Chapter 2

DATA HANDLING IN THE SMART GRID: DO WE KNOW ENOUGH?

Richard Chow, Alvaro Cardenas and Emiliano De Cristofaro

Abstract Data privacy in the smart grid is an important requirement for consumers. Central to the data privacy issue is the handling of energy-usage data, in particular, data retention, aggregation and anonymization. Government and industry groups have formulated various policies in this area, mostly based on fair information practice principles. This paper argues that the current policy-level work is insufficient – scientific work is needed to fully develop and implement privacy policies. A research agenda is proposed that balances the advantages of fine-grained energy-usage data with the associated privacy risks. For comparison purposes, the paper describes analogous policies and implementations related to telecommunications, web search and medical data.

Keywords: Smart grid, data handling, privacy

1. Introduction

The smart grid is being designed to enable utilities, customers and third-party providers to monitor and control energy use. Data collected by the smart grid will provide several advantages to all parties, including better decisions regarding energy-usage, enhanced understanding of consumer demands and increased energy distribution efficiency.

The smart grid also raises the issue of data privacy – especially for consumers – because smart meters will allow large-scale data collection, making individual household data available at unprecedented levels of granularity. Monitoring energy consumption at low levels of granularity can facilitate the inference of detailed information about consumer behavior. The behavioral information is highly valuable to advertising companies, law enforcement and criminals. The potential for egregious invasions of personal privacy makes it imperative to ensure that proper controls are in place.

Responding to these concerns, governments and standards organizations are developing privacy standards and policies to guide smart grid deployments.

J. Butts and S. Shenoi (Eds.): Critical Infrastructure Protection VII, IFIP AICT 417, pp. 21–32, 2013.

Recommendations for privacy controls in smart grid deployments have been provided by the North American Energy Standards Board (NAESB) [35], National Institute of Standards and Technology (NIST) [29], U.S. Department of Energy (DOE) [34], Texas Legislature and Public Utility Commission [25] and California Public Utilities Commission (CPUC) [4], among others. The Smart Grid Policy Framework [22] released in 2011 by the Executive Office of the President specifically recommends that, as a starting point, state and federal regulators must consider methods to ensure that detailed energy-usage data is protected in a manner consistent with the fair information practice principles drafted by the Federal Trade Commission (FTC) [8].

Recent work in the area of smart grid privacy has been necessarily broad in its coverage of systems and use cases. This paper concentrates on the handling of energy-usage data. Data handling issues cut across multiple privacy use cases and include data retention, aggregation and anonymization. Gaps in existing work are highlighted and a roadmap is proposed for areas that require exhaustive research. In particular, this paper describes technical issues that must be addressed in order to implement government and industry policies.

Data handling in the smart grid is a relatively immature issue. Consequently, this paper considers precedents with regard to handling potentially sensitive telecommunications, web search and medical data. The analysis of privacy strategies in these domains provides useful lessons that can guide research efforts for securing smart grid data.

2. Related Work

This section describes data handling issues covered in existing standards, guidelines and regulations. Much of the relevant prior work focuses on policy-level issues. Although there appears to be general agreement on data handling principles and goals, little work has been done on concrete approaches to achieve the data handling objectives.

2.1 Standards Bodies and Industry Guidelines

NIST's NISTIR 7628 document [29] provides general, policy-level guidance on data handling. For instance, the guidance on data collection recommends:

> *"Limit the collection of data to only that necessary for smart grid operations, including planning and management, improving energy use and efficiency, account management, and billing."*

The guidance on retention recommends:

> *"Limit information retention. Data, and subsequently created information that reveals personal information or activities from and about a specific consumer location, should be retained only for as long as necessary to fulfill the purposes that have been communicated to the energy consumers."*

The guidance on aggregation recommends:

> *"Energy data and any resulting information, such as monthly charges for service, collected as a result of smart grid operations should be aggregated and anonymized by removing personal information elements wherever possible to ensure that energy data from specific consumer locations is limited appropriately."*

The guidance on deletion recommends:

> *"When no longer necessary, consistent with data retention and destruction requirements, the data and information, in all forms, should be irreversibly destroyed."*

The general guidance is similar to the computer security principle of least privilege [27]. In practice, however, it may not be clear what privileges are actually required. Indeed, utilities are unable to turn policy to practice. For example, utilities are advised to collect "only ... necessary [data] for smart grid operations, including planning and management." More efficient planning and management are possible with more data. Similarly, utilities are advised to retain data "only for as long as necessary." However, more data maintained for a longer period of time typically supports better trend analysis and forecasting.

The NAESB provides business practices to follow with respect to third-party access and disclosure of energy-usage data [35]. The guidelines apply to data access directly from utilities but, like NISTIR 7638 [29], the guidelines are at a high level.

The U.S. Department of Energy's Advanced Security Acceleration Project for the Smart Grid (ASAP-SG) has produced two formal documents that outline the security requirements that a third party should meet in order to access electricity-usage data: (i) Security Profile for Advanced Metering Infrastructure [1]; and (ii) Security Profile for Third Party Data Access [2]. However, these documents primarily focus on general security principles rather than privacy concerns.

The document from the National Regulatory Research Institute entitled, Must There Be Conflict Between Energy Management and Consumer Privacy? [18], notes that categorizing usage data can reduce the potential for privacy compromise associated with personal information, billing data, operational data and event data. The document also mentions that aggregating data can make it safe to share; however, the method for aggregating data is not provided. Similarly, an early policy paper from the Office of the Information and Privacy Commissioner of Ontario [5] discusses the privacy implications of data mining and, in particular, fair information practices, but again, the guidance is at a high level. Privacy organizations such as the Electronic Privacy Information Center, Center for Democracy and Technology, and Electronic Frontier Foundation have also weighed in on data handling policies for the smart grid.

2.2 State and Utility Commission Regulations

The CPUC Rulemaking 08-12-009 document [4] contains general rules for utilities, contractors and other third parties that are modeled on the FTC's fair

information practice principles [8]. Like the work produced by the standards bodies, the guidance is policy-oriented rather than implementation-oriented.

Another example is the Colorado Public Utilities Commission's Decision No. R11-0922 [26], which describes the so-called "15/15 Rule:"

> *"In aggregating customer data to create an aggregated data report, a utility must take steps to ensure the report is sufficiently anonymous in its aggregated form so that any individual customer data or reasonable approximation thereof cannot be determined from the aggregated amount. At a minimum, a particular aggregation must contain: (i) at least fifteen customers or premises; and (ii) within any customer class, no single customer's customer data or premise associated with a single customer's customer data may comprise 15 percent or more of the total customer data aggregated per customer class to generate the aggregated data report (the "15/15 Rule"). Notwithstanding, the 15/15 Rule, the utility shall not be required to disclose aggregated data if such disclosure would compromise the individual customer's privacy or the security of the utility's system."*

This is one of a few guidelines that define the meaning of safe aggregation. However, the reasoning and privacy guarantees underlying the 15/15 Rule are not clear. Other examples of high-level guidance are provided in Oklahoma HB 1079 [23], the Illinois Statewide Smart Grid Collaborative: Collaborative Report [6], and the National Association of Regulatory Utility Commissioners Resolution on Smart Grid Principles [21].

Regulations have also considered other issues such as data ownership and third-party data access. While utility companies are responsible for protecting electricity consumption records from unauthorized use, the data may be shared with third parties by consumers who are most likely unaware of the privacy risks. While some regulations give consumers "ownership" of their electricity consumption data, others (e.g., Oklahoma HB 1079 [23]) stipulate that the utility company owns the smart meter data and that the utility company is "authorized to share customer data without customer consent with third parties who assist the utility in its business and services, as required by law, in emergency situations, or in a business transaction such as a merger" [23].

2.3 Research Literature

The research literature focuses on technical approaches that can improve privacy – examples include cryptographic mechanisms for facilitating data aggregation [28], differential privacy for sharing aggregated smart grid data with third parties (while preventing the identification of patterns about a single consumer) [30], and the use of batteries to mask electricity consumption [17]. While these mechanisms are promising, they do not answer the questions about data handling and data governance posed in this paper. In addition, most of these approaches are not yet ready for deployment. For example, the "zero-knowledge proofs of knowledge" that are required by the work on cryptographic commitments are computationally impractical. Similarly, using batteries to

mask electricity consumption will pose challenges due to the additional costs and the operational requirements involved in managing battery lifetime.

The research literature related to data handling primarily focuses on the information that can be inferred from very fine-grained electricity consumption data in the scale of seconds. However, most advanced metering infrastructures will collect data at fifteen-minute intervals or longer. There is a need to understand the trade-offs existing between data collection intervals and the associated privacy risks.

3. Comparison with Other Domains

This section discusses the approaches used to implement similar privacy policies in other domains. The focus is on the treatment, disclosure and retention of user data in the context of: (i) telephone logs; (ii) web search data; and (iii) health data.

3.1 Telephone Logs

In the United States, privacy policies for telecommunications-related data (e.g., data about customer usage stored by carriers) have mainly been the responsibility of the Federal Communications Commission (FCC). In 1996, the FCC was granted the authority to regulate how customer proprietary network information (CPNI) is treated. In 2007, the FCC began to regulate how data collected by telecommunications companies about customer telephone calls may be used [7]. The key points of the regulation are: (i) limiting information that carriers may provide to third parties without customer consent; (ii) defining how customer service representatives may share call details; and (iii) requiring notification obligations on the part of carriers. Like energy-usage data, telecommunications data has value beyond telecommunications companies (e.g., law enforcement and intelligence). This complicates the formulation and implementation of privacy policies.

At this time, there are no strict regulations governing the length of time that carriers may retain telephone logs. This has contributed to the variety of data retention policies that are implemented across the telecommunications industry [15]. Some U.S. carriers store call activity logs for days, some for months, some even retain call and text message content. Data retention policies are supposed to be opaque; information about these policies is available to the public because of a recent leak from the U.S. Department of Justice.

3.2 Web Search Data

Most Internet providers maintain logs of user search queries. This practice has raised serious privacy concerns because search data contains potentially sensitive information about the interests and web behavior of Internet users. Anonymization is a natural solution, but historical examples highlight the pitfalls. For instance, when AOL released anonymized user search queries,

researchers were able to reconstruct the identities of some users [13]. Most Internet search engine providers have independently created the data retention policies they implement. According to a 2011 *New York Times* article [14], Google maintains search records for nine months, Microsoft implements a six-month retention policy and Yahoo! has extended its retention time for detailed user records from three month to eighteen months. The privacy measures implemented by search engine providers are unclear because details about their anonymization techniques have not been released.

Another reason for implementing a data retention policy is to address liability issues. Because security breaches often result in the loss of user information, it is prudent for companies to minimize the financial risk they incur when handling sensitive data.

No U.S. legislation specifically addresses the control that users have over personal information related to their online activities, although guidelines are being developed. The FTC has encouraged the protection of online user privacy through several initiatives, such as through its report entitled, Protecting Consumer Privacy in an Era of Rapid Change [9]. Additionally, the FTC's fair information practice principles [8] help control personal information in the electronic marketplace by introducing notice, awareness, consent, access and security principles. The U.S. Department of Commerce's Internet Policy Task Force has released a report entitled, Commercial Data Privacy and Innovation in the Internet Economy: A Dynamic Policy Framework [33], which describes policy guidelines and regulations regarding commercial data privacy. Another effort named Do Not Track Us [19], which is spearheaded by academia, enables Internet users to opt out of website tracking.

3.3 Health Data

The sensitivity of health-related data has prompted privacy concerns that have been the focus of regulation as well as research efforts. In 1996, the U.S. Government enacted the Health Insurance Portability and Accountability Act (HIPAA) [31], which establishes national standards for electronic health care transactions and addresses the security and privacy of health data. In 2003, the Privacy Rule was added, which requires companies and providers to notify individuals about the use of protected health information and to keep track of disclosures. HIPAA also grants an individual the right to file complaints with the Office for Civil Rights of the Department of Health and Human Services.

The extreme sensitivity of genetic information has also been recognized. Protection is provided by the Genetic Information Nondiscrimination Act (GINA) of 2008 [32]. Among other points, GINA strictly regulates the retention, disclosure and treatment of data collected by genetic testing companies.

Privacy laws such as HIPAA and GINA provide much greater protection than laws and regulations associated with telephone calls and web search data. Indeed, the need to enforce health-related regulations is now almost universally recognized [12].

3.4 Outlook for the Smart Grid

Based on the treatment, disclosure and retention of telephone logs, web search data and health data, we make some projections regarding energy-usage data. First, privacy regulations and policies for most electronic user data are specified at a high level, if at all. As a result, it is likely that implementation practices for handling smart grid data will be left to utilities and their meter data management partners and vendors. The concern is, of course, that data handling issues are subtle. For example, the implications of employing even seemingly simple anonymization and aggregation techniques need to be considered carefully.

Second, companies often self-regulate their data handling even in the absence of regulations. However, as seen with telephone logs and web search data, each company generally has different policies and procedures for data handling, most of which are opaque to the consumer. The same trend will likely be seen in the smart grid where utilities will implement their own policies and procedures for energy-usage data. Indeed, without proper guidance, privacy analysis would require case-by-case evaluation.

4. Open Issues Related to Data Handling

This section discusses important issues related to the handling of energy-usage data in the context of the smart grid. These issues are excellent candidates for research in the area of smart grid privacy.

4.1 Safe Data Intervals

A key fair information practice is data minimization, which stipulates that data should not be collected unless it is needed and should not be kept longer than necessary. Many smart grid applications such as demand-response analysis will require the collection and storage of fine-grained data. However, any fine-grained data that is collected could be coarsened after it is no longer needed. This coarse data could be preserved in long-term storage or it could be released in various contexts after first checking the privacy implications.

Releasing even "coarse" monthly energy-usage data can result in an invasion of privacy, as demonstrated by the controversy surrounding the public disclosure of Al Gore's utility bills for his 10,000-square-foot Nashville mansion [36]. Perhaps more interesting is whether a two-hour collection interval or a four-hour interval would be considered safe. Clearly, this question depends on the perspectives of the end-users and the information that can be inferred from the collected data. Two-hour or four-hour data would not be safe if it reveals that the resident is on vacation. However, usage data at this level of granularity may not reveal specific appliance usage or other fine-grained usage patterns and could, therefore, be stored by utilities for the long term.

One of the key factors when considering safe data collection and retention policies is the ability to identify electricity appliances via non-intrusive load

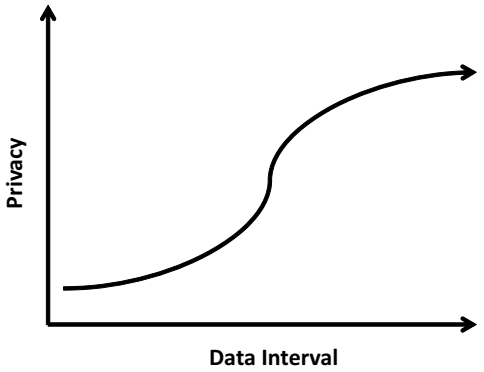

Figure 1. Conceptual plot of privacy versus data interval size.

monitoring (NILM) [3, 11]. Most of the research in this field has considered residential electricity consumption monitors, such as The Energy Detective (TED) and Current Cost, instead of residential smart meters provided by electric utilities. The residential electricity consumption monitors collect and analyze consumption data at sub-second intervals. However, current smart meter deployments do not have these levels of granularity. The vast majority of smart meter deployments collect data in intervals of fifteen minutes or more.

In general, there is a notable lack of research on the types of inferences that can be made with different levels of data granularity. Figure 1 shows the perceived relationship between privacy and data interval size. Clearly, privacy increases as the size of the data interval increases, but very little is known about the precise nature of this tradeoff. Prudenzi [24] is one of a few researchers who has examined energy-usage data in intervals ranging from a few minutes to a few hours.

In order to implement the data minimization principle, it is important to understand exactly what can be inferred from smart meter data as a function of various data intervals. Further research is needed to explore the trade-off between data collection intervals and the associated levels of privacy that are attained.

4.2 Data Aggregation Across People

Aggregation of energy-usage date across people – instead of over time – must also be considered. The privacy risks associated with fine-grained electricity-usage data can be reduced by computing aggregates across people (e.g., neighborhoods) and deleting individual usage data. However, the release of such aggregated data may still have some privacy implications. It is, therefore, important to conduct scientific studies on the conditions under which aggregations of energy-usage data could leak information about individual users and to devise appropriate strategies for mitigating the privacy risks.

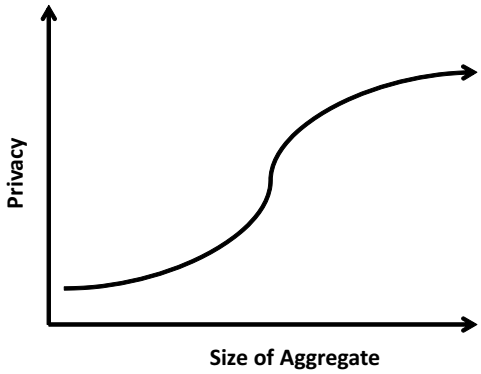

Figure 2. Conceptual plot of privacy versus aggregation size.

Figure 2 shows a conceptual plot of privacy versus aggregation size. In the representation, privacy increases as the size of the aggregate increases. However, the precise nature of the relationship has not been studied by the research community. Providing a scientific backing for Colorado's "15-15 Rule" [26], which intuitively postulates a safe aggregation principle, would be highly desirable. More research is needed to identify sound aggregation rules instead of simply specifying rules based on intuition.

Another issue is that aggregation does not protect an individual when all the individuals whose data is aggregated are similar in some way. For example, usage metrics associated with high-income communities are likely to be different from low-income communities. While this particular example may not raise privacy concerns, there may be situations where aggregated data compromises the privacy of the individuals who are part of the aggregate. In-depth studies are needed that focus on the privacy properties of aggregated smart grid data.

4.3 Consequences of Latency for Data Access

One concern about real-time electricity consumption data is that it could facilitate burglaries – prospective burglars could tell whether or not residents are home. Latency can help mitigate this privacy threat. For example, consumption data could be released one week or one month later. While some data might be required in real-time, algorithms such as those used for energy disaggregation to identify faulty or inefficient devices or appliances can be performed days, weeks or months later instead of in real time or near real time.

4.4 Safe Anonymized Data

Simply removing identifying information, such as a name or address, is insufficient to anonymize many kinds of personal data. For example, location traces can easily reveal home and work locations from which the identities of individuals can be deduced [10, 16]. As in the case of the AOL incident, user

identities can be discerned from web search data [13]. A linkage attack may also be used to de-anonymize data. For example, electricity-usage patterns could be de-anonymized by linking them to electricity-usage patterns of users in another database. An instantiation of this idea is discussed by Narayanan, *et al.* [20], where anonymized Netflix users were identified by matching their data to the IMDb database.

5. Conclusions

The vast majority of the work on handling energy-usage data has been at the policy level and is based primarily on fair information practices. Very little research has focused on strategies for implementing these policies and identifying the vulnerabilities that remain. Based on a comparison with other industries, it appears that implementation strategies for smart grid privacy will not be regulated in the near term. Thus, each utility will be forced to develop its own data handling strategy and architecture. Therefore, significant technical research is required in order to provide concrete guidance to designers. This guidance would bridge the gap between policy and implementation.

An important research agenda item is to understand the tradeoff between privacy and data interval size. Another is the development of safe anonymization, sanitization and aggregation strategies. Yet another is the examination of the privacy effects of data latency.

Clearly, all these issues are intertwined. A working system may very well involve the combination of a particular data interval, aggregation and latency strategy. Acquiring a sophisticated understanding of these concepts will be difficult as it will cover technical, legal and human factors issues; however, it is an important first step to enhancing privacy in the smart grid.

References

[1] Advanced Security Acceleration Project for the Smart Grid (ASAP-SG), Security Profile for Advanced Metering Infrastructure, EnerNex Corporation, Knoxville, Tennessee, 2009.

[2] Advanced Security Acceleration Project for the Smart Grid (ASAP-SG), Security Profile for Third Party Data Access, EnerNex Corporation, Knoxville, Tennessee, 2011.

[3] D. Bergman, D. Jin, J. Juen, N. Tanaka, C. Gunter and A. Wright, Distributed non-intrusive load monitoring, *Proceedings of the IEEE Power Engineering Society Conference on Innovative Smart Grid Technologies*, 2011.

[4] California Public Utilities Commission, 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 and Electric Company, Rulemaking 08-12-009, Sacramento, California, 2011.

[5] A. Cavoukian, Data Mining: Staking a Claim on Your Privacy, Office of the Information and Privacy Commissioner of Ontario, Ontario, Canada, 1998.

[6] EnerNex Corporation, Illinois Statewide Smart Grid Collaborative: Collaborative Report, Knoxville, Tennessee (`www.ilgridplan.org/Shared%20Documents/ISSGC%20Collaborative%20Report.pdf`), 2010.

[7] Federal Communications Commission, Report and Order and Further Notice of Proposed Rulemaking, FCC 07-22, Washington, DC, 2007.

[8] Federal Trade Commission, Fair Information Practice Principles, Washington, DC, 2012.

[9] Federal Trade Commission, Protecting Consumer Privacy in an Era of Rapid Change, FTC Report, Washington, DC, 2012.

[10] P. Golle and K. Partridge, On the anonymity of home/work location pairs, *Proceedings of the Seventh International Conference on Pervasive Computing*, pp. 390–397, 2009.

[11] G. Hart, Nonintrusive appliance load monitoring, *Proceedings of the IEEE*, vol. 80(12), pp. 1870–1891, 1992.

[12] Hewlett-Packard, HIPAA goes hitech, Palo Alto, California (`h20195.www2.hp.com/v2/GetPDF.aspx/4AA1-4056ENW.pdf`), 2010.

[13] D. Kawamoto and E. Mills, AOL apologizes for release of user search data, *CNET*, August 7, 2006.

[14] V. Kopytoff, Yahoo! will keep search queries for 18 months, *New York Times*, April 18, 2011.

[15] D. Kravets, Which telecoms store your data the longest? Secret memo tells all, *Wired*, September 28, 2011.

[16] J. Krumm, Inference attacks on location tracks, *Proceedings of the Fifth International Conference on Pervasive Computing*, pp. 127–143, 2007.

[17] F. Li, B. Luo and P. Liu, Secure information aggregation for smart grids using homomorphic encryption, *Proceedings of the First IEEE International Conference on Smart Grid Communications*, pp. 327–332, 2010.

[18] S. Lichtenberg, Smart Grid Data: Must There Be Conflict Between Energy Management and Consumer Privacy? National Regulatory Research Institute, Silver Spring, Maryland, 2010.

[19] J. Mayer and A. Narayanan, Do Not Track, Universal Web Tracking Opt Out, Center for Internet and Society, Stanford Law School, Stanford, California (`donnottrack.us`).

[20] A. Narayanan and V. Shmatikov, Robust de-anonymization of large sparse datasets, *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 111–125, 2008.

[21] National Association of Regulatory Utility Commissioners, Resolution on Smart Grid Principles, Washington, DC (`www.naruc.org/smartgrid`), 2011.

[22] National Science and Technology Council, A Policy Framework for the 21st Century Grid: Enabling Our Secure Energy Future, Executive Office of the President, Washington, DC, 2011.

[23] Oklahoma State Legislature, House Bill 1079, Oklahoma City, Oklahoma, 2011.

[24] A. Prudenzi, A neuron nets based procedure for identifying domestic appliances pattern-of-use from energy recordings at meter panel, *Proceedings of the IEEE Power Engineering Society Winter Meeting*, vol. 2, pp. 941–946, 2002.

[25] Public Utility Commission of Texas, Electric Substantive Rules – Chapter 25, Austin, Texas 2007.

[26] Public Utility Commission of the State of Colorado, Proposed Rules Relating to Smart Grid Data Privacy for Electric Utilities, Decision No. R11-0922, Denver, Colorado, 2011.

[27] J. Saltzer and M. Schroeder, The protection of information computer systems, *Proceedings of the IEEE*, vol. 63(9), pp. 1278–1308, 1975.

[28] E. Shi, T. Chan, E. Rieffel, R. Chow and D. Song, Privacy-preserving aggregation of time-series data, *Proceedings of the Network and Distributed System Security Symposium*, 2011.

[29] Smart Grid Interoperability Panel – Cyber Security Working Group, Guidelines for Smart Grid Cyber Security: Vol. 2, Privacy and the Smart Grid, NISTIR 7628, National Institute of Standards and Technology, Gaithersburg, Maryland, 2010.

[30] G. Taban and V. Gligor, Privacy-preserving integrity-assured data aggregation in sensor networks, *Proceedings of the International Conference on Computational Science and Engineering*, vol. 3, pp. 168–175, 2009.

[31] United States Congress, Health Insurance Portability and Accountability Act of 1996, Public Law 104-191, 104th Congress, Washington, DC, 1996.

[32] United States Congress, Genetic Information Nondiscrimination Act of 2008, H.R. 493, 110th Congress, Washington, DC, 2008.

[33] U.S. Department of Commerce, Commercial Data Privacy and Innovation in the Internet Economy: A Dynamic Policy Framework, Internet Policy Task Force Green Paper, Washington, DC, 2010.

[34] U.S. Department of Energy, Data Access and Privacy Issues Related to Smart Grid Technologies, Washington, DC, 2010.

[35] R. Varela, NAESB is developing smart grid data privacy standards, American Public Power Association, Washington, DC, March 2, 2011.

[36] T. Zeller, An inconveniently easy headline: Gore's electric bills spark debate, *New York Times*, February 28, 2007.