to the development of privacy friendly solutions in the long term. The assumption of unlimited availability of data leads to lazy design, where such access becomes a necessity.

[53] Rial, Alfredo, and George Danezis. "Privacy-preserving smart metering." Proceedings of the 10th annual ACM workshop on Privacy in the electronic society. ACM, 2011.

Limiting access of researchers to only small sample rich datasets for exploration, and then services for privacy friendly processing of bulk data, incentivises the design of both privacy friendly research methods but also privacy friendly final products, business models, and long term operations.

We have seen that for small focused exploratory research projects, mechanisms based on anonymization, sampling load profiles and opt-in can be used to provide researchers with high quality datasets. For the provision of statistics, privacy friendly query services can provide aggregates or results of arbitrary computations on very large datasets without leaking information about any household. Finally, a proper framework for authorization, authentication and data access by users can enable an ecosystem of privacy friendly third party applications. These facilitate competition, can enable privacy friendly alternatives, and allow the user to have control over who is processing their data as they do in other on-line services.

## SHORT BIO

George Danezis is a researcher at Microsoft Research on the topic of computer security and privacy. Before joining Microsoft in 2007 he was a visiting scholar at KU Leuven and a research associate at the University of Cambridge where he completed this PhD in 2004. George Danezis has been the program chair of the Privacy Enhancing Technologies Symposium (PETS) in 2005 and 2006, the conference on Financial Cryptography and Data Security in 2011, and the ACM conference on computer and communications security (CCS) in 2011 and 2012. He has published over 50 peer-reviewed scientific articles on the topics of privacy and security in international conferences and journals, and serves on the board of ACM CCS, PETS and ACM Information Hiding and Multimedia Security.