Title
Guidance on the usability-privacy tradeoff for utility customer data aggregation

Authors
Ruddell, Benjamin L
Cheng, Dan
Fournier, Eric Daniel
et al.

Publication Date
2020-12-01

DOI
10.1016/j.jup.2020.101106

Peer reviewed

Utilities Policy 67 (2020) 101106

Contents lists available at ScienceDirect

Utilities Policy

journal homepage: http://www.elsevier.com/locate/jup

Guidance on the usability-privacy tradeoff for utility customer data aggregation

Benjamin L. Ruddell [a,*], Dan Cheng [b], Eric Daniel Fournier [b], Stephanie Pincetl [b], Caryn Potter [a], Richard Rushforth [a]

[a] Northern Arizona University, School of Informatics Computing and Cyber Systems, Flagstaff, AZ, USA
[b] University of California Los Angeles, California Center for Sustainable Communities at the Institute of the Environment and Sustainability, Los Angeles, CA, USA

ARTICLE INFO

Keywords:
Utility customer data
Data privacy
Data de-identification
Data aggregation
Data ethics
Utility data policy
Regulation

ABSTRACT

Modern cities, along with their researchers and innovators can benefit from applying " big data" to their sustainability and infrastructure problems and policies, e.g., water and energy consumption. Unfortunately, current utility customer data (UCD) privacy rulemaking fails to ensure safe release of these data for the public benefit and does not currently strike a sound balance between the competing values of usability and privacy. This paper presents a statistical analysis of the tradeoff between usability and privacy for UCD in Los Angeles. The tradeoffs vary by economic sector (residential vs. commercial/industrial) and by utility type (water, electricity, natural gas). This paper provides guidance for safer and more ethically balanced aggregation and release of utility customer data.

1. Introduction

1.1. Public versus private benefits in the use of utility customer data

Utilities are publicly or privately-owned infrastructure operators and monopolistic providers of water, natural gas, and electricity. Within dense urban environments, at the individual account-level, energy and water utility customer data (UCD) tend to be non-uniformly distributed with respect to space, time, and various other categories of use [Fournier et al., 2019, Porse et al., 2016, Gurney et al., 2015, Pincetl et al., 2015, Rushforth and Ruddell, 2015, Mini et al., 2014]. Understanding the causes and consequences of these patterns requires the ability to match historical consumption data with socio-demographic attributes, building attributes, climate attributes, as well as information about the level and distribution of energy and water efficiency or conservation program participation.

Gaining access to account-level UCD enables researchers and innovators to develop solutions, including new products, demand-side energy and water management measures, as well as insight into issues of equity and justice. The information can also be abused in ways that do not benefit the public or the individual customer holding the account. In today's dramatically changing energy and water utility landscape, including water stress, Community Choice Aggregators (CCAs), concerns over greenhouse gas emissions and air pollution, and customer preferences around water and energy, the potential for using UCD to improve policy for public benefit- and to prevent the abuse of UCD-has become more urgent [Zipper et al., 2019, Szulecki, 2018, Becker and Naumann, 2017, Sweeney, 2014, Weinrub, 2017, Schoor and Scholtens, 2015, Hoffman and High-Pippert, 2005].

In the absence of accessible and granular energy and water UCD, insights into the relationships between urban form (such as building vintage, size, and zoned use) as well as community socio-demographics (such as age, race, income, poverty) and resource usage, are hindered. There are equity and social justice problems created by this lack of access to data [Hayashi, 2013]. Furthermore, without granular UCD, municipalities, and adjacent-industry regulators charged with developing resource usage efficiency or conservation programs must do so in ways that are ignorant of the detailed patterns of consumption of the jurisdiction's various users, from heavy users to vulnerable populations. Data gaps enhance the likelihood that the policies in question may end up being either ineffective or even counter-productive, relative to their stated objectives. For example, in California, investor-owned utility (IOU) ratepayers annually fund $1 billion to the financing of energy efficiency programs. The use of UCD *ex -post*, to empirically verify the accuracy of estimated efficiency program savings or to assess the equity the policy, continues to be rare and inadequate to justify the scale of

* Corresponding author.
E-mail address: benjamin.ruddell@nau.edu (B.L. Ruddell).

https://doi.org/10.1016/j.jup.2020.101106
Received 30 October 2019; Received in revised form 2 August 2020; Accepted 2 August 2020
Available online 9 October 2020
0957-1787/

investment in the programs-because the necessary observational performance data is frequently unavailable (Liang et al., 2017).

1.2. Classification of UCD as private and personally identifiable information

In most U.S. states, as in most other parts of the world, UCD at the granularity of the account, customer name, or address is presumed to be personally identifiable information (PII, e.g. McCallister, 2010) and is often considered " protected" PII data (WPUDA, 2016, PUC-CO, 2015). Some common examples of UCD-PII include water, wastewater, electricity, and natural gas use. As the need to better understand energy and water use has grown, utilities and regulators have responded by developing data aggregation and customer data privacy rules to protect UCD-PII. Concerns about releasing utility customer data are similar to the privacy issues in gathering data from private landowners in conservation easement contracts, alongside many other private-land conservation practices (Rissman et al., 2017), and also resemble concerns over the release of student data or personal health data. Data fusion using GIS, remote sensing, and digital trace data create added concerns (Zipper et al., 2019). Unfortunately, UCD rules to date have been developed without sufficient examination of the risks or benefits created through the release of UCD, and strike a poor balance between the public benefit in the analytical usefulness of the data and the individual interest in privacy. In some cases, a lack of nuanced distinction and statistical rigor in the rules leads to outright mistakes where sensitive data is released, and non-sensitive data is withheld.

Researchers and regulators have both sought to strike a workable and acceptable balance between the need to protect the privacy of their subjects and the desire to extract value from account-level data. This balance has been most elaborately struck in the field of public health research, but now the balance must also be forged by a number of other fields as well. For example, economic studies involving the use of confidential financial and business transactions, and social and political science research, where investigators have administered large scale surveys involving questions on potentially sensitive subjects, are now widespread. Our study concerns balancing the rights of individuals to privacy and the desire to enhance public benefit through useful application and study of personally identifiable information (PII) within the context of UCD-specifically in this paper's example, account level water, electricity, and natural gas data.

The balance of analytical usefulness and privacy has a statistical and regulatory manifestation but is ultimately an ethical exercise. The ethics are nuanced. In contrast to individual health, educational, or transactional data, water and energy UCD pose a different usability-privacy tradeoff. In our opinion, UCD may pose lower risk to the individual and may involve a more compelling public interest as compared with many other contexts. This is because it is difficult to argue that individual users- or the public benefit-would be significantly harmed if account level utility usage data were to become public knowledge, and court cases demonstrating this harm are absent. Of course, the details matter a great deal, and (perhaps out of an abundance of caution) rules limiting the release of raw account-level data are nevertheless widespread. Harms are possible in theory: utility customers could, in theory, be more exposed to shaming, relentless sales pitches, theft of trade secrets, or fraud, if too much of their data is made public (Zipper et al., 2019). We are not advocating the release of individuals' customer data, but rather for a more nuanced, expertly determined, and statistically precise, and competent process for determining what data is released and at what level of aggregation for de-identification. It makes sense to allow broader access to UCD, if we can balance the policy objectives of usability and privacy. This paper attempts to address the current deficit in rules for releasing UCD-PII [Sirkiä et al., 2017] to make as much information available as possible without creating risk of excessive disclosure.

Fundamental tradeoffs are involved in accessing and using personal data in research for public benefit, as has been widely discussed in the academic literature [Jensen 2013; Li and Li 2009; Loukides and Shao 2008; Ramanayake and Zayatz 2010; Rastogi et al., 2007; Sankar and Rajagopalan 2013; Wu 2013]. These tradeoffs matter because data has economic and research value, and this value increases along with analytical usefulness and with the granular identifiability of the subject [Acquisti, 2010]. Data aggregation tends to undermine the usefulness of the data [Brickell and Shmatikov, 2008; Emam et al., 2009]. However, optimization of the usability -privacy tradeoff is possible in each context, and researchers and regulators are motivated by the public interest to seek this optimization [Ghosh et al., 2012; Lane, 2005; Zhong et al., 2005].

Differential privacy and algorithmic privacy are promising solutions to the privacy dilemma. Differential privacy allows the calibration of usabilty and privacy to suit the user and use case [Eigner et al., 2014; Soria-Comas et al., 2014]. Aggarwal and Philip (2008) survey algorithms that can extract data from individual data without de-identification in the final results. Applebaum et al. (2010) demonstrate that algorithms can facilitate anonymous data sharing. Privacy can be designed into raw -sensor data flow, but by default is not [Cavoukian et al., 2010, Chakraborty et al., 2012, Efthymiou and Kalogridis, 2010, Erkin et al., 2013, Groat et al., 2011, Rajagopalan et al., 2011]. However, this paper does not focus on the promising future solutions of differential or algorithmic privacy, but rather on the present-day problem of determining how to aggregate and de-identify a given type of UCD without risking excessive disclosure.

Many techniques for de-identification and aggregation of PII have been developed by computer science and database researchers. Some of these methods include randomization, k-anonymity, l-diversity models, and distributed privacy preservation [Aggarwal and Philip, 2008]. However, the vast majority of these techniques have been developed for data that either lack geographic identifiers or upon which geographically explicit analyses are not intended to be performed. When geographical identifiers are present and necessary for useful analysis, as in the case of account-level UCD, it creates new and different challenges for usability -privacy tradeoffs. In the instance of spatial groups of individuals-which is the relevant instance for utility customers linked to addresses, aggregation of small numbers of customers dramatically increases the risk of re-identification due to the potential for spatial linking of datasets among other threats [Emam et al., 2010, 2011; Young et al., 2009; Zayatz, 2007]. Spatial linking and re-identification is a more severe technical problem for UCD because spatial usage patterns and spatial linking are essential to the analytical utility of UCD, unlike more typical PII data (e.g., personal health information).

In summary, UCD a special class of PII and requires a unique balance between usefulness and privacy. Having established the need for better guidance, we next examine problems with current rules and practices for UCD-PII management in the broader context of U.S. PII law and research data ethics. We argue that current UCD-PII rules could be dramatically improved within the existing framework of law, precedent, and rules. In the methods and results sections, we conduct a quantitative analysis of usability -privacy tradeoffs in the context of Los Angeles UCD aggregation for water, electricity, and natural gas data. We conclude with specific recommendations for how regulators can improve rules for balancing of usability and privacy for UCD-PII applications.

1.3. Existing PII De-Identification Rules and the Rule of Fifteen

The prevailing rulemaking frameworks for UCD use in the United States draw heavily from the experience of California as the first state to create relevant regulations for UCD aggregation, including the Rule of Fifteen (R15). R15 mandates that UCD may not be released unless aggregated into group size N of at least fifteen customers ($N \geq 15$), with no single customer comprising a maximum fraction F more than fifteen percent of the group's total volume ($F \leq 15\%$ or 0.15). See Appendix A for a detailed discussion of this historical context for the "Rule of Fifteen"

(R15) and similar regulations that narrowly prescribe UCD aggregation practices. As we will demonstrate below, both N and F are critical variables in relation to UCD and require a more nuanced treatment of PII than R15 allows.

Schwartz and Solove (2011) argue that the meaning of PII is unclear without detailed clarification of the use case and processes for that data [Schwartz and Solove 2011]. They observe three existing concepts covered by the term "PII" which jointly identify PII; (1) tautological, that is, sensitive PII is self-explanatory and is common sense to the typical person, (2) non-public, that is, PII is information that is not already available through publicly accessible sources, and (3) individually specific, such as name, address, or phone number. They also argue that the concept of PII should be extended to explicitly consider whether de-identified information could be re-identified through modern database linking techniques or computer analysis. Based on these arguments, it is clear that a blanket application of R15 to all utility account data, without nuance, is not sufficient to govern modern UCD-PII practices, because this blanket application does not explicitly consider which aspects of the "private" UCD is in fact already public (and thus by definition, not PII), or whether it could be made so through linking and quantitative analysis post-aggregation.

Unfortunately, the UCD aggregation rules in California and subsequent did not follow the well-established guidance, precedent, and rules governing the usage of other sensitive PII in the areas of health or finance. Health and financial transaction records are heavily regulated at the federal level by the Health Insurance Portability and Accountability Act (HIPAA) and the Financial Institution Privacy Protection Act (FIPPA) [HIPAA, U.S. Congress, 1996, FIPPA, U.S. Congress, 1999a,b]. The ad-hoc and historically unrooted regulatory environment for UCD-PII rulemaking means that many well-established and sophisticated guidelines and protections that exist for other PII are absent or underdeveloped for UCD-PII, namely, the expert-determination method, institutional review boards (IRB), rigorous data usage monitoring requirements, and reporting protocols and penalties for potential data breaches. HIPAA was developed for a heightened individual risk context compared with UCD. In spite of this- and perhaps because of this-the HIPAA process is an excellent template for improved UCD privacy rules.

R15 and HIPAA's "safe harbor" definition of PII are both examples of Schwartz and Solove (2011) "specific-types" definition. That is, R15 is a kind of safe -harbor method. HIPAA-compliant de-identification of private health records can be accomplished through the application of two methods: "expert determination" or "safe harbor". [HHS, 2003, 2012]. A recent proposed rule by the Arizona Corporation Commission prescribes a safe -harbor method used to maintain confidentiality in the release of grouped data [Arizona Corporation Commission Docket R14-2-2215, 2014]. After a dataset is de-identified by applying R15, data can be released to external parties, aggregated and joined to other datasets, and used for research purposes beyond the originally intended and primary use of the data without obtaining customer permission, because the aggregated data is no longer PII.

From a technical standpoint, the safe -harbor de-identification method is straightforward and simply involves the removal of specifically named attributes from a dataset prior to its release to a third party, including public and private entities. In Arizona's proposed law, for example, PII attributes would have included; Customer name, geographic location below the State level, dates that are indicative of the customer's age, phone numbers, email addresses, IP addresses, physical addresses, SSN's, all kinds of serial and account numbers, license numbers, URL's, photos, beneficiaries, biometrics, and any other uniquely identifying characteristics of the customer. The blanket exclusion of the physical address is problematic because it compromises the usefulness of the data for research that links to an address's physical attributes like square footage etc., but the other exclusions can be made without compromising address-linked utility consumption analyses. In practice, implementation of R15 for UCD-PII results in the masking of the resource consumption of many groups of users, especially those in the large commercial and large industrial groups that have relatively few and relatively high-volume (" whale") users. Small changes in the wording and interpretation of R15 have dramatic consequences for both the usability and privacy of aggregated datasets.

From an ethical standpoint, the safe harbor method is compromised because it may not allow an optimal tradeoff between usability and privacy to be struck and may allow serious mistakes in de-identification at the same time. For this reason, a more nuanced expert determination is better, at least where expert determination is worth the added effort. The safe harbor method (e.g., R15) is strict, blunt, fragile, and suboptimal. By contrast, the more nuanced expert-determination method requires the data aggregator to obtain an opinion from a qualified statistician that the risk of re-identifying an individual from the data set is relatively small. Then an independent, qualified, and authorized reviewer, such as a university's institutional review board (IRB), reviews and approves the expertly recommended (a) statistical aggregation method, (b) data protection practices (throwing away the keys linking original with de-identified data), and (c) ethical consideration of both benefit and risk. HIPAA provides the following guidance on how an expert should flexibly construct groups to prevent individual identification:

> "A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable: (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination." [HHS, 2012; HHS 2015]

We provide more detailed guidance on both the expert-determination and safe-harbor methods in the unique context of UCD-PII de-identification so that both processes can be applied without making serious errors, and so that future rules can be improved to allow for expert determination.

1.4. Common problems with safe harbor de-identification rules for UCD

1.4.1. The whale problem

Of the two parallel requirements of R15 (N and F), experience has shown that the stronger constraint on the masking of aggregated data tends to be the F $\leq$ 15% limit. At the customer account-level, water and energy resource usage data tend to be log-normally distributed, or small/left-skewed, with a "long tail" comprised of a small number of high-volume accounts (e.g., Kwac et al., [2013]). In the industry jargon, these high-volume accounts in the long tail of the distribution are called "whales." It is a logical error to craft a data aggregation rule that does not handle whales, because whales are ubiquitous in UCD.

Within the constraints posed by R15, there are at least two possible strategies for reducing this masking. The first involves defining a " whale group" across a diverse set of geographies or categories. The second involves simply dropping whales from their groups; this is standard practice at present. If we create a whale group, we are explicitly identifying the accounts within that group as such. Historically, this kind of account designation has proven to be politically sensitive and has been strongly opposed by large corporations on the grounds of protecting trade secrets.

However, if we drop whales completely, where should exceptionally large usage be accounted for in the data? If we exclude it entirely, we are deliberately introducing a large underreporting error. Worse, dropping whales normally fails to de-identify them due to the vulnerability of whales to the subtraction attack: we often know the total usage of a group from separate reports and summaries, so dropping the whale but reporting the group's total allows the accurate guessing of the whale's

usage by subtracting the group's reported total (sans whale) from the true total [Felten 2012].

Separately, if the N or F is small as in R15, this creates an added identifiability problem for whales: statistical certainty that there is no whale in a group. Assuming a long-tailed distribution of usage, F»15% as N→15. In a zero-sum game, identifying whale-free groups is equivalent to identifying groups containing the whales. This is a dilemma that must be avoided because it forces us to either identify the whales or mask all the groups.

1.4.2. Small group sizes are not a good tradeoff

When paired with the whale limit F, any small group size N guarantees relatively uniform usage within the group and compromises privacy. Consider the following hypothetical scenario where N = 15 utility customers form an aggregated group of total usage 100 that obeys an unusually bimodal distribution, such as if two homogenous but very different subgroups are aggregated together. In this extreme and limiting instance, six individuals each account for the maximum F = 15% of the group's total consumption, one customer uses 10% of the total, and eight use the minimum of F = 0% apiece. The average customer accounts for 1/15 = 6.7% of the group's total consumption; the largest customer uses 8.3% more than average; the smallest customer uses 6.7% less than average. If the group is log-normal, the central tendency guarantees far greater uniformity than this extreme bimodal instance. In other words, R15 ensures that customers are guessable within half or double the group average. This is a problem. By contrast, N = 50 gives a factor of roughly ten or 1000% separating the greatest and least user in the group and is statistically more private.

Separately, small N is a problem if statistical summaries are calculated and disclosed for a group, because a sample size of 15 does not guarantee sufficient accuracy when estimating various group statistics like mean, standard deviation, skew, etc. While the precise sample size required to estimate a given statistic at a desired threshold of power depends largely on the specific shape of the underlying population distribution, it has been observed that significantly larger sample sizes (consisting of as many as 50 to 100 individuals) would likely be required to achieve levels of statistical power in excess of 0.8 [Amatya et al., 2013, Dulal et al., 2013], at least for log-normal distribution types.

Nor does N = 15 provide enough granular detail for routine research questions in many cases. In the State of Illinois, R15 was found too strict for the purposes of evaluating groups of large power users and generators, because there are relatively few power plants, and a large percentage of them are whales. As a replacement for R15, the 4/80 rule was adopted for power plants in Illinois, which allows groups as small as N = 4 with a single facility responsible for not more than 80% of the whole group's total [Livingston et al., 2018]. N = 4 is the smallest group size rule we have encountered, but this is apparently sufficient for at least one U.S. state regulatory agency's purposes to preserve industrial users' privacy. The 4/80 rule may, paradoxically, de-identify more thoroughly than R15 in an important sense: 4/80 creates a 220% spread between the largest possible whale and the group's average, whereas 15/15 creates only a 124% spread. One issue faced by regulators in rulemaking is the counter-intuitive ways in which the aggregation rules may work and statistical expert determination may be useful.

1.4.3. Group construction mistake: concentrating a sensitive class within one group

Utility customers fall into different classes, some of which may be "sensitive classes" requiring greater privacy protection than others under ethical and legal guidelines. R15 fails to clarify the difference between sensitive and non-sensitive classes, subclasses, or geographical areas, or to recognize that forming a group from members of a sensitive class may be unethical and unallowable. Examples of potentially sensitive or protected classes are many, and include;

- Critical infrastructure system components,
- Law enforcement or public health facilities,
- Secure government facilities,
- Secret facilities,
- Facilities where controversial medical research is taking place,
- Facilities where dangerous chemicals or biological agents are in use,
- Universities or K-12 Institutions,
- Specialized industrial facilities with trade secret processes,
- Individuals with a common medical condition,
- Individuals who have explicitly denied consent for the use of their data,
- Members of a political, religious, or other voluntary affiliation, and
- Any vulnerable class that could otherwise be put at risk of harm if identified as a member of the class.

This problem should be solvable because many utility usage research inquiries are decorrelated and unconcerned with these types of users, making possible the construction of groups that reveal sensitive information merely by identifying a user as a member of a group. But it is easy to make this group construction mistake accidently if one is not vigilant.

1.4.4. Group construction mistake: self-similar stratification

When constructing groups from account data, it may be tempting to segment all accounts into categories based on common characteristics, such as volume usage, due to the method's simplicity. However, constructing groups that obey R15 by sorting the population of accounts by utility usage and then creating groups of N adjacent low-usage accounts is a mistake. For example, the smallest 15 users form the first group, then the next 15 smallest users, and so on until the largest users form the final group. We will call this the self-similar stratification method of group construction. Such an approach guarantees that most groups contain entirely accounts with usage extremely close to the group's mean usage. If one knows that self-similar stratification was used to create the group, it is possible to guess most accounts' usage within a few percentage points. It is possible to accidently create self-similar groups by performing spatial clustering of adjacent accounts in neighborhoods or zones that are coincidentally very similar to each other in terms of usage. However, spatial grouping does not create the same mistake, because one does not know a priori that the adjacent customers are similar in usage, so individual users are not guessable. For these reasons, self-similar stratification is considered an inappropriate method for specifically categorizing low-usage volume accounts.

1.4.5. Group construction mistake: equal-interval stratification

It is usually not optimal to construct groups from equal intervals in terms of spatial distance or facility age, as very few account attributes for possible use for aggregation are uniformly distributed across the intervals, so this strategy results in too few groups and too much masking, and thus suboptimal usefulness. For example, if we are aggregating groups of accounts sharing the same square-mile block of a city's area, there are far more residences in some square mile blocks than in others; some groups will have thousands of accounts, others will have dozens, and still others will be masked due to N < 15. By contrast, it results in finer detail, more groups, less masking, and more utility from the data if we allow our group boundaries to change so that all groups have similar numbers of accounts-for instance by searching for spatial boundaries of groups that yield target numbers of individuals (a kind of gerrymandering). More detail on this problem is provided in Appendix C.

1.4.6. Spatial location is not (normally) PII because account addresses are public information

Safe-harbor rules that prohibit the release of account addresses err for two reasons in the UCD context. First, because utility accounts are by definition (normally) affixed to specific facilities and addresses that are already on maps and in phone records and in assessors' and other public records, the existence of that facility or address and its connectivity to

the utility grid is already public information and cannot thus be considered protected PII by Schwartz & Solove's [2011] "non-public" definition. For example, it is very different to disclose that a home in Flagstaff, Arizona is connected to the City of Flagstaff's water services (as are nearly 100% of homes in the city) as compared with disclosing that a person of a specific name is currently living there- and perhaps is also the customer of a medical center that treats cancer. The former is easily guessable, inconsequential and, importantly, already public information, whereas the latter is (perhaps) unguessable, laden with risk, and considered private information. We need to more carefully and expertly distinguish between UCD that is PII (e.g., names) and that which is not (e.g., addresses)- or, if you prefer, between "customer" and "account" UCD. Customers must be protected, but some aspects of their accounts need not be.

2. Materials and methods

2.1. Materials

The precise tradeoff between usability and privacy can be measured in several ways. One can measure disclosure risk versus utility [Duncan and Lynne Stokes, 2004]. R15 expressed the tradeoff as a politically determined minimum-privacy threshold, specifically the minimum group size needed to achieve an acceptable level of anonymity to the individual customer. In the following discussion, we aim to more precisely measure the usability -privacy tradeoff in the specific context of utility customer data. We will measure this tradeoff as a tension between (first) minimum group size N, (second) maximum use percentage of an individual within the group F, and (third) the percentage M of accounts whose usage becomes masked in the process.

To accomplish this, a large body of real-world UCD is required. The California Center for Sustainable Communities (CCSC) at the UCLA Institute of the Environment and Sustainability has obtained monthly account level billing records for electricity, natural gas, and water consumption for a very large number of customer accounts within the Southern California region. In the case of electricity and natural gas, CCSC's data coverage includes 20+ million accounts located within the region's IOUs' service territories as well as several million additional accounts located within the region's MOUs' service territories. These records span the time period between 2006 and 2017. It is worth noting that the majority of utilities themselves do not retain longitudinal data, thus change over time cannot be evaluated unless another party, such as UCLA, takes custody of both the raw and the aggregated data. Longitudinal data enhances policy understanding of program implementation, rate changes, and more, and is increasingly critical for sound management of services. Therefore, UCD rules must enable this use case.

The addresses for each account in the CCSC backend database are pushed through a multi-stage address geocoding pipeline. For the >90% of accounts whose addresses can successfully be geocoded the parcel level, building attribute information is matched from the local county assessor's parcel database. These attributes include building size, vintage, and use. Similar procedures are used to link parcel geocoded accounts to data from the US Census Bureau on aggregated de-identified socio-demographic attributes such as household income levels, percentages of renter owners, race, and age distributions.

Such context -enhanced account-level energy and water consumption data have been used by the CCSC to develop the UCLA Energy Atlas and Water Hub. These public-facing websites present historical energy and water consumption data for groups of customers that have been aggregated to various reference geographies and zoned-use categories of interest to researchers, policymakers, local governments, and the public at large. The aggregations are frequently subject to data masking due to the requirements of the privacy rules in effect, and thus, they represent an important test for the tradeoff between usability and privacy associated with alternative aggregation rules.

2.2. Testing aggregation methods

All of the analyses developed with CCSC account-level consumption data were aggregated in accordance with California's state utility data aggregation guidelines, as specified in CPUC decision 14-05-016 [California Public Utilities Commission Decision 14-05-016, 2014]. In instances where the requirements of these aggregation guidelines prevented the disclosure of detailed information that is relevant to the present analysis, synthetic data were generated by randomly drawing those data from synthetic distributions based on those observed within actual CCSC data.

For the quantitative analysis of the tradeoffs between usability and privacy associated with different privacy rules, 400 different unique pairwise combinations of minimum group size (N) and maximum individual use percentage (F) thresholds were generated. These combinations were produced by incrementing N by units of 5 over the interval range 0–100, and F by units of 5% over the interval range 0%–100%. Each pairwise combination was then applied to the customer grouping used to develop the UCLA Energy Atlas website. These groupings consist of the combination of the 240+ individual geographies, each disaggregated into five different building zoned-use categories. For each pairwise combination of N and F, we record the percentage of total consumption that would have to be masked and the fraction of the total number of account groups that would have to be masked to comply with the N/F rule under consideration. The results of this analysis are discussed in Section 3.2.

For this paper's quantitative analysis of tradeoffs between usability and privacy associated with the creation of a dedicated whale group, a synthetic dataset of 10,000 individual accounts was generated so that the distribution of individual account consumption levels could be revealed in detail without compromising any real-world account's data. This simulation procedure worked by repeatedly randomly sampling a bi-variate log-normal distribution parameterized to reflect the known real-world correlation between the volume of an individual account's consumption and the identity of its zoned use. The results of this analysis are discussed in Section 3.3.

3. Results

3.1. Optimizing account number N, whale fraction F, and masking fraction M

A larger N improves privacy because it means that it is harder to guess usage associated with an account, and if N»10 it is very hard to guess usage accurately (unless there is a whale). Larger N_m, the average number of accounts contained in the set of groups, is always preferable to enhance privacy, but N is also the enemy of usabilty because it reduces the number of groups G. We want to maximize both N and G, but cannot do both simultaneously. By contrast, the average whale fraction F_m is not something we want to maximize or minimize, but rather we choose F_m to allow optimization of N_m and G. A is the number of accounts, and $A = N_m \times G$. In an R15 style privacy framework, $1 \geq F_m \geq 1/N_m$, so F_m is constrained by N_m. As $F_m \rightarrow 1$ (100%), $G \rightarrow A$, and $N_m \rightarrow 1$, which eventually violates the N-minimum of the rule. As $F_m \rightarrow 1/N_m$, we eventually converge to the self-stratified group where all members have the same usage, which violates privacy principles (albeit mistakenly satisfies R15, see section 1.4.4). The solution to maximizing privacy is, therefore, to maximize N and choose a moderate F. This maximizes in-group heterogeneity, minimizes the probability of a randomly correct guess at identity, and safeguards against identifying a whale. But, maximizing N also minimizes usabilty by minimizing G, which violates the purpose of balancing usability and privacy. So, what would an ideal (N–F) rule look like in terms of balancing usability and privacy?

Because F dominates the masking considerations under log-normal account distributions, the approach we take is to maximize G (and therefore minimize N) subject to an assumed maximum F and

additionally subject to a maximum allowable masked fraction of groups M. Usability is maximized by minimizing M and N, and privacy is maximized by maximizing M and N. There is no "correct" value of N, F, or M for all applications, but we are able to observe the tradeoff in real-world UCD drawn from our experience with data from Southern California and make recommendations based on this experience in the subsequent discussion.

3.2. Effects of different N and F rules on masking fraction M in Los Angeles

The effect of N and F on M can be difficult to assess analytically because no real-world utility system obeys idealized rules of account structure and geography. Masking rates, M, emerge from the complex interaction between the spatial correlation structure of account-level consumption and the geographic distribution of designated reporting boundaries.

Fig. 1 illustrates the results of applying a large number of different masking rules, formed by unique pairwise combinations of different values for N and F, to the masking of account groupings used in the UCLA Energy Atlas front end website. As masking increases and as N increases, the analytical usefulness is lost. To balance usabilty and privacy, we want rules that achieve low masking fractions in combination with large numbers of accounts, N, per group. Fig. 1 shows results for electricity, natural gas, and water utility accounts, and characterizes masking fraction with respect to both the fraction of groups masked and the fraction of total consumption masked. The results of this empirical evaluation of Los Angeles UCD-PII aggregation provide several important insights.

Three findings are apparent. First, the masking rates are much more sensitive to the rule's maximum whale fraction F than they are to the minimum number of accounts N. A second important finding is that masking rates for using a single aggregation rule can differ dramatically for electricity, natural gas, and water, owing to sector-specific distributions of customers. Masking rates for electricity and water are lower than those for natural gas-probably because natural gas usage is more heavily concentrated and spatially autocorrelated within Los Angeles. A large fraction of the natural gas in Los Angeles is used by large facilities as an industrial scale fuel for manufacturing and power plants, so users near one of these facilities become masked due to whale masking requirements. This increased natural gas usage masking limits, for example, spatially detailed greenhouse gas emissions accounting, or the understanding of variation by building size, vintage, zoning, or inhabitant in the rest of the geography.

Third, we find that a Rule of Fifty (R50, $N \geq 50$, $F \leq 50$) would significantly reduce both the percentage of total consumption masked and the percentage of geographies masked for both energy and water UCD, while simultaneously (and ironically) increasing N_m and thereby privacy, as compared with R15. R50 generates more privacy and usability by these measures, while still keeping a relatively large group number G. Relative to both geographic and consumption-based definitions of masking severity, transitioning to the proposed new Rule of 50 would result in roughly an order of magnitude less masking for the reporting of both electricity and natural gas consumption and as much as two orders of magnitude less masking for the reporting of water consumption. An ideal rule for this specific city and utility data type might be closer to 10/40 ($N \geq 10$, $F \leq 40$)- but R50 is a good general choice for all UCD because it is both conservative and simple.

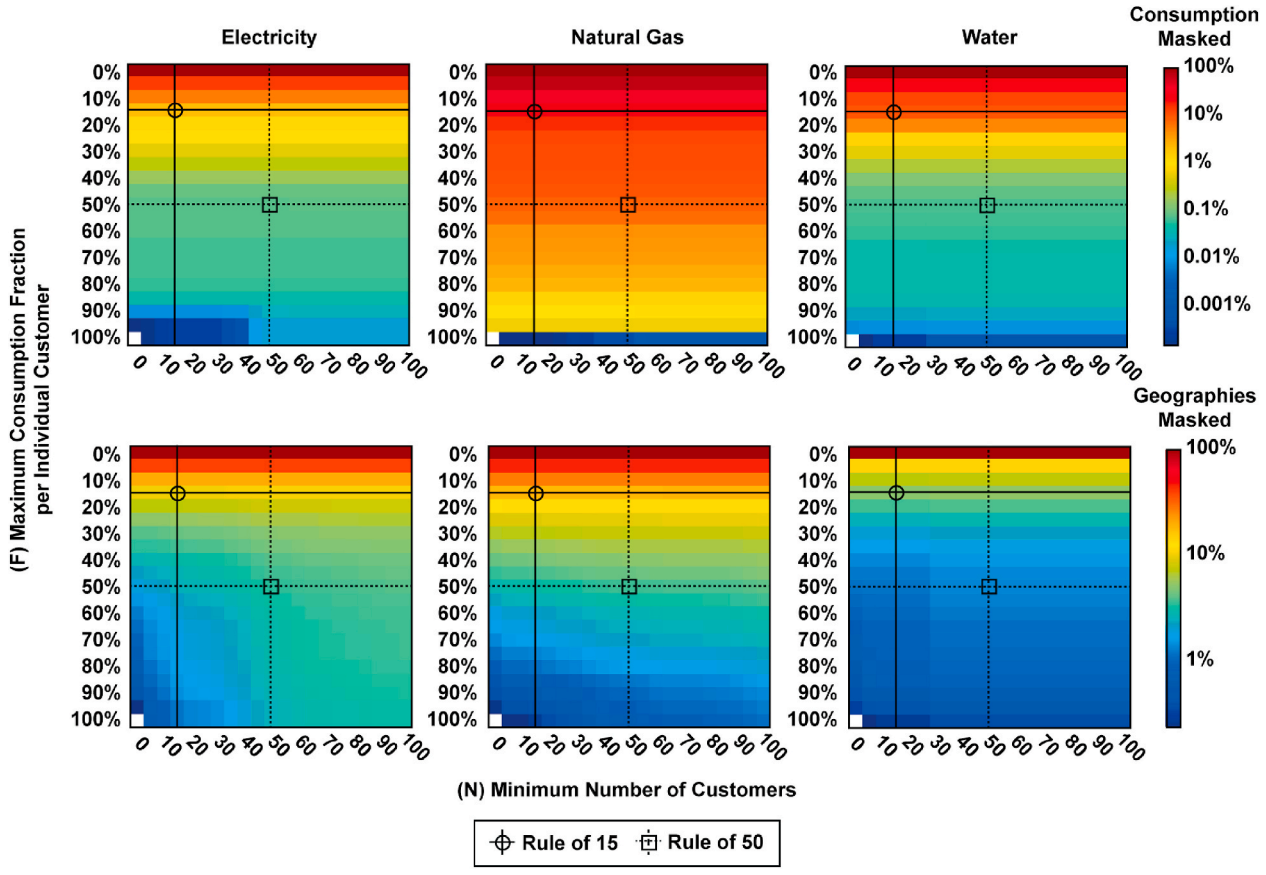

Fig. 1. Empirical results from the application of a large number of masking rules to a large sample of real-world UCD taken from the UCLA Energy Atlas and Water Hub. Each rule represents a different pairwise combination of incremental values of N, the minimum number of accounts (horizontal axis) and F, the maximum fraction of consumption associated with a single account (vertical axis) The masking rate associated with each rule is plotted using a gradient colormap: blue = low, red = high. Masking is assessed both in terms of the percentage of consumption masked (top row) and the percentage of geographies masked (bottom row).

3.3. Efficacy of a whale group to reduce masking rates

Whales are ubiquitous in UCD, so group construction methods must specifically handle them. R15 is poor at handling whales. A solution is to use a whale group (WG). We developed a greedy algorithm, UNMASK, capable of systematically adding large individual accounts to a whale group subject to the constraint that the whale group would itself not be subject to masking. This algorithm works by iteratively selecting individuals to be placed into the whale group in such a way so that the group's total consumption is minimized (see SI 1 for details).

To illustrate this, consider the hypothetical example of a city utility comprised of 10,000 total accounts for which consumption is to be reported for four different account groups: Single-Family Residential, Multi-Family Residential, Commercial, and Industrial (Fig. 2). The application of R15 to these account groups requires that all of the Industrial group's accounts be masked due to the presence of individual whales whose consumption levels are three orders of magnitude higher than the median. The algorithm iteratively removes the largest account within each masked group and places it into the whale group until the aggregation rule's requirements are satisfied, and the initially masked group becomes unmasked. In this example, under the R15 rule, this initial step is successful in unmasking the original Industrial customer group, but the new whale group which has been created would itself be subject to masking due to its having an insufficient number of accounts ($N < 15$). In order to make the newly constructed whale group viable, UNMASK applies an additional step sequentially cannibalizes the largest accounts from other groups and adds them into the whale group, until the point at which R15 is satisfied for all, or as many as possible. The net result of this two-step UNMASK algorithm is summarized in Fig. 2 and Table 1, presented as a contrast before and after the creation of the whale group.

Remember, however, that the benefits of the construction of a whale group come at two costs. First, a whale group costs knowledge of the original group membership of the whales (e.g., in this case, was the whale a single-family, multi-family, commercial, or industrial account type?). Second, a whale group identifies its members as whales, which might be politically inconvenient where that membership is not already obvious.

4. Discussion

4.1. Recommendation: an HIPAA-style dual-rule approach that allows for statistical expert determination in addition to a safe harbor should be adopted

We recommend that regulators move from the currently problematic safe -harbor implementation (e.g., R15), to allow instead both (1) a more nuanced safe -harbor implementation of a Rule of 50 (R50) to improve both usefulness and privacy, and alternatively (2) the expert determination method using a process of authorized Institutional Review Board approval. The expert determination method is superior owing to its flexibility and optimality under various use cases but requires much more expertise and effort. This dual approach is already implemented by HIPAA rules as a very clear template and precedent that routinely protects far more sensitive PII data in the health context.

4.2. Recommendation: follow four sound practices for safe group construction

This article has demonstrated four relatively safe group construction practices that should be considered in order to optimize the usability-privacy tradeoff for UCD, and which can be applied within R15, R50, expert determination, or any other rule. These practices include:

- Use a whale group to avoid masking groups containing large accounts (Section 1.4.1, 3.3)
- Avoid groups that identify a protected class (Section 1.4.3)
- Avoid self-similar stratification which compromises privacy (Section 1.4.4)
- Construct groups containing equal account frequencies, not equal intervals (Appendix C, Section 1.4.5)

4.3. Recommendation: an account's group membership and location (address) should not be considered PII

One of the principal beneficial use cases of historical account-level UCD is for the production of maps that reveal geographic differences in the volume of resource consumption by city, neighborhood, census tract, and so on. A fundamental prerequisite to the usefulness of this analysis is, however, the disclosure of the group's geolocation information. We argue that the group's spatial location, including spatial boundaries of the group or a metadata list of addresses belonging to the group, must be released in order to strike a balance between usability

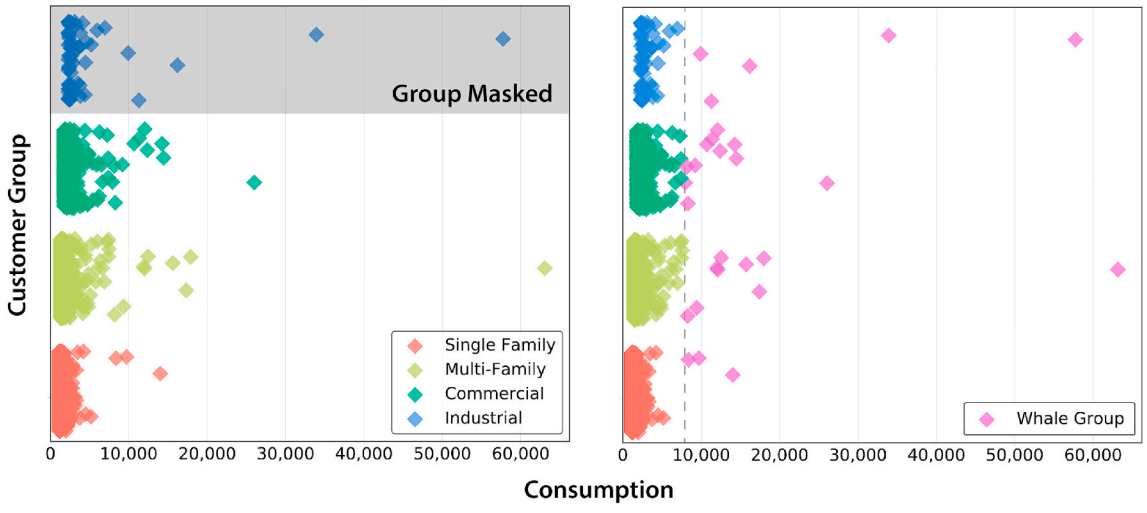

Fig. 2. Illustration of a synthetic experiment on four log-normally distributed account groups, before and after the construction of a fifth whale group using the UNMASK algorithm (x -axis in arbitrary units of consumption).

Table 1
Summary statistics for a synthetic experiment on four log-normally distributed account groups, before and after the construction of a fifth whale group using the UNMASK algorithm (arbitrary units of consumption).

Customer Group	Single-Family		Multi-Family		Commercial		Industrial		Whale	
Before-After	Before	After	Before	After	Before	After	Before	After	Before	After
Group Masked	No	No	No	No	No	No	Yes	No	N/A	No
Number of Customers N	6,882 (68.82%)	6,879 (68.79%)	2,461 (24.61%)	2,452 (24.52%)	591 (5.91%)	580 (5.80%)	65 (0.65%)	60 (0.60%)	N/A	28 (0.28%)
Total Consumption	294,035 (22.43%)	265,386 (20.24%)	529,186 (40.37%)	373,648 (28.51%)	328,599 (25.07%)	213,532 (16.29%)	157,715 (12.03%)	40,035 (3.05%)	N/A	416,952 (31.81%)

and privacy. Moreover, we argue that group location information and address lists should not be considered PII in this UCD context because, with exceptions, an address's existence is already public information. By legal definition and by common sense, PII is not protected when it is already routinely and legitimately public knowledge. Therefore, location metadata is allowable to release under existing safe -harbor rules, especially when location is aggregated to the level of a group that otherwise complies with safe harbor rules. See Appendix B for more discussion.

4.4. Recommendation: disclosure rules and informed consent should differentiate between data user types

Both usability and privacy are relative to the use case and the identity of the user of the data. Not all third-party data users (neither utilities nor customers) are the same in terms of their motivations for requesting data access and their commitment or competence and trustworthiness for its safe handling and publicly beneficial use. We recommend that rules and processes for UCD-PII disclosure and informed-consent-seeking explicitly differentiate between three use cases and groups of data users: (1) government agencies and regulators with jurisdictional overlap with a utility's service area (inclusive of those agencies' nondisclosure-bound contractors); (2) organizations (such as publicly funded universities) engaged in publicly beneficial and publicly published research, research that is controlled by a regulator-authorized institutional review board (IRB) that reviews and approves the detailed methods for their competence, compliance, and ethics; and (3) all other parties, such as unrelated governments, not-for-profit organizations, for-profits, or the general public. Additionally, utilities should be required to disclose the customer's own data back to the account owner of record and also the property owner (if they are different). The regulator would be responsible for the process categorizing data users into one of these three categories.

The first two user groups pose minimal risks to the customer if they have clearly established and approved protocols for data protection, as most (but not all) already do. User types (1) and (2) above would promptly receive from the utility the data they need after completing a successful review of their safe-harbor or expertly determined privacy protection methods, e.g., by an IRB. The first two groups should be guaranteed access to a formal process through which their UCD access requests may be submitted and evaluated. This process should specify which entities are eligible to request data, which data are eligible to be requested, the time period over which requests are guaranteed to be processed, and the mechanisms for transferring data if a request is approved. A useful model for this process is the precedent which has been set with the State of California's Energy Data Request Program, which was created as an outcome of the CPUC Decision 14-05-016.

Utilities' standard-practice informed consent procedures should be expanded to include by default these first two groups, out of an abundance of transparency toward their customers. However, the lack of this informed consent should not prohibit disclosure to these first two groups of users absent that informed consent. This is a particularly important consideration for historical UCD for which informed consent can never be collected retroactively.

The release of granular data to the third group creates added risks. This diverse third user group should, therefore, only receive UCD-PII granular data only after informed consent to that specific release and use is explicitly given by the customer.

Corollary to this recommendation is that (a) all customers should be asked to grant informed consent for the release of their granular data to all kinds of users (e.g. ICPSR, 2018) despite the inconvenience of this (McDonald and Cranor, 2008), and that (b) utilities should be required to archive and preserve all their granular account data- or pay for it to be archived and preserved by a trusted third party bound by the same rules as the utility so that the UCD can be disclosed when it is needed by an authorized party in the future.

4.5. Recommendations for high-frequency UCD

In addition to space, time (both resolution and lag) is a key dimension in balancing usability and privacy. The gravity of the privacy risks posed to an individual by the potential disclosure of their utility consumption data are inversely proportional to both (a) the temporal frequency at which the data has been collected and (b) the temporal delay between the period of data collection and the date of its disclosure. High temporal resolution consumption data (i.e., <15-min interval data) provides more granular insight into the individual customer's activity and behavioral patterns associated with a given account. Moreover, consumption data that is available with little or no-delay (i.e., real-time data) provides a greater degree of certainty of the current whereabouts or activities of the individual customer in question, thus making the information far more actionable and consequential (Shi et al., 2011). Consideration should be given to these two factors when an expert -determination method is used. In general, because UCD varies dramatically with seasonal cycles, interval UCD loses most of its sensitivity after a season (e.g., roughly three months) have passed since it was collected. Recommendation 4.3 supports recommendation 4.5. It would be concerning indeed for account-level granular usage data to be available in tandem with high-frequency UCD. If only one of these, or neither of these, is available, the surveillance risk to the customer holding the account decreases dramatically.

5. Conclusions

Privacy rules limit the extent to which public benefits can be generated from private UCD. Consequently, the rules concerning UCD privacy must be ethically justified in terms of their public benefits and private risks (Zipper et al., 2019; Flaherty, 1990). For monopolistic

utilities and infrastructure operators that utilize natural resources and provide basic services such as electricity, natural gas, and water, there is a compelling public interest in knowing how these systems and resources are utilized. Ethical considerations call for balancing the analytical usefulness of this data against the risks created by the release of this data. In the present era of resource scarcity and big data, the public also has a heightened interest in the efficient and transparent management of its utility systems. Better practices and rules are needed for the disclosure of UCD.

The fact that U.S. states currently allow their utilities to disclose UCD to their consultants and contractors without strict HIPAA-style or IRB safeguards is telling about the relatively small risks and significant public benefits of UCD disclosure. A number of U.S. States already allow access by university researchers and all government agencies.

California's Public Utilities Commission (CPUC) approved rules to provide energy usage and usage-related data to various stakeholders, including local governments, researchers, as well as state and federal agencies, in accordance with CPUC procedure. These procedures order utilities to make customers' electric and natural gas usage available quarterly. Identities are anonymized by only allowing a customers' zip code, customer class, and the number of customers in each zip code available (Energy Data Access Decision, (D.)14-05-016). Another important outcome from this rulemaking was the establishment of an Energy Data Access Committee (EDAC), which briefly served to advise utilities on practices for data access until it was discontinued. State such as Connecticut and Oregon have variations of customer data sharing. Connecticut requires the written consent of the customer before data can be released (Connecticut General Assembly, Gen. Stats. §16-245°(d) and § 16-244h-4). Electric distribution, electric, and gas companies are also required to make consumption data of all non-residential building available for benchmarking purposes while preserving confidentiality (State of Connecticut, Public Act No. 11–80). Oregon's utilities, including Portland General Electric and Pacific Power, utilize platforms such as Green Button to share residential and small business customer data. The only third-party organization that regularly received energy usage data is the Energy Trust of Oregon, as per the Electric Company Transfer of Data Rule (Oregon PUC, Order No. 12323).

We recommend an expanded role for HIPAA-style expert determination methods of data aggregation and disclosure practices and for revised usage of safe harbor methods. Safeguards for UCD disclosure include those enumerated in Section 4.4. Central to these safeguards is the creation of an authorized and specialized IRB for UCD-PII in each State, with members from utilities, utility regulators, academic institutions, and ratepayer advocacy groups, serving as an arbiter of public trust. This recommendation is similar to the call by Zipper et al., for open and ethical data management (2019). These IRBs would vet those requesting UCD, evaluate UCD requests in terms of purpose, competence, and ethics, put in place user-specific data use restrictions, and ensure that UCD are used in the public interest. IRBs were established as part of the 1974 National Research Act and have been instrumental in ensuring the ethical treatment of human subjects in medical, nutritional, and social science research (Cseko and Tremaine, 2013).

The widely adopted safe harbor R15 rule is inadequately nuanced to ensure competent de-identification. Our statistical results demonstrate that it is possible to simultaneously improve the minimum guaranteed level of customer privacy and the usefulness of statistical analyses conducted using UCD if R15 was amended to R50 (where R50 is $N \leq 50$ and $F \leq 50$). By raising both F and N it is possible to competently de-identify customers and reduce masking, simultaneously addressing analytical usefulness and privacy for common and important real-world UCD applications in the water, power, and natural gas sectors. R50 provides a superior safe harbor rule to R15 and should be adopted for (at least) the water, electricity, and natural gas utility data applications.

The current status quo for UCD disclosure and privacy is characterized by ambiguity in the applicable methods and rules, which creates significant potential downside risk by all concerned and guarantees that the analytical usefulness of these critical data is not fully realized. Under this status quo, most utilities to err on the side of privacy at the expense of usability when evaluating requests for access to UCD-PII. Gaining access to granular UCD can thus, at present, be a risky, lengthy, and arduous process that may be cost-prohibitive for some potential users. This study recommends practices for how regulators can strike a workable and acceptable balance between usefulness and privacy in rules for utility customer data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding is provided by Northern Arizona University and by the National Science Foundation ACI-1639529, INFEWS/T1: Mesoscale Data Fusion to Map and Model the U.S. Food, Energy, and Water (FEW) System. The research and opinion are those of the authors and not necessarily that of the funding agencies.

Appendix A. Historical origins of the Rule of Fifteen for UCD-PII de-identification

The principal barrier to researcher access to UCD is the fact that account-level UCD is legally considered to be protected as sensitive PII, much like a person's medical records, student records, or private transactions. Moreover, in practice, UCD is often (and arguably incorrectly) treated by the utilities themselves as if it were their proprietary information rather than the account holder's proprietary information. The legal designation of account-level UCD as PII means that it is subject to restrictions in terms of (first) both the types of entities which are legally permitted to access it, as well as (second) the de-identification measures that must be applied in order to make any derived data products publicly available [California Public Utilities Commission Decision 14-05-016, 2014]. In the latter case, these public sharing restrictions are normally expressed in the form of data aggregation rules.

The typical data aggregation rule states that before release PII data must be aggregated into sets containing a minimum number of accounts "N." Sometimes the rule also specifies that the consumption of any individual account within that group must not exceed some maximum fraction "F" of the group's overall total consumption. If these rules are not satisfied for any particular group, then the consumption data for that group cannot be publicly disclosed and the data for that group is therefore said to be "masked." By definition, masked data cannot provide any usefulness, and coarsely aggregated groups provide less usefulness (but more privacy). The most common safe harbor aggregation rule used in U.S. States at present is "the Rule of Fifteen" or "the 15/15 Rule" (R15), where ($N \geq 15 \bigcup F \leq 15$).

R15 was first implemented in California, and it is useful to understand the history of how and why the rule came to be. California has been at the forefront of efforts to increase public access to granular utility data for several decades. In California, account-level UCD (e.g., electricity, natural gas, and water services) have been legally designated as PII since 1997, when State Senator Byron Sher introduced UCD privacy regulations into the state legislature [California Senate Bill 448, 1997]. This introduction was a direct consequence of the use of address -level water billing data to publicly

shame large users in his district during a drought. Water retailers in California, both public, private, and not-forprofit (such as mutual companies), backed this prohibition, as did the State's various energy utility providers. In California, as in many states, many utilities are regulated by a Public Utilities Commission (PUC). In 1997 the California Public Utilities Commission (CPUC) ruled to set a standard. The relevant language of the original rule reads:

> "Customer information shall be confidential unless the customer consents in writing. This shall encompass confidentiality of customer specific billing, credit or usage information. This requirement shall not extend to disclosure of generic information regarding the usage, load shape, or other general characteristics of a group or rate classification, unless the release of that information would reveal customer specific information because of the size of the group, rate classification, or nature of the information." [California Public Utilities Commission Code Section 394.4(a), 1997]

As the State of California continued to develop increasingly aggressive efficiency and conservation standards and investments, local governments – which are often responsible for the implementation of the standards and assessment of their performance– found themselves prevented by this rule from accessing UCD for their own jurisdictions. This situation resulted in numerous complaints to the CPUC about inappropriately limited usage data. In response to these complaints, in 2014, Administrative Law Judge Sullivan of the CPUC, initiated a year-long investigation into data access and ruled in favor of greater data access. This new rule established minimum-group-size aggregation thresholds that differ by the user and use case of the data. This ruling attempted to balance usability and privacy by providing greater access to lower-risk use cases and to use cases that are more clearly for the public benefit. For example, academic institutions operating under a non-disclosure agreement (NDA) with the utilities, could access account-level UCD-PII, subject to a utility decision about the appropriateness of the request. For all public-facing disclosures, however, the ruling dictated that data must be aggregated according to R15.

> *"In aggregating customer data to create an aggregated data report, a Utility must take steps to ensure the report is sufficiently anonymous in its aggregated form so that* any *individual customer data or reasonable approximation thereof cannot be determined from the aggregated amount. At a minimum, a particular aggregation must contain: 1) at least fifteen customers or premises, and 2) within any customer class, no single customer's data on or premise associated with a single Customer's data may comprise 15 percent or more of the total customer data aggregated per Customer class to generate the aggregated data report (the "15/15 Rule"). A Utility shall not be required to disclose aggregated data if such disclosure would compromise the individual Customer's privacy or the security of the Utility's system"* [California Public Utilities Commission Decision 14-05-016, 2014]

In the intervening years since the delivery of the CPUC's original 1997 ruling, other states including, Illinois and Colorado have followed suit by proposing or adopting similar rules [Illinois Commerce Commission Decision 13–0506, 2014, Colorado Public Utilities Commission Decision R15-0406, 2015].

Appendix B. Address & geolocation metadata of accounts should not be considered PII

It is demonstrable that the release of publicly available metadata information (such as address) regarding accounts should not be considered PII. It is also demonstrable that group member identities, including especially their addresses or geolocations, must be released to provide useful analysis using most UCD. Publishing an alphabetized list of account addresses included in an aggregated group is not the same as linking an individual customer's name and usage data to their address. The former is not the information that is intended by R15 or other HIPAA-style safe -harbor rules (i.e., Schwartz and Solove's " specific-types" definition); it is rather already-public information and not PII using Schwartz and Solove (2011) " non-public" definition of PII. Those operating under R15 or other safe -harbor rules should feel confident in releasing alphabetized metadata lists describing membership in aggregated groups, because this interpretation already satisfies the spirit of the rulemaking (unless the letter of the rules explicitly forbids this release).

However, the letter of R15 in some States may currently be problematic, depending on who is making the judgment. Based on the plain wording of 394.4(a) and R14-2-2215 and similar rules there may be confusion about whether it is legally acceptable to release metadata associated with those groups. Although the proposed implementation of R15 in Arizona (for instance) specifically lists the customer's address and name in its safe harbor list, this is logically inconsistent with the reality of what is and is not private information (name might be PII, address is not PII). The plain wording of the rule is illogical because it is inconsistent with Schwartz & Solove's definition; that locational metadata is already public information, and therefore cannot be considered private by definition. The plain wording of the rule is also unethical because it destroys most of the usability of the aggregated group data and without providing any privacy benefit or risk reduction to the customer.

Metadata defining the group's location in space and time is essential in order to make use of the aggregated UCD, because this utility often comes from mapping, address or census tract linking, or from space-time trend analysis. Timestamps such as month or year for which the group is valid are essential. Either a list of addresses belonging, or a GIS boundary, is also essential. Armed with this information it is possible to link the group's water or energy use with the group's acreage, indoor square footage, grass lawn area, number of pools, type of heaters and air conditioning, tax records, and other data to perform research on the causes and effects of utility usage. In principle it is similarly acceptable to release any metadata that is already public information, such as a list of phone-book customer names included in the group, but this other metadata may not be strictly necessary.

To prove the point that metadata lists of addresses belonging to a group do not constitute PII for the purposes of aggregated UCD, consider the corresponding author's property and associated utility accounts in Flagstaff, Arizona (Figure A1, red boundary). By linking the group location and

boundary with publicly available data, all of this is readily visible. The Coconino County Assessor's Office makes it legitimately disclosed public information that Ben Ruddell owns 4399 E Coburn Dr., along with what he paid for the property, who he bought the property from, what he pays in taxes, the precise legal definition of the property including square footage, type of heating, and many other details. It is public information that this is a single -family residence by zone and construction and not an apartment building, factory, or storefront. It is possible to use publicly available street -level imagery and satellite photography to map the number of trees on the property, the color of paint, the amount of grass being grown, the absence of a pool (in this case), and the impervious area of the rooftop and driveway-without visiting in person or conducting surveillance. It is guessable that this property is not a whale customer based on its moderate size and absence of lawn or pool. Because this property is in the incorporated limits of the City of Flagstaff, it is guessable that this property is connected to the city's water utility service, and it is equally guessable that this property is served by Arizona Public Service (electricity) and Southwest Gas (natural gas). It is guessable that either Ben Ruddell or someone who rents from or cohabitates with Ben Ruddell is paying these utility bills. All of this is easily accessible public record, and is therefore excluded from Schwartz and Solove (2011) " non-public" definition of protected PII, because further identification of these factors, and further risk of identification of Ben Ruddell's location, is not possible through the release of Ben Ruddell's personal UCD.

By contrast, what is not already public information is who is living at this property at the moment, or Ben Ruddell's social security number on his account, the exact high-frequency amount and timing of water, gas, and electricity consumed by Ben Ruddell (et al.) and the property at 4399 E Coburn Dr., the amount paid, who is paying, or the means of payment. These later data are the UCD-PII with which we must concern our rulemaking.

Consider the application of R15 to construct a compliant aggregated group of residential water accounts (not customers!) containing 4399 E. Coburn Dr. A utility might group together fifteen adjacent residences on E. Coburn Dr. The utility would release the total amount of water or energy used by this group of 15 residences, the number N = 15 of members in the group, the name of the group (perhaps it is named "E Coburn Dr Residential Group 6"), the year of the grouping, and the geospatial boundary describing the spatial extent and location of the group (blue outline). Based on the information presented in Figure A1, it is easy to use the GIS boundary or name of the group to identify the list of parcel addresses associated with the group, along with the owners of those parcels (who are responsible for the utility bills).

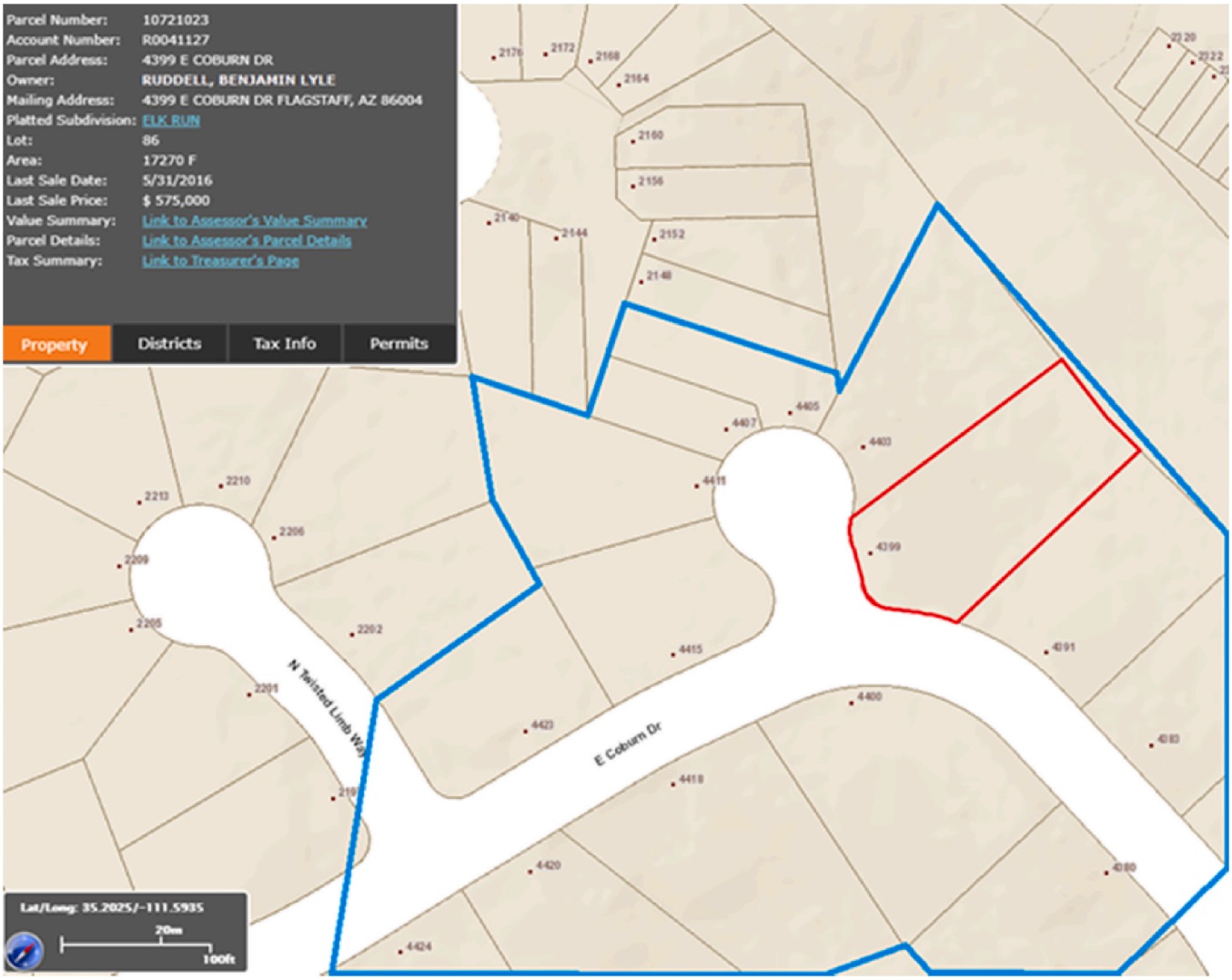

Fig. A1. The GIS boundary (blue line) of a group of fifteen residential water accounts on Coburn Drive in Flagstaff, Arizona including the corresponding author's property and account at 4399 E Coburn Dr. (red boundary). Overlay credit: Coconino County Assessor's office parcel viewer system, accessed April 24th, 2018. The name, address, and geolocation of this group's members is already public information, and is therefore by definition not protected PII, despite being PII.

Appendix C. Assign groups with equal frequencies, not equal intervals

Assuming that the expert -determination approach is sanctioned as a method for de-identification of utility customer consumption data, it is likely, and indeed desirable, that individual account records will come to be joined to other relevant attributes. As our experience has shown, these attributes provide essential context for the interpretation of raw consumption figures both for researchers and policymakers alike. Assembling a group of accounts on the basis of a common geographic area tends to be fairly straightforward in terms of the process of construction. The set of geographic boundaries being used has typically pre-determined (that is, a city's administrative territory or zipcode) and all that needs to be done is to perform a spatial intersection query to determine whether or not each account is contained within the relevant geographic area.

However, assembling a group of accounts on the basis of non-spatial attributes is a fundamentally different proposition. When doing so the individual responsible for defining the groups is faced with many more options in terms of how the group can potentially be defined. How these choices are made can have significant impacts on the levels of masking which must be applied to the de-identified data products.

To illustrate this idea, consider a hypothetical example based on Fig. 1 in the paper's body but in which the geographic location of accounts have been spatially joined to attributes for the buildings with which they are associated. Plotted in both figure panels is the frequency distribution of accounts sorted on the basis of the construction vintage of their associated buildings. In the upper plot within the figure, the vertical lines divide these accounts into six different groups using a regular 20 year time interval. This equal interval approach to this problem results in a situation where, under R15, de-identified consumption data for three of the six groups would have to be masked due to there being an insufficient number of accounts (shaded in gray, Figure B1).

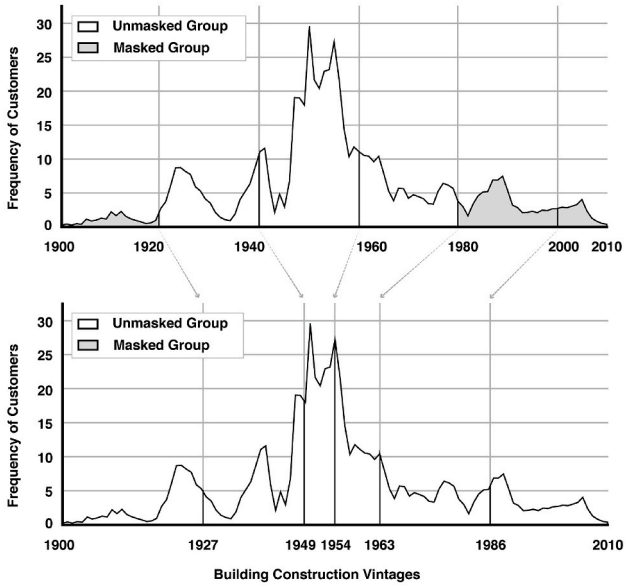

Fig. B1. Graphical illustration of how the definition of account groups can be modified to reduce masking. The plot at the top of the figure shows accounts aggregated into six groups on the basis of a set of simple equal interval breaks applied to their associated buildings' construction vintage years. The plot at the bottom shows the impacts of re-aggregating the same set of accounts into six groups on the basis of irregularly spaced interval breaks, positioned in response to the count frequency distribution of their associated buildings' construction vintage years. The frequency distribution approach eliminates masking of the tails of the distribution.

The underlying cause for the elevated levels of masking in this example is the non-uniform frequency distribution of buildings when sorted by their construction vintage. Such irregularities are quite common among attributes that relate to land use planning patterns or cycles of economic growth. In this example, one way to reduce the effective masking rates, without violating R15 or compromising customer privacy, is to use an irregular time interval as the basis for defining the groups. Doing so ensures that roughly equivalent numbers of accounts are captured within each, as shown in the lower plot within Figure B1.

This frequency distribution process is similar to the process of constructing a whale group, and for a similar reason: reduction in masking (Figure B2). The principal difference is that in the former the number of account groups was predetermined and it was only the boundaries between the groups that were being manipulated, whereas in the later, both the number of account groups and the boundaries between them which are being deliberately specified.

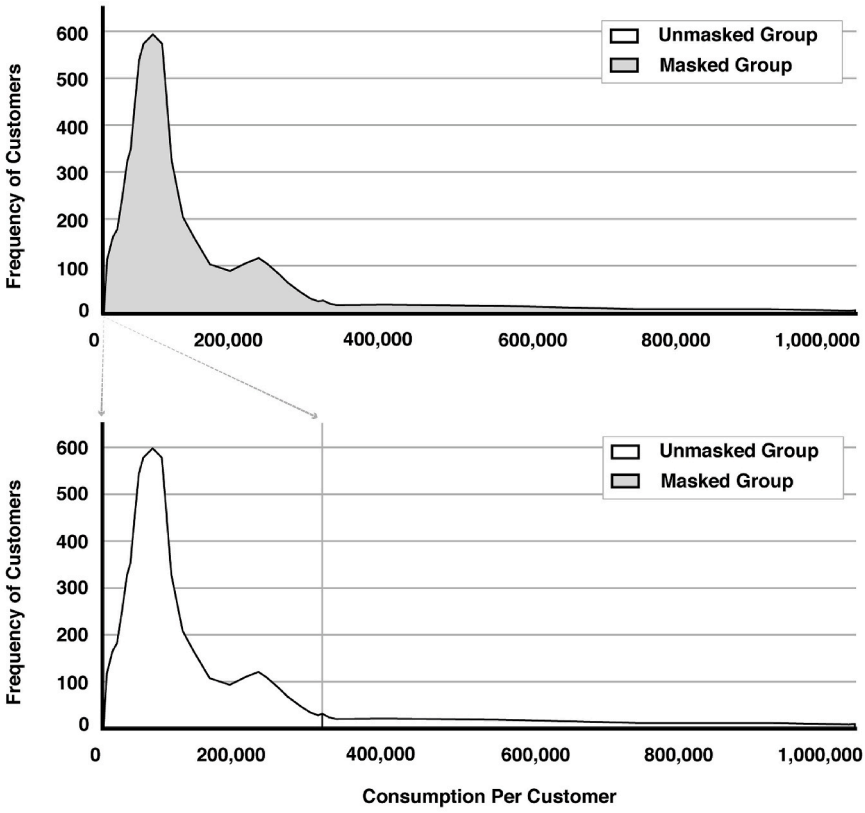

Fig. B2. Graphical illustration of how a new group of accounts can be defined to reduce masking. The plot at the top of the figure shows accounts aggregated into a single group. The plot at the bottom shows the impacts of re-aggregating the same set of accounts into two groups, with the new group being defined on the basis of the count frequency distribution of accounts sorted by their associated consumption volume.

When applied in space, the frequency distribution method means that several smaller neighborhoods with fewer accounts would be grouped together into one group, whereas a dense neighborhood with many accounts would be split into multiple groups.

References

Acquisti, Alessandro, 2010. The Economics of Personal Data and the Economics of Privacy.

Aggarwal, Charu C., Philip, S Yu, 2008. A general survey of privacy-preserving data mining models and algorithms. In: Privacy-Preserving Data Mining. Springer, pp. 11–52.

Amatya, Anup, Bhaumik, Dulal, Gibbons, Robert D., 2013. Sample size determination for clustered count data. Stat. Med. 32 (24), 4162–4179.

Applebaum, Benny, Ringberg, Haakon, Michael, J., 2010. Freedman, matthew caesar, and jennifer rexford. "Collaborative, privacy-preserving data aggregation at scale. https://doi.org/10.1007/978-3-642-14527-8_4, 56–74.

Arizona Corporation Commission, Docket No. RU-00000A-14-0014. http://images.edocket.azcc.gov/docketpdf/0000156945.pdf.

Becker, Sören, Naumann, Matthias, August 2017. Energy democracy: mapping the debate on energy alternatives. Geography Compass 11 (8), e12321. https://doi.org/10.1111/gec3.12321.

Brickell, Justin, Shmatikov, Vitaly, 2008. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vols. 70–78.

California Public Utilities Commission (CPUC), 2014. Decision [D. 14-05-016] Adopting Rules to Provide Access to Energy Usage and Usage-Related Data while Protecting Privacy of Personal Data.

Cavoukian, Ann, Polonetsky, Jules, Wolf, Christopher, 2010. Smart privacy for the smart grid: embedding privacy into the design of electricity conservation. Identity in the Information Society 3 (2), 275–294.

Chakraborty, Supriyo, Charbiwala, Zainul, Choi, Haksoo, Raghavan, Rangan, Mani, B., 2012. Srivasta. "Balancing behavioral privacy and information utility in sensory data flows. Pervasive Mob. Comput. 8, 331–345. https://doi.org/10.1016/j.pmcj.2012.03.002.

Cseko, G.C., Tremaine, W.J., 2013. The role of the institutional review board in the oversight of the ethical aspects of human studies research. Nutr. Clin. Pract. 28 (2), 177–181.

Dulal, K Bhaumik, Kapur, Kush, Bhaumik, Runa, Domenic, J Reda, 2013. Sample size determination and hypothesis testing for the mean of a lognormal distribution. Journal of Environmental Statistics 5 (1), 1–21. http://www.jenvstat.org/v05/i01/paper.

Duncan, George T., Lynne Stokes, S., 2004. Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding. Chance 17 (3), 16–20. https://doi.org/10.1080/09332480.2004.10554908.

Efthymiou, Costas, Kalogridis, Georgios, 2010. "Smart grid privacy via anonymization of smart metering data. In: 2010 First IEEE International Conference on, vols. 238–43. Smart Grid Communications (SmartGridComm).

Eigner, Fabienne, Kate, Aniket, Maffei, Matteo, Pampaloni, Francesca, Pryvalov, Ivan, 2014. Differentially private data aggregation with optimal utility. In: Proceedings of the 30th Annual Computer Security Applications Conference, vols. 316–25.

Emam, K El, Brown, A., Journal of the American, and Undefined, 2009. Evaluating Predictors of Geographic Area Population Size Cut-Offs to Manage Re-identification Risk. Academic.Oup.Com. https://academic.oup.com/jamia/article-abstract/16/2/256/960457. (Accessed 22 January 2019).

Emam, Khaled El, Brown, Ann, AbdelMalik, Philip, Neisa, Angelica, Walker, Mark, Bottomley, Jim, Roffey, Tyson, 2010. A method for managing Re-identification risk from small geographic areas in Canada. BMC Med. Inf. Decis. Making 10 (1), 18.

Emam, Khaled El, Jonker, Elizabeth, Luk, Arbuckle, Bradley, Malin, December 2, 2011. "A systematic review of Re-identification attacks on health data." edited by roberta W. Scherer. PloS One 6 (12), e28071. https://doi.org/10.1371/journal.pone.0028071.

Erkin, Zekeriya, Troncoso-Pastoriza, Juan Ramón, Lagendijk, Reginald L., Pérez-González, Fernando, 2013. Privacy-preserving data aggregation in smart metering systems: an overview. IEEE Signal Process. Mag. 30 (2), 75–86.

Felten (Ed.), 2012. Is Aggregate Data Always Private?. Federal Trade Commission. https://www.ftc.gov/news-events/blogs/techftc/2012/05/aggregate-data-always-private. (Accessed 20 April 2018).

Flaherty, David H., 1990. On the utility of constitutional rights to privacy and data protection. Case West. Reserv. Law Rev. 41, 831.

Fournier, Eric D., Federico, Felicia, Porse, Erik, Pincetl, Stephanie, 2019. Effects of building size growth on residential energy efficiency and conservation in California. Appl. Energy 240 (no. June 2018), 446–452. https://doi.org/10.1016/j.apenergy.2019.02.072.

Ghosh, Arpita, Roughgarden, Tim, Sundararajan, Mukund, January 2012. Universally utility-maximizing privacy mechanisms. SIAM J. Comput. 41 (6), 1673–1693. https://doi.org/10.1137/09076828X.

Groat, Michael M., Hey, Wenbo, Forrest, Stephanie, 2011. "KIPDA: K-indistinguishable privacy-preserving data aggregation in wireless sensor networks." in INFOCOM. In: Proceedings IEEE, 2024–32, p. 2011.
Gurney, K.R., Romero-Lankao, P., Seto, K.C., Hutyra, L.R., Duren, R., Kennedy, C., Grimm, N.B., Ehleringer, J.R., Marcotullio, P., Hughes, S., Pincetl, S., Chester, M.V., Runfola, D.M., Fedderma, J.J., Sperling, J., 2015. Track urban emissions on a human scale. Nature 525 (7568), 179–181. https://doi.org/10.1038/525179a.
Hayashi, Koichiro, 2013. Social issues of big data and cloud: privacy, confidentiality, and public utility. In: Proceedings - 2013 International Conference on Availability, Reliability and Security, vol. 2013. ARES. https://doi.org/10.1109/ARES.2013.66, 506–11.
HHS, 2003. HIPAA Privacy Rule for Research. April 24th, 2018. https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/special/research/research.pdf?language=es.
HHS, 2012. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability Act (HIPAA) privacy rule. November 26 2012. https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf. (Accessed 1 May 2018).
Hoffman, Steven M., High-Pippert, Angela, October 26, 2005. Community energy: a social architecture for an alternative energy future. Bull. Sci. Technol. Soc. 25 (5), 387–401. https://doi.org/10.1177/0270467605278880.
ICPSR, 2018. Recommended Informed Consent Language for Data Sharing. Institute for Social Research at the University of Michigan. https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/conf-language.html. (Accessed 1 May 2018).
Illinois Commerce Commission, Decision 13-0506, https://www.icc.illinois.gov/downloads/public/edocket/367604.pdf.
Jensen, Meiko, 2013. Challenges of privacy protection in big data analytics. In: 2013 IEEE International Congress on Big Data, 235–38. IEEE. https://doi.org/10.1109/BigData.Congress.2013.39.
Kwac, Jungsuk, Tan, Chin-Woo, Sintov, Nicole, Flora, June, Rajagopal, Ram, 2013. Utility customer segmentation based on smart meter data: empirical study. IEEE SmartGridComm 2013. https://doi.org/10.1109/SmartGridComm.2013.6688044.
Lane, Julia, 2005. Optimizing the use of micro-data: an overview of the issues. In: SSRN Electronic Journal, August 1. https://doi.org/10.2139/ssrn.807624.
Li, Tiancheng, Li, Ninghui, 2009. On the tradeoff between privacy and utility in data publishing. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 517–26.
Liang, Jing, Qiu, Yueming, James, Timothy, Ruddell, Benjamin L., Dalrymple, Michael, Earl, Stevan, Castelazo, Alex, 2017. Do energy retrofits work? Evidence from commercial and residential buildings in Phoenix. In: Journal of Environmental Economics and Management. https://doi.org/10.1016/j.jeem.2017.09.001. ISSN 0095–0696.
Livingston, V., Olga, Pulsipher, C., Trenton, Anderson, David, Vlachokostas, Alex, Wang, Na, 2018. An analysis of utility meter data aggregation and tenant privacy to support energy use disclosure in commercial buildings. Energy 159. https://doi.org/10.1016/j.energy.2018.06.133.
Loukides, Grigorios, Shao, Jianhua, 2008. Data utility and privacy protection trade-off in k-anonymisation. In: Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, vols. 36–45.
McCallister, Erika, 2010. Guide to Protecting the Confidentiality of Personally Identifiable Information. Diane Publishing.
McDonald, Aleecia M., Cranor, Lorrie Faith, 2008. The cost of reading privacy policies. ISJLP 4, 543.
Mini, C., Hogue, T.S., Pincetl, S., 2014. Estimation of residential outdoor water use in Los Angeles, California. Landsc. Urban Plann. 127, 124–135.
Pincetl, S., Graham, R., Murphy, S., Sivaraman, D., 2015. Analysis of high-resolution utility data for understanding energy use in urban systems: the case of Los Angeles, California. J. Ind. Ecol. https://doi.org/10.1111/jiec.12299.
Porse, E., Derenski, J., Gustafson, H., Elizabeth, Z., Pincetl, S., 2016. Structural, geographic and social factors in urban building energy use: analysis of aggregated account-level consumption data in a megacity. Energy Pol. 96, 179–192.
Public Utilities Commission of the State of Colorado, Decision No. R15-0406, https://www.sos.state.co.us/CCR/Upload/AGORequest/BasisandPurposeAttachment2014-00436.pdf.
Rajagopalan, S Raj, Sankar, Lalitha, Mohajer, Soheil, Vincent Poor, H., 2011. Smart meter privacy: a utility-privacy framework. ArXiv Preprint ArXiv:1108.2234.
Ramanayake, Asoka, Zayatz, Laura, 2010. Balancing disclosure risk with data quality. Statistics 4.
Rastogi, Vibhor, Dan, Suciu, Hong, Sungho, 2007. The boundary between privacy and utility in data publishing. In: Proceedings of the 33rd International Conference on Very Large Data Bases, 531–42.
Rissman, A.R., Owley, J., L'Roe, A.W., Morris, A.W., Wardropper, C.B., 2017. Public access to spatial data on private-land conservation. Ecol. Soc. 22 (2), 1–13. https://doi.org/10.5751/ES-09330-220224.
Rushforth, Richard R., Ruddell, Benjamin L., 2015. The hydro-economic interdependency of cities: Virtual water connections of the Phoenix, Arizona Metropolitan Area. Sustainability 7 (7), 8522–8547.
Sankar, L., Rajagopalan, S.R., 2013. Utility-privacy tradeoffs in databases: an information-theoretic approach. In: IEEE Transactions on. https://ieeexplore.ieee.org/iel7/10206/4358835/06482222.pdf.
Schoor, Tineke Van Der, Scholtens, Bert, 2015. Power to the people: local community initiatives and the transition to sustainable energy. Renew. Sustain. Energy Rev. 43, 666–675. https://doi.org/10.1016/j.rser.2014.10.089.
Schwartz, Paul M., Solove, Daniel J., 2011. The PII Problem: Privacy and a New Concept of Personally Identifiable Information, vol. 86. New York University Law Review. https://doi.org/10.1525/sp.2007.54.1.23, 1814–94.
Shi, Elaine, Chan, H.T.H., Rieffel, Eleanor, Chow, Richard, Song, Dawn, 2011. Privacy-preserving aggregation of time-series data. In: Annual Network & Distributed System Security Symposium. NDSS).
Sirkiä, Jukka, Laakso, Tuija, Ahopelto, Suvi, Ylijoki, Ossi, Porras, Jari, Vahala, Riku, 2017. Data utilization at Finnish water and wastewater utilities: current practices vs. State of the art. Util. Pol. 45, 69–75.
Soria-Comas, Jordi, Domingo-Ferrer, Josep, Sánchez, David, Martinez, Sergio, 2014. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. The VLDB Journal—The International Journal on Very Large Data Bases 23 (5), 771–794.
State of Washington: Data Privacy Guidelines for Large Utilities. http://www.wpuda.org/assets/Energydocs/model%20data%20privacy%20guideline%20for%20large%20utilities%2009%2008%2016.pdf.
Sweeney, Sean, 2014. Working toward Energy Democracy." in State of the World 2014, 215–27. State of the World. Island Press/Center for Resource Economics, Washington, DC. https://doi.org/10.5822/978-1-61091-542-7_20.
Szulecki, Kacper, January 2, 2018. Conceptualizing energy democracy. Environ. Polit. 27 (1), 21–41. https://doi.org/10.1080/09644016.2017.1387294.
U.S. Congress, 1999a. Gramm–leach–bliley Act of 1999, S.900. In: 106th U.S. Congress, November 4th, 1999.
U.S. Congress, 1999b. The health insurance portability and accountability Act of 1996, public law 104-191. In: 104th U.S. Congress, August 21st, 1996.
U.S. Department of Health and Human Services (HHS), 2015. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.
Weinrub, Al, 2017. Democratizing municipal-scale power. In: Energy Democracy, pp. 139–171.
Wu, Felix T., 2013. Defining privacy and utility in data sets. U. Colo. L. Rev. 84, 1117.
Young, Caroline, Martin, David, Skinner, Chris, 2009. Geographically intelligent disclosure control for flexible aggregation of census data. Int. J. Geogr. Inf. Sci. 23 (4), 457–482.
Zayatz, L., 2007. Disclosure avoidance practices and research at the US census Bureau: an update. J. Off. Stat. 23 (2), 253–265. https://doi.org/10.1109/TED.2007.909043.
Zhong, Sheng, Yang, Zhiqiang, Wright, Rebecca N., 2005. Privacy-enhancing k-anonymization of customer data. In: Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 139–47.
Zipper, S.C., Stack, W., Deines, J.M., Befus, K.M., Bhatia, U., Albers, S.J., et al., 2019. Balancing open science and pdata privacy in the water sciences. Water Resour. Res. 55, 5202–5211.