



FILED

05-13-13
01:34 PM

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
(Filed December 18, 2008)
Phase III Energy Data Center

ADMINISTRATIVE LAW JUDGE'S RULING ADDING TECHNICAL MEMOS TO THE RECORD, AND INVITING COMMENTS AND REPLIES; REVISING SCHEDULE FOR FILING USE CASES, COMMENTS AND REPLIES

1. Summary

This ruling adds two technical memos to the record of this proceeding and invites comments and replies on the matters addressed.

In addition, this ruling revises the schedule for the working group report and comment cycle established in Administrative Law Judge Ruling of February 27, 2013 in this proceeding.¹ The new schedule is as follows:

¹ *Administrative Law Judge's Ruling Setting Schedule to Establish "Data Use Cases," Timelines for Provision of Data, and Model Non-Disclosure Agreements (ALJ Ruling).*

Schedule	Date
Report on results of collaborative work groups, including use cases, filed and served	June 10, 2013
Opening comments	June 20, 2013
Reply Comments	June 27, 2013
Proposed Decision anticipated	August 2013
Commission Decision anticipated	September 2013

2. Technical Memos Prepared by Electronic Frontier Foundation

In conjunction with collaborative efforts to develop “data use cases” to ensure the timely provision of energy data to requestors of data interested in topics of policy interest to California ratepayers, utilities, and policy makers, the Electronic Frontier Foundation provided two memos to the service list in this phase of this proceeding on April 1, 2013.

The first memo is titled “Legal Considerations for Smart Grid Energy Data Sharing” and is Attachment A to this document. This memo states that it “covers legal background relevant to this proceeding, providing a brief explanation of important laws that apply to energy usage data sharing, as well as a brief background of the legal landscape covered in the proceeding to date.”

The second memo is titled “Technical Issues with Anonymization & Aggregation of Detailed Energy Usage Data as Methods for Protecting Customer Privacy” and is Attachment B to this document. The memo states that it “addresses the technical issues surrounding aggregation and anonymization of customer data.” The memorandum contains, as Appendix A, a paper titled *Privacy Technology Options for Protecting and Processing Utility Readings*.

In order to promote the development of the record in this proceeding, this ruling moves both items, which are attached, into the record of this proceeding. Furthermore, this ruling invites parties to respond to these memos through comments and replies.

Concerning the second memo, there appears to be an inherent tension between the technical memo, which focuses on the failings of techniques for protecting the privacy of data, and Appendix A, which proposes “Robust Privacy Technology Options.” In particular, this ruling invites comments on the “Laplacian mechanism” and “the *Subsample and Aggregate* mechanism” for incorporating “noise” from a specific noise distribution. How robust are these techniques for protecting the privacy of a particular statistic? Does the addition of “noise” dilute the power of subsequent statistical analyses of the data, or does the fact that the noise is generated by a known distribution enable adjustments that eliminate bias? What effect, if any, does the addition of noise have on the variance of statistical estimators? Is software to add “noise” readily available and commonly used? How costly is this software? In what settings should these mechanisms be used and where are they not needed?

In addition, this ruling invites comments on other techniques for protecting the privacy of data, including but not limited to, those discussed in the Appendix A to Attachment B.

Comments and reply comments addressing these two memorandums should be provided in a separate section of the comments and reply comments to the working group report, which was established in an ALJ Ruling on February 27, 2013. The next section of this ruling revises the due dates for these comments and reply comments

3. Schedule Revisions

The ALJ Ruling issued on February 27, 2013 set a due date of May 15, 2013 for a working group report that summarizes the results of the collaborative working groups. That Ruling also invited opening comments on June 5, 2013 and reply comments on June 19, 2013.

In late April, Pacific Gas and Electric (PG&E) suggested a two-month extension of this schedule. This suggestion triggered a wide range of responses. The Administrative Law Judge (ALJ), via e-mail, asked that parties weigh in on the schedule by noon, May 3, 2013.

San Diego Gas & Electric Company, Southern California Edison, The Utility Reform Network, the Division of Ratepayer Advocates, and Southern California Gas Company supported PG&E's request for a two-month extension.

PG&E made six arguments on behalf of its request in its e-mail of May 1, 2013 to the service list.

1. The interested parties have had several working group sessions, with excellent facilitation by Judge Hecht, and as a result have agreed that further small group working group meetings are necessary in order to provide more specific, precise recommendations on complex technical issues regarding how to define "personally identifiable information" and "anonymized" or "aggregated information." These small working group sessions have not yet taken place.
2. Since the last working group sessions, several parties have submitted revised or new "use cases" for discussion by interested parties. These new or revised "use cases" need time to be discussed by interested parties.
3. Following the small working group sessions, a further working group session of all interested parties is scheduled for the last week in May.

4. Separate from the small working group sessions, the investor-owned utilities (IOUs) are jointly developing a proposed protocol for “processing” and “streamlining” the review and response to data access requests. This joint IOU protocol still needs to be reviewed and discussed among all interested parties.
5. In addition to the technical issues, use case discussions, and IOU protocols, the interested parties still need to discuss the “Model Non-Disclosure Agreement” attached to your ALJ Ruling and seek consensus on the text and use of that model agreement. Also pursuant to the ALJ Ruling, the parties and IOUs also need to discuss the costs associated with processing and disclosing customer data, and who pays the costs.
6. Following these discussions, the IOUs need to take into account the results of the discussions and the interested parties’ views, and then draft and circulate for comment a consensus working group report, as required by the ALJ Ruling. PG&E expects that at least two rounds of consensus-building comments on the draft report may be required and possibly direct discussions among the parties to resolve any disputes regarding consensus language in the draft.

MEA stated that it did not oppose the extension request.

Distributed Energy Consumer Advocate (DECA), California Center for Sustainable Communities at UCLA, Energy Institute at Haas, SolarCity, and California Center for Sustainable Energy, Sunible and the Consumer Federation of California supported a much shorter extension instead. DECA also constructively points out that the detailed work of the collaborative groups should decrease the time needed to prepare comments on the report.

These arguments convince us that more time is needed, and this ruling revises the schedule to extend deadlines by approximately one month.

The one month extension should enable interested parties to work through May. The ALJ Ruling of February 27, 2013 expected that the parties would work collaboratively and accomplish much, but the ALJ Ruling did not expect that parties would reach a consensus on all use cases and on a uniform process for providing data. Parties should work to agree on as much as possible, but the working group report, now due on June 10, should, if no consensus is reached, simply report areas of agreement and disagreement. The ALJ Ruling anticipated that the comment cycle will enable parties to correct any errors made in the working group report and to provide the information needed for the Commission to decide matters, particularly those situations where no consensus emerged.

We therefore amend the schedule to allow more time for the work of the collaborative groups, to keep a comment cycle, and to produce a timely decision regarding open issues. We adopt a revised schedule as follows:

Schedule	Date
Report on results of collaborative work groups including use cases, filed and served	June 10, 2013
Opening Comments	June 20, 2013
Reply Comments	June 27, 2013

Proposed Decision anticipated	August 2013
Commission Decision anticipated	September 2013

IT IS RULED that:

1. The memo prepared by the Electronic Frontier Foundation titled “Legal Considerations for Smart Grid Energy Data Sharing,” which is Attachment A to this ruling, is incorporated into the record of this proceeding.
2. The memo prepared by the Electronic Frontier Foundation titled “Technical Issues with Anonymization & Aggregation of Detailed Energy Usage Data as Methods for Protecting Customer Privacy,” which is Attachment B to this ruling, is incorporated into the record of this proceeding.
3. Interested parties may address these memos and the questions posed in this ruling as part of their comments and replies pertaining to the working group report required by the February 27, 2013 *Administrative Law Judge’s Ruling Setting Schedule to Establish “Data Use Cases,” Timelines for Provision of Data, and Model Non-Disclosure Agreements* in this proceeding.
4. The schedule set forth in the February 27, 2013 Administrative Law Judge Ruling is revised. The working group report set forth in ruling paragraph 4 is now due on June 10, 2013. The comments invited in ruling paragraph 6 are now due on June 20, 2013. The reply comments invited in ruling paragraph 6 are now due on June 27, 2013.

Dated May 13, 2013, at San Francisco, California.

/s/ TIMOTHY J. SULLIVAN
Timothy J. Sullivan
Administrative Law Judge

ATTACHMENT A

**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
(Filed December 18, 2008)
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: Legal Considerations for Smart Grid Energy Data Sharing

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 the parties' discussions during the Working Group meetings outlined in Judge Sullivan's February 27, 2013 ruling, titled *Administrative Law Judge's Ruling Setting Schedule to Establish "Data Use Cases," Timelines for Provision of Data, and Model Non-Disclosure Agreements* ("Ruling").

This memorandum covers legal background relevant to this proceeding, providing a brief explanation of important laws that apply to energy usage data sharing, as well as a brief background of the legal landscape covered in the proceeding to date. The other memorandum, titled *Technical Issues with Anonymization & Aggregation of Detailed Energy Usage Data as Methods for Protecting Customer Privacy*, offers some technical background on aggregation and

anonymization models for protecting privacy.

The proceeding thus far has established both basic principles and a targeted legal 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. In 2012 the Privacy Rules were extended to customers of gas corporations, community choice aggregators, as well as residential and small commercial customers of electric service providers.³ It now presents an opportunity to apply this framework in establishing effective, secure protocols for more streamlined access to the rich and highly sensitive information captured by smart meters.

Following the Ruling, the Working Group is expected to discuss definitions of “aggregate” and “anonymous” data, as well as standards for achieving optimal aggregation or anonymization and reasonable protocols for sharing those categories of data. In order to fulfill these goals, Working Group participants must have the legal landscape on which we are operating firmly in hand. Further, understanding the legal contours of smart grid data sharing will enable more productive discussions of the validity and/or scope of the proposed “use cases” set out in the Ruling.

DISCUSSION

During this proceeding, the Commission has established that smart grid data can reveal a great deal of private information about life inside a premises, including: how many inhabitants are home or away at a given time; when those inhabitants go to bed, wake up, take showers, or cook dinner; and what devices inhabitants use, including personal medical devices.⁴ Known privacy and security risks include, among others:

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) [“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) [“2011 Decision”].

³ D. 12-08-045 (August 23, 2012).

⁴ See Statement 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>. See also

- Data breach (hacking) or data leaks (inadvertent disclosure to the public);
- Re-identification of aggregated and/or anonymized data to reveal personally-identifying information; and
- “Mission creep,” the potential future expansion of access to energy usage data to include additional users or uses of the data beyond what was initially contemplated (e.g., for law enforcement).

This proceeding has also already established the applicability of a variety of laws intended to protect Californians’ data privacy interests. Many of these laws are already discussed in the 2011 Decision and are reflected in the Privacy Rules. In the Privacy Rules phase of the proceeding and in his presentation at the January 15th Workshop, Chris Warner of Pacific Gas & Electric provided a list of the laws and regulations relevant to the collection, maintenance, use, and disclosure of smart grid data.⁵ Additionally, in its Opening Comment on the Proposed Energy Data Center (“EDC”), EFF raised questions regarding the applicability of existing state law, including the Information Practices Act of 1977 (“IPA”),⁶ to EDC proposals. Parties participating in the January 15th and 16th Workshops identified as the IPA as a relevant topic for further review.⁷

To aid this phase of the proceeding, this memorandum further discusses some of these laws as applied to the disclosure of customer energy usage data. Specifically, it briefly reviews the California Constitution, the Fair Information Practices Principles (“FIPPs”), and Public Utilities Code Section 8380 (commonly referred to as “SB 1476”) as important foundations for the Privacy Rules. It then provides further review of the IPA and its applicability to agency sharing of energy usage data. Finally, the memorandum reviews for the Working Groups the key provisions of the Privacy Rules themselves, which implement SB 1476, other relevant law, and the FIPPs for smart meter data. With a foundational understanding of these laws, the Working Groups will be better equipped to devise solutions for smart grid data sharing that comply with these existing laws.

Mikhail A. Lisovich, Deirdre K. Mulligan & Stephen B. Wicker, *Inferring Personal Information from Demand-Response Systems*, IEEE SECURITY & PRIVACY (Jan.–Feb. 2010).

⁵ *Appendix A: List of Current Statutes, Regulations, Decisions and Protocols Related to Customer Privacy Applicable to California Energy Utilities*, Attachment B from Ruling D. 11-07-056; Slide presentation by Christopher J. Warner, *Existing Energy Data Sharing Protocols: A Potential Consensus Approach*, CPUC Workshop (Jan. 15, 2013), available at ftp://ftp.cpuc.ca.gov/13011516_EgyDataWorkshop/.

⁶ Opening Comments of the Electronic Frontier Foundation, at 10–11 (Dec. 17, 2012) [hereinafter EFF Opening Comment].

⁷ Slide presentation by Christopher J. Warner, *Existing Energy Data Sharing Protocols: A Potential Consensus Approach*, CPUC Workshop (Jan. 15, 2013), available at ftp://ftp.cpuc.ca.gov/13011516_EgyDataWorkshop/.

Before commencing the Working Groups, participants should understand that these laws require us to propose definitions and implement “use case” solutions that are dynamic and adaptable. This is because the legal landscape governing data sharing varies—and can change dramatically—depending on a number of factors: (1) the identity of the data custodian; (2) the identity of the data requester; (3) the purpose of the data disclosure; and (4) the level of granularity of the data requested. The proposed use cases represent different permutations of these variables, so the law necessarily treats them differently. Understanding the legal obligations that attach to each data-sharing scenario will enable more accurate evaluation and more effective problem-solving.

A. California Law

1. The California Constitution

Article I, Section 1 of the California Constitution recognizes each individual’s right to privacy. There is general agreement among the judicial, scholarly, legislative, and regulatory communities that the data collected by smart meters reveals intimate details about the lives of California citizens. As such, the California Constitution establishes a baseline obligation to protect energy usage data from harmful disclosure or use.

The same interests that motivated California citizens to enact Section 1 by ballot amendment in 1972 still apply today: (1) the overbroad collection and retention of unnecessary personal information by government and business interests; and (2) the improper use of information properly obtained for a specific purpose, for example, the use of it for another purpose or the disclosure of it to some third party.⁸

Representative of the high value the California public places on privacy, the California Constitution imposes an obligation to protect consumer privacy on all parties—including private parties—engaging in smart grid data sharing. As such, addressing privacy issues are necessarily central to this proceeding, and Working Group participants should bear in mind adequate protections against unauthorized use or disclosure of personal information when addressing definitions and use cases.

/

⁸ *White v. Davis*, 13 Cal. 3d 757, 775 (1975).

2. *Information Practices Act*

The IPA (California Civil Code section 1798 *et seq.*) governs the manner in which state agencies, as defined in the IPA, disclose personally identifiable data that they collect and maintain. The statute applies to state-wide agencies, including the Commission and the California Energy Commission (CEC).⁹ Should the Commission designate one of these agencies as a custodian of smart grid data, the IPA will apply to that agency's disclosure of the data.

The IPA protects energy usage data that “identifies or describes an individual”—in this context, an individual utility customer.¹⁰ The IPA offers a non-exhaustive list of example types of “personal information” that might be used to identify or describe an individual, including an individual's “name, social security number, physical description, home address, home telephone number, education, financial matters, and medical or employment history.”¹¹ At the January Workshop, Professor Ashwin Machanavajjhala asserted that additional types of information, such as sex, birthdate, and zip code, operate as “quasi-identifiers,” capable of re-identifying an individual when linked to other available data. The IPA's open-ended list of identifiers would include that information as well.

As a general rule, state agencies are not permitted to disclose any personal information “in a manner that would link the information disclosed to the individual to whom it pertains.”¹² However, a number of exceptions apply, subject to varying protocols and approval procedures depending on the data recipient. For example, Section 1798.24 authorizes disclosure of an individual's personal data in the following pertinent scenarios, among others:

- With the prior written voluntary consent of the individual, Cal. Civ. Code § 1798.24(b);
- To persons, or another state agency, such as the CEC, for whom the information is necessary to fulfill statutory duties, Cal. Civ. Code § 1798.24(e);
- Where the CPUC is required by law to disclose the information to a local government (or federal government) entity,¹³ Cal. Civ. Code § 1798.24(f);
- Disclosure to a researcher, if (1) he provides assurance that the information will be used solely for statistical research or reporting purposes, and (2) he does not

⁹ Cal. Civ. Code § 1798.3.

¹⁰ Cal. Civ. Code § 1798.3(a).

¹¹ The IPA also includes “statements made by, or attributed to, the individual” within its list of identifiers. Cal. Civ. Code § 1798.3(a).

¹² Cal. Civ. Code § 1798.24.

¹³ We note that there are two separate exceptions relating to warrant and subpoena requirements.

receive the information in a form that will identify the individual, Cal. Civ. Code § 1798.24(h); and

- Disclosure to a researcher within the University of California system, provided that the request is approved by the Committee for the Protection of Human Subjects, Cal. Civ. Code § 1798.24(t).

Of particular relevance to Working Group discussion is Section 1798.24(h), which specifically addresses disclosure for research purposes. This provision underscores the California legislature's commitment to protecting the privacy of the individual(s) to whom the data pertains by explicitly limiting disclosure of personally identifiable information to researchers, while allowing research. We additionally note that Section 1798.24(e) also practically limits the scope of agency disclosures to only those specifically and directly authorized by statute, lest the exception swallow the rule.

One of the fundamental privacy concerns motivating the enactment of the IPA was the risk of data breach, a problem that is prevalent and well-documented among all institutions, including California institutions. An important obligation the IPA imposes on third party data recipients working within the University of California system is that requests for disclosure of personal information must first be approved by the Committee for the Protection of Human Subjects (CPHS), or another institutional review board that has written authorization from the CPHS. Although Section 1798(t) appeared in the original 1977 version of the statute, the specific language requiring approval from the CPHS was added in 2005 to ensure that the UC satisfies minimum standards for data security.¹⁴

This amendment responds to a high-profile computer hacking incident and data breach that occurred in August 2004, in which a UC Berkeley researcher inadvertently disclosed names, addresses, social security numbers, birthdates, and phone numbers for nearly 1.3 million people residing in California.¹⁵ Data breaches continue to plague the UC system, giving credence to the state legislature's concern about security protocols at public research institutions. For example, in December 2006, UCLA alerted approximately 800,000 current and former students, faculty,

¹⁴ See Stats. 2005, c. 241 (S.B. 13) § 1 (“The Legislature recognizes the research community has legitimate needs to access personal information to carry out research . . . the provisions of this bill are not intended to impede research but rather to require and set minimum standards for careful review and approval of requests.”).

¹⁵ EFF Opening Comment, at 11. See also Senate Bill Analysis, Third Reading, Stats. 2005, c. 241 (S.B. 13) (Aug. 17, 2005). In that case, the researcher requested data from the Department of Social Services (DSS) about participants in the In-Home Supportive Services (IHSS). Although the researcher needed only a random sample of IHSS data, the DSS made the entire IHSS database available for download. Shortly thereafter, a hacker broke into the researcher's computer system, causing a massive data breach.

and staff that a sophisticated computer hacker had broken into its systems and accessed a restricted database containing their personal information.¹⁶ More recently, in 2011, the UCLA Health System notified over 16,000 patients that their names, birthdates, addresses, and medical information had been stolen during the burglary of a physician's home.¹⁷ Although the physician had stored the data on an encrypted external hard drive, the password for the hard drive was written on a piece of paper kept near the computer that was found missing after the incident.

As such, the IPA provides both legal requirements binding on relevant agencies and overall guidance as to how California has thus far approached data risks for California citizens. Accordingly, although the IPA is not binding on utility companies, academic or local government researchers, or other parties who cannot be characterized as state agencies, it nevertheless provides useful guidance in this situation because it approximates how California law might treat the disclosure of energy usage data more generally.

B. The Privacy Rules

In the smart grid context, statewide concern in California with consumer privacy has culminated in the Commission's adoption of the Privacy Rules, which specifically address the sharing of energy usage data held by investor-owned utilities ("IOUs"). The Privacy Rules most directly address the type of data sharing at issue in this phase of the proceeding: (1) they specifically regulate energy usage data collected by smart meters, and (2) they concern disclosure by the IOUs to third party data requesters. As such, they provide the governing general authority on energy usage data sharing by the IOUs.

Accordingly, the Privacy Rules are the primary source of legal guidance as the Working Groups determine how to manage any disclosure of such data, and comprise the central feature of our discussion on relevant law. Part 1 of this section provides a brief background to the Privacy Rules, adopted in 2011, and their implementation of the provisions of SB 1476 and the FIPPs. This background provides a fuller understanding of the Privacy Rules for those participants not previously involved in the proceeding. Part 2 explains the standards and requirements for disclosure of covered information set forth in the Privacy Rules.

¹⁶ *UCLA Warns of Unauthorized Access to Restricted Database*, UCLA NEWSROOM (Dec. 12, 2006), <http://newsroom.ucla.edu/portal/ucla/UCLA-Warns-of-Unauthorized-Access-7571.aspx?RelNum=7571>.

¹⁷ *UCLA Medical Officials Say Patient Information Data Stolen*, L.A. TIMES BLOG (Nov. 4, 2011), <http://latimesblogs.latimes.com/lanow/2011/11/ucla-patient-identification-stolen.html>.

1. Brief Background to the Privacy Rules: SB 1476 and the FIPPs

In 2010 the California legislature passed **SB 1476**, now codified as Public Utilities Code Section 8380, to regulate the use and disclosure of utility customer data collected by smart meters. SB 1476 applies both to “electrical corporations and gas corporations.” Subject to some exceptions, SB 1476 generally prohibits disclosure of “electrical or gas consumption data . . . available as part of an advanced metering infrastructure, [including] the name, account number, or residence of the customer.”¹⁸ Under Section 8380 (b)(1) “an electrical corporation or gas corporation shall not share, disclose, or otherwise make accessible to any third party a customer’s electrical or gas consumption data, except as provided in subdivision (e) or upon the consent of the customer.” The Privacy Rules implement these restrictions and their exceptions with regard to the IOUs.

In addition to implementing the requirements of SB 1476, the Commission established that the sharing of energy usage data should follow **Fair Information Practice Principles** (FIPPs), a widely accepted international framework for handling electronic information in a privacy-protective manner. In the 2011 Decision, the Commission explicitly adopted the FIPPs as California’s policy for smart grid privacy. Thus, the foundational principles set forth in the FIPPs provide guidance to the Working Groups as participants determine how to most effectively implement the Privacy Rules.

The eight principles embodied in the FIPPs can inform privacy discussions in the upcoming Working Groups in a number of ways. For example:

1. *Transparency*: Any new repository of data that is separate from the IOUs would make it more difficult to provide notice to individual utility customers about the use or dissemination of their personal information
2. *Individual Participation*: The Commission should continue to use consent measures to involve individual utility customers in processes for data collection, use, dissemination and maintenance. Unlike typical consumers, many utility customers have no choice when buying energy. As a result, foregoing consent for disclosure is not bargained for in the relationship with the utility.
3. *Purpose Specification*: Requesting parties must be required to specify the purpose underlying the request prior to authorization for disclosure.
4. *Data Minimization*: Only the data actually necessary for the particular purpose identified should be disclosed. The FIPPs’ minimization principle helps in developing

¹⁸ Pub. Util. Code § 8380(a).

data handling practices that limit data breach and other risks before they happen, and helps data handlers decide on data needs in an efficient manner.

5. *Use Limitation*: There must be mechanisms to ensure that the disclosure of information is used solely for the specified purpose(s).
6. *Data Quality and Integrity*: If multiple parties were permitted to collect and store energy usage data, it would be harder to ensure that the data is accurate, relevant, timely, and complete. The problems associated with one data set may be multiplied across parallel data sets.
7. *Security*: Any data collected from the IOUs and stored pursuant to security protocols that are less rigorous than those utilized by the IOUs may be susceptible to loss, unauthorized access, destruction, modification, or unintended disclosure.
8. *Accountability and Auditing*: Mechanisms are already in place to enforce IOUs compliance with the FIPPs. It will be of utmost importance during the Working Groups to ensure that any other entity collecting and maintaining smart grid data be accountable for customer privacy in the same manner.

Both the FIPPs and SB 1476 were at the forefront when the Commission ultimately decided to adopt the Privacy Rules.

2. Privacy Rules, adopted in D. 11-07-056 (Attachment D)

Recognizing the need to more directly operationalize the FIPPs and the requirements of SB 1476 to protect consumer privacy in smart meter data,¹⁹ the Commission adopted the Privacy Rules, which regulate the disclosure of energy usage data by IOUs. As noted above, last year the Privacy Rules were extended to cover gas utilizes, community choice aggregators, electric service providers, and other “load serving” entities.²⁰ The Privacy Rules determine the extent to which an IOU may disclose energy usage data to third parties, depending on the purpose for which the data will be used. It covers all energy usage data captured by smart meters that, “when associated with any information . . . can reasonably be used to identify an individual [utility customer]”²¹ Data that cannot reasonably be re-identified are excluded from the Privacy Rules.²²

¹⁹ 2011 Decision, at 19–21.

²⁰ D. 12-08-045 (August 23, 2012).

²¹ The exact language of the Privacy Rules reads:

“Covered information” does not include usage information from which identifying information has been removed such that an individual, family, household or residence, or nonresidential customer cannot reasonably be identified or re-identified. Covered information, however, does not include information provided to the Commission pursuant to its oversight responsibilities.

The Privacy Rules categorize various potential uses into two categories. “Primary purposes” are uses of the data that directly serve utility operations, are specifically authorized by the utility company or the Commission in connection with an energy-related program, or are for services required by state or federal law. “Secondary purposes,” cover all other uses. Each category comes with its own list of obligations and security protocols relating to data transfer. The Rules impose these obligations on both the IOU disclosing the data and the third party recipients of the data.²³

a. Primary Purpose

Under the Privacy Rules, a covered entity may only disclose covered information without customer consent if the data will be used for a “primary purpose.” The Privacy Rules identify four limited purposes that fit within this category:

- (1) [to] provide or bill for electrical power or gas,
- (2) [to] provide for system, grid, or operational needs,
- (3) [to] provide services as required by state or federal law or as specifically authorized by an order of the Commission, or
- (4) [to] plan, implement, or evaluate demand response, energy management, or energy efficiency programs under contract with an electrical corporation, under contract with the Commission, or as part of a Commission authorized program conducted by a governmental entity under the supervision of the Commission.²⁴

Privacy Rules § 1(b). Further, for the purposes of “analysis, reporting or program management,” disclosure of “aggregated usage data that is removed of all personally-identifiable information” is permissible, “provided that the release of that data does not disclose or reveal specific customer information because of the size of the group, rate classification, or nature of the information.” Privacy Rules § 6(g).

²² As explained in our accompanying memo titled *Technical Issues with Anonymization & Aggregation of Detailed Energy Usage Data as Methods for Protecting Customer Privacy*, which covers recent scientific advancements in re-identification, no level of basic anonymization and aggregation provides a guarantee against re-identification. The Commission should pursue more robust solutions.

²³ The Privacy Rules govern “covered entities,” a category that includes:

- (1) [A]ny electrical corporation, or any third party that provides services to an electrical corporation under contract, (2) any third party who accesses, collects, stores, uses or discloses covered information pursuant to an order of the Commission, unless specifically exempted, who obtains this information from an electrical corporation, or (3) any third party, when authorized by the customer, that accesses, collects, stores, uses, or discloses covered information relating to 11 or more customers who obtains this information from an electrical corporation.

Privacy Rules § 1(a). The Commission’s authority to create regulations binding on third parties derives from the language of SB 1476, which conferred upon the Commission “broad powers and a legislative mandate” to take regulatory action to protect consumer interests. 2011 Decision, at 33–35.

²⁴ Privacy Rules § 1(c).

Section 6(b) further clarifies which entities may access, collect, store and use covered information for primary purposes without customer consent:

- An electrical corporation
- A third party acting under contract with the Commission to provide energy efficiency or energy efficiency evaluation services authorized pursuant to an order or resolution of the Commission
- A governmental entity providing energy efficiency or energy efficiency evaluation services pursuant to an order or resolution of the Commission.²⁵

According to the 2011 Decision, “[t]o the extent other governmental organizations, such as the California Energy Commission or local governments, may seek Covered Information in a manner not provided in these rules, the Commission will determine such access in the context of the program for which information is being sought absent specific Legislative direction.”²⁶

Accordingly, where the Privacy Rules do not explicitly provide for a certain form of disclosure, the Commission will determine on a case-by-case basis whether the disclosure is appropriate, and whether it is permissible under relevant legislation, such as the IPA. Please see above for more information about the IPA.

Sections 6(c)(1)(a–b) provides additional insight as to what qualifies as a “primary purpose,” and how disclosures must be carried out. Under these provisions, an IOU may share covered information with a third party without customer consent (a) if “explicitly ordered to do so by the Commission” or (b) if the disclosure serves “a primary purpose being carried out under contract with and on behalf of the electrical corporation disclosing the data.”²⁷ These provisions indicate that the Commission intended for the “primary purpose” category to cover a fairly narrow selection of disclosure scenarios, largely directed to IOU operations (such as billing, maintenance, and the like by contractors), along with the noted services, when under direct Commission oversight.

“Primary purpose” disclosures create a chain of obligations that carry down to subsequent custodians of “covered information.” When disclosure occurs for a “primary purpose,” the covered entity disclosing the data “shall, by contract, require the third party to agree to access, collect, store, use, and disclose the covered information under policies, practices and notification requirements no less protective than those under which the covered entity itself operates as

²⁵ Privacy Rules § 6(b).

²⁶ See 2011 Decision at 47-48.

²⁷ Privacy Rules §§ 6(c)(1)(a–b).

required under this Rule, unless otherwise directed by the Commission.” Thus, a “primary purpose” recipient of covered information must employ at least the same privacy and security measures as those implemented within the IOU from which it collected the data. The Privacy Rules attach to all data that originates with the IOUs, regardless as to whom ultimately takes possession of it.²⁸

b. Secondary Purpose

Any purpose that does not fall within one of the above categories is considered a “secondary purpose” under the Privacy Rules.²⁹ IOUs are prohibited from disclosing covered information for any secondary purpose without the “prior, express, written authorization” of each utility customer represented in the data.

Three limited exceptions to this requirement exist. A covered entity may only disclose smart grid data without customer consent in the following situations: (1) disclosure pursuant to a certain types of legal process (such as a warrant or court order); (2) disclosure in “situations of imminent threat to life or property; and (3) disclosure “authorized by the Commission pursuant to its jurisdiction and control.”³⁰ Again, without an authorization order from the Commission, third parties not working on behalf of the utility company likely cannot obtain covered information without the prior, express, written authorization from utility customers.

c. Data Minimization Requirements

Under Section 5(c), covered entities must limit the disclosure of smart grid data to only that which is “reasonably necessary or as authorized by the Commission” to carry out the specific purpose permitted under the Privacy Rules. For data uses constituting “secondary purposes,” this means that the covered entity may not disclose more information than is

²⁸ Privacy Rules § 6(c)(1). Rule 6(c)(2) reinforces the recursive nature of the Privacy Rules:

Any entity that receives covered information derived initially from a covered entity may disclose such covered information to another entity without customer consent for a primary purpose, provided that the entity disclosing the covered information shall, by contract, require the entity receiving the covered information to use the covered information only for such primary purpose and to agree to store, use, and disclose the covered information under policies, practices and notification requirements no less protective than those under which the covered entity from which the covered information was initially derived operates as required by this rule, unless otherwise directed by the Commission.

Privacy Rules § 6(c)(2).

²⁹ Privacy Rules § 1(e).

³⁰ Privacy Rules §§ 6(d)(1–3).

reasonably necessary to carry out the specific purpose authorized by the customer in writing. As noted above, data minimization requires entities to consider, in advance of disclosure, what data is reasonably necessary for the agreed-upon purpose before disclosing the data.

d. Data Security and Breaches

Section 8 of the Privacy Rule establishes the minimum security requirements that covered entities must employ when in possession of covered information. “Covered entities shall implement reasonable administrative, technical, and physical safeguards to protect covered information from unauthorized access, destruction, use, modification, or disclosure.”³¹ Furthermore, when a breach has been detected, a covered third party must notify the disclosing IOU within one week, and the utility must notify the Commission of all breaches affecting one thousand or more customers.³² Utility companies are additionally obligated to file an annual report at the end of the each calendar year, chronicling all security breaches affecting covered information that year.

e. Enforcement and Recourse for Privacy Rule Violations

If a recipient party fails to comply with its contractual obligations to handle the covered information in a manner “no less protective” than those under which the originating entity operates—a “material breach” under the Privacy Rule—“the disclosing entity shall promptly cease disclosing covered information to such third party.”³³

CONCLUSION

The laws and regulations described above each bear heavily on the data sharing scenarios contemplated within this proceeding. As such, it will be important for participants to enter the Working Group discussions with a firm understanding of their relevant provisions, with the Privacy Rules front and center.

Among the California state Constitution, the IPA, the FIPPs, SB 1476, and the Privacy Rules, utility customers receive legal protections for the privacy of their energy usage data.

³¹ Privacy Rules § 8(a).

³² Privacy Rules § 8(b). The Commission may also request that the utility company provide notification of any other breach for which notification is not already compulsory.

³³ Privacy Rules § 6(c)(3).

These protections, in various ways, bind the IOUs, the Commission, and other state agencies handling smart meter data, as well as third parties who obtain energy usage data from the utilities. At this stage of the proceeding, keeping these laws and regulations in mind will better position the Working Groups to devise solutions that are appropriately tailored to each disclosure scenario and are consistent with applicable law.

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

(END OF ATTACHMENT A)

ATTACHMENT B

**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
(Filed December 18, 2008)
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.cpsc.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

¹⁰Presentation 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 Shmatikov, 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 system²² 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 is necessary at all stages of this proceeding to determine:
 - What types of data can be released or should not be released under the requirements of the Privacy Rules;
 - What privacy solutions have been shown from experience to adequately or inadequately protect customers' private information; and
 - 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 US²⁷ and Europe²⁸ 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. NIST²⁹, 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 provisions³⁰. 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 terrorists³¹. 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 monitoring³² 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 watched³³.
- 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 practices³⁴ 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 Netherlands³⁵ 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.shtml>)

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 DECC³⁶ 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. Anonymization³⁷ 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 household³⁸. 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 used³⁹: 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 sources⁴⁰. 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 mechanism⁴¹. 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 mature⁴³ and rich⁴⁴ 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 system⁴⁵, and appropriate privacy technologies can be deployed to support privacy policies⁴⁶.

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 private⁴⁹, 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-reduce⁵². 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 readings⁵³. 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.

(END OF ATTACHMENT B)