Energy Data Access:
A Guide to Leveraging Differential Privacy

December 30, 2021

Prepared by:
Carmen Best & Mariano Teehan in collaboration with Technical Advisory Group

Prepared for:
Janghyun Kim, National Renewable Energy Lab, Department of Energy

RECURVE
SHAPE THE FUTURE OF ENERGY

Abstract:
The ***Energy Data Access: Guide to Leveraging Differential Privacy*** provides a practical framework for using differential privacy in a risk-based data access framework. The ultimate goal of this document is to help demystify differential privacy in the context of current energy data access frameworks. The guide provides historical context, a structure for assessing risk in use-case development, and tools to orient decision-makers and practitioners to the opportunities for expanding data access frameworks to include differential privacy techniques. The user guide is intended for use by stakeholders and decision-makers to help inform the development of risk-based data access frameworks appropriate to a jurisdiction's specific policy goals and objectives and provide tools to properly balance and communicate the trade-offs of privacy and usability of datasets.

Acknowledgments:
This work would not have been possible without the contributions of the voluntary Technical Advisory Group. Members contributed their time to discuss the issues covered in this guide to develop this final product. We thank them for their input and note that while the content reflects the views of the experts convened, it does not imply their endorsement of all of the content.

Jason Harville	California Energy Commission
Abhilasha Wadhwa	California Public Utilities Commission
Harry Bergman	Department of Energy
Jon Callas and Lee Tien	Electronic Frontier Foundation
Marc Paré	Independent Consultant
Chris Conley	Independent Consultant
Jeffrey Deason	Policy Group - Lawrence Berkeley National Lab
Tianzhen Hong	Energy Technologies - Lawrence Berkeley National Lab
Bill Rus	MCE
Michael Murray	Mission:data
Lin Ainsworth	National Renewable Energy Lab
Ben Ruddell	Northern Arizona University
Chris Villarreal	Plugged In Strategies
Barry Hooper	San Francisco Environment
Eric Fournier	University of California Los Angeles

This material is based on work supported by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office under NREL Subcontract No. SUB-2021-10433. The views expressed in the article do not necessarily represent the views of NREL, the DOE, or the U.S. Government.

Table of Contents

History and Context of Energy Consumption Data Access **4**
Early Origins of Energy Data Access Proceedings 4
What is differential privacy? 6

Existing Frameworks for Energy Consumption Data Access **8**
Currently Accepted Practices 8
Common Attributes 8
Customer Access and Direct Authorization 10
Aggregation Practices to Enable Third Party Access 11
Limits of Current Data Sharing Practices 12
Conclusion 13

Fundamentals for Assessing Risk and Harm **15**
Why (re)Discuss Risk and Harm? 15
What Harms and Risks have been uncovered so far? 17
Climbing the Ladder of Privacy Risk 19
Putting A Risk on a Rung 21
Filters for Mitigating and Managing Risk: The 5 Safes 21
Using Privacy Factors To Explain Risk Tradeoffs (data & outputs) 23

Defining Use Cases **28**
How have "Use Cases" been used in energy data frameworks? 28
Universal Aggregation Rule in Lieu of Use Cases: Illinois 28
Highly Specified Use Cases - California 29
Breaking up the Considerations for a Use Case: New York 31
Use Cases Translated to Risk-Based Groups: New Hampshire 32

Choosing a Privacy Factor (ε) **35**
Using the Five Safes to Assess the Situation 35
Safe Project: Demand Response Market Settlement 35
Safe People: 3rd Party DRP, ISO, Settlement Platform 36
Safe Settings: Settlement Platform Vendor ++ 37
Safe Data: Hourly Load Shapes of Participants and Non-Participants 38
Safe Outputs: Appropriate Privacy Factor (ε) 38
Using the Energy Data Privacy Explorer to Consider Trade-Offs 39

History and Context of Energy Consumption Data Access

LEARNING OBJECTIVES:

- Understand the primary origins of data access frameworks for customer utility data
- Typical motivations and responsibilities of key actors at play in data access frameworks
- Definition of differential privacy and its relevance to this history

Early Origins of Energy Data Access Proceedings

While the origins of aggregation strategies for utility data can be traced back to disputes over access to water consumption data in 1997,[1] it was with the first large deployments of advanced metering infrastructure (AMI) in approximately 2008-2010 that states such as California began seeking a comprehensive way to balance the interests of distributed energy resource service providers, who wanted access to their customers' smart meter data, and consumer privacy. With smart meters able to provide data on electric power usage at 5-minute, 15-minute, or 60-minute intervals, it became critical for both individual customers and third party demand response and energy efficiency providers to access this data to avoid having to install thousands of dollars of redundant metering and telemetry systems per site. State commissions such as California and Illinois established rules in 2013-2014 by which customers could authorize their chosen distributed energy resource providers to access smart meter usage data electronically via standards such as Green Button, and also frameworks for qualified third parties to access data in the public interest.

As time passed, however, additional distributed energy requirements and new use cases rendered the narrow focus on permission-based access to AMI data inadequate and, in some respects, obsolete. The most current use cases for accessing a broader range of energy-related customer information have evolved into the following:

1) Distributed energy resource providers want individual customer usage and billing data electronically (with customer permission) to qualify customers and deliver products and services, ideally in standardized formats across all U.S. utilities.
2) Building owners and managers want whole-building monthly energy usage from their utilities to track energy usage and attain an EnergyStar score without arduous consent processes.

[1] B.L. Ruddell et al. Utilities Policy 67 (2020) 101106

3) Academics and researchers want access to as much detailed data as possible, including usage, cost, demographics, home/building characteristics, etc., without individual consent.

4) Distributed energy resource aggregators want anonymized consumption trends, ideally without individual consent.

5) Local governments are implementing and monitoring decarbonization initiatives that affect constituent energy consumption patterns and need consumption data to target initiatives and monitor progress toward goals. They are also implementing state and local energy benchmarking regulations and for enforcement may need data without individual customer consent.

In each case, the utilities, as the most common custodian of customer energy usage information, have control over the information that requesters seek. This is true even when the customer approves the sharing of their data with an authorized third party. The constraints on utilities' release of information have evolved via technical processes, commission-ordered restrictions, or *de facto*, utility-determined practices which may or may not have an explicit basis in law. The clarity and ability to operationalize procedures for accessing data with customer consent continue to prove challenging.

Cases authorizing data release without customer consent have been more restrictive and presented additional challenges to generating useful data. Processes like the Energy Data Request Program (EDRP)[2] in California have proven useful as a primary mechanism by which utilities are compelled to field and review third party data requests from eligible parties like academic institutions. For other cases involving access to aggregate data, requesters have tried to push the envelope to acquire as much detailed information as possible without incurring the burden of seeking each individual's permission.

In both cases, balancing the right to privacy, usability of the data, and public benefit from analysis or utilization of the data are at the core of the decision, as well as the nature of risk if data is exposed. In a relatively balkanized regulatory world, state agencies, utilities, and stakeholders are deliberating rules and guidance on what is ultimately a social and ethical decision and boiling it down to a technical specification. As fast-changing mathematical and "Big Data" approaches to reidentification and protection have emerged in recent years, state utility regulatory commissions have been understandably hesitant to wade into questions such as: "how can we balance the interests of legitimate market activity, academic research, energy-related public policy and customer privacy?" and "how do we ensure that today's approaches don't prove to be ineffective tomorrow?"

Setting aside the daunting nature of the task, it is inevitable that practices will continue to evolve. Keeping data access rules and protocols fresh will be essential in a distributed energy

[2] The Energy Data Request Program (EDRP) was established by the California Public Utilities Commission in 2011 and continues to function today. Example from PG&E's portal: https://pge-energydatarequest.com/

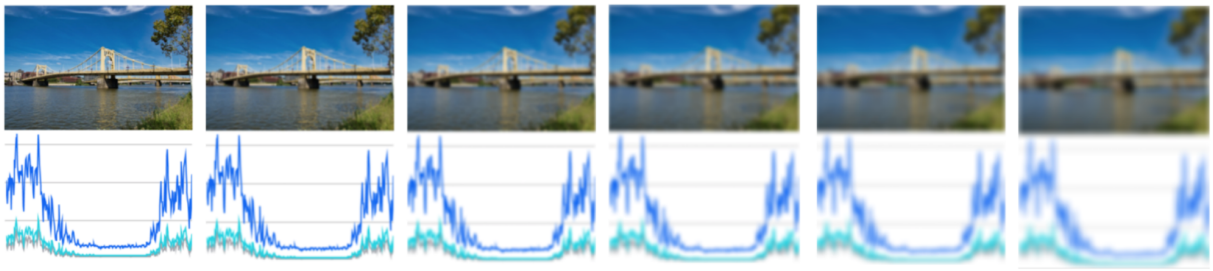
future in which more and more actors will either be enabled or thwarted by data access rules. This guide is not intended to provide a comprehensive history or thorough comparative analysis of data access frameworks in the US, much less the rest of the globe, but instead offers insights into how one strategy, differential privacy, can be added to the range of options to balance privacy and usability appropriately.

What is differential privacy?

Differential privacy is a mathematical technique to introduce noise into a data set to reduce the risk of any individual being re-identified within the data set, and potentially to reduce information exposed if an individual is re-identified.

> *Differential privacy is a system for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset. The idea behind differential privacy is that if the effect of making an arbitrary single substitution in the database is small enough, the query result cannot be used to infer much about any single individual, and therefore provides privacy*. Wikipedia

Differential privacy is typically achieved by adding random noise to query results. Adding more noise makes the data more anonymous, but it also makes it less useful in certain situations. Imagine it like pixelating an image of a bridge – as noise is added, it is still evident that the image is a bridge, and even that it's a suspension bridge, but it is no longer possible to identify which specific bridge it is. However, this analogy is imperfect; ultimately we don't necessarily want end users to see the bridge either. Successful differential privacy would generate a different picture entirely, but the mean of the pixel values would be identical to the original.



Since differential privacy is applied with a mathematical expression, it can be reported with a discrete parameter called "epsilon." This value can be interpreted as a privacy factor to describe exactly how much noise has been added to the dataset. Unlike the current convention of adopting aggregation rules like 15/15, which creates a binary compliance structure, a privacy factor (ε) can be selected along a scale to dial in the right privacy/usability trade-off point for a particular use case.

With appropriate orientation and standard tools, stakeholders and decision-makers can more precisely consider the tradeoffs of risk that may apply to a particular use case. With a

shared understanding of what the privacy factor means, practitioners can more specifically communicate how they have protected data sets and build trust and confidence in sharing data with qualified actors. Providing education and tools to support all market actors in this discussion is the primary focus of this guide.

Conclusion

Energy data access is likely to grow in importance as distributed energy resources become a greater part of our energy infrastructure. Data access is core to integrating grid management, local climate action and solutions for customers. Data access frameworks need to consider new solutions that may offer greater security and help balance the privacy and usability of energy consumption data.

ADDITIONAL REFERENCES

Data Access Guidance State and Local Energy Efficiency Action Network (2012) *A Regulator's Privacy Guide to Third-Party Data Access for Energy Efficiency.* Prepared by M. Dworkin, K. Johnson, D. Kreis, C. Rosser, J. Voegele, Vermont Law School; S. Weissman, UC Berkeley; M. Billingsley, C. Goldman, Lawrence Berkeley National Laboratory.

Energy Data Unlocking Innovation With Smart Policy (December, 2017) Michael Murray, Laura Kier And Bob King, P.E. Mission:data Coalition http://www.missiondata.io/s/Energy-data-unlocking-innovationwith-smart-policy.pdf.

Remaking Energy: The Critical Role of Energy Consumption Data, Alexandra Klass and Elizabeth Wilson104 CAL. L. REV. 1095 (2016), available at https://scholarship.law.umn.edu/faculty_articles/557

Knowledge Is Power: How Improved Energy Data Access Can Bolster Clean Energy Technologies & Save Money Ethan N. Elkind, January 2015

Sharing Load Profile Data: Best Practices and Examples, Northeast Energy Efficiency Project, May 2021

Energy Data Accelerator: Guide to Data Access and Utility Customer Confidentiality, Better Buildings, US Department of Energy, January 2016

Existing Frameworks for Energy Consumption Data Access

Existing Data Access Frameworks - inline edits

LEARNING OBJECTIVES:

- Understand some of the common attributes of data access frameworks (principles)
- Understand the common practices, like aggregation, embedded in the guidance
- Explore some of the challenges and opportunities to refresh data access frameworks

Currently Accepted Practices

As advanced metering infrastructure makes more and more granular data available and the importance of optimizing demand-side energy resources grows, utilities, policymakers, and the public must reevaluate the tradeoffs between data privacy and the need to provide broader access to energy consumption information. The significant public benefits of reducing overall energy costs and the need to transition to reliable, high-renewable electricity supply are clear. The ever-more obvious risk of failing to drive down carbon emissions potentially creates a new set point in balancing the need to aggressively decarbonize energy consumption against the importance of protecting individual privacy.

To date, states with slightly different stakeholder groups, different regulatory mandates, and varying political contexts have adopted a variety of data access frameworks across the United States. Many more are likely on the horizon. Despite these differences, these frameworks do contain some common principles and practices. Understanding these existing data access frameworks provides insights on which elements may be replicated, which are ripe for updating, and how we might chart a path to include alternative approaches for balancing privacy and usability such as that offered by differential privacy. At a minimum it will be necessary to weave differential privacy into these frameworks for it to gain traction and acceptance as a best practice.

Common Attributes

Data access frameworks adopted by regulatory bodies or other state agencies are usually specifically dedicated to customer energy usage data:

> *Customer Energy Usage Data (CEUD) reflects an individual customer's measured energy usage but does not identify the customer.*

Other data that helps explain energy usage, like the customer's location, rate tariff, demographics etc. are usually a proximate consideration as well in these proceedings.

One preliminary litmus test for allowing ANY access to this data is to classify uses as either primary or secondary purposes wherein:

> ***Primary Purpose:*** *The use of Account Data or CEUD that is reasonably expected by the customer: (1) to provide or reliably maintain customer-initiated service; and (2) including compatible uses in features and services to the customer that do not materially change reasonable expectations of customer control and third-party data sharing.*
>
> AND
> ***Secondary Purpose****: The use of Account Data and CEUD that is materially different from the Primary Purpose and is not reasonably expected by the customer relative to the transactions or ongoing services provided to the customer by the Service Provider or their contracted agent.*

Primary and secondary purpose definitions may vary by jurisdiction but, in essence, provide a first line of defense against frivolous and gratuitous access to customers' energy consumption.

Data custodians such as utilities or third parties, can also adopt principles and guidelines in lieu of specific directives. The Data Guard Energy Data Privacy Program, initiated and supported by the U.S. Department of Energy, outlines a voluntary code of conduct (VCC) for utilities and third parties in protecting and providing access to energy consumption data including methodologies for creating aggregated or anonymized data.

The voluntary code of conduct and formally adopted data access frameworks typically have similar requirements, including specific applications for enabling the protection and the reasonable provision of data including:

- Notice
- Purpose Statement
- Access and Control
- Data Minimization
- Use and Disclosure Limitation
- Data Quality and Integrity
- Data Security (including breach notification)
- Accountability and Auditing

Many are organized around or grounded in the Fair Information Practice (FIP) Principles, developed by the FTC in the 1970s, and later adopted by the Department of Homeland Security.[3] For example, California's data privacy proceeding, one of the first ones in the U.S., explicitly called out the FIP principles as their guiding framework.

> *"…this decision adopts the FIP principles as the framework for developing specific regulations to protect consumer privacy…" California PUC, D.11-07-056 (July 2011)*

Debates in proceedings adopting data access frameworks frequently have revolved around the perceived trade off of privacy versus the potential benefits of access for parties beyond the customer and the custodian. In most jurisdictions, a core tenet is that control of data is in the hands of customers, with the exception of justifiable primary use cases, where consent

[3] See: https://www.dhs.gov/xlibrary/assets/privacy/privacy_policyguide_2008-01.pdf

may not be required. Utilities or a third party are considered custodians of the data on behalf of (consenting or non-consenting) customers; hence, they hold the liability to protect the privacy of the customer. As such, custodians of the data may have access to identifiable information such as customer usage and other metadata that enhance that data's usefulness and offer potential insights that would otherwise be unavailable.

Data custodians are typically required to protect outputs to ensure data is non-identifiable (back to the customer) by anonymizing and/or aggregating data sets. They are also typically responsible for ensuring the processes and formats to share data are consistent across utilities, secure and enabled with standards like NAESB REQ 21 and Green Button[4], but these protocols are not consistently followed in practice.

Most data access frameworks also include a commitment to a risk-based approach to assess which entities are authorized to get what data. In reality, the sophistication of risk assessment entertained in any given framework development process or in the resulting requirements can vary significantly.

Customer Access and Direct Authorization

Given that most data access frameworks support the core tenet that control of data is in the hands of customers, these frameworks typically articulate pathways for customers to consent to use their data or directly transfer it to third parties as well. These mechanisms are largely not in scope of this paper, but we provide a brief summary of some considerations, because if consent protocols are highly functional then other aggregation approaches may not be necessary for certain data sets.

Debates in data access framework development are often set around the assumption that customers are unwilling to share their data for public interest research under all scenarios. However, many precedents can be found in other industries where consumers voluntarily share their data as a public good. In fact, certain private companies have incorporated this voluntary sharing of customer data into their business models. Some examples include non-anonymized reviews submitted by consumers on platforms such as Yelp and Amazon. In some cases, the consumer is compensated for sharing their individual data through discount coupons or other rewards (e.g. grocery stores offer discounts for consumers to sign-up for 'free' memberships for regular discounts. The terms and conditions typically give the store the customer's approval to share the customer's purchase habits, phone number and address with other vendors). While these practices are common, many privacy advocates would also argue that the bar for what qualifies as "informed consent" from customers is too low and that many customers don't understand the implications of what they're signing up for.

While approaches for streamlining customer consent are found in many business models, they have not been translated to publicly funded demand-side management programs. In

[4] https://www.energy.gov/data/green-button

these programs, participating customers are frequently getting discounted equipment or services subsidized by the program. These customers could be inclined to authorize use of their data for evaluation of the program in which they are participating, but most programs do not ask the customer at the time of signing up if they would be willing to share their energy usage data generally (opt-in) in exchange for the service being provided.

The California Public Utilities Commission recently authorized this pathway via a conclusion of law to potentially allow a customer to donate their energy use information for public interest research.[5] To date, this authorization has not been put into practice.

Aggregation Practices to Enable Third Party Access

Direct customer consent is not always practical or feasible for many valid use cases. Data access frameworks therefore also have established rules and procedures for third parties to access energy consumption and other utility data. Aggregation thresholds for authorized parties to receive or present customer consumption are a common output of most energy data access frameworks.

Aggregation thresholds have been adopted as a common practice as a means of protecting customer privacy and still attempt to enable certain use cases that don't need identifiable information or customer consent. Many unique wrinkles exist on the definition of identifiability, especially for location-specific use cases where physical assets are inherently identifiable, i.e. a building exists on a map and one can see it on the street. The threshold requirements may apply to the receipt of data from a custodian to an authorized party or are part of the legal requirements to present data. It almost exclusively applies to data shared without customer consent.

Probably the most famous, and frequently villainized, "rule" that emerges from data access framework deliberations is the 15/15 aggregation threshold. The 15/15 aggregation threshold means that any aggregation of data must have at least 15 customers (numerator) and no one customer is more than 15% of the load (denominator). This is a common threshold set for use cases using residential data but also commonly applied in commercial use cases too. Given the homogeneity of residential populations, data sets can be shared that provide basic trending information without a big loss in the usefulness of the data. For industrial situations, a common aggregation threshold is 4/80 which means it must have 4 customers and no one customer is more than 80% of the load.

Many jurisdictions have adopted universal aggregation thresholds (also sometimes called "screens"). Some exceptions include Colorado, California, New York and New Hampshire. In these cases, decision-makers recognized or were compelled to accommodate variation in use

[5] D.20-03-027 Conclusion of Law 30 states that “It is reasonable to provide IOU customers the option of voluntary public donation of their energy use data rather than assume that every customer is unwilling to share their individual energy use data for public interest decarbonization-related research.”

cases or at a minimum by customer type.[6] In some situations, aggregation thresholds have also been differentiated to recognize the effect of time (15 minutes, hourly, daily, monthly, quarterly, annually) and geography (zip+4, zip, census block, city, county, state).

As noted, adopting an aggregation threshold is a common point of significant debate in the stakeholder processes considering data access frameworks representing where the technical and social trade-off between value and privacy is negotiated. As the crux of any data access framework, it will inevitably need to be addressed regardless of the privacy protection methods adopted. For example, annual statewide consumption data is somewhat useful and provides fairly straightforward privacy protection with aggregation; the risk of exposing private information is negligible, so the public benefit required to justify the disclosure need not be great to justify provision of data. Hourly residential consumption data on a feeder could be very useful for a variety of use cases, but given the small size of the population pool, privacy may be more difficult to protect without compromising the value of the data, so the potential benefit may need to be more specific and substantial to justify risk of exposure.

Introducing the use of differential privacy, where noise or blurring can be mathematically added to the data set and thereby limit the ability to re-identify a customer, can help. While not immune to the same debates about the trade-offs, differential privacy can help refine the choices available and more clearly communicate the appropriate trade-offs, but comparisons with existing aggregation would only be applicable on a case-by-case basis after each method has been applied to a real world usage dataset.

The choice of metric for establishing such an equivalence may be useful to consider. For example under 15/15 aggregation, for a given dataset, we find that the data for X% of customers, or total consumption, or geographic reporting areas etc. would not be able to be disclosed. The size of X would constitute a margin of error that could be equated to the levels of noise that would be introduced by the selection of a given privacy factor (epsilon) value. We recommend further exploration of these comparisons in the future.

Because differential privacy offers a scalar and mathematical description of privacy, stakeholders and practitioners can dial and communicate a discrete balance point with usability on a sliding scale of a privacy factor for any given use case. One of the reasons aggregation thresholds can be problematic is that they create a binary compliance framework and a "cliff" of non-compliance and risk exposure once the threshold is exceeded. Stakeholders and practitioners will need to build intuition and "fluency" in adapting differential privacy approaches and understanding this metric in context.

Limits of Current Data Sharing Practices

In addition to some specific issues of security that can be achieved with aggregation-only (discussed in the next chapter), with more experience, the limits of usability have been revealed. Practitioners and academics have experienced these limitations directly. In 2020,

[6] Littel, David (Regulatory Assistance Project) "MiPower Grid: Customer Education and Participation Session 2: Data Privacy, Sharing, and Customer Consent" Michigan Public Service Commission. 22 June 2021. Presentation.

researchers published a statistical analysis of the tradeoffs of usability and privacy for utility consumption data. The researchers note that the lack of sufficient consideration of risk and benefits may be striking "poor balance of between the public benefit in the analytical usefulness of the data and the individual interest in privacy." They also discuss the practical value of aggregation in geographic applications. Based on their statistical analysis, they recommend adoption of a 50/50 aggregation threshold over the commonly used 15/15 to improve both the minimum level of customer privacy and the usefulness of statistical analysis.[7]

The researchers also make several recommendations to improve the status quo for data access frameworks to help address other perceived challenges. Most of these recommendations are tied to the fact that rules for data access can be ambiguous and variable across jurisdictions and regulatory bodies. Their recommendations include:

- Adoption of a HIPAA-style dual-rule approach that allows for statistical expert determination of privacy protection
- Adopting specific practices for safe group construction
- Not treating account group membership and location (address) as personally identifiable information
- Differentiating disclosure rules and informed consent between data user types

Differential privacy practices in data access frameworks would likewise benefit from these recommendations. As an emerging solution, it will also need knowledgeable decision-makers and practitioners to expand options for secure data access.

Maintenance of data access frameworks is necessary to ensure that they adapt to the latest technology and capabilities and continue to strike the right balance of usability and privacy in a changing world. Contemporary capabilities around differential privacy and the urgency of optimizing demand side resources are two factors that may, in and of themselves, warrant re-consideration of data access frameworks. Many jurisdictions are still in the nascent stages of making energy consumption data available and are formulating their own data access frameworks. The information in this guide can support those decisions and ensure all options are on the table.

Conclusion

Standard practices have emerged for data access frameworks, but there is also some variability in how they are structured. Aggregation thresholds are a common output of data access frameworks and the inclusion of strategies or guidelines enabling differential privacy as an option where appropriate are not yet available. Including a role for differential privacy within data access frameworks could enhance the ability to find the appropriate balance points for data sharing among data users, data custodians and regulators.

[7] Benjamin L. Ruddell, Dan Cheng, Eric Daniel Fournier, Stephanie Pincetl, Caryn Potter, Richard Rushforth, Guidance on the usability-privacy tradeoff for utility customer data aggregation, Utilities Policy, Volume 67, 2020, 101106, ISSN 0957-1787, https://doi.org/10.1016/j.jup.2020.101106

ADDITIONAL REFERENCES

Data Privacy and Smart Grid: A Voluntary Code of Conduct (VCC) (January, 2015) U.S. Department of Energy.

Driving Building Efficiency with Aggregated Customer Data: A Brief Review of Selected State Practices in the U.S. (2013) Regulatory Assistance Project

Energy Data Access: Blueprint for Action Toolkit website (2021) United States Department of Energy

Benjamin L. Ruddell, Dan Cheng, Eric Daniel Fournier, Stephanie Pincetl, Caryn Potter, Richard Rushforth, Guidance on the usability-privacy tradeoff for utility customer data aggregation, Utilities Policy, Volume 67, 2020, 101106, ISSN 0957-1787, https://doi.org/10.1016/j.jup.2020.101106

The Conversation Explainer: what is differential privacy and how can it protect your data? March 18, 2018.

Villereal, Chris (Plugged In Strategies). "Advanced Metering Infrastructure (AMI) Work Session." New Jersey New Jersey Board of Public Utilities Docket No. EO20110716. 23 November 2020. Presentation.

Littel, David (Regulatory Assistance Project) "MiPower Grid: Customer Education and Participation Session 2: Data Privacy, Sharing, and Customer Consent" Michigan Public Service Commission. 22 June 2021. Presentation.

Fundamentals for Assessing Risk and Harm

LEARNING OBJECTIVES:

- Understand the typical dimensions of risk and harm that come to play in energy data frameworks
- Understand the progression of risk from individual identification through to social benefit/cost trade-offs
- Understand the 5 safes that need consideration; including mathematical treatment of data for privacy.
- Understand the relationship between the privacy of the data and usability of data reflected in ε (epsilon) or the privacy factor.

Why (re)Discuss Risk and Harm?

At the core of data access frameworks lies an interest in protecting individuals and society from harm. Exposing private information of individuals can lead to harm to those individuals in certain situations. Inaction with respect to social challenges may also lead to harm individually and collectively. In the context of exposing energy consumption patterns, real and perceived harms have been brought forward in nearly all proceedings on the topic of energy data access. If there wasn't some real or perceived harm from having private information energy consumption specifically exposed or a real benefit in using individual consumption data, this guide would be obsolete.

The balance point for risk and harm is also a dynamic one. As the urgency and impact of a changing climate are revealed in tangible ways, the cost of protecting privacy in lieu of action may be too high. For example, in 2014 Pacific Northwest National Lab published a report on the importance of sharing data for the purpose of supporting building benchmarking policies.[8] This report was leveraged to pass legislation in California that enabled a state agency, the California Energy Commission, to collect energy consumption data from all utilities in the state for this purpose as well as to support the broader agenda of energy management, a core mission of the agency.[9] Similarly, using data on excessive water

[8] Livingston O.V., T.C. Pulsipher, D.M. Anderson, and N. Wang. 2014. Commercial Building Tenant Energy Usage Aggregation and Privacy. PNNL-23786. Richland, WA: Pacific Northwest National Laboratory. Available at: https://www.pnnl.gov/publications/commercial-building-tenant-energy-usage-aggregation-and-privacy

[9] AB 802 Energy Efficiency (2015 - 2016) https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201520160AB802

consumption, once taboo, has shifted into the space of being perceived as a common good by many.[10]

Using gut-check or anecdotal assessments of risk and harm from exposing energy consumption data can generate outsized fear or an underappreciated sense of risk and potential harm. Anecdotal examples may get outweighed attention because they are scary rather than because they are likely to happen. That doesn't mean that real harm eventualities that are rare should not be considered, but rather that in a societal decision-making framework the likelihood of harm should be weighed against the severity of that harm. When and where well-defined criteria for assessing actual harm or risk exist, the collaborative conversations can more meaningfully drive toward effective solutions to mitigate the harm or minimize the risk.

> *The purpose of a risk-based approach is to replace an otherwise subjective gut check with a more guided decision-making approach that is scalable and proportionate, resulting in solutions that ensure data is useful while being sufficiently protected. Statistical estimators are used to provide objective support, with greater emphasis placed on empirical evidence to drive decision making.*
> – [Arbuckle, El Emam. Building an Anonymization Pipeline]

Throughout this guide we will lean heavily on a framework for assessing risk, put forth in a practical handbook called "Building and Anonymization Pipeline." Originally developed for health care industry considerations of data protection, this handbook offers a primary reference and framework that can streamline and simplify the otherwise vast world of risk assessment and provides some clear and discrete checkpoints and classifications that mirror those that have emerged in data access frameworks that exist today.

Looking at energy data access frameworks in effect today, all have had to consider risk and harm in some way. There is no such thing as zero risk in data anonymization or privacy protection. Stakeholder processes are undertaken to sift and balance the tradeoffs between the risk of re-identification and the potential resulting harm with the usefulness of the data for social good. It is rare to see an explicit discussion (much less modeling) of the weighting of risks and benefits to determine anonymization criteria for energy data. Such proceedings, new or revisited, may be strengthened if they were to include more structured treatment of assessing risk and harm.

With a robust discussion of risks and harm within a guided framework, more use cases may be accommodated, and the appropriate techniques can be vetted and chosen to protect privacy while maintaining usability. As technology and society shift, perceived and real risk and value also shift and the decision point on the tradeoffs needs to be revisited from time to time.

[10] Article: "South Africa: Cape Town's Map of Water Usage Has Residents Seeing Red", 17 January 2018 available at https://allafrica.com/stories/201801180110.html

What Harms and Risks have been uncovered so far?

The National Renewable Energy Lab (NREL) is preparing a comprehensive literature review of peer reviewed papers exposing and discussing perceived and real examples of risk and harms from exposing individual energy consumption data. The study begins with an inventory of some of the most common research on re-identifying individuals within a data set. One of the most famous studies of harms is from Dr. Sweeny, where public anonymized health data and voter information was able to reveal individuals.[11] In two other cases, Netflix and IMDB were combined to reveal voting tendencies.[12] In another study, the likelihood of re-identification increased if a person had participated in a human genome study.[13]

For buildings, the core risk of harm from energy consumption data comes down to being able to reveal aspects of lifestyle, such as occupancy, use of certain classes of end-use energy appliances or other technologies. It may also include things like revealing which tv channels individuals watch or if they are engaged in certain activities, illicit or not. That was the focus of the majority of studies that identified privacy leaks for residential situations. In terms of research, tensions are mostly on smart meter data and not other measurements like sub-meter data or IoT devices.

The literature primarily considers theoretical risks rather than actual risk or harm incurred. Little work exists that reveals any of the potential risks and harms that have been identified in theory, with some notable exceptions. One real life example included using energy data to find terrorist safe houses in Europe by seeing outsized usage signalling more than typical occupancy.[14] In one other case, in Naperville IL, the resident could not opt-out of having a smart meter installed on the basis of their concern around the risk of revealing their privacy.[15] Other articles highlight a situation in which public shaming for using too much water was reviewed for the harm to the customer versus the benefit of a water conservation program.[16] These examples and a more comprehensive overview of the literature will be found in the full study pending release.

---

[11] L. Sweeney, "Weaving Technology and Policy Together to Maintain Confidentiality," The Journal of Law, Medicine & Ethics, vol. 25, no. 2–3, pp. 98–110, 1997.
[12] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in 2008 IEEE Symposium on Security and Privacy (sp 2008), 2008, pp. 111–125.
[13] N. Homer et al., "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays," PLOS Genetics, vol. 4, no. 8, p. e1000167, 29 2008.
[14] B. S. Amador, "The Federal Republic of Germancy and Left Wing Terrorism," NAVAL POSTGRADUATE SCHOOL MONTEREY CA, 2003.
[15] "Naperville Smart Meter Awareness v. City of Naperville, No. 16-3766 (7th Cir. 2018)," Justia Law. [Online]. Available: https://law.justia.com/cases/federal/appellate-courts/ca7/16-3766/16-3766-2018-08-16.html.
[16] J. Glionna, "Las Vegas outs its water hogs -- at least when asked - Los Angeles Times," 2015. [Online]. Available: https://www.latimes.com/nation/la-na-vegas-water-hogs-20151030-story.html. [Accessed: 25-Mar-2021]. AND J. Horwath, "Top 10 WATER GUZZLERS," Santa Fe Reporter, 2015. [Online]. Available: http://www.sfreporter.com/news/coverstories/2015/03/31/top-10-water-guzzlers/. [Accessed: 25-Mar-2021].

Literature Review of Harms / Risks

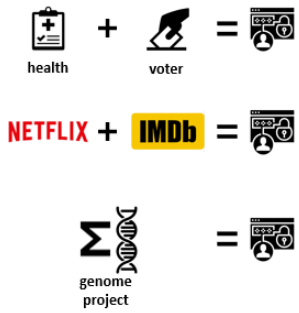
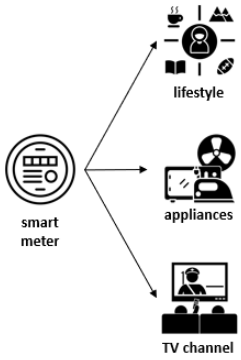
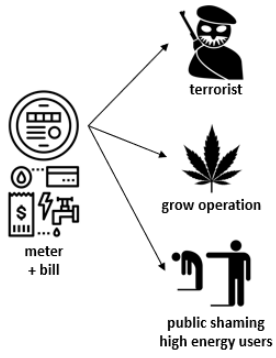
Research Examples General (popular ones)	**Research Examples** Buildings	**Risks** in theory?	**Risks** in real life?
		• **activity/lifestyle:** location of occupants in building, occupancy in household or building, religious belief of occupants, shower patterns, sleeping patterns in household, smoker or non-smoker of occupant, meal time, working hours, economic/employment status of occupant in household, homeowners TV watching pattern and preferred channel, time of usage of appliances in household • **electricity generation timings of renewable energy sources** • **equipment not operating in desired efficiency** • **health/medical status of occupants** • **number of residents in a household** • **whereabouts of an electric vehicle**	
Sweeney (1997) Narayanan and Shmatikov (2008) Homer et al. (2008)	Hart (1992) Molina-Markham et al. (2010) Lisovish et al. (2010) Rouf et al. (2012) Greveler et al. (2012)	Zhao 2014, Barbosa 2016, Eibl 2017, Zhang 2017, Yang 2017, Cao 2019, Finster 2015, Desai 2019, Pappachan 2017, Jia 2017, Ghayyur 2018, Savi 2016, Barbosa 2014, Baloglu 2018, Liao 2019, Xu 2018, Jelasity 2014, Laforet 2015, Zipper 2019, Lou 2020, Hassan 2020, Ou 2020, Guan 2020, Soykan 2019, Bao 2015, Pohls 2014, Ni et al. 2017, Tudor et al. 2016, Wang et al. 2016, Parker et al. 2021, Lou et al. 2017, Tian et al. 2018, Gohar et al. 2019, Rouf 2012, Ny and Mohammady 2014, Hassan 2019, Asghar 2017, Hassan 2018, Schwee 2020, Xiao 2018, Gulisano et al. 2016, Guo et al. 2021, Jonsson and Nelson 2015, Acs 2011, Pillitteri 2014, Backes 2014, Wang and Wu 2020, etc.	Amador (2003) Douris (2017) Glionna (2015) Horwath (2015)

Image prepared by Janghyun Kim, National Renewable Energy Lab, 2021

When assessing risk, it is important to remember that it takes time to learn things. Availability of the data from smart meters is relatively new, and it is clear that they provide valuable information. The ability to learn how to use that information is another matter entirely. This is true for instances of using the data for intended[17] or seemingly untoward purposes like police surveillance. It is clear that new technologies create the ability to learn new ways to exploit people -- so it is not sufficient to assume if a risk is not evident today that it won't emerge tomorrow.

It is also likely that violations of personal privacy may be happening that are not widely known. For example, detecting marijuana grow operations has been a focal point of the privacy debates in the past. Power data could be used to allow for the identification of the places to "point" thermal detectors without warrants violating individual privacy. Smart meter data is also useful in identifying high users mining crypto currency. While not illegal, cryptocurrency mining has come under public scrutiny.[18] From one perspective, knowing which users are consuming large amounts of energy for marijana growing or cryptocurrency mining creates an opportunity to create targeted demand response programs. From another point of view, it may represent an unacceptable violation of personal privacy. Ultimately questions of privacy are driven, rightly so, because energy consumption data by its nature reveals activities in one's home and existing law (the 4th Amendment) provides extra

[17] ACEEE's study in 2020 illustrated the significant underutilization of AMI data for energy savings programs. Noting "ACEEE surveyed 52 large utilities and found that most of them are greatly underutilizing this technology." *Leveraging Advanced Metering Infrastructure to Save Energy*

[18] J. Huang, "Bitcoin uses more electricity than many countries - New York Times", 2021. [Online]. Available: https://www.nytimes.com/interactive/2021/09/03/climate/bitcoin-carbon-footprint-electricity.html [Accessed: 21-Sept-2021]

protection when considering the boundary between one's home and the outside world. Clearly, all applicable laws must be considered in the assessment of risks and harms for releasing energy consumption data.

Given the inevitability of harms and risks identified in the literature and those yet unknown, it is important to ensure that approaches to mitigating risk are up to the challenge. For example, a 2020 paper explored some of the unique characteristics of differential privacy for the protection of building energy use data. The paper noting that every anonymization technique, from k-anonymity, to the 15/15 rule, to differential privacy, reveals some amount of information about the individuals in the dataset.[19] If one were able to find out the consumption information for 14 of the 15 individuals in a dataset anonymized by the 15/15 rule, for example, it would be possible to then deduce that final participant's energy usage; these are the same kind of breaches that were revealed in the Netflix and health data papers cited above. In contrast to other privacy techniques, differential privacy quantifies the risk to each individual contained in a dataset, no matter how much additional data is released about them. This risk is quantified as ε, the privacy parameter or factor.

Given the vulnerabilities of existing privacy practices, the guarantees offered by differential privacy mechanisms have a number of highly desirable properties.

- They are composable: it is possible to compute an exact bound on privacy loss from multiple statistical releases. Multiple queries against a database gracefully degrade privacy guarantees instead of catastrophically failing (non-binary outcomes are possible).
- The privacy protections are not dependent on an attacker's existing level of knowledge.
- The outputs of a differentially private mechanism are robust to post-processing, not compromising privacy no matter how much additional computation is performed on them.[20]

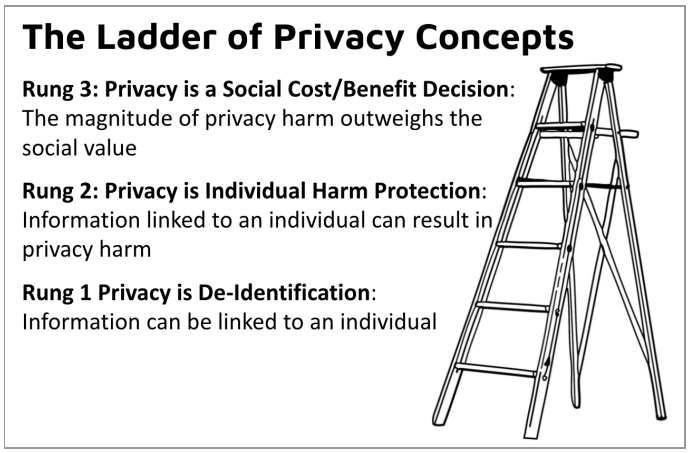
Differential privacy as a means of mitigating the risk of re-identification will be explored in more detail later in the chapter, but should not be mis-construed to be the only means of mitigating risk.

Climbing the Ladder of Privacy Risk

One of the challenging aspects of assessing privacy risk is reaching a common understanding of what privacy is in the first place. Privacy risk is highly context-dependent, with stakeholders often referring to quite different aspects with the same term of "privacy" or "risk". One conceptual framing that has proven useful in discussions so far is The Ladder of Privacy Concepts.

[19] Differential Privacy for Expanding Access to Building Energy Data (2020) Young, Paré; ACEEE Conference paper
[20] Ibid.

The simplest concept of privacy ("Privacy is De-Identification") is at the bottom of the ladder. In this concept, privacy is the notion that it is not possible to identify an individual's data in a given anonymized data release. Moving up the ladder broadens this concept to claim that re-identification is not a privacy harm on its own, but, instead, must also result in a meaningful privacy harm to an individual. Finally, at the top of the ladder is the concept that we may still deem an anonymization approach sufficiently "private" if it results in significant social value, even if some individuals are subject to privacy harm along the way.



These concepts build on one another. For example, to discuss the cost/benefit trade-off of privacy, we must first accept that re-identification of individual data is possible. This is not always trivial. Current privacy regulations are primarily written in terms of Rung 1, meaning that any re-identification at all is considered an illegal privacy violation; but in reality it may not lead to an actual harm and therefore may represent unnecessary limitations to the release of data useful to meeting other social objectives.

This rudimentary conceptual framework for privacy harm supports stakeholders in defining privacy in a concrete way. The idea of putting it in a ladder is that those engaged in the conversation have to come to terms with the first rungs of the ladder before climbing above that.

In the context of current guidelines or rules for aggregation of energy consumption data the "Rung 1" premise is that an individual cannot be isolated from the aggregation. When that happens, privacy has been breached and privacy has not been protected. It presents a cliff of protection. Aggregation is "you have it or don't." In reality, not all breaches have the same privacy harms, but if you've assumed that they do, then protecting privacy at the first rung is moot. This is why most access frameworks include classifications of the risks of various data types or other classifications to augment an aggregation threshold.

Protecting privacy is ultimately about avoiding privacy harms -- not just absolute protection. Rung 2 introduces the important nuance that in the data access construct we are making a tradeoff between the harms and the benefits of providing data access. This is also where a privacy parameter, rather than an aggregation threshold, starts to become useful because it offers some scalability in communicating and applying the ultimate choice. The potential variability of privacy parameters, including appropriate bounding, is considered alongside the specificity that may be included in a use case for the data, and the measures one may need to include in considering risk, harms and benefits. Aggregation thresholds could also be variable by sector, as they are in some jurisdictions, to address this need but given other qualities of utilizing a privacy factor may offer more flexibility by use case.

Putting A Risk on a Rung

Meaningful conversations around risk will be challenging. The "ladder" construct is offered as a means to organize the discussion and hone in on where the balance points may exist. Since every jurisdiction will also have variation in the collective understanding of harm or benefits, it is important to level-set the nature of the risk being addressed through particular rules, rather than getting hung up on debates around appropriate privacy filters that could be better addressed by access or other legal constructs (such as prohibiting sharing).

For example, the risk of energy consumption data getting into the hands of law enforcement could be addressed with legal constructs limiting or prohibiting this access. It may be true that many of the harms can just be mitigated by prohibiting law enforcement agencies from getting this information. If conversations are only focused on who can get data to cause harm, then laws to protect against those issues are the appropriate remedy which allows the data access framework to focus instead on the necessary conversations about the actual sort of risks and rewards, cost and benefits that come with the public purpose goals. Such laws and rules about access can be coupled with privacy protections (like differential privacy) to create space to have meaningful, actionable guidance.

While the overarching risks do need to be addressed in context (i.e. what powers do law enforcement have to get the data) such examples also illuminate the need for the mathematical solutions to be coupled with the social construct and that harms will vary by residential or commercial classes as well as other dimensions. By putting "risk on a rung," stakeholders can navigate these discussions with sharper intention and hopefully more clarity in the resulting guidance.

Filters for Mitigating and Managing Risk: The 5 Safes

Given the variation in understanding risk and attempting to classify it, at some point guidance and rules need to be clarified. What kinds of projects will we allow to get energy consumption data? Who will have access to energy consumption data? What capabilities will they need to have to protect data? How will they be expected to protect data? What are the boundaries for sharing the data? These are all common questions that have been tackled in data access proceedings around the United States, but have been taken on various forms.

One way to organize these questions was put forward by Arbucke and El Emam in "Building an Anonymization Pipeline." They present the consideration of use cases with respect to the "5 safes'', shown in the graphic below. This structure offers a sequential assessment of potential risk related to the core questions of "for what purpose is the data being shared"; who will have access, what capabilities to they need, what security and privacy practices are they expected to have, how are they expected to handle the data, and finally what requirements for data outputs will they be bound to. Adapting this list to a stakeholder proceeding process for assessing individual use cases, or for establishing high-level universal boundaries can simplify the discussion of risk and practical steps to address it.

[Arbuckle, El Emam. *Building an Anonymization Pipeline*]

An example of defining "Safe Projects" can be found in California's landmark Smart Grid Decision[21], where eight use cases that were submitted by stakeholders and further refined and deliberated in stakeholder working groups, are presented in the next chapter on "use cases." As noted in the earlier chapter, each use case also attempted to handle the 2/4 remaining safes as well: data, and output expectations defined in the regulations.

Another example of classifying "Safe People" and "Safe Settings" comes from New York. The criteria for third parties' data access includes a certification and audit process to ensure that entities who are getting access to data are qualified to protect it. In essence, this guidance covered both "safe people' and "safe settings". In the April 2021 Order the New York Department of Public Service adopted criteria for establishing a new "Data Ready Certification." The review of qualifications will be centralized to reduce the number of touch points or security agreements that a third party may need to acquire and allow streamlined compliance to be applicable across all utilities in the state. It is intended to replace existing rules for Data Security Agreement (DSA) and Self Attestation (SA) process for an energy service entity to validate that it has the necessary cybersecurity and can meet the privacy requirements of any utility providing data. It is essential that cybersecurity guidelines such as these are clear and specific so that those authorized to request data or access barriers can be prepared and those overseeing the process can enforce it fairly and consistently.

Given the typical bounds of data access framework deliberations, the "5 Safes" offer a useful starting point to look critically at the existing rules and frameworks to ensure that each of

[21] The decision *D.14-05-016* later adopted the working group report with modifications: https://docs.cpuc.ca.gov/PublishedDocs/Efile/G000/M076/K995/76995999.PDF

the 5 are handled appropriately and specifically where differential privacy may be an appropriate solution. The first two categories, projects and people, fall largely outside of any mathematical solutions that differential privacy may offer but are essential to determine social value, chain of custody and liability. With respect to settings, data, and outputs, there is significant space to consider the mathematical settings for differential privacy.

Using Privacy Factors To Explain Risk Tradeoffs (data & outputs)

Once the basic decisions on who may have access to data for what purposes are settled (with consideration of the technical possibilities for protecting the data), it is time to consider specific protections for that data when handled and particularly when shared. One option is using differential privacy or "blurring" techniques to protect the data. As noted in the first section of this chapter, this approach offers many advantages, but to date, differential privacy has not been at the forefront of considerations for protecting shared data. One of the key reasons is a lack of basic intuition around what a differential "privacy factor" is "saying".

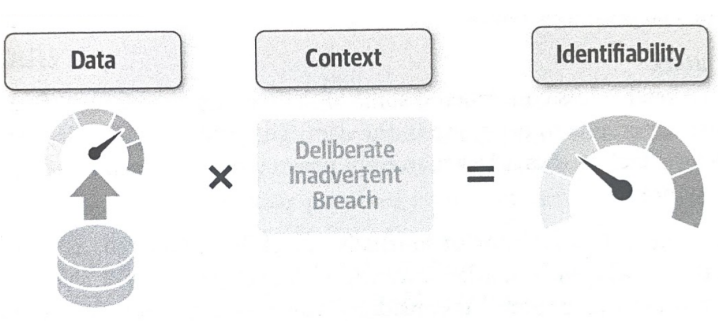
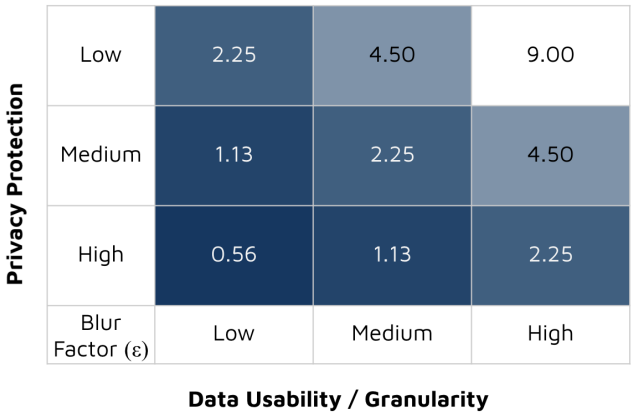


Figure 3-4. Overall identifiability is a combination of the probability of re-identification in the data given an attack times the probability of an attack in the first place (determined through threat modeling).



Improvised**: A risk matrix provides a visual demonstration of risks to assist decision making, and in this case the likelihood of an attempt to re-identify **v. the need for granularity.

Protecting against re-identification risk, Rung 1, is the combination of the probability of re-identification in the data given an attack times the probability of an attack in the first place. The probability of an attack is considered is partly mitigated in the rules around "safe people" and "safe projects" but the security against re-identification given an attack is where differential privacy can provide extra value.

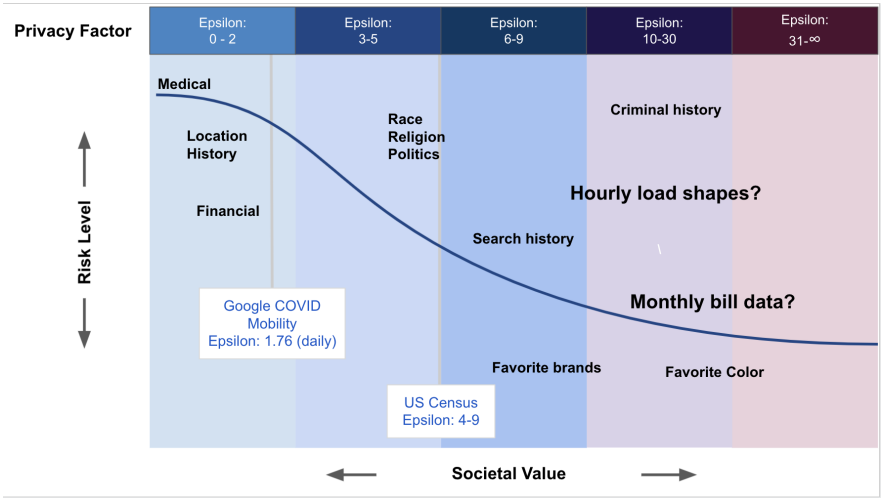
Aggregation-only thresholds face the primary limitation that when they're breached the data is insecure; it's a cliff of security. Differential privacy offers the ability to scale the level of security against the clarity of the data itself. The privacy factor or epsilon (ε) is the mathematical representation of that trade off between re-identification risk and the accuracy or clarity of the data.

ε is the exponent of the ratio of probabilities
between two databases that differ by a single element.[22]

$$Pr[A(D_1) \in S] \leq exp(\epsilon) \cdot Pr[A(D_2) \in S]$$

In the matrix (figure x.), the inverse relationship of the trade off and the resultant "privacy factor" or epsilon value is illustrated. Note that offering the highest resolution of the data offers the lowest data privacy protection; and the lowest data privacy protection offers the highest resolution in the data. Hence a privacy factor of 0.56 will be very secure but not offer high resolution of the data. A privacy factor of 9.0 offers higher resolution with lower re-identification protection (the actual values are for illustrative purposes only). The other combinations illustrate the range of privacy factors available to achieve a certain end; it is entirely possible that a privacy factor of 2.25 could offer high privacy protection and high granularity for one data set; and simultaneously offer low privacy protection and low granularity in another.

It may also be helpful to understand risk and data resolution trade offs for energy consumption data in the context of application of privacy factors for other situations. We caution and remind, however, that the "appropriate" privacy factor is a function of the dataset and how it is used and presented. Nonetheless, referencial heuristics are a helpful means of getting familiar with the metric and hone in on relative scales of privacy protection using a privacy factor (ε) as the metric. For example, Google protected COVID mobility data of individuals by using a privacy factor of 1.76 (daily); and the Census and LinkedIn typically use a privacy factor in the range of 4-9. While not a definitive means to establish the "right" privacy factor for a given use case or situation, these examples offer some boundaries to consider how energy consumption data may be reasonably protected. As is discussed later, the presentation and access components to the data also factor into the choice of the "right" privacy factor for a given situation. In some ways, it's most beneficial asset is as a communication tool to explain how any given data set has been protected.



Since aggregation thresholds factor so prominently in current data access frameworks, decision makers may want to know how a privacy factor (ε) would compare. Direct

[22] Differential Privacy Wikipedia page https://en.wikipedia.org/wiki/Differential_privacy

comparison is only possible if each is translated to a "margin of privacy error" for any given data set. Since each uses a fundamentally different mathematical construct for protecting privacy, one cannot say a particular privacy factor is offering more or less privacy than a given aggregation threshold in general terms. Key differentiating characteristics of the methods are presented in Table 1.

Each category in the table below illustrates various capabilities of the different approaches to protect privacy. The definition is provided first to explain what each approach means. The ability to adjust privacy thresholds by unique to a project or people is true for either privatization strategy, but accomplished in slightly different ways. Both can be applied in a risk based framework but have different characteristics to protect against re-construction attacks. As noted earlier aggregation is a binary test, and differential privacy offers a scaled test. The ability to describe how one applied each also varies, in that an aggregation threshold can only be truly known if it's compliant by revealing the data. For differential privacy the calculations used to derive the "noise" and how it is applied can be published alongside the data. As noted earlier, differential privacy is becoming widely recognized as a security best practice whereas aggregation has come under scrutiny in the literature for various vulnerabilities. Finally, in the context of data access frameworks the nature of a technical review for either approach may need to rely on data scientists or privacy experts to ensure compliance while aggregation thresholds may be enforceable by a more traditional stakeholder or advocate role.

Table 1. Differential Privacy and Aggregation Support Privacy Protection in Different Ways

	Aggregation Threshold	**Privacy Factor (ε)** (differential privatization)
Definition	*Representation of the maximum energy consumption (i.e. 15%) of a discrete number of individuals (i.e. 15) in a data aggregation needed to protect privacy of all individuals in the data set.*	*ε is the exponent of the ratio of probabilities between two databases that differ by a single element.[23] It is the indicator of the amount of "noise" added to a data set.* $Pr[A(D_1) \in S] \leq exp(\epsilon) \cdot Pr[A(D_2) \in S$
Adjustable based on Project or People	*Yes: 15/15 commonly used for commercial; 10/100 for residential*	*Yes: $0 < \varepsilon < \infty$ may be appropriate for variable use cases*
Risk-based Framework	*Must be understood in context of multiple factors to understand risk*	*Value directly represents the trade-off between clarity of data and re-identification risk for a given data set.*
Future-proof privacy bounds	*Does not protect against the introduction of current or future*	*Protects rigorously against linkage attacks and puts bounds on privacy*

[23] Differential Privacy Wikipedia page https://en.wikipedia.org/wiki/Differential_privacy

	side-information or reconstruction attacks	*leaks.*
Testable	*binary compliance status (yes/no)*	*scaled compliance status with testable assumptions*
Clear, traceable documentation of application	*cannot publish details of privacy protecting mechanism without compromising privacy by publishing the data itself*	*details of privacy protecting mechanism (instructions for generating the noise) can be published alongside results.*
Security best practice	*documented failures in the literature*	*prominently used for sensitive data and used by public entities (like U.S. Census)*
Technical review network	*consultant or advocate pool*	*privacy experts*

The difference between any given aggregation threshold versus a privacy factor could only be fully understood with a threat assessment of the potential risks of the aggregation. The valuable qualities and limitations of either approach are a more useful point of comparison. For example, differential privacy offers a clear articulation of the privacy "bought" with a privacy factor derived, whereas aggregation may have historically offered simplicity in inclusion in the regulatory process.

Conclusion

Risk and harm is a foundational element of energy data access frameworks. The ladder framework will help stakeholders take stock of the types of risks being considered from the basic risk of re-identification up to and through the trade-off of risks for capturing social value for data sharing. The "5 Safes" offer a consistent construct to consider the sequencing of data access considerations and ensure that all have been accounted for in any particular situation and that they mirror structures used in existing data access frameworks.

In any given data access framework, decisions must be made about how data may or may not be shared with third parties and how they in turn are expected to protect privacy. Historically, aggregation thresholds have been the "go-to" as precedent and simplicity, without much consideration for alternative approaches. In this chapter, the basics of differential privacy were presented to expand the options available of finding the right balance of risk, usability and benefits of providing data. Familiarity with the use of a privacy factor (ε) will take some practice, but many industry heuristics are emerging.

The key value-add for integrating differential privacy is by offering a way to more clearly calibrate and communicate risk for particular data outputs. Differential privacy guarantees the ability to disclose "some" information about an arbitrary set of reporting entities (groups, geographies, etc.) where aggregation-based rules do not. Moreover, the error associated with

these reported values can be tuned based upon a decision maker group's assessment of the re-identification risk. No such tuning is possible with aggregation. Once the aggregation rule has been chosen, the error properties are fixed relative to the statistical distribution of the customers in the reporting groups, geographies, etc.

ADDITIONAL REFERENCES:

Building an Anonymization Pipeline: Creating Safe Data; by Luk Arbuckle, Khaled El Emam ; (April 2020) Published by O'Reilly Media, Inc. ISBN: 9781492053439I; https://www.oreilly.com/library/view/building-an-anonymization/9781492053422/

Differential Privacy for Expanding Access to Building Energy Data (2020) Young, Paré; ACEEE Conference paper

Electronic Frontier Foundation Technical Memo, 2013. *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*

A Precautionary Approach to Big Data Privacy; Arvind Narayanan, Joanna Huey,Edward W. Felten March 19, 2015 *https://link.springer.com/chapter/10.1007/978-94-017-7376-8_13*

Leveraging Advanced Metering Infrastructure to Save Energy 2020, ACEEE Policy Paper

New York Data Access Framework: CASE 20-M-0082 - In the Matter of the Strategic Use of Energy-Related Data ORDER ADOPTING A DATA ACCESS FRAMEWORK AND ESTABLISHING FURTHER PROCESS, April 15, 2021

Defining Use Cases

LEARNING OBJECTIVES:

- Understand why defining use cases has been an important strategy in establishing energy data frameworks
- Consider how definitions of use cases may affect particular solutions for mitigating re-identification risk
- Understand how use cases could be enabled with differential privacy

How have "Use Cases" been used in energy data frameworks?

Tackling re-identification risk and access to energy data requires breaking the discussion down into more tangible parts like "who will get the data" and "how will the data be used?" These core questions allow stakeholders to hone in on what the actual risks may be and animate debate around who should have access to data and why they should be authorized to have access to certain types of data. It is an essential step in framing the boundaries of selecting appropriate privacy protections and access criteria.

The vignettes in this section span situations in which use cases are NOT defined, specifically going with universal aggregation "rule" for public facing data (Illinois), to defining use cases VERY specifically (California) to an even broader construct of defining risk-based groups (New Hampshire). This is not an exhaustive analysis of best practice or even preferred practice for defining use cases but rather helps illustrate some of the considerations and trade offs in defining use cases that can potentially affect the rules around how data is ultimately protected and how differential privacy may or may not be useful.

Universal Aggregation Rule in Lieu of Use Cases: Illinois

Illinois does not have a statewide standard to access aggregated energy use data for specified user groups.[24] The Commission did adopt a 15/15 rule for utilities when they release aggregated whole-building usage data which are compiled data sets of individual customer usage. This universal aggregation approach is also used in Colorado. Parties have noted that the effect of adopting such a blunt instrument has potentially limited the consideration of use cases and more detailed considerations of the tradeoffs of potential value from usability

[24] "*There are no statewide standards for access to aggregated energy use data. The provision of aggregated usage data is described in Docket No. 13-0506. The Commission adopted a 15/15 Rule when a utility releases an anonymized, compiled data set of individual customer usage. It means that a utility is allowed to provide customer usage data when there are at least 15 customers (within a delivery class) within the same geographic area and a single customer's load must not comprise more than 15% of the customer group's total load." Illinois Commerce Commission Docket No. 13-0506, Final Order at 17*

of the data for more discrete situations. The possibility of applying techniques such as differential privacy to help create this balance is also blocked.

Highly Specified Use Cases - California

In 2013, the California Public Utilities Commission solicited parties in the Smart Grid Proceeding to describe and propose high priority energy data use cases to *". . . add clarity to the current situation in ways that would help both utilities and requestors of data.*"

The Commission provided a template to enable parties to present their case consistently and to enable comparisons across use cases, a practice that has been repeated in many jurisdictions considering data access frameworks. From this request eight very specific use cases emerged for further consideration in a working group.[25] The recognized use cases were:

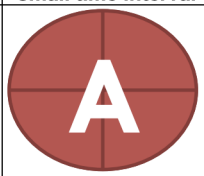
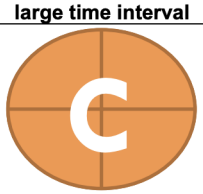
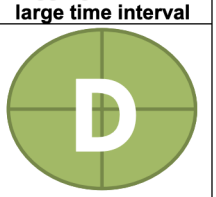
> ***Use Case 1:*** *Local Governments seeking access to aggregate data for use in creating legislatively required Climate Action Plans and implementation of energy efficiency programs.*
>
> ***Use Case 2:*** *Research institutions seeking monthly billing data, which may be PII, to evaluate energy policies, including energy efficiency policies, and publishing results in aggregate, non-PII form.*
>
> ***Use Case 3:*** *Research institutions seeking anonymous, individual hourly energy consumption data with other energy-related characteristics to evaluate energy policies, including energy efficiency programs and rate design, and publishing results as statistical coefficients. Thus, the data could be PII if it contained sufficient characteristics to permit reverse engineering, but the published results that describe the influence of energy-related attributes on consumption, would not be PII.*
>
> ***Use Case 4:*** *Other governmental entities, like the CEC's Energy Upgrade California Program, seeking energy efficiency program participation data by customer identification number in order to cross-reference this data with other program data, and thereby evaluate government-sponsored, legislatively mandated programs, while publishing results in aggregate, non-PII form. Thus, this data is highly granular, but non-PII, while may be "reversed engineered," but the published results would be non-PII.*
>
> ***Use Case 5****: Environmental non-governmental organizations, like the NRDC, requesting PII customer repayment history and energy consumption pre and post-retrofit for energy efficiency, to support general financial decision-making on energy-efficiency investments through on-bill financing, and produce results that provide aggregate, non-PII findings that link energy usage to other relevant characteristics (e.g. geography, building characteristics, customer financial characteristics, and financing vehicle). In this case, the data is definitely PII, but the results – a decision whether a particular area, type of building, type of customer, or type of financing is viable – in non-PII.*
>
> ***Use Case 6:*** *Solar installation company requesting monthly energy consumption data energy efficiency and participation in the net energy metering program, aggregated to a geographic area that protects PII, to reduce the product development and engineering costs in order to advance residential and commercial solar installations. In this case, the data, prior to aggregation, is PII, while the results – the identification of areas where solar power is financially feasible – is non-PII.*
>
> ***Use Case 7:*** *Building owners and managers seeking monthly energy consumption by building to conduct building benchmarking analyses pursuant to AB 758 and AB1103, and publishing aggregate, non-PII results. In this case, raw data that is PII would likely be needed, but the results concerning the efficacy of the program, are not PII. Moreover, it may prove possible to anonymize such data via an algorithm.*
>
> ***Use Case 8****: Energy efficiency contractor seeking CPUC-released aggregate data, similar to what the California Solar Statistics program releases, but using Energy Upgrade California data and other aggregate energy consumption data, to help validate the quality and value of energy efficiency work. Here, the raw data studied is likely PII but the*

[25] Administrative Law Judge's Ruling Setting Schedule to Establish "Data Use Cases," timelines for Provision of Data, and Model Non-Disclosure Agreements, February 2013.

program result – the validation of the energy efficiency work – does not necessarily reveal PII. Once again, it may prove possible to apply an algorithm that provides anonymization that cannot be reverse engineered.

Each use case includes the identification of the **party** that will receive the data, the **purpose** for receiving the data (sometimes multiple purposes in a single-use case), and the **nature of the data** they will be receiving. As parties discussed these use cases, they first classified risk and reward components based on the type of data that would be shared as defined in the use case. This was intended to limit the request to the minimum data needed to conduct analysis and fulfill the purpose.

With the benefit of hindsight, some authorized parties cited in these use cases have found that they cannot fulfill the basic purpose or intent of the use case within the given data requirements. Having such detail in the use case may have had the unintended consequence of precluding access to more granular data for valid purposes. It may also have limited the consideration of particular methods to protect privacy (such as differential privacy) because an aggregation threshold was already assumed. However, the thoroughness with which location, temporal, sensitivity and public value were considered in the process also illustrates that risks and rewards were carefully assessed in the stakeholder process.

	Specific location and small time interval	**Geographic aggregation and small time interval**	**Specific location and large time interval**	**Geographic aggregation and large time interval**
Quadrant Label				
Sensitivity	**High** Clearly personally identifiable, includes details of timing, and specific activities can be exposed.	**Moderate** Location is not personally identifiable.	**Moderate** Location is identifiable. Monthly (or annual) data masks timing of specific activities, such as startup or occupancy.	**Low** Not personally identifiable. Monthly or annual interval masks specific activities.
Public Policy Value	**Limited** Contains more data than necessary for uses other than academic research or services provided with consent.	**Moderate** Illuminates load shape, limited use in efficiency program delivery.	**High** Informs priorities for investment and service delivery.	**High** Essential for greenhouse gas emissions tracking and city planning.

Source: Working Group Report - California Smart Grid Decision

For the Local Government use case, it specifies both who is handling the data and the purpose for which they are using it. If the regulation would have instead distinguished between the objective for having the data from who will be handling the data it is possible that other viable data protections (like aggregation) could have been used. For example, Local Governments could have been granted access to raw data, or designated as a trusted handler of private data if they could ensure protection from public exposure, or mitigate risk of internal sharing (i.e. energy data getting in the hands of police for example). In hindsight, a hybrid approach may have proven helpful - allowing the Local Government to be a trusted handler of private data to afford policymakers the clarity and analytic power of the private

data, while recognizing that public policy in a democracy requires some information to be public.

In energy consumption data frameworks, utilities are by default the data custodians (handlers and providers) of the data from the original source - utility meters. Therefore, it is logical that data frameworks primarily contemplate applying protections prior to the default custodian releasing the data. It is important however to separately consider "who" may have access to data from "what" they may do with it. If data custodians or other actors can demonstrate adequate protections, then limitations on access may not be as necessary; however rigorous requirements for them to protect the data can remain intact.

The detail on defining use cases in this example also points to the challenge of highly specific rules being both comprehensive and omniscient at a single point in time. It is a good idea to adopt systems, or clearly call out procedures to modify, maintain and refresh the guidance as experience provides insight and as technology and conditions evolve.

Breaking up the Considerations for a Use Case: New York

Instead of combining qualifications, data structure, and protections in a single use case, New York has established a separate process for considering who is qualified to handle data.[26] In the April 2021 Order, the New York Department of Public Service adopted a new "Data Ready Certification" process that was intended to resolve many of the issues that "*have hampered data access up to this point.*" (see Risk chapter) This clear separation of the consideration of who is qualified to have data and the end uses is a helpful way to ensure that unintended barriers are minimized.

New York is simultaneously soliciting use cases from stakeholders to prioritize in the development of public data outputs from an Integrated Energy Data Repository (IEDR) as is common. This process is similarly offering stakeholders the opportunity to define use cases and identifying the "Minimum Necessary Data Attributes'' is part of the requirements. This includes specifying the precision, accuracy, granularity of the data needed. New York envisions that analytics may be embedded within the centralized data repository system. This derivative data could be made available to the public or authorized entities. The IEDR could also simply serve as a portal for access to "raw" data to authorized third parties who would then be required to apply explicit privacy protections when presenting or releasing data. Since the IEDR could potentially serve both purposes, privacy protections will need to be considered for each use case.

While New York adopted a universal statewide aggregated data set privacy screen of 4/50 to be applied generally to all aggregated data sets reporting monthly or annual energy usage totals, they also recognized that use-case specific data screens (including differential privacy) may also be considered for special applications or initiatives. This more contemporary

[26] California considers this in the context of non-disclosure agreements required for certain use cases that allow for sharing of private data without consent for primary purposes.

example of considering use-cases may offer flexibility to find the appropriate application of differential privacy.

Use Cases Translated to Risk-Based Groups: New Hampshire

New Hampshire offers an example of translating stakeholder use cases into a relatively simplified set of "risk-based groups." Privacy requirements are assigned to each group. The risk-based groups were the synthesized result of multiple use cases prepared for the Commission's consideration in developing the parameters of a centralized data hub (DE-19-197). In the final settlement, instead of articulating classes of users and associated rules, they classified four risk-based user groups based on the type of data they would likely use and established the guidelines for cybersecurity requirements for those data.

> *Platform users shall register as one of 4 risk-based groups depending on the expected data that will be downloaded from the Platform. The risk-based groups are:*
>
> *i. User of anonymized and aggregated data or municipal-level energy usage data. No customer permission required. Anonymized and aggregated data assumes municipal threshold level of 100 or more customers, or for pursuit of energy benchmarking with a contractual relationship at 4/50 (4 customers and no one customer is more than 50% of the data).*
>
> *ii. User with customer permissioned access to fewer than 100 customers' data at any time.*
>
> *iii. User with customer permissioned access to 100-1,000 customers' data at any time.*
>
> *iv. User with access to greater than 1,000 customer records at any time*

The categories illustrate discrete consideration of an end use (project) from a user (people). By specifying the aggregation threshold, it is also taking on the specification of how the data should be protected. It leaves little room for alternative methods to provide potentially better protection without sacrificing usability or enhancing usability without sacrificing protection. It could be fairly easily modified to add an exemption for differential privacy to be used as an option, with appropriate validation and documentation by qualified data handlers.

Conclusion

These four examples demonstrate the value of use cases to frame and limit the range of risks that may be inherent in choosing the appropriate privacy protections for a data set. They identify who may be using data, and the purposes for which it may be used and then decide how the data should be handled and how any outputs should be likewise be protected.

With no use cases, universal protections (like 15/15 rule) risk leaving a large amount of data unusable and therefore not accessed. Alternatively, very specific use cases may lack clear pathways for recourse after the practical challenges are revealed for specific use cases. More

synthesized interpretations can boil down to core use cases with clear protection criteria transferable to any data custodian.

It is impossible to fully anticipate every potential "end in mind" for the universe of use cases. By isolating the considerations, and then building in some flexibility and process for proposing and defending case-specific privacy protections, data access frameworks can be better prepared for the future.

In the example of the "Local Government" use case in California, the classification of a multi-pronged use case with a user may have created extra challenges with usability and access issues. Mechanisms that can authorize users (safe people) distinctly from projects may help mitigate these unintentional barriers, or at least isolate the potential risks of each factor. Perhaps specifying qualifications for being entrusted with data allows for more discrete guidelines for protecting privacy (with aggregation or with differential privacy). As such, entities such as local governments may be entrusted to manage and protect sensitive PII data (internally and externally) at the most granular level to more successfully meet their obligations.

Recommendations:

- Utilize use cases to assess and define key risk parameters for projects
- Consider qualifications of users independently from the end-use or project to ensure prosper consideration of each in the risk profile (i.e. do not confuse use case with user)
- Including explicit privacy protections may not be required if a path for review and approval of use-case specific protections is defined.
- Adopt an on-going process to review the functionality of use cases and associated requirements to ensure they comport with risk management best practice and a pathway to make corrections as needed.

ADDITIONAL REFERENCES:

State and Local Energy Efficiency Action Network. 2012. A Regulator's Privacy Guide to Third-Party Data Access for Energy Efficiency. Prepared by M. Dworkin, K. Johnson, D. Kreis, C. Rosser, J. Voegele, Vermont Law School; S. Weissman, UC Berkeley; M. Billingsley, C. Goldman, Lawrence Berkeley National Laboratory.

Illinois:

- ICC Docket 13-0506 (January 2014 Order)

California:

- Smart Grid Proceeding R.08-12-009 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.
- Administrative Law Judge's Ruling Setting Schedule to Establish "Data Use Cases," timelines for Provision of Data, and Model Non-Disclosure Agreements, February 2013.

New York:

- IEDR Proceeding Home Page (circa 2020-21)
- IEDR Proceeding (DPS Docket) : New York State Public Service Commission Case 20-M-0082: Proceeding on Motion of the Commission Regarding Strategic Use of Energy Related Data
- NY Data Access Framework Order (April 2021)

New Hampshire:

- New Hampshire Central Repository Docket No. DE 19-197
- State of New Hampshire Before the Public Utilities Commission Electric and Natural Gas Utilities Development of a Statewide, Multi-Use Online Energy Data Platform Docket No. DE 19-197 SETTLEMENT AGREEMENT
- Comments informing specific use cases:
    - Greentel Group Use Cases
    - Local Government Coalition Use Cases Proposals
    - Mission:data Use Cases Proposals
    - Office of Consumer Advocate Use Cases Proposals
    - Liberty Utilities request for service list changes
    - Joint Utility Comments - Greentel Use Cases
    - Joint Utility Comments - Mission Data Use Cases
    - Joint Utility Comments - OCA Use Cases
    - Joint Utility Comments - Local Government Coalition Use Cases

Choosing a Privacy Factor (ε)

LEARNING OBJECTIVES:

- Use the five "safes" to define a specific use case and document and validate an appropriate privacy factor for the relevant data set.
- Utilize the Energy Data Privacy Explorer to examine the tradeoffs for a particular data set and communicate the chosen privacy factor (ε)
- Understand possible privacy factor (ε) values for energy consumption data in relation to other contemporary applications of differential privacy

Using the Five Safes to Assess the Situation

The five "safes" construct borrowed from Arbuckle and El Emam allows for systematic consideration of a use case. It is also helpful in considering the proper privacy factor (ε) for protecting a data set using differential privacy or "blurring" techniques. In this chapter, we will use the structure to consider a specific use case and select an appropriate privacy factor (ε).

While any given jurisdiction may have different use cases and different boundaries of data access, almost all of them consider the five safes in one way or another. The proper application of differential privacy techniques is dependent on the data set one is privatizing. Hence, it is essential to consider the nature of the project, who will be handling and seeing the data, and the nature of the data in assessing what privacy factor will be applied to the outputs. Just as a universal aggregation is problematic, so too is a universal privacy factor (ε).

It is also possible to apply differential privacy or blurring techniques to various parts of the data flow. In assessing this use case, we'll simply consider privatization at the endpoint of presenting final data and results to a public agency and the public as a whole.

Safe Project: Demand Response Market Settlement

What are the legal and ethical boundaries of the data sharing scenario, and is privatization needed as a privacy-protective measure?

Defining a project provides an opportunity to describe both the scenario for which the data is needed and the context for using that data. As in most use cases, this is inherently where the "moral justification' has to come in as a threshold question. In some jurisdictions, this will validate or invalidate a particular use or clarify the specific requirement needed for aggregation.

In this case, an independent system operator (ISO) has a method for validating demand response impacts described in their tariff but wants to protect the privacy of non-participant customers in the final data sets. To implement this method, the energy consumption of non-participant energy consumers is compared to the energy consumption of participant energy consumers. This comparison is used to quantify the aggregate impact of an intentional demand response intervention and inform appropriate payment.

This type of use case may become more prominent with greater deployment of distributed energy resources among ISOs nationally.[27] Comparison groups using non-participant energy consumption patterns is recognized as a methodological best practice for assessing demand-side impacts. It may also be one of the best options for isolating the effects of dual participation or potential overlap of retail and wholesale interventions.

The relative impact of a demand-side intervention to the grid would be considered a "primary purpose" or core information for managing and operating the grid in a DER / renewable stacked grid. However, that doesn't negate the potential boundaries for data sharing. These may include: the impracticality for non-participant energy consumers to provide individual consent, the possibility that customers have a right to privacy against being identified in the data set, and even that the ISO may have limited authorization for seeing individual energy consumption data.

Safe People: 3rd Party DRP, ISO, Settlement Platform

Who are the anticipated data recipients, what are their motivations and capacity to re-identify (or obligation and capacity to protect) and who may they know in the data?

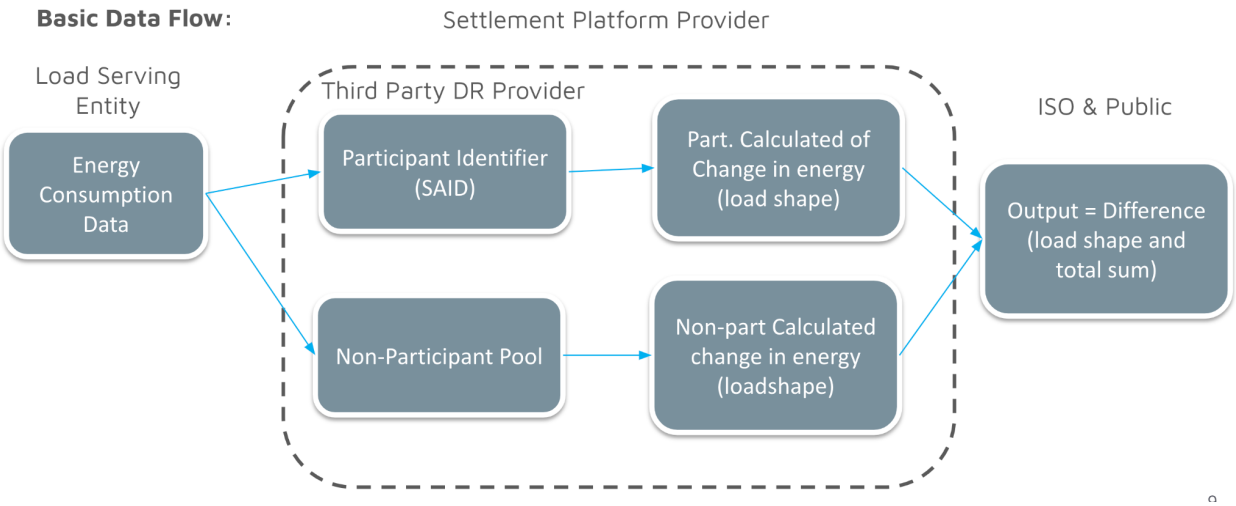
In the consideration of safe people, we need to catalogue who's in the system as well assessing whether they are trustworthy actors.

In this scenario, the source of the energy consumption data and other important metadata about participant and non-participant customers is the load serving entity (LSE) and the demand response provider(s). A settlement platform provider may also be involved as a data handler and the third party conducting the measurement and verification. In some cases the DRP could or would be considered the recipient and be responsible to conduct the calculations. The calculations are then translated to results and outputs for the independent system operator (ISO) or potentially the public, depending on the level of transparency necessary or expected.

In the diagram of data flows below, the nature of the data, responsible parties, and the treatment of the data for the analysis are shown. In this case, it is the savings load shapes (i.e. the difference in people's actual load shapes = the savings load shape). Since the final information is a derivative calculated value, it is not clear that this data is even sensitive at this point. For purposes of this exercise, we'll assume that at least the change in energy

[27] FERC Order 2222 called for consideration of DERs in all ISO/RTO markets and cites expectations for having means of avoiding double counting and robust telemetry and communications both of which could be partially addressed with measurement and verification approaches.

consumption of non-participants in the data set (used without consent) still calls for some threshold of protection.



We also know that individuals within organizations can pose risks, and the organization itself can pose risks. Given this set of actors, the nature of data they handle and protect would already require some level of protection against those threats as an ordinary course of business. This use case introduces no additional risks or threats on that front.

Safe Settings: Settlement Platform Vendor ++

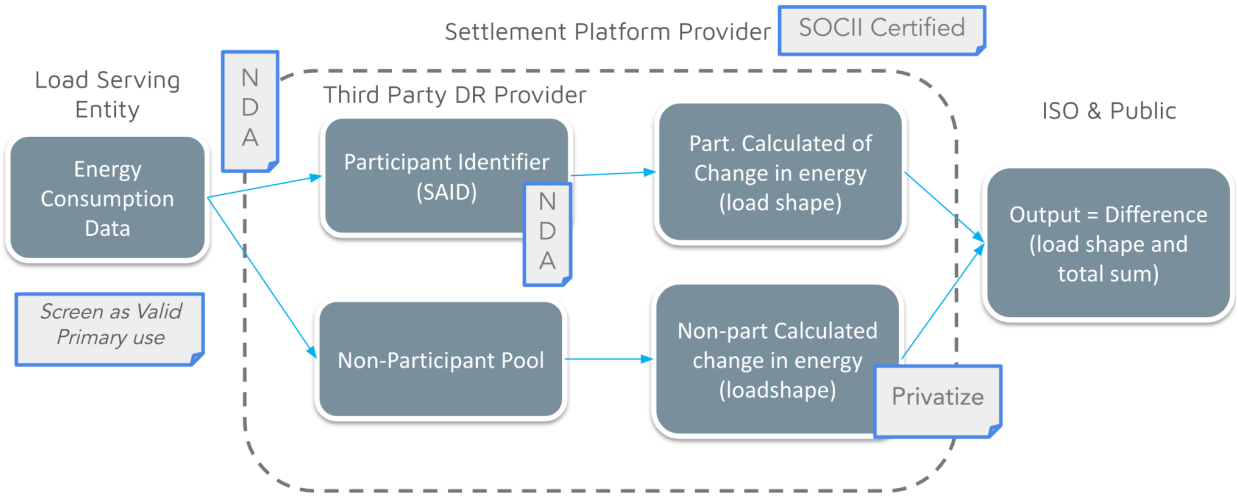
What are the technical and organizational controls in place to prevent a deliberate attempt to re-identify or prevent a data breach?

Even if the parties have been deemed trustworthy, it is necessary to ensure some protection in the system. This protection comes both in the form of an obligation to protect and a means of ensuring the entity has the proper qualifications to protect the data. Without either, the customers' data would be left with no protection.

As illustrated in the graphic of the data flow, several settings would be in place to transfer and protect the data. The first is screening for the valid use of the data for a primary purpose (i.e., managing the grid's operations). In most jurisdictions, any data sharing with a third party would require a non-disclosure agreement (NDA) to cover the protections and parameters necessary for handling the data. It would also likely be tied to a contract that articulates minimum qualifications to secure and protect that data (not just sign the contract). SOC 2[28] Certification is one example of that validation, as was described in the New York credentialing process.

Additional NDA's may also be present between participants and the demand response provider. The NDA would likely include provisions or consent to share the data (and/or limitations of sharing). Privatization or application of differential privacy can be considered as a specific "setting" and could be where an aggregation threshold is defined.

[28] SOC 2 *"Developed by the American Institute of CPAs (AICPA), SOC 2 defines criteria for managing customer data based on five "trust service principles"—security, availability, processing integrity, confidentiality and privacy."*



Safe Data: Hourly Load Shapes of Participants and Non-Participants

What is the level of identifiability, considering the people and settings of the data environment, and what threats to the data need to be managed?

The data in consideration for this scenario is primarily energy consumption data. Other identifying data like building type, location, and business may also be included, increasing the re-identification risk, but is essential to usability in this context.

Typical Data in the System:

- Hourly load shapes (i.e. 8760 energy consumption data)
- Derivative savings load shapes
- Service Account ID (for participants)
- Participant gives consent
- Zip Code or other locational information
- (Commercial) NAICs code (for matching)

Given the organizations in the system, we can have confidence that the people are qualified in handling the data and are going to comply with NDAs with secure transfer protocols and protections. The risk of re-identification based on load shape would depend on the breadth of identifiers in the data set (size, NAICs geographic breakdown) and the presence of outliers.

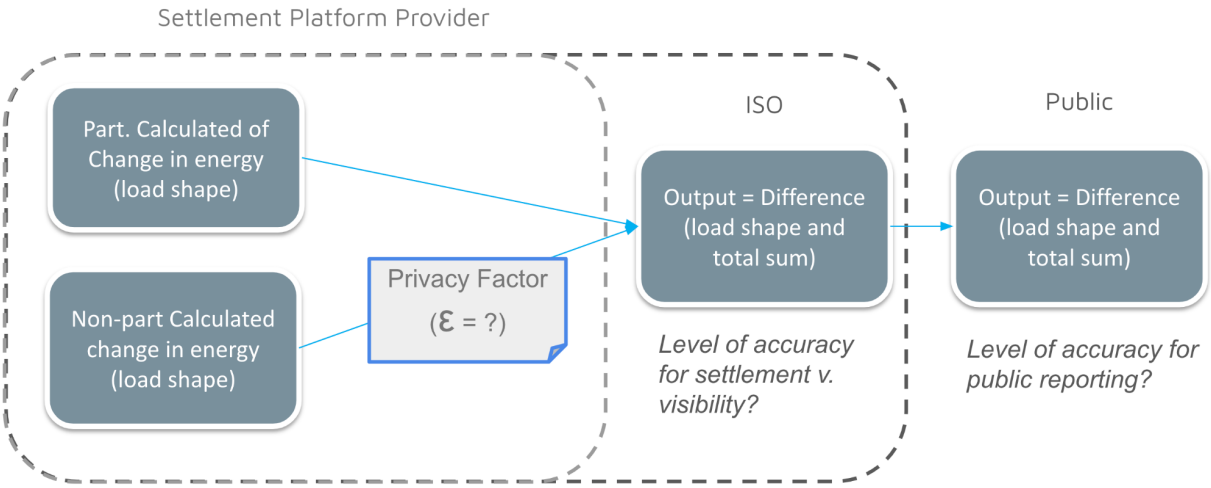
Safe Outputs: Appropriate Privacy Factor (ε)

What are the concerns in using the privatized data for the intended and other purposes and what is a suitable identifiability threshold (v. a usability/clarity/accuracy need)?

By looking at the end of the data flow and quantification process, we can focus on where the privacy factor (ε) would be applied. After the quantification using the raw data, the outputs of the non-participants would be privatized before it goes to the endpoint of the ISO and

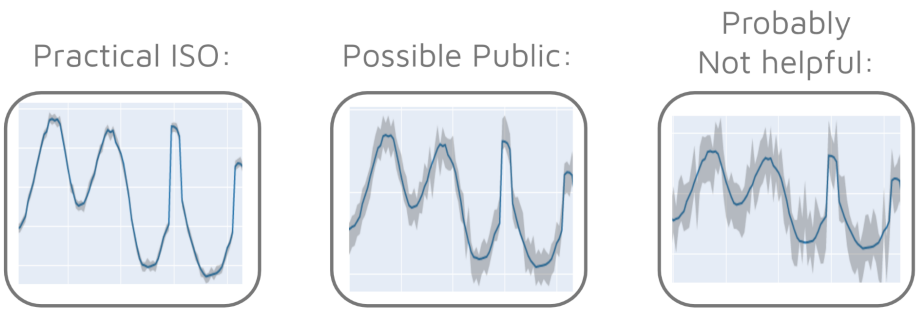
potentially the entire public. This is an example of "global application" of differential privacy wherein the output is privatized rather than the raw data coming into a system. At this juncture, a decision needs to be made regarding the "right" level of privacy that needs to be applied to the data set. Given all the steps to get there, it should be abundantly clear that this is a contextual decision. Considering data output is the same place an aggregation threshold would likely be applied to any public-facing data set.

It may be the case that the ISO would be authorized to see the raw data, but the public version would be privatized, or other combinations are envisioned. Given that the privacy factor (ε) reflects the granularity of the results and the privacy afforded, the trade-off needs to be resolved at this point in the process. The key factors beyond ethical considerations are the size of the data set and the presence of outliers.



Using the Energy Data Privacy Explorer to Consider Trade-Offs

To aid in the decision-making, the Energy Data Privacy Explorer can help describe the trade-offs statistically and visually present the outcomes of the options. Then they can be considered in either an optimization context (usability v. privacy), a heuristic context or based on existing regulations or obligations (if they exist). Most generally, a range of options may exist to privatize the outputs:



The Energy Data Privacy Explorer is an interactive software interface designed to help users consider the trade-off of precision and privacy protections in a given data set in partnership with DOE and NREL. The tool is intended to support the conversation around the selection of or perhaps the defense of an appropriate privacy factor and more effectively communicate a selected privacy factor decision to stakeholders.

The key factors that will affect the selection of the privacy factor (ε) are:

- the **size** of the data set (more is easier to protect - less loss of granularity)
- the **variance** of the energy consumption patterns within the data set is likely driven by the presence of outliers and (less variance means easier to protect - less loss of granularity)
- the **number of data points** that are going to be presented (fewer data points mean it's easier to protect - may lose granularity on longitudinal value)

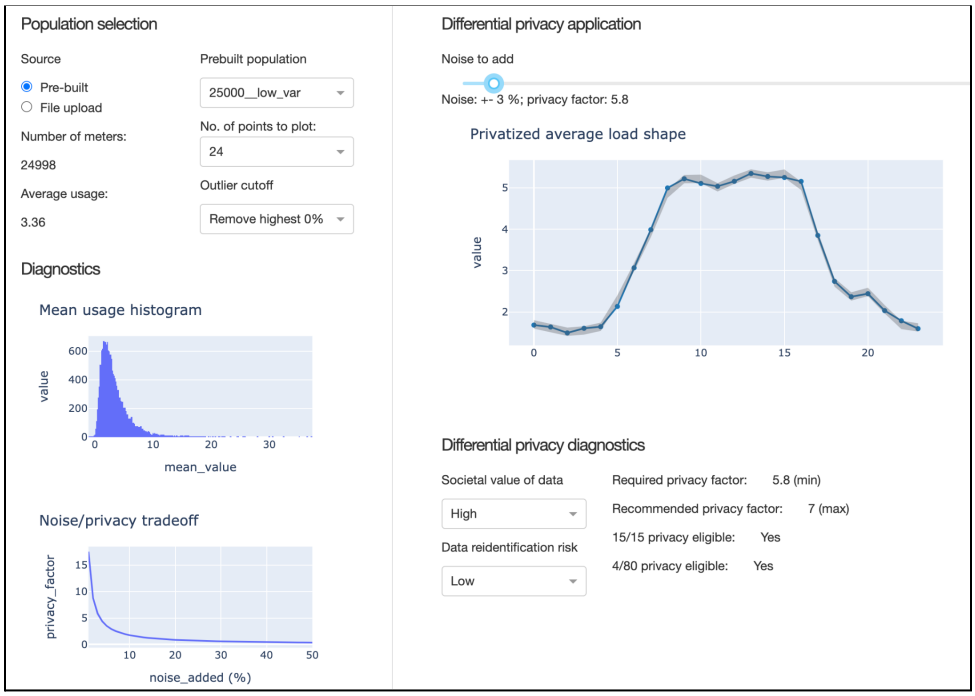
The Energy Differential Privacy Explorer allows users to test the sensitivity of these key parameters to build a more intuitive sense of what acceptable privacy factors (ε) may be when discussing a given use case. The source code is available[29] to enable users to test their own data sets and communicate, and potentially justify or negotiate, the privacy factor for a given use case.

Returning to the ISO settlement use case scenario, a primary goal is to provide a relatively tight error band for the integrity of the reported outputs and still provide some protection in the re-identification of non-participants in the dataset. It is a high value use case for the data to the electrical system operations and therefore society. The risk of re-identification is quite low given the nature of the anonymized load shape data, especially if the data is presented in a static report versus a downloadable data set.

Let's consider some options. Imagine a data set with 25,000 individual meters in the population with low variance in their energy use. A low level of error can be obtained without having to remove outliers. The recommended privacy factor (ε) of 7 is based on the heuristic assessment of the risk of re-identification versus the societal value of the data (see the next section for more detail on likely boundaries). If the goal for this data set is to keep the error in the range of 1-3 percent, the required privacy factor may need to be set at a liberal value. The size of the privacy factor (ε) is inverse of the level of re-identification privacy protection. A privacy factor of 5.8 offers **more** privacy protection than a privacy factor of 7.

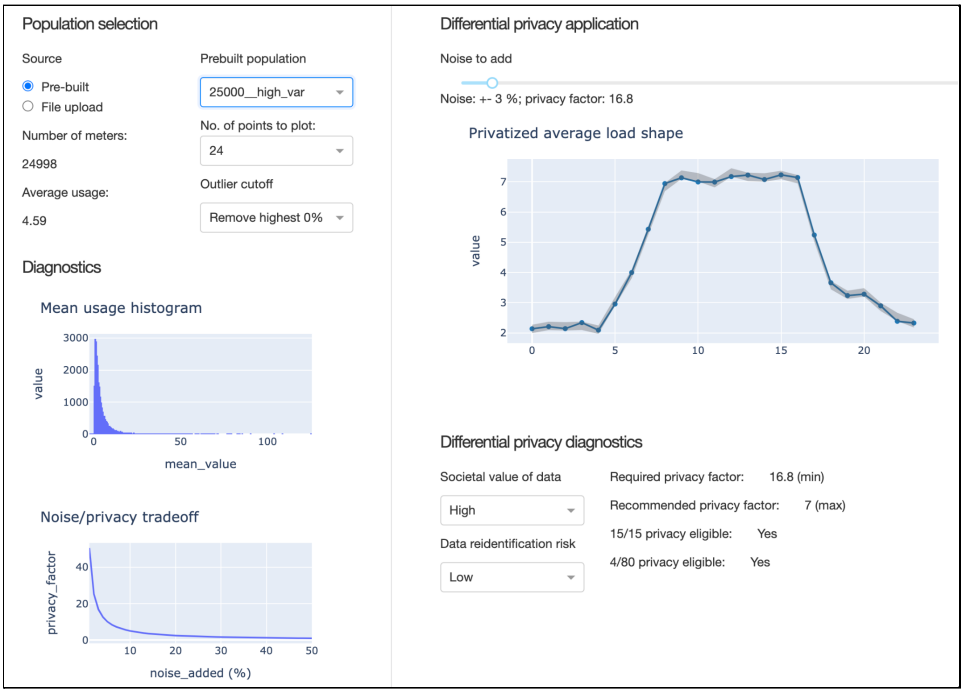
[29] eeprivacy contains the numerical methods to support differential privacy computations: https://github.com/recurve-inc/eeprivacy
edp_explorer is the source code for the web application:
https://github.com/recurve-methods/edp_explorer

Image 1. Large Sample; Low Variance; Low Error Range; Privacy factor (ε) = 5.8



Note that with the same number in the population (25,000) but with a high variance within the population, the required privacy factor (ε) is higher (16.8) if the low error band needs to stay in place.

Image 2. Large Sample; High Variance; Low Error Range; Privacy factor (ε) = 16.8



Simply removing 1% of the outliers can allow the privacy factor to come back into range:

Image 3. Large Sample; High Variance (remove 1% outliers); Low Error Range; Privacy factor (ε) = 8.7

Stakeholders or practitioners can configure many other scenarios with the Energy Differential Privacy Explorer to consider smaller sample sizes, higher error bounds, and the number of data points shared. If a debate is centering on the societal value of the data, it is possible to toggle the options and discuss the scale of the impacts with other stakeholders or clients.

It is important to note that while this synthetic data set would also be compliant with 15/15 aggregation thresholds (see "diagnostics" in the lower right hand category), there are likely instances where real data sets cannot meet both criteria. More importantly, as we've noted throughout this guide, differential privacy can offer many benefits, including greater protection since differential privacy techniques augment aggregation by adding random noise to the outputs.[30] The privacy factor (ε) is also a helpful mathematical expression to communicate precisely the protection applied to the data, once a little intuition is built around what it means.

Given that there is little application of differential privacy in the energy consumption data literature, and current regulations do not prescribe an acceptable range for epsilon (which we don't recommend) we can consider other heuristics to see how values compare. The application of differential privacy for the U.S. census is an excellent example of how the

[30] Differential Privacy: A Sound Way to Protect Private Data

selection of the privacy factor can become controversial.[31] Contemporary applications in COVID mobility data is another great example.

Google's COVID-19 mobility data subjects users to a daily ε=1.76, while Facebook's COVID-19 Movement Range reports users to daily ε=2.0. Other applications, such as LinkedIn and the US Census, set ε values in the range of 4 to 9.[32]

	Comparison Group ε
Consumption Histogram	0.1
SVT for Clamping Bound	0.2
Comparison Group Load Shape	4.0
Total Predicted Comparison	1.25
Total Observed Comparison	1.25
Total Privacy Impact (Pre-amplification)	6.8
Amplification factor[2]	0.124
Total Privacy Impact	**0.843**

Privacy Impact of OhmConnect Comparison Group Use Case

One example of applying differential privacy comes from an NREL/DOE report completed with Recurve. In this project, the platform provider had a large homogeneous data set. As a result the privacy factor (ε) offered a high degree of privacy protection with a non-participating comparison group loadshape with an ε=0.843. The results had a very low error factor (±0.5% on estimated savings of 19.3%). As noted the data set was large, over 60,000 meters, and the savings estimate was for a limited number of data points related to a demand response event.

As illustrated in the table on the left, the privacy impact (epsilon) prior to applying an amplification factor (which accounted for use of the original full population of customers) was 6.8 which is right in a similar range as privacy protection offered in the U.S. Census data.[33] Differential privacy was chosen as a more robust means of protecting the data, in lieu of specific aggregation rules or regulations for sharing hourly consumption data.[34] Smaller data sets with greater variability in load shapes will not likely have such luck in establishing a tight error band and high level of privacy protection.

Conclusion

Since differential privacy is used to provide privacy protection defined by a specific data set, and data sets need to be protected in different ways for different conditions, the final privacy factor (ε) appropriate for the situation may be different. Heuristic ranges relative to other social situations, and perhaps ranges or boundaries for given use cases, serve as a foundation for a defensible privacy factor. The Energy Data Privacy Explorer is a tool to understand the trade offs and communicate the choice internally and to others. Coupled with the Energy Efficiency Privacy Library, practitioners and decision-makers can operationalize the inclusion of differential privacy into revisions to or new data access frameworks.

[31] Differential Privacy for Census Data Explained
[32] See links in the Additional References section.
[33] For more information: *Applying Energy Differential Privacy To Enable Measurement of the OhmConnect Virtual Power Plant: A study of Demand Response during the California August 2020 blackouts* .
[34] California's aggregation requirements only cover monthly data sets as of the writing of this document.

ADDITIONAL REFERENCES

Energy Differential Privacy Explorer (tool)

Energy Efficiency Data Privacy Library (github code repo)

Applying Energy Differential Privacy To Enable Measurement of the OhmConnect Virtual Power Plant: A study of Demand Response during the California August 2020 blackouts
2020, Recurve Analytics Inc.

US Census:

- 2020 Census Data Products: Disclosure Avoidance Modernization (webpage)
- Differential Privacy for Census Data Explained (National Council of State Legislatures)

Google COVID Community Mobility Reports

- COVID Community Mobility Reports (Google)
- Google COVID-19 Community Mobility Reports: Anonymization Process Description (version 1.1) (Cornell University)

Facebook Movement Range Maps

- Protecting privacy in Facebook mobility data during the COVID-19 response (Facebook Research)