Reconciling privacy and efficient utility management in smart cities

David Rebollo-Monedero[1]*, Andrea Bartoli[2], Juan Hernández-Serrano[1], Jordi Forné[1] and Miguel Soriano[1,2]

[1] Department of Telematics Engineering, Universitat Politècnica de Catalunya (UPC), C. Jordi Girona 1-3, E-08034 Barcelona, Spain

[2] Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Av. Carl Friedrich Gauss 7, E-08860 Castelldefels, Barcelona, Spain

ABSTRACT

A key aspect in the design of smart cities is, undoubtedly, a plan for the efficient management of utilities, enabled by technologies such as those entailing smart metering of the residential consumption of electricity, water or gas. While one cannot object to the appealing advantages of smart metering, the privacy risks posed by the submission of frequent, data-rich measurements cannot simply remain overlooked. The objective of this paper is to provide a general perspective on the contrasting issues of privacy and efficient utility management, by surveying the main requirements and tools, and by establishing exploitable connections.

***Correspondence**

D. Rebollo-Monedero, Universitat Politècnica de Catalunya (UPC), Campus Nord, Módulo C3, C. Jordi Girona 1-3, E-08034 Barcelona, Spain.

E-mail: david.rebollo@entel.upc.edu

1. INTRODUCTION

Strongly increasing demographic flows towards cities call into question the environmental, economic and social sustainability of current urban models. Smarter utility management emerges as the answer to continue providing citizens with services as essential as electricity, water or gas, to the point that the very future of human prosperity and well-being may hinge on how we as a people recognise the necessity for efficient use or resources [1]. Fortunately, such unfaltering trends are matched by a breathtaking technological progress that may prove of great assistance in attaining the efficiency improvements we seek in utility management and many other aspects of smart cities.

While one cannot object to the appealing improvements enabled by technologies such as those entailing smart metering of utility consumption, the privacy risks posed by the submission of frequent, data-rich measurements cannot simply remain overlooked. The extensive literature on privacy for the smart grid abounds with very real examples of issues derived from smart metering [2–5]. Figure 1, adapted from [6], plots the consumption of power over time in an individual household, recorded on a minute-by-minute basis, presented to support the fact that a surprising amount of individual and social behavioural patterns may be inferred from it.

Having succinctly motivated the contrasting aspects of privacy and smart utility management, the object of this manuscript is to offer a quick glance at the connection between privacy requirements in smart metering and the privacy-enhancing technologies available. More specifically, our goals are as follows.

- Perhaps, the most extensively studied aspects of privacy for smart metering, or any information system, deal with unauthorised access to sensitive data, by means of authentication, data access control policies and *confidentiality*, implemented as cryptographic protocols [2, 3, 7]. In our exposition, we adopt a fresh perspective that attempts to go beyond well-established solutions for confidentiality. Indeed, we describe fairly diverse genres of privacy-enhancing technologies originally conceived for a wide variety of information systems, contributing to illustrate their applicability to issues pertaining to the efficient, smart management of utilities in the cities of the foreseeable future.

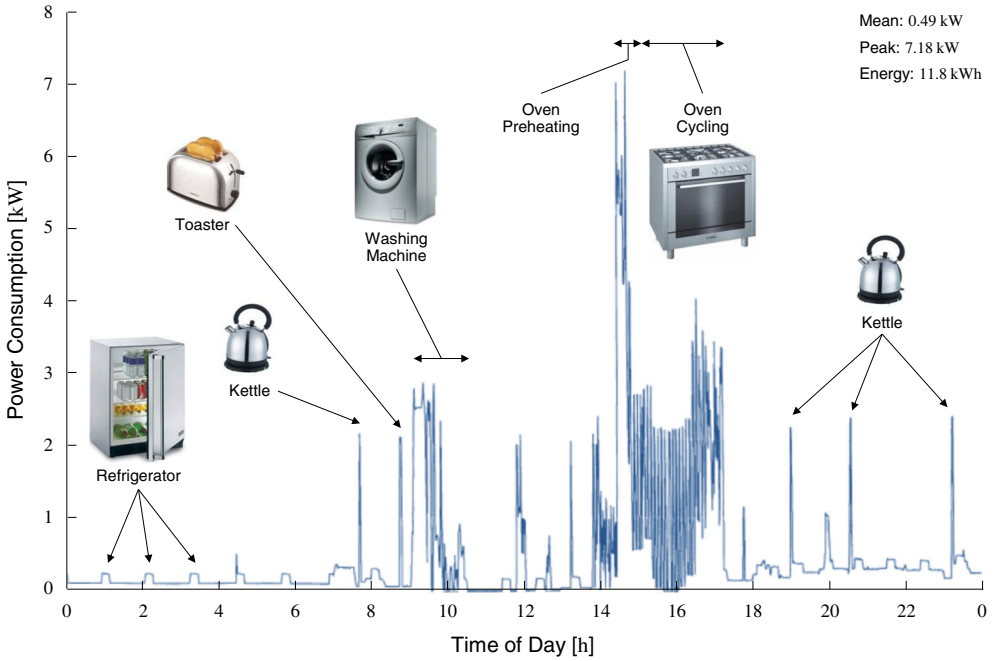


Figure 1. Example of electricity consumption over time for an individual household, adapted from [6].

- However, we strive to make a brief, accessible presentation focusing on notional clarity over technical details or thoroughness. Concordantly, we do not pretend to provide an exhaustive review of the state of the art on privacy technologies. Instead, we aim to reach a wider audience of researchers in the field of smart utility management, not necessarily familiar with the intricacies of privacy-enhancing mechanisms, seeking an introductory overview with emphasis on conceptual breadth over technical depth.
- Last but not least, we hope that the conceptual connections drawn here between requirements and tools serve as a valuable start point for the researcher making preliminary acquaintance with the subject.

The remainder of the paper enumerates privacy requirements, analyses related privacy technologies, delves into privacy metrics, and concludes with remarks on the interplay between privacy and system usability.

2. PRIVACY RISKS IN SMART METERING

Further to the well-established threats pertaining to confidentiality in advanced metering infrastructures, we would like to briefly enumerate privacy risks addressed by a less traditional type of privacy solutions. For the sake of clarity, along with the enumeration of risks, we obligatorily make a succinct introductory allusion to the corresponding privacy-enhancing technologies, which will be explored in the next section in matching order.

- **Eavesdropping.** We argued in the introductory section that a surprising amount of confidential information may be inferred from precise, frequent measurements of utility consumption. Consequently, the unintended disclosure of such measurements poses a serious privacy risk, especially in wireless networks, exposed to illicit eavesdropping. The problem of controlling the access to the contents of sensitive data in communication systems is traditionally accomplished by means of cryptographic mechanisms, also applicable to the more specific case of smart utility management. *Confidentiality* may be defined in a general manner as the protection of data from unauthorised disclosure [8]. In this paper, however, we shall prefer the more restrictive semantics referring only to the protection of the contents of the data via encryption. Traffic analysis, essentially inferring information merely from the flow of encrypted packets through a network, is a problem related to confidentiality whose importance warrants a separate subsection in the present manuscript.
- **Traffic analysis.** Communication of utility-related data between smart metres and utility companies may conveniently reuse pre-existing general-purpose networks, such as the Internet, sharing a variety of traffic exchanged among multiple parties. In addition, the frequency and size of such utility data may be efficiently adapted to the consumption patterns and requirements of different customers and companies at different times. Under these practical assumptions, message encryption is insufficient to mitigate all possible kinds of privacy risks derived from network

eavesdropping. Concealing the content of data packets hinders attackers in their efforts to learn the information exchanged, but does not prevent those attackers from unveiling who is communicating with whom, when or how frequently. We shall see that *anonymous-communication systems* encompass a large family of solutions designed to address the latter problem.

- **Statistical disclosure.** A variety of scientific studies, particularly research on efficient utility management in smart cities, may benefit greatly from the publication of statistical databases containing utility consumption information of customers, related to demographic information. Unfortunately, wide disclosure of such sensitive information poses serious privacy concerns. This relevant dilemma is commonly tackled under the mild assumption that the slight perturbation of certain attributes, exploitable to infer the identity of the individuals involved, should not severely affect underlying statistical correlations. In the next section, we shall illustrate the fundamental principles of such perturbative techniques, within the field of *statistical disclosure control (SDC)*, in the context of smart metering.
- **Consumption profiling.** Precise, frequent measuring of utility consumption is dangerously prone to user profiling, in the sense of analysis of personal behavioural patterns. The concept of *hard privacy* refers to the preservation of privacy by the user itself, simply by minimising, obfuscating or perturbing the information released, without the requirement of trusted intermediaries [9]. In the context of smart metering, we must address the practical case when the intended recipient of the data, namely the utility company, may not be fully trusted by the customers, but a minimum amount of accurate data is needed for billing. Cryptographic *data aggregation* may be employed to reveal only the aggregated consumption of a collective users to the utility company, preventing consumption profiling on an individual basis.

The flow of utility-related information and the privacy risks derived, according to the aforementioned list, are visually summarised in Figure 2.

3. RELATED PRIVACY-ENHANCING TECHNOLOGIES

We now proceed to offer a brief overview of privacy-enhancing technologies addressing the requirements itemised in the previous section. To make our exposition more concrete from the standpoint of system design, an example of architecture containing the main elements involved in the protected communication of utility measurements is diagrammatically represented in Figure 3.

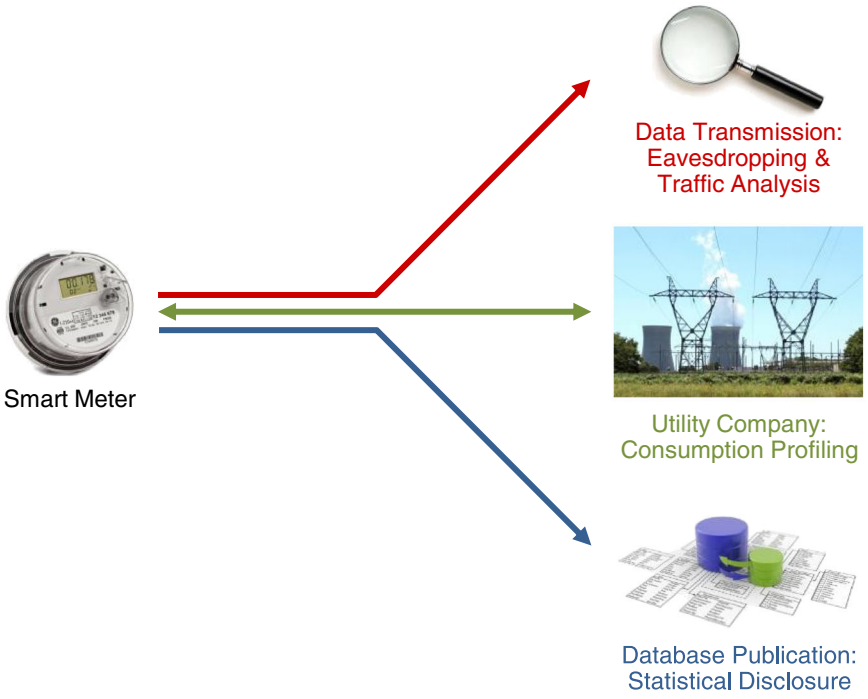


Figure 2. The utility-related information flow leads to several sources of privacy risks, including confidentiality violation, traffic analysis, statistical disclosure and user profiling based on consumption.

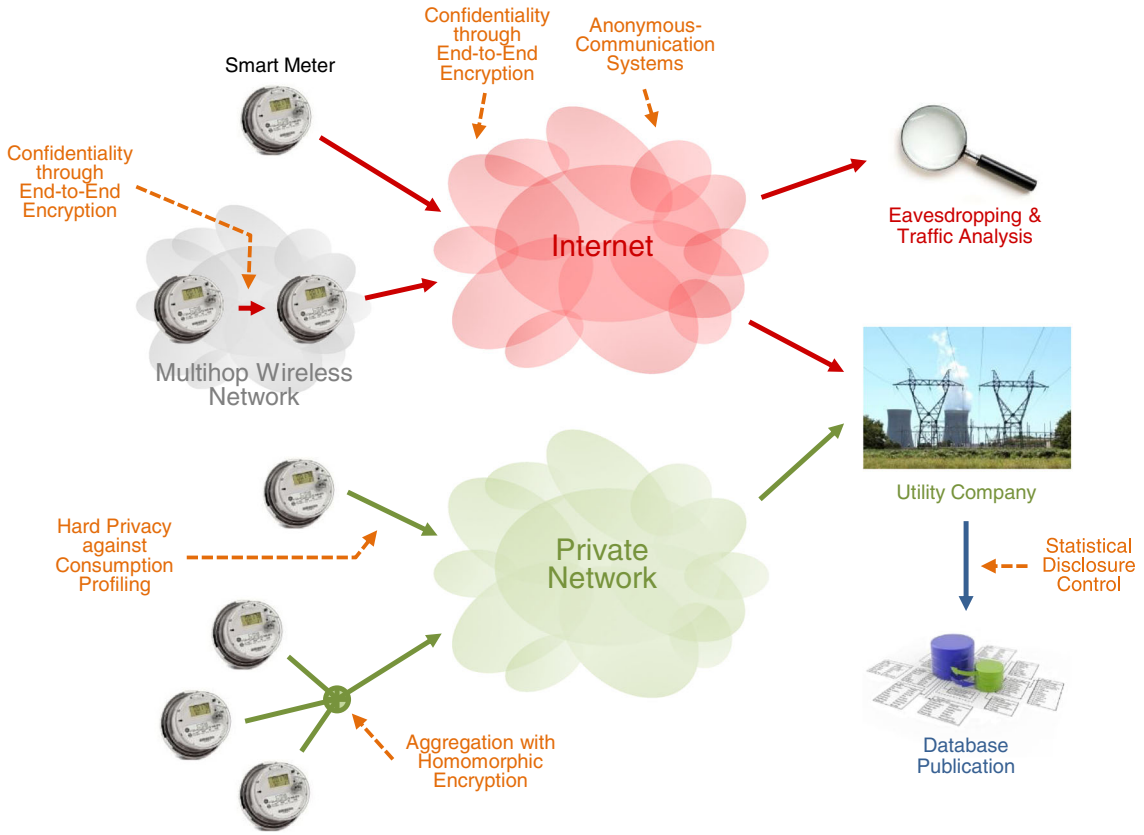


Figure 3. Example of architectural elements involved in the privacy-enhanced communication between smart metres and utility companies.

This diagram illustrates more concretely the information flow of Figure 2, pointing out specific privacy vulnerabilities in an advanced metering infrastructure and the types of mechanisms capable of addressing them. We would like to recommend bearing this depiction in mind when reading the upcoming subsections on the respective classes of privacy-enhancing technologies.

Before we proceed, we should hasten to point out that the adjective *usability* is used in the context of data accuracy or system functionality, instead of the more common qualifier *utility*. The reason is to prevent confusion with the term utility in reference to electricity, water or gas provision.

3.1. Confidentiality via cryptographic mechanisms

We have stressed the serious privacy risks posed by the unintended disclosure of frequent, data-rich measurements of utility consumption in smart metering. These risks are clearly manifest in the abundant literature [2, 3], and graphically evidenced in the introductory section with Figure 1, adapted from [6]. In this subsection, we review confidentiality issues solved through the implementation of cryptographic mechanisms [8], where by confidentiality we mean the restriction of the access to the contents of any sensitive information related to smart utility management to its intended recipients. The related issue of traffic analysis, even when encrypted, is treated separately in Section 3.2. Our necessarily brief exposition intends to serve merely as a conceptual introduction to the subject, partly because of our focus on less traditional privacy-enhancing mechanisms in subsequent subsections. More extensive studies on such type of mechanisms from a more technical perspective include [2, 3].

Complete provision of confidentiality must address a number of issues that might be presented along three dimensions, space, time and layer, which we depict in Figure 4 and proceed to describe next.

As far as *space* or network topology is concerned, encryption must enforce *end-to-end* confidentiality, precluding any intermediate nodes or entities from eavesdropping, along the path from a sender to a receiver, typically the smart metre and the utility company, respectively. *Hop-by-hop* confidentiality refers to the principle that not only the contents, but also certain headers with routing

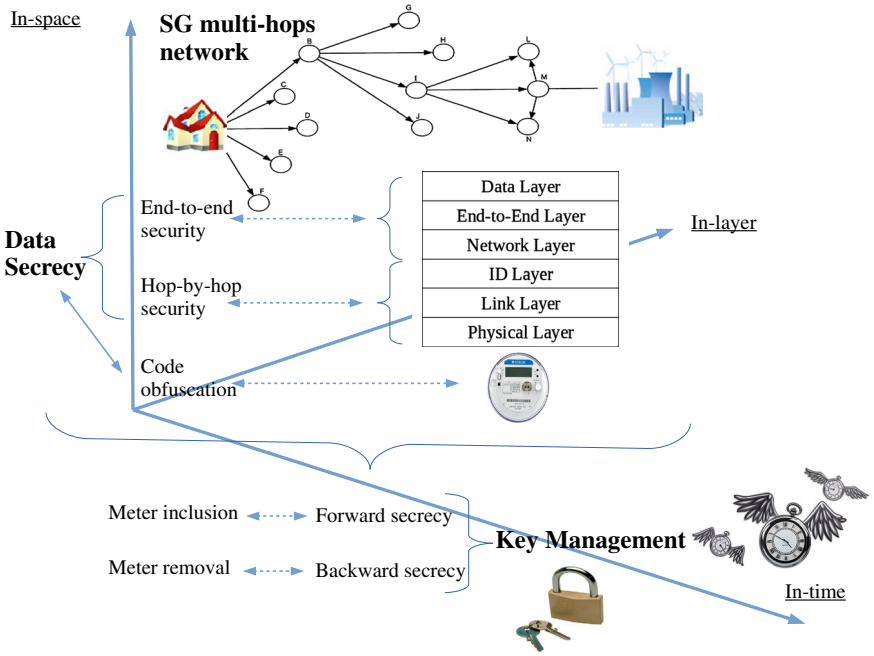


Figure 4. Provision of confidentiality along the dimensions of space, time and layer.

information may be encrypted, in such a manner that a specific piece of information may be accessed only by the routing nodes requiring it. Even if routing headers are encrypted, traffic analysis may exploit size and timing to infer the path followed by a packet through the network, an additional problem whose solutions are discussed in Section 3.2.

Because of the low-power, low-complexity constraints present across most of the topology of a typical communication network for smart utility management, symmetric block cyphers such as Skipjack, RC5, AES and Twofish, are most suitable for *data secrecy* [10]. Although the computational cost of asymmetric cryptographic operations has hindered its application in constrained devices, recent improvements in elliptic-curve cryptography have enabled proposals for the smart grid [11]. *Obfuscating* not only the data but also the *code* contained in a potentially compromised device, such as a smart metre, will undoubtedly make it harder for an attacker to gain access to sensitive information such as security credentials [12], especially if the obfuscation methods vary across devices.

Secondly, the dimension of *time* obeys to the requirement of maintaining confidentiality over the entire lifespan of the communication network. Specifically, *forward secrecy* is the requirement that no node should be able to read any future messages after leaving the network. *Backwards secrecy* requires that a newly joining entity be unable to read any previously transmitted messages. In order to fulfil such requirements, reliable key-management protocols must be implemented throughout the communication network in the advanced metering infrastructure [13, 14]. An example of *key-management* platform specifically recommended for smart grids is Zigbee PRO (ZigBee Alliance 2400 Camino Ramon Suite 375 San Ramon, CA 94583 USA) [15].

Thirdly and lastly, the combination of cryptographic mechanisms operating within different *layers* of the communication protocol reinforces the overall communication security, fulfilling the requirements of end-to-end and hop-by-hop confidentiality aforementioned. Encryption at the *physical* layer protects headers right at the lowest level of the protocol stack thereby being the most conservative alternative. However, security in wireless networks is most commonly implemented at the next layer, specifically at the *media access control* sublayer, part of the data link layer. This type of security, in addition to protecting the contents of the data exchanged at this level, contributes to hinder attackers in their efforts to perform traffic analysis. Obviously, traffic analysis can also be hindered to a certain extent by encryption at the *network* layer. Encryption only at the *application* layer, while protecting the contents of the information actually exchanged between sender and receiver, is clearly not as resilient against such analysis.

3.2. Anonymous-communication systems

We have mentioned that message encryption is insufficient to mitigate privacy risks derived from traffic analysis based on routing, size, timing and frequency patterns of messages

between customers and utility companies. Motivated by these risks in a general communication context, numerous privacy-protecting technologies referred to as *anonymous-communication systems* emerged. The fundamental strategies to counter traffic analysis based on message routing and size involve header encryption, message padding and splitting and even the insertion of dummy traffic. However, such countermeasures fail to address the risks posed by the analysis of the time instants in which messages are sent, routed and received. The first anonymous-communication system attempting to also counter timing analysis was the *Chaum mix*, essentially a trusted node that delays and re-orders messages with the purpose of providing unlinkability between the incoming and outgoing messages.

A wide range of sophisticated variations on the original mix shortly ensued [16], with the same purpose. One of the most relevant varieties is a family of mixes known as *threshold pool mixes*. The leading idea is for the mix to collect a number of incoming messages, store them in the internal memory of the mix and output some of them when the number of messages kept in its memory reaches a certain threshold. In order to reduce the correlation between outgoing and incoming messages, the mix modifies the flow of messages by resorting to two strategies, namely the delay and re-ordering of messages.

The precise manner in which the mix operates is the following. Consider a threshold pool mix in steady state, forwarding batches of k messages at a time, with a buffer of $n > k$ messages thus keeping at least

$$m = n - k > 0$$

messages at a given time. At some point, the mix contains its minimum of m messages and waits for k additional messages to come in, for a total of $n = m + k$. Once this threshold n is reached, k messages are drawn randomly, independently and uniformly among all stored messages, regardless of the order in which they arrived, and sent out simultaneously. This leaves the mix again at its minimum m, and the process is repeated from that point on. Figure 5 is a snapshot of the operation of a threshold pool mix with $k = 2$ and $n = 5$.

Naturally, chains of mixes can be implemented to distribute trust, and, certainly, delaying messages affects the usability of these systems. Nevertheless, higher delays provide users with a higher degree of message unlinkability. In short, mix systems pose an inherent trade-off between anonymity and delay, in addition to the overheads derived from any encryption or padding.

Alternative low-latency anonymous-communication systems appeared later to provide routing anonymity on the Internet to a certain extent, without the price of message delay. Onion routing and subsequent improvements termed *the* second-generation version of *onion routing (Tor)* [17], consist in networks of trusted routing nodes that, unlike mixes, do not introduce additional delays. In a nutshell, a user wishing to send a message chooses a chain of onion routers and encrypts the message in a multilayered manner—hence the onion metaphor. This multilayered encryption is such that each router, after decrypting—peeling off a layer of encryption—, retrieves the address in plaintext of the node immediately subsequent in the path, along with an encrypted portion meant for said next node, all the way to the final recipient. We would like to stress that, as these systems boil down to anonymously relying messages without introducing delays, they are susceptible to traffic analysis based on timing comparisons.

Yet another type of anonymous communication systems builds upon the principle of user collaboration with a limited degree of trust. In the *Crowds* protocol [18, 19], for instance, a group of users will collaborate to submit their messages to a specified recipient. Simply put, when sending a message, a user flips a biassed coin to decide whether to submit it directly to the recipient or to send it to another user, who will then repeat the randomised decision.

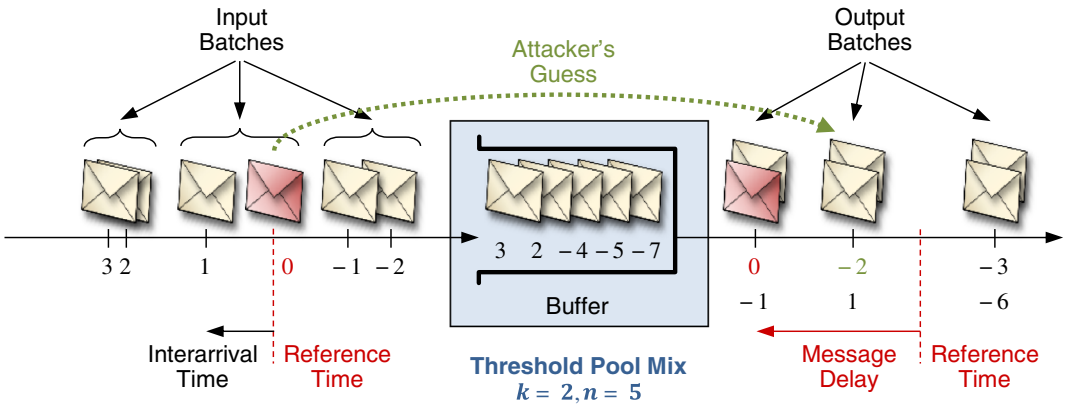


Figure 5. Possible state of a k-input, k-output threshold pool mix, with $k = 2$ and with a buffer of $n = 5$ messages. When the number of messages stored in the mix reaches $n = 5$, the mix chooses $k = 2$ messages at random and forwards them simultaneously in a batch.

Adding an initial forwarding step substantially increases the uncertainty of the first sender from the point of view of the final receiver, at the cost of an additional hop. In the end, anonymity comes at the expense of traffic overhead and delay.

3.3. Statistical disclosure control

We have pointed out the convenience of the publication of confidential statistics such as utility consumption, for a variety of demographic studies, particularly those dealing with smart grid efficiency. The objective of *SDC* [20] is to control the risk that information about specific individuals can be extracted from amongst statistical summary results. In the SDC terminology, a *microdata* set is a database table whose records carry information concerning individual respondents, either people or companies. This database commonly contains a set of attributes that may be classified into identifiers, quasi-identifiers and confidential attributes. Firstly, *identifiers* allow unequivocal identification of individuals. This is the case of social security numbers or full names, which would be removed before the publication of the microdata set. Secondly, *quasi-identifiers*, also called *key attributes*, are attributes that, in combination, may be linked to external information to re-identify the respondents to whom the records in the microdata set refer. Examples include address, gender, age, job type, height and weight. A notorious fact is that 87% of the population in the USA may be re-identified solely on the basis of the triple consisting of their ZIP code, gender, and date of birth, according to 1990 census data [21]. Finally, the dataset contains *confidential attributes* with sensitive information on the respondent, such as salary, religion, political affiliation, health condition and electricity consumption. The classification of attributes as key or confidential may ultimately rely on the specific application and the privacy requirements the microdata set is intended for, and in fact consumption patterns could be conceivably construed as key attributes as well.

Hence, from the standpoint of privacy protection, mere removal of identifiers is in general insufficient in the publication of microdata sets for statistical studies. Intuitively, perturbation of numerical or categorical key attributes enables us to preserve privacy to a certain extent, at the cost of losing some of the *data usability*, in the sense of accuracy with respect to the unperturbed version. *k-Anonymity* [22] is the requirement that each tuple of key-attribute values be shared by at least k records in the dataset. This may be achieved through the *microaggregation* approach illustrated by the simple example depicted in Figure 6, where gender, age and ZIP code are regarded as key attributes and monthly average utility bill as a confidential attribute. Rather than making the original table available, we publish a k-anonymous version containing aggregated records, in the sense that all key-attribute values within each group are replaced by a common representative tuple. As a result, a record cannot be unambiguously linked to the corresponding record in any additional database assigning identifiers to key attributes. In principle, this prevents a privacy attacker from ascertaining the identity of an individual for a given record in the microaggregated database, which contains confidential information.

Even though the simplicity and algorithmic tractability of k-anonymity as a measure of privacy makes it a widely popular criterion in the SDC literature, it is not without shortcomings. Indeed, while this criterion prevents identity disclosure, it may fail against the full disclosure of the confidential attribute. Concretely, suppose that a privacy attacker knows Alice's key attribute values. If the attacker learns that she is included in the released table depicted in Figure 6 and has access to external information

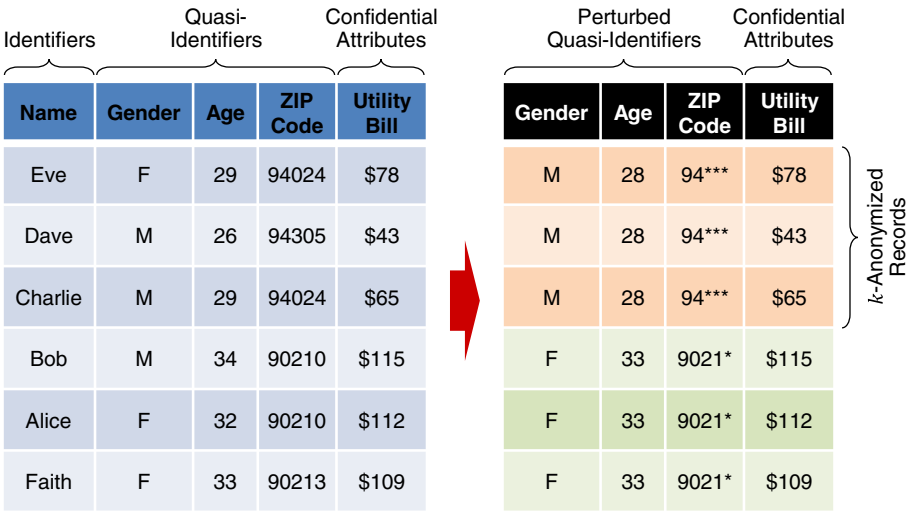


Figure 6. Hypothetical example of k-anonymous micro-aggregation of published data relating demographic information with utility consumption for a population of individuals, with $k = 3$.

on her key attributes, then the attacker may conclude that her monthly average bill amounts to a figure between \$109 and \$115, fairly similar values. This is known as *similarity* or *homogeneity attack*, meaning that values of confidential attributes within a group may still be quantitatively or qualitatively similar.

From a more general, probabilistic perspective, the *skewness attack* exploits the difference between the prior distribution of confidential attributes in the entire population, and the posterior distribution of those attributes within a specific group of the table. Assume, in the example of Figure 6, that the average utility bill across the population of study is \$87, and that this is known by the privacy attacker. With the information revealed by the published table and the knowledge of Dave's key attributes, this privacy attacker will gain further knowledge in a statistical sense. In particular, the attacker will deduce that Dave's monthly bill is in the range from \$43 to \$78, well below the population's average, or at most \$65 with 67% likelihood. A *background-knowledge attack* exploits additional side information to further refine statistical inferences. For instance, if Dave were the only member in the group listed as a graduate student living in a university dorm and the rest were known to live in family detached houses, the attacker might infer that Dave is likely to have the lowest utility bill, \$43.

Intuitively, we seek to introduce the smallest perturbation in the key attributes, thereby preserving as much as possible the statistical quality of the published data. A number of algorithms for microaggregation have been developed, with the goal of minimising the perturbation of the key attributes with accordance to a variety of distortion measures, while meeting a given k-anonymity constraint [20]. Once again, we encounter privacy-enhancing mechanisms incurring a cost, this time in terms of data usability.

Among the best known algorithms in the SDC community, the *maximum distance (MD)* algorithm [23] and its less computationally demanding variation and the *MD to average vector (MDAV)* algorithm [24, 25], are fixed cluster size algorithms that perform particularly well in terms of the distortion they introduce, for many data distributions. The *probability-constrained Lloyd* algorithm generalises certain optimality conditions originally devised for quantizer design [26] and is capable of outperforming MDAV in a variety of synthetic and standardised datasets [27], albeit at the expense of increased computational complexity and mathematical sophistication.

3.4. Hard Privacy against consumption profiling

We have introduced the concept of hard privacy [9, 28], apropos of which, we would like to complete the privacy-enhancing scenarios covered thus far with the compelling case of protecting the data on utility consumption revealed by customers. Specifically, we contemplate the protection against a potentially untrusted utility company, carried out directly by the users themselves with no third parties involved and by means of data-perturbative strategies.

Among the privacy technologies explored in this work, perhaps the avenue of perturbation of consumption profiles, present the greatest opportunities for open-ended research. Mathematically formal studies of hard privacy of user profiles include [29, 30], which investigate the submission of search queries to an information provider and the tagging of resources in the semantic Web, respectively. Loosely speaking, the cited work defines *user profile* as a histogram of relative frequencies of activity across predefined categories of interest, for example, business, science and sports in a news context. Then, it investigates the effects of forging bogus queries and of suppressing certain tags in terms of privacy gains because of an effective perturbation of the apparent profile from the point of view of an external observer. Costs in system usability are defined as the relative amount of forged queries and suppressed tags, respectively, on the account of traffic and processing overheads and degradation of semantic functionality. Finally, the cited work proceeds to analyse the maximisation of the privacy gains for given usability constraints resorting to convex optimisation techniques.

The perturbation of profile of interests from the standpoint of an untrusted observer may be accomplished without forgery and suppression of online activity, so long as users trust and collaborate with each other. This is illustrated by [31], which contemplates the exchange of queries among users prior to its submission to a common information provider.

In order to relate privacy risks in the smart grid with the literature on antiprofiling, and more concretely [29,30], we provide an illustrative example in which behavioural patterns in electricity consumption are mathematically modelled by means of simple histograms. Figure 7 depicts a hypothetical example in which power consumption data for a particular season (e.g. winter) is aggregated in the form of a histogram of relative values across six four-hour periods during the day. This histogram could, of course, be accompanied by the total overall consumption during the season, so that absolute values could be recovered.

Naturally, any perturbation of the user's accurate profile with the goal of attaining a higher degree of privacy must conform to any billing requirements or at least be accompanied with any unaltered data needed for accurate billing. Beyond such billing constraints, profile perturbation can, in principle, be conducted directly in any fashion unlike the indirect query forgery or tag suppression mechanisms in [29, 30]. An exciting connection with the work cited is that the measure of privacy proposed involves a reference profile, commonly representing the average consumption patterns across the population of interest. Although details are postponed until the next section on privacy metrics; in essence, the anonymity of a perturbed profile is quantified as an information-theoretic

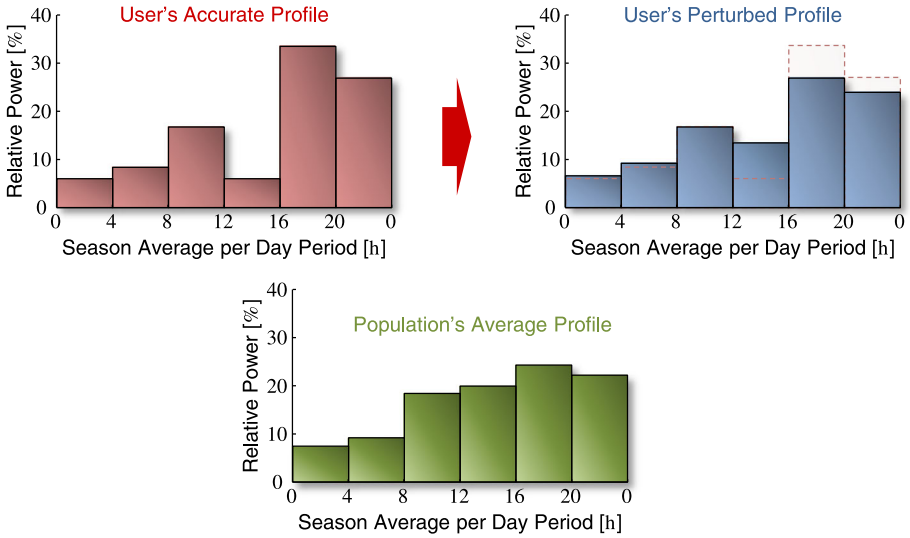


Figure 7. Perturbation of an individual user profile, capturing behavioural patterns in electricity consumption, to approach the population's average profile.

measure of discrepancy with respect to the population's average profile. The more similar to the average consumption pattern, the more anonymous a user's behaviour may be considered.

3.5. Data aggregation against consumption profiling

A somewhat laxer form of hard privacy, where individuals transform or perturb sensitive data prior to sending it to any untrusted external party, relies on the collaboration with neighbouring parties. More precisely, assuming that such data can be numerically aggregated, *homomorphic encryption* [32] enables several individuals to exchange and aggregate data in such a manner that the recipient is able to decipher the aggregated value, but no party involved is able to unveil each of the additive parts. Assuming that those groups of collaborating users persist through extended periods of time, profiling of consumption habits cannot pinpoint individual users but only characterise small collectives. Data aggregation has been extensively studied in the context of smart metering [33].

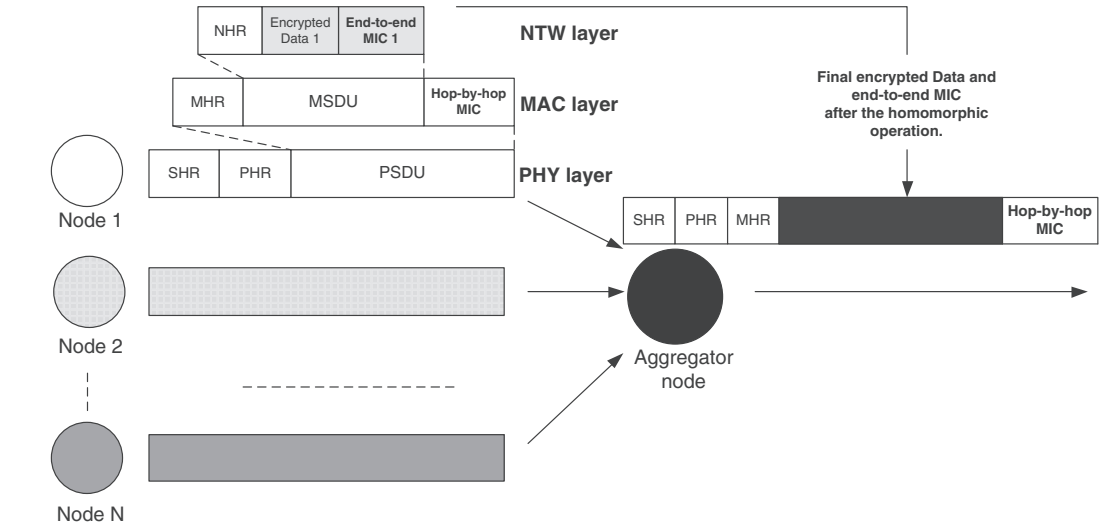
It must be pointed out that data aggregation is effectively a form of perturbation, where each of the individual profiles participating in the aggregation is replaced by a common aggregated profile. An important difference with respect to the forgery and suppression strategies described in Section 3.4, is that unless hard privacy and data aggregation are combined, simple lossless data aggregation by itself does not permit the strategies of parameter optimisation exploited in [29, 30]. Notwithstanding this difference, the number of profiles aggregated may be construed as a tuning parameter directly impacting anonymity and inversely affecting data usability.

In general, data aggregation may be either lossy or lossless. Strictly speaking, it is only the former mode of operation, *lossy*; the one that resorts to adding or averaging consumption measurements, as assumed previously. *Lossless* data aggregation merely consists in the concatenation of several data payloads into a single, larger packet, thereby reducing protocol overhead.

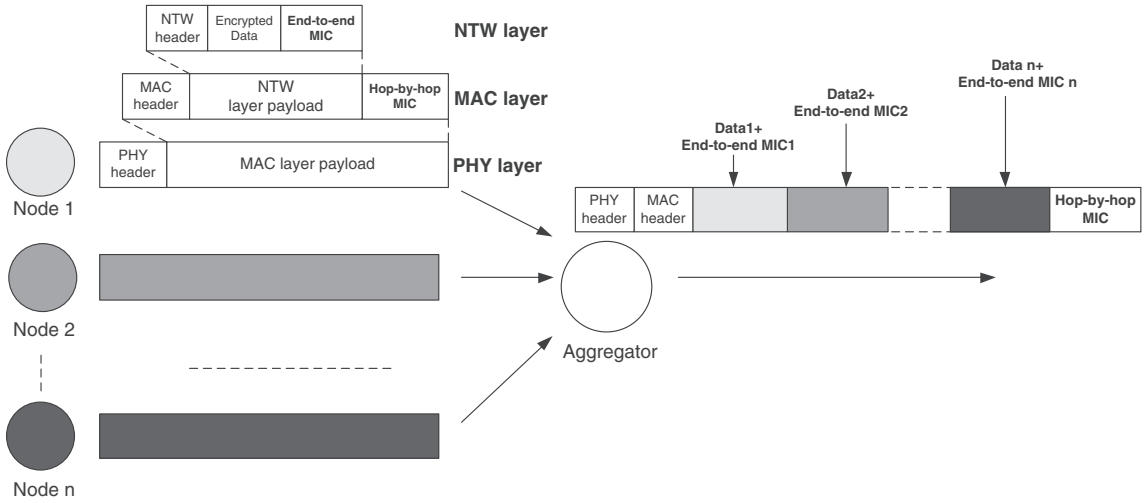
Despite the fact that lossless aggregation employs neither homomorphic encryption nor mathematical addition, a degree of anonymity may be attained provided that the concatenation of consumption measurements is decipherable by the utility company, for instance by means of a group key, but the correspondence of those measurements with the users within the group aggregated remains unknown. If individual keys are used instead, lossless aggregation is merely a traffic-efficient solution for confidentiality.

A typical secure lossy data-aggregation solution where end-to-end and hop-by-hop security is provided with homomorphic encryption, is shown in Figure 8(a). A concrete example of lossy data aggregation is [34], where all smart metres are connected to a substation in charge of collecting and aggregating electricity measurements. The solution proposed is secure as long as two smart metres remain uncompromised.

Figure 8(b) represents a secure lossless data-aggregation method where end-to-end and hop-by-hop security are guaranteed at the network and link layers, respectively. Because consumption measurements typically contain just a few bytes of data, its concatenation contributes to reduce the relative overhead of packet headers. However, overly long packets may translate into frequent retransmissions because of more likely packet errors, which in



(a) Lossy.



(b) Lossless.

Figure 8. Data aggregation with end-to-end and hop-by hop-securities.

turn may cause larger energy expenditures. Taking this into consideration, appropriate packet sizes for lossy data aggregation are computed experimentally in [35].

4. QUANTITATIVE ASSESSMENT OF THE FULFILMENT OF PRIVACY REQUIREMENTS

Quantifiable measures of privacy [36] and usability are undoubtedly essential to the assessment, comparison, improvement and optimisation of privacy-enhancing mechanisms for advanced measurement infrastructures such as the smart grid, in terms of both their privacy and usability, from both theoretical and numerical perspectives. In this work, we shall focus on the aspect of privacy, for which we may capitalise on the existing measures of information, uncertainty, (attacker's) estimation error and diversity, which abound in the fields of information theory [37], statistics and engineering.

Of particular significance is the quantity known as *Shannon's entropy*, a measure of the uncertainty of a random event, associated with a probability distribution across the set of possible outcomes. A well-known interpretation of this entropy refers to the game of 20 questions, in which one player must guess what the other is thinking through a series of yes/no questions, as quickly as possible. Informally, Shannon's entropy is a lower bound on—and often good approximation to the minimum of—the average

number of binary questions regarding the nature of possible outcomes of an event, to determine which one in fact has come to pass, intelligently exploiting their known probabilities.

The following subsections touch upon the measurement of anonymity and privacy for the technologies succinctly reviewed in §3, except for those based on encryption, given the absence of a continuous trade-off between confidentiality and usability for the intended recipient of the data—the data is either made confidential or not.

4.1. Anonymous-communication systems

In the special case of anonymous-communication systems, described in Section 3.2 (see also Figure 5), the knowledge of the privacy attacker may be modelled by a probability distribution on the possible senders of a given message. Certainly, one could measure the degree of anonymity attained by the mere cardinality of the set of candidate senders. The logarithm of such cardinality is in fact called *Hartley's entropy*. Loosely speaking, Hartley's entropy may be regarded as a *best-case* metric from an optimistic point of view of users (worst or pessimistic for adversaries), in the sense that it represents a privacy attacker's thorough effort in considering any and all possibilities, regardless of their likelihood.

A metric oriented towards an *average-case* scenario would take into consideration the probability distribution of candidate senders, or recipients, of a given message, thereby exploiting its potential skewness. Inspired by the aforementioned interpretation of Shannon entropy as the effective uncertainty within a set endowed with a probability distribution, Serjantov [38] proposed it as a measure of anonymity.

An alternative interpretation of Shannon's entropy in the context of privacy is offered in [36] on the basis of the *asymptotic equipartition property* and the concept of *typical set* [37]. Precisely, for an adversary joint estimating sequences of uncertain outcomes, rather than individually guessing single occurrences, Shannon's entropy is a measure of the effective cardinality of the set of candidate sequences.

At the other extreme in the family of entropies lies the *min-entropy* of a distribution, defined as the negative logarithm of the most likely outcome. This third example of anonymity measure associated with a set of candidate identities may be construed as a *worst-case* metric in the sense that users are only concerned with the most vulnerable statistical link between messages and senders or recipients [36].

Possible measures of usability comprise average delay and the probability that such delay exceeds a given threshold tolerated by the application at hand.

4.2. Statistical disclosure control

In the particular case of SDC, reviewed in Section 3.3, (see also Figure 6), we already defined k-anonymity as a measure of privacy and briefly commented on several attacks. Common measures of usability loss for numerical attributes include the *mean squared error* between the original and the perturbed tuples. The vulnerabilities of k-anonymity motivated the appearance of a number of enhancements [39], some of which we proceed to review. For example, p-sensitive k-anonymity incorporates the additional restriction that there be at least p distinct values for each confidential attribute within each k-anonymous group. While this overcomes the similarity attack to a certain extent, the vulnerability to the skewness attack still remains.

Measures to address skewness compare the prior distribution of the confidential attributes in the population with the posterior distribution within each group of the microaggregated data. *t-Closeness* is the requirement that for each group, a specific metric of discrepancy between those distributions should not exceed a threshold t. A particularly useful, information-theoretic metric of discrepancy between probability distributions is the *Kullback–Leibler (KL) divergence*, also called *relative entropy* for its relationship with Shannon's entropy. Both Shannon's entropy and KL divergence are also tightly related to the information-theoretic quantity known as *mutual information*, a measure of the uncertainty in one random event unveiled by the outcome of a second, related event [37]. Partly inspired by t-closeness, Rebollo *et al*. [39] define privacy risk as the average KL divergence between the posterior and the prior distributions for each group, a measure that may be regarded as an average-case version of t-closeness. This *average privacy risk* is then shown to be equal to the mutual information between the confidential attributes and the observed, perturbed key attributes.

4.3. Hard privacy and data aggregation against consumption profiling

Last but not least, in the distinct case of hard privacy of consumption profiles modelled as probability distributions, presented in Section 3.4, (see also Figure 7), the concept of KL divergence plays once more a key role. Concretely, Rebollo *et al*. and Parra-Arnau *et al*. [29, 30] proposed to measure the anonymity of a user's consumption profile, possibly perturbed, as its KL divergence with respect to the average profile across the entire population. When the population's profile is taken to be the uniform distribution, this divergence boils down to being equivalent to Shannon's entropy.

Leveraging on a celebrated information-theoretic rationale by Jaynes, the KL divergence is interpreted as an (inverse) indicator of the commonness of similar profiles in said population [40]. As such, we should hasten to stress that under this interpretation, the KL divergence is a measure of anonymity rather than privacy, in the sense that the obfuscated information is the uniqueness of the identity behind the online activity, rather than the actual

Table I. Main privacy risks, related privacy-enhancing technologies, and quantifiable metrics of privacy discussed in this manuscript.

Consumption-Information Flow	Privacy Risks	Privacy-Enhancing Technologies	Privacy Metrics
Network transmission 	Eavesdropping	Confidentiality through encryption • Symmetric block cyphers • Elliptic-curve cryptography	(no privacy-usability trade-off)
	Traffic analysis	Anonymous communication systems • Pool mixes • Onion routing • Crowds protocol	• Hartley entropy • Shannon entropy • Min-entropy
Database publication 	Statistical disclosure	Statistical disclosure control • Microaggregation	• *k*-Anonymity • *p*-Sensitive *k*-anonymity • *t*-Closeness • Average privacy risk
Utility company 	Consumption profiling	Hard privacy • Profile perturbation • Data aggregation	• Kullback-Leibler divergence • Shannon entropy

profile of interests. Indeed, a profile of interests already matching the population would not require perturbation. Alternative privacy requirements might be formulated to protect inferences of the original user profile from the observed perturbation.

In this last case of hard privacy against consumption profiling, appropriate metrics of usability should carefully contemplate the quantitative impact of the perturbation of consumption information in terms of the intricate series of benefits of smart metering. Example of mathematically tractable measures include redundancy and suppression rate in query forgery and tag suppression, the respective technologies studied in the cited work and, effectively, the amount of perturbed activity.

We remarked in Section 3.5 (see also Figure 8) that lossy data aggregation is effectively a form of perturbation, where each of the individual profiles participating in the aggregation is replaced by a common aggregated profile. This abstract equivalence permits the direct application of the same metrics of anonymity proposed in [29, 30]. The number of users being aggregated is thus viewed as a tuning parameter of the compromise between anonymity and data usability, the former measurable by means of the KL

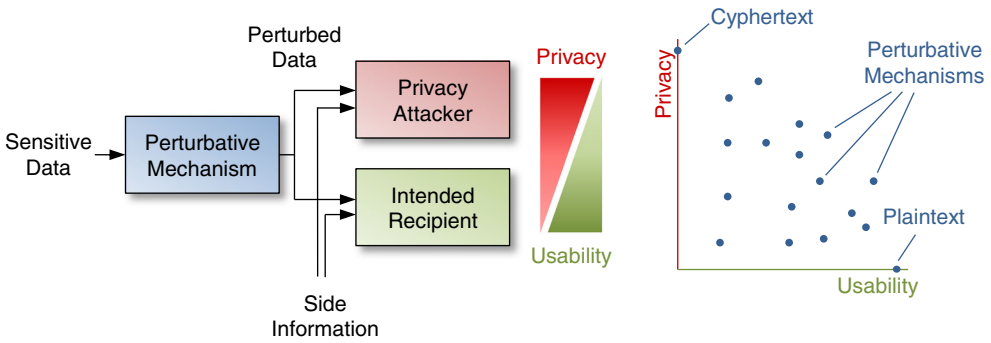


Figure 9. We investigate privacy in smart metering beyond the traditional approaches of access control and confidentiality through encryption, considering privacy-enhancing mechanisms based on data-perturbative strategies. Such perturbation poses a trade-off between the privacy and usability of the data.

divergence between the aggregated, apparent profile, and the population.

5. CONCLUDING REMARKS

The protection of user privacy in advanced measuring infrastructures is inextricably tied to the sustainable development of information and communication technologies for the efficient management of utilities.

In our introductory overview of various privacy-enhancing solutions, the emphasis on conceptual breath over technical depth allows us not only to reach a wider audience, but also to look more comprehensively beyond the more traditional approaches of access control and confidentiality through encryption. Indeed, fairly diverse genres of more recent technologies, originally conceived for a variety of information systems, are illustrated here in the context of smart metering. The main privacy risks, related privacy-enhancing technologies, and quantifiable metrics of privacy discussed in this manuscript are visually summarised in Table I.

Of particular importance is the compelling case when the intended recipient of sensitive information is not fully trusted and may thus be construed as a privacy attacker as well. Traditional encryption offers the possibilities of either fully delivering or completely obfuscating data, by either providing or not a cryptographic key permitting its deciphering. In the case of untrusted recipients, however, we are faced with a dilemma of great practical relevance. Most of the novel privacy mechanisms explored in this paper resort to perturbing or obfuscating certain information released, but only to a certain degree, in lieu of simply making it either completely available or unavailable. This fundamental notion is captured by the diagrammatic representation in Figure 9.

In addition, we have established that privacy often comes at a price in terms of usability of the underlying data or information system. In anonymous-communication systems, this price takes the form of message delay or traffic overhead; in SDC, the cost relates to the distortion introduced in the key attributes; and finally, hard privacy and lossy data aggregation come at the expense of accuracy in the user consumption profile.

Further, both privacy and usability may be attained to various quantifiable degrees, constituting contrasting quantities in a *privacy-usability trade-off*. The existence of this inherent price is a strong motivation to develop adequate privacy metrics and ultimately to design practical privacy tools achieving the maximum privacy for a desired usability level, or vice versa; ideally, we seek the optimal privacy-usability trade-off.

Future research aimed at reconciling privacy and efficient utility management in smart cities must therefore build upon a widely interdisciplinary variety of fields. These include not only extensive work on more traditional, cryptographically based security, but also a number of privacy-enhancing technologies intersecting with the fields of information theory and engineering optimisation. We hope that this work succeeds in offering a quick, fresh glance at such compelling aspect of the future of our society.

ACKNOWLEDGEMENTS

We would like to thank J. Parra-Arnau for the helpful comments on the state of the art of privacy technologies. This work was partly supported by the Spanish Government through projects Consolider Ingenio 2010 CSD2007-00004 “ARES" and TEC2010-20572-C02-02 “Consequence", and by the Government of Catalonia under grant 2009 SGR 1362. D. Rebollo-Monedero is the recipient of a Juan de la Cierva post-doctoral fellowship, JCI-2009-05259, from the Spanish Ministry of Science and Innovation.

REFERENCES

1. Karnouskos S, Silva PGD, Ilic D. Energy services for the smart grid city, In *Proceedings IEEE International Conference on Digital Ecosystem Technologies-Complex Environment Engineering (DEST-CEE)*, Campione d’Italia, Italy, June 2012; 1–6.
2. Smart Grid Interoperability Panel, Cyber Security Working Group, Guidelines for smart grid cyber security: Vol. 1, smart grid cyber security strategy, architecture, and high-level requirements, and Vol. 2, privacy and the smart grid, National Institute of Standards and Technology (NIST). *Interagency Rep. 7628*, August 2010. Available from: http://csrc.nist.gov/publications/nistir/ir7628/nistir-7628_vol1.pdf [4 September 2013].
3. Quinn EL. Privacy and the new energy infrastructure. *Center for Energy and Environmental Security (CEES),* working paper no. 09 - 001, 2008, 1–47. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1370731.[4 September 2013].
4. McDaniel P, McLaughlin S. Security and privacy challenges in the smart grid. *IEEE Security and Privacy Magazine* 2009; **7**(3): 75–77.
5. Cavoukian A, Polonetsky J, Wolf C. Smartprivacy for the smart grid: Embedding privacy into the design of electricity conservation. *Identity, Information Society* 2010; **3**(2): 275–294.
6. Wood G, Newborough M. Dynamic energy-consumption indicators for domestic appliances: Environment, behaviour and design. *Energy, Building* 2003; **35**(8): 821–841.
7. Adams C, Lloyd S. *Understanding PKI: Concepts, Standards, and Deployment Considerations*, 2nd ed. Addison-Wesley Professional: Boston, MA, 2002.

8. Stallings W. *Cryptography and Network Security: Principles and Practices*, 5th ed. Prentice Hall: Boston, MA, 2005.
9. Danezis G. Introduction to privacy technology, Katholieke Universiteit Leuven, Computer Security and Industrial Cryptography (COSIC) Res. Talk, 2007. Available from: http://research.microsoft.com/en-us/um/people/gdane/talks/Privacy_Technology_cosic.pdf [4 September 2013].
10. Cheng H, Ding Q. Overview of the block cipher, In *Proceedings International Conference on Instrumentation, Measurement, Computer, Communication and Control (IMCCC)*, Hangzhou, China, October 2012; 1628–1631.
11. Qing-hai B, Wen-bo Z, Peng J, Xu L. Research on design principles of elliptic curve public key cryptography and its implementation, In *Proceedings IEEE International Conference on Computer Science and Service System (CSSS)*, Nanjing, China, August 2012; 1224–1227.
12. Costa R, Pirmez L, Boccardo D, Rust LF, Machado R. TinyObf: code obfuscation framework for wireless sensor networks, In *Proceedings International Conference on Wireless Networks (ICWN)*, Las Vegas, NV, July 2012; 68–74.
13. Das S, Ohba Y, Kanda M, Famolari D, Das SK. A key management framework for AMI networks in smart grid. *IEEE Communications Magazine* 2012; **50**(8): 30–37.
14. Liu N, Chen J, Zhu L, Zhang J, He Y. A key management scheme for secure communications of advanced metering infrastructure in smart grid. *IEEE Transactions on Industrial Electronics* 2013; **60**(10): 4746–4756.
15. Radmand P, Domingo M, Singh J, Arnedo J, Talevski A, Petersen S, Carlsen S. Zigbee/Zigbee PRO security assessment based on compromised cryptographic keys, In *Proceedings International Conference on P2P, Parallel Grid, Cloud and Internet Computing (3PGCIC)*, Fukuoka, Japan, November 2010; 465–470.
16. Serjantov A, Dingledine R, Syverson P. From a trickle to a flood: Active attacks on several mix types, In *Proceedings Information Hiding Workshop (IH),* Springer-Verlag, Berlin, Heidelberg, 2002; 36–52.
17. Dingledine R, Mathewson N, Syverson P. Tor: the second-generation onion router, In *Proceedings Conference USENIX Security Symposium*, Berkeley, CA, 2004; 21–21.
18. Reiter MK, Rubin AD. Crowds: anonymity for Web transactions. *ACM Transactions Information System Security* 1998; **1**(1): 66–92.
19. Rebollo-Monedero D, Forné J, Pallarès E, Parra-Arnau J, Tripp C, Urquiza L, Aguilar M. On collaborative anonymous communications in lossy networks. *Security and Communication Network, Special Issue Secure. Completely Interconnected World* 2013 in press. Available from: http://dx.doi.org/10.1002/sec.793 [4 September 2013].
20. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, de Wolf PP. *Statistical Disclosure Control*, Survey Methodology. John Wiley & Sons: West Sussex, UK, 2012.
21. Sweeney L. Uniqueness of simple demographics in the U.S. population. *Tech. Rep. LIDAP-WP4*, Carnegie Mellon Univ., Sch. Comput. Sci., Data Priv. Lab., Pittsburgh, PA, 2000.
22. Samarati P. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering* 2001; **13**(6): 1010–1027.
23. Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering* 2002; **14**(1): 189–201.
24. Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogenerous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery* 2005; **11**(2): 195–212.
25. Domingo-Ferrer J, Martínez-Ballesté A, Mateo-Sanz JM, Sebé F. Efficient multivariate data-oriented microaggregation. *VLDB Journal* 2006; **15**(4): 355–369.
26. Rebollo-Monedero D, Forné J, Pallarès E, Parra-Arnau J. A modification of the Lloyd algorithm for k-anonymous quantization. *Information Sciences* 2013; **222**: 185–202. Available from: http://dx.doi.org/10.1016/j.ins.2012.08.022 [4 September 2013].
27. Rebollo-Monedero D, Forné J, Soriano M. An algorithm for k-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers. *Data & Knowledge Engineering* 2011; **70**(10): 892–921. Available from: http://dx.doi.org/10.1016/j.datak.2011.06.005.
28. Deng M. Privacy preserving content protection, *Ph.D. dissertation*, Katholieke Univ. Leuven, Dept. Elect. Eng. (ESAT), 2010.
29. Rebollo-Monedero D, Forné J. Optimal query forgery for private information retrieval. *IEEE Transactions on Information Theory* 2010; **56**(9): 4631–4642. Available from: http://dx.doi.org/10.1109/TIT.2010.2054471 [4 September 2013].
30. Parra-Arnau J, Rebollo-Monedero D, Forné J. A privacy-protecting architecture for recommendation systems via the suppression of ratings. *Journal: International Journal of Security and Its Applications (IJSIA), Science & Engineering Research Support Society (SERSS)* 2012; **6**(2): 61–80.
31. Rebollo-Monedero D, Forné J, Domingo-Ferrer J. Coprivate query profile obfuscation by means of opti-

mal query exchange between users. *Journal: IEEE Transactions on Dependable and Secure Computing, Special Issue on Data and Application Security (DAS)* 2012; **9**(5): 641–654. Available from: http://doi.ieeecomputersociety.org/10.1109/TDSC.2012.16 [4 September 2013].

32. DeMillo RA. *Foundations of Secure Computation* (Lipton RJ, Dobkin DP, Jones AK, eds). Academic Press, Inc.: Orlando, FL, 1978.
33. Mármol F, Sorge C, Petrlic R, Ugus O, Westhoff D, Pérez G. Privacy-enhanced architecture for smart metering. *International Journal of Information Security* 2013; **12**(2): 67–82.
34. Garcia FD, Jacobs B. Privacy-friendly energy-metering via homomorphic encryption, In *Proc. Int. Workshop Secur., Trust Mgmt. (STM)*, Athens, Greece, 2010; 226–238.
35. Bartoli A, Hernández-Serrano J, Soriano M, Dohler M, Kountouris A, Barthel D. Secure lossless aggregation for smart grid M2M networks, In *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Gaithersburg, MD, 2010; 333–338.
36. Rebollo-Monedero D, Parra-Arnau J, Diaz C, Forné J. On the measurement of privacy as an attacker's estimation error. *International Journal of Information Security* 2013; **12**(2): 129–149. Available from: http://dx.doi.org/10.1007/s10207-012-0182-5 [4 September 2013].
37. Cover TM, Thomas JA. *Elements of Information Theory*, Second. Wiley: New York, 2006.
38. Serjantov A, Danezis G. Towards an information theoretic metric for anonymity. In *Proc. Workshop Priv. Enhanc. Technol. (PET)*, Vol. 2482. Springer-Verlag: Berlin and Heidelberg, 2002; 41–53.
39. Rebollo-Monedero D, Forné J, Domingo-Ferrer J. From t-closeness-like privacy to postrandomization via information theory. *IEEE Transactions on Knowledge and Data Engineering* 2010; **22**(11): 1623–1636. Available from: http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190 [4 September 2013].
40. Parra-Arnau J, Rebollo-Monedero D, Forné J. Measuring the privacy of user profiles in personalized information systems. *Future Gen. Comput. Syst.* 2013. Available from: http://dx.doi.org/10.1016/j.future.2013.01.001 [4 September 2013], In press.