Using Differential Privacy for the Internet of Things

Carlos Rodrigo Gómez Rodríguez, Elena Gabriela Barrantes S.

Using Differential Privacy for the Internet of Things

Carlos Rodrigo Gómez Rodríguez , Elena Gabriela Barrantes S.

Universidad de Costa Rica, Ciudad Universitaria Rodrigo Facio, Montes de Oca, San José, Costa Rica

carlos.gomezrodriguez@ucr.ac.cr,
gabriela.barrantes@ecci.ucr.ac.cr

Abstract. In this paper we propose a hybrid privacy-protection model for the Internet of Things (IoT) with the ultimate purpose of balancing privacy restrictions and usability in data delivery services. Our model uses traditional de-identification methods (such as *k*-anonymity) under low-privacy requirements, but allows for the transmission of aggregate statistical results (calculated with a privacy-preserving method such as Differential Privacy) as an alternative if the privacy requirements exceed a threshold. We show a prototype implementation for this model, and present a small step-by-step example.
Keywords: Privacy Negotiation– Internet of Things - Differential Privacy

1 Introduction

The data collected in the context of the Internet of Things (IoT), is being incorporated in the business model of many companies, as it takes advantage of the millions of devices freely collecting and distributing data on a daily basis. CISCO states that in year 2015, 563 millions of new mobile devices and connections were added to the 7.3 billion that were already there in 2014 [1]. Data traffic reached 3.7 exabytes per month at the end of 2015, 1.6 times more than 2014, originating in 97 million of wearable devices that generated 15 petabytes of monthly traffic. It is projected that by 2020 the monthly mobile data traffic will reach 20.6 exabytes per month. All this ecosystem composed by people, smart devices, sensors, data collectors, data analyzers, predictors and applications might cause an inappropriate exposure of personal information that the user is not aware of [2]. One possibility is to allow data consumers and producers to negotiate privacy agreements in advance, and then use these agreements to try to satisfy, in the best possible way their requirements.

We are interested in the following scenario: IoT data producers (for example, residential users of smart meters) and IoT data consumers (for example, recommender or safety applications) negotiate a privacy agreement through a third trusted party (TTP), a hypothetical privacy broker. The broker will administer the aggregate full data-base of records generated by all data producers. The existence of such a third party is not a new concept [3] and it could work even on a decentralized environment [4].

The privacy broker will use some sort of negotiation protocol between the interested parties, and will record privacy agreements as rules between each user and each consumer. For example, producer requirements could be coded as permissions over some data attributes, and consumer requirements could be given by one or more information loss metrics. Producer requirements will always have a higher priority, given that a violation could expose sensitive data. Although there are one-to-one agreements between consumers and producers, when a consumer requests data corresponding to several producers, the calculation of a joint result that satisfies all individual producer privacy requirements is not trivial, and can cause severe information loss.

We propose a hybrid model where the broker has two privacy-preserving methods to define what to communicate to a requesting consumer: (1) a default, record-by-record descriptor method [5, 6] (also called syntactical de-identification [7]) that will be used when the satisfaction of producer requirements generates an information loss within the threshold required by the data consumer; and (2) an alternative ("statistical") method that will create a set of statistical descriptors for the original data [8]. These aggregate results are the ones that will be used to answer the consumer query if the information loss exceeds the agreed threshold. The statistical method has to be also privacy-preserving, such as Differential Privacy [9].

The motivation for the proposed model is the fact that data consumers are also an important part of the IoT ecosystem, so for cases when satisfying joint producer requirements result in a large information loss, our model could at least provide an aggregate result that might be useful to the consumer, instead of returning a useless set of records. For example, a recommender app might be able to continue functioning even in the presence of intermittent aggregate statistics instead of more granular data.

To further specify the model, we decided to use a privacy-negotiation environment such as the one proposed by Ukil et al in [2], and take some other elements from their model. The constraints resulting from the application of negotiated parameters to a given set of data to be published provide us with a natural source of varying requirements. For our proof of concept we use *k*-anonimity as the syntactic de-identification method [10], and Differential Privacy (DP)[6] as the statistical privacy-preserving method. Finally, as information loss metric we use the one proposed by Iyengar in 2002 [11].

The rest of the paper is organized as follows: Section 2 describes the proposed model; Section 3 presents the proof of concept prototype; Section 4 describes an IoT scenario for the model; and Section 5 presents two examples of prototype use. Finally, in Section 6 there is a discussion and possibilities for future work are outlined.

2 Proposed model

Following Ukil et al.[2], our scenario considers two actors: (1) the data producer, that is, the individual that is exposing information gathered automatically by devices like sensors; and (2) the data consumer, which requests gathered IoT data for analysis.

The actors communicate through a third party broker that stores consumer data and receives, evaluates, and answers requests for access, originating with data producers. The broker is also responsible for negotiating and reaching privacy agreements between producers and consumers. This third party must have global rules incorporating local laws and good practices. For example, the release of any personally identifiable information (PII) should be excluded from any privacy agreement by default. Furthermore, depending on the privacy-preserving release method the broker must use the settings most appropriate to defend consumer privacy. For example, the model in [2] uses the rules given in Lodha et al. [12] to define the type and degree of de-identification to be applied for the syntactic anonymization methods used.

The particular negotiation agreements between each producer and consumer could be coded as complex rules over data semantics, or be as simple as permissions given over data attributes, which is the approach used in [2]. Their rule representation is an access matrix that relates all the relevant attributes to specific data consumers using a single permission flag.

After privacy agreements have been set up, the high-level view of our model is that data consumers will query the broker for data belonging to a group of data producers. The broker will first check if the data requested could be provided under any global rules, and, if so, will query its aggregate producer database. The table resulting from this query will then be syntactically anonymized [7] (for example, with *k*-anonymity) using the particular rules triggered by the consumers that are part of this particular response. Once the anonymization is complete, an information loss metric would be calculated. If the data loss is within the threshold negotiated with the consumer, the anonymized table is sent. Otherwise, a set of aggregate statistical descriptors is calculated from the original resulting table using a privacy-preserving method (for example Differential Privacy), and the set of statistical descriptors is sent as response to the data consumer. This way, some useful information is still returned to the consumer.

3 Proof of concept prototype

As proof of concept for the model we implemented a prototype in which we use *k*-anonymization [10] as the syntactical model, and Differential Privacy (DP) [9] as the statistical anonymity model. In particular, for *k*-anonymization, we used a simple implementation locally developed, and for DP, we used PINQ, the McSherry implementation [13]. As the information loss metric we used the one proposed by Iyengar [11] because it works on generalization hierarchies.

The prototype does not currently collect data from producers and consumers, and neither executes negotiation processes. Instead, it feeds on a given dataset, a given agreement set, and reads the parameters for the global rules. Given the restrictions imposed by the McSherry implementation, the dataset is stored in a Microsoft SQL Server. The logical architecture for the prototype is shown in Figure 1.

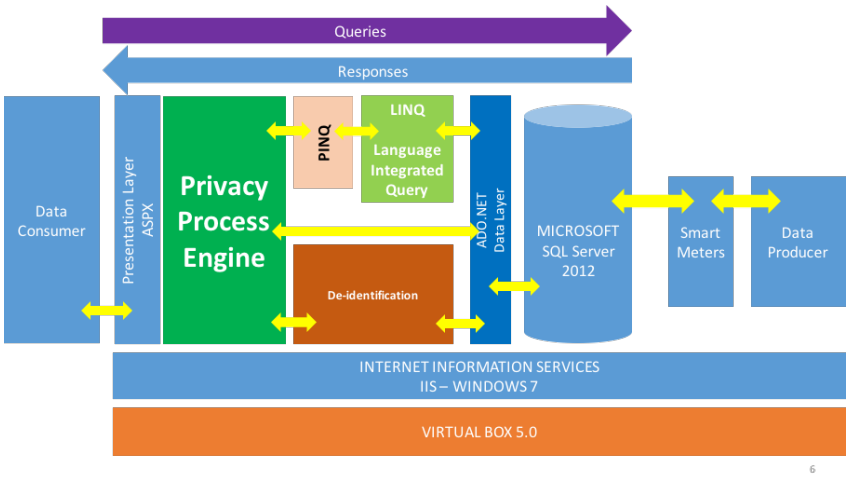

Figure 1. Prototype logical architecture

To represent privacy agreements between consumers and producers we use a matrix similar to the one in [2], but the semantics of the permission flag is taken as a permission to include in a syntactical result. That is, if a producer sets the flag to 0 on a specific attribute for a given consumer, that attribute for that producer could not be included in an anonymized table independently of the degree of k. The matrix is filled using a simple user interface, as the negotiation process in itself is outside the scope of this paper. Given the matrix semantic, the information loss is calculated in two phases: the first one after eliminating all records of users that were not willing to share an attribute required by the query, and the second one after applying k-anonymization to the remaining records. For the first one the information loss is calculated as the percentage of deleted records with regard to the total returned by the query. In the second phase, Iyengar´s information loss is applied as if the remaining records were the original dataset. Somewhat arbitrarily, both losses are added to calculate total loss before deciding what to return to the consumer.

The global privacy rules included were: (1) The elimination of PIIs; (2) A given classification of non-PII attributes in quasi-identifiers and sensitive data; (3) A choice of k by the prototype operator for *k*-anonymization; (4) A choice of DP parameters by the operator (epsilon and privacy budget); and (5) A choice of information loss threshold for each consumer. Rule 1 is executed manually. For rule 2, the respective columns

are marked in advance. Values for rules 3, 4, and 5 are read through several interfaces.

The interfaces were implemented in ASP.NET, using Microsoft Internet Information Services 7 on Microsoft Windows 7 Professional and running on an Oracle Virtual Box engine version 5.0.16. The whole process is encapsulated in the same ASP.NET application, using the C# class library as well and invoking PINQ 0.1.0.0 and LINQ 4.0.0.0 libraries and ADO.NET 4 to finally access the MS SQL Server 2008 R2

The process followed by our proof of concept is depicted in Figure 2. It starts with a data consumer making a data request (1). This request generates a query, and the data resulting from the application of the query is then analyzed using the attribute access matrix, and filtering all the records that must be removed from the query as determined by the application of the privacy rules generated during the negotiation (2). After this reduction, the data loss is computed as a percentage of records lost with respect to the original query answer (3). If the data loss threshold is reached (4), a set of predefined statistical descriptors are calculated on the original query response using a DP implementation (5) and delivered to the Data Consumer as response (11). Otherwise (4), the filtered response is de-identified using –again- the applicable privacy-preserving rules (6) and the data loss (Iyengar) for this process is calculated (7) added to the loss of the first filtering process (8), and compared against the data loss threshold (9). If this integrated loss estimation reaches the threshold, the set of statistical descriptors of the original query is calculated using a DP implementation (5) and delivered to the Data Consumer as response (10).

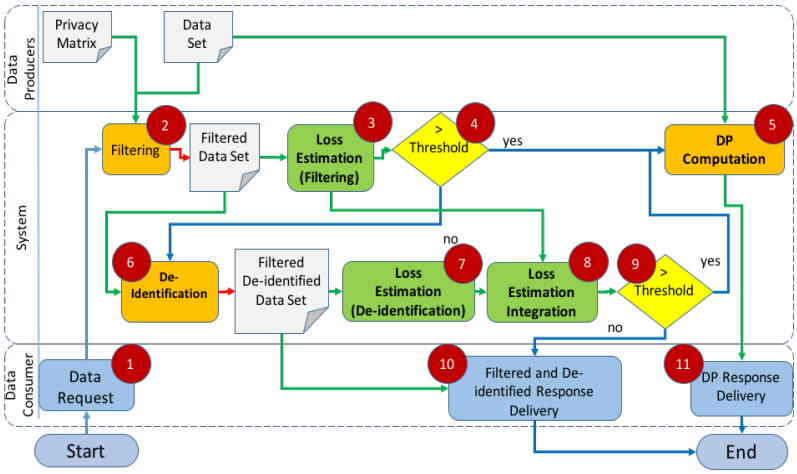

Figure 2. Protection method selection process.

To determine the feasibility of the proposed model, we describe a possible IoT application scenario, then choose a real dataset that fits the proposed scenario, implement a prototype and show its use for a simple, and then a larger case.

4 An IoT scenario for the model

The composite data collected from IoT sensors and devices in different households has the potential to create unexpected (and unwanted) disclosures. For example, Greveler et al. [14], presents a case where energy consumption data transmitted to the service and application support layer, allows intrusive identification about devices located inside the households of data producers (TV set, refrigerator, toaster, oven). For example, depending on the frequency of analysis of the electricity usage profile (for instance at a 0.5−1s sample rate) the data can reveal what channel a TV set in the household is displaying. Therefore, something that at first sight looks inoffensive, like taking part into an electricity consumption survey, could reveal information that the house owner might not have wanted to share. Despite the huge importance and knowledge that can be generated understanding consumption patterns for power at home, data collected from these devices can also be used as a surveillance tool.

Given the findings in [14], the scenario in which we situate the model is electricity consumption data. Each sensor sends its data to servers over the Internet, which in turn uses the data to show users their current energy consumption and estimate the monthly bill on a real-time basis. The data producer is the application that reads the sensor and sends the data it produces over the Internet. The data owner, and by extension the producer, is the user that installed the sensor: possibly the inhabitant of the house. The value for the producer of sending the data is the possibility to use the energy resources more efficiently or to get access to new technologies that could improve living. To execute calculations, the application requires permission to sense the data of all appliances in the household. Analysis of the sensed data could be used to recommend improvements to other users of the system. Furthermore, the application will need to process information gathered from different households on a particular region. The data could be very detailed and enable the profiling of people living in the house.

This paper only intends to provide evidence that the proposed model could be instantiated in a working prototype, but not yet investigate its usefulness in real environments. However, given that next step, to test the prototype, we used the collection published by the Department of Industry, Innovation and Science of Australian Government (DIIS) under the project name “Sample household electricity time of use data”, that can be accessed in https://data.gov.au/dataset/sample-household-electricity-time-of-use-data. The objective of DIIS for this project was to ensure that households have enough information to improve the use of electricity network. The dataset contains 10,828,120 energy readings for 808 several types of users.

The collection includes electricity use (in kwh) measured approximately every 30 minutes for a year using a smart meter to collect this information. It also provides basic demographic information, that could be used to infer customers. Customer demographic information includes customer ID (that allows to create a relationship between consumption and customer), address region, income range, appliances in house, people in house, and others. As mentioned before, with just the electricity consumption, it can be determined what appliances are being used in the household.

5 Examples of model application

The first example is presented to provide a step-by-step description of the model application. Table 1 shows a 18-record hypothetical result table (we proceed directly to the processing after the Privacy Broker has extracted the raw data corresponding to the consumer request).

Suppose we also have a previously negotiated matrix of data access for a given data consumer from each of the data providers. The matrix contains a column for each quasi-identifier (QI), and a value of 0 or 1 indicating whether the user is willing to share the QI in a release or not. The matrix is shown in Table 2.

Table 1. Example result table.

Cust. Id	County	Sex	Age	Reading Value	Outlet
1	101	M	47	393	TV
2	101	F	42	423	TV
3	101	F	55	231	TV
4	101	M	49	554	TV
5	103	M	31	334	TV
6	102	M	42	676	TV
7	103	F	29	445	TV
8	101	F	25	332	TV
9	102	M	43	553	TV
10	103	M	44	445	TV
11	103	F	29	765	TV
12	102	F	57	432	TV
13	101	F	47	113	TV
14	101	M	31	455	TV
15	103	M	57	321	TV
16	102	M	33	334	TV
17	103	F	34	654	TV
18	103	F	54	442	TV

Table 2. Example privacy matrix

Customer Id	County	Sex	Age
1	1	1	1
2	1	1	1
3	1	0	1
4	0	1	0
5	1	1	1
6	1	1	0
7	1	1	1
8	0	0	0
9	1	1	1
10	1	1	1
11	1	1	1
12	1	0	0
13	0	0	1
14	1	1	1
15	1	1	0
16	1	1	1
17	1	1	1
18	1	0	0

The resulting table has to be filtered against the preset values from the privacy matrix shown in Table 2. In this case, a total of 8 records must be removed from the result table, since those data producers will not allow to show information for at least one of the requested columns. The records that must be removed are shown with a shadow on them.

The first information loss calculation is just the number of removed records divided by the total number of records that should be considered. For this particular case the loss for the first pass is 0.444. For a 0.5 threshold, the de-identification process has to proceed.

To provide k=2 k-anonymity the quasi-identifiers were categorized as follows: County: 100 to 101, and 102 to 103; Sex: * (suppressed); Age (years): less than 32, 32 to less than 40, 40 to less than 48, 48 to less than 56, 56 to less than 64, and 64 and more. Any combination of these quasi-identifiers ranges provides 2 or more records, accomplishing the *k*=2 requirement. The information loss metric was calculated for this as 0.44. Table 3 shows the *2*-anonymized data.

Table 3. Result of filtering the data in Table 1 given the restrictions in Table 2. Using k=2 k-anonymity. ID is an identifier value, QI are quasi-identifiers.

ID	QI	QI	QI	DATA	DATA
Customer Id	County	Sex	Age	Reading_Value	Outlet
*	[100,101]	*	[40-48[	393	TV
*	[100,101]	*	[40-48[	423	TV
*	[100,101]	*	[0-32[	334	TV
*	[102,103]	*	[0-32[	445	TV
*	[102,103]	*	[40-48[	553	TV
*	[102,103]	*	[40-48[	445	TV
*	[102,103]	*	[0-32[	765	TV
*	[100,101]	*	[0-32[	455	TV
*	[102,103]	*	[32-40[	334	TV
*	[102,103]	*	[32-40[	654	TV

Adding both information loss metrics, it results in a value of 0.95, exceeding the pre-set threshold of 0.5. So in this particular case, the privacy selection process had to deliver DP aggregated functions to the data consumer. For the purpose of showing the noise induced by DP, six consecutive answers are shown in Table 4.

Table 4. Results from DP computation for 6 identical requests performed by the data consumer. Data Request = Information from all records that County equals {101, 102,103} and Outlet='TV'

•	Request Nbr	1	2	3	4	5	6	Actual Data
•	**Record Count**	16.00	20.00	15.00	16.00	19.00	15.00	**18.00**
Age	**Percentile 25**	28.00	26.00	33.00	30.00	26.00	27.00	**31.00**
	Percentile 50	43.50	42.50	41.50	47.50	43.50	37.50	**42.50**
	Percentile 75	48.25	53.25	50.25	48.25	50.25	45.25	**50.25**
Reading Value	**Percentile 25**	330.50	331.50	335.50	330.50	335.50	332.50	**333.50**
	Percentile 50	437.00	436.00	435.00	439.00	435.00	438.00	**333.50**
	Percentile 75	554.25	552.25	551.25	554.25	556.25	551.25	**333.50**
•	**Sum**	7,794.00	7,545.00	7,068.00	6,945.00	6,481.00	5,787.00	**7,902.00**

As it could be seen, instead of delivering a nearly useless de-identified dataset to the consumer, a set of statistical descriptors was provided, partially preserving utility for a hypothetical consumer application.

For the second example, we chose to use the energy readings for the retail/domestic users of the DIIS database (a total of 222 users and 8,908,454 registers), with the attributes Customer Identification, Region Name, Reading Time, Appliance Name, and Reading Value. An example record is (10036802, CANTERBURY, 2013-11-14 03:37:59, TV, 202.242). The DIIS dataset information is published in CSV format, and the data was imported to a Microsoft SQL Server for the prototype.

Attribute Customer Identification is a direct identifier, so it is not included in the negotiation, and it is automatically suppressed from any result. Attributes Reading Time and Reading Value are assumed to be the sensitive values, and Region Name and Appliance Name are the ones taken to be the quasi-identifiers and therefore included in the negotiation.

As an approximation to the negotiation values reflected by the matrix, we used the classification of privacy concern given by Humphrey Taylor based on a telephone survey by Harris Poll, conducted on a USA nationwide cross section of 1,010 adults [15]. Taylor concludes that 26% of respondents were "privacy fundamentalists", 10% were "privacy unconcerned", and 64% were "privacy pragmatists". Based on this classification, around 26% of the 222 users (58) were randomly assigned with a negative decision on the sharing of both quasi-identifiers, around 10% of the users had a positive sharing decision on both quasi-identifiers (22), and the remaining approximately 64% of the population (142 users) were assigned random decision values for the quasi-identifiers (all positive, one of them positive, both negative).

For the test, we assumed a data consumer query for the electrical readings of 22 users (10%) randomly chosen users in the week between the 8 and 15 of February of 2014. The privacy matrix of each of those users revealed 8 users with a do-not-share preference, 4 asserting a use-everything preference, and 10 users that allowed access to one attribute but not the other. In this test the filtering interprets the matrix as "do not communicate anything from users with both attributes hidden", so we remove all the records belonging to the do-not-share users which correspond to 12,863 out of a total of 23,610 readings for this query. Therefore, the first-step total data loss was slightly over 35%, which is still within the preset threshold. Afterwards, a k=60 anonymization was performed, and the loss metric (LM) was calculated as 0.118. Combining this with the first-step, the combined loss is still below 50% so we communicate the anonymized results to the data consumer.

6 Discussion and Future Work

We have proposed a hybrid model to deliver data with privacy protections to consumers, and presented a particular implementation just to test its feasibility. Both the model and the prototype provide us with a flexibility in the semantic of the negotiation, the application of either semantic or statistical anonymization.

Taking into account the high volume of data generated by IoT, and that this data could have traces of private information that producers are not aware of, it is important to provide privacy protection. Privacy negotiation might provide usable methods of accomplishing this goal. To satisfy not only data producer but also data consumer requirements, in this paper DP is added to the negotiation model to be able to provide some utility in constrained settings, without sacrificing privacy. An implementation of this design was shown.

Extending DP implementations like PINQ could help to adhere this privacy protection layers to different contexts. So far, PINQ has few aggregated computations (count, sum, average, statistic order and median), and even when DP seems to have increased its application in different scenarios recently, Microsoft left behind new implementations for PINQ, and even when current version 0.1 is fully functional, it lacks of many other functions to provide even more utility to the data consumers. It was last released back in 2009, but it is still downloadable from the Microsoft official web site and it is supported in Windows 10, their latest operating system.

It would be very valuable if the attribute two-dimensional matrix, could be improved to a multidimensional matrix that can represent not only data fields and consumers with binary resolutions, but also consider time, purpose of queries, data age, etc. Also, settings could be more than just 1's and 0's, but adding different levels, for instance, k values (k-anonymity parameter) and/or epsilon and privacy budget (DP parameters) values.

There are many complexities involved in the application of hybrid models. It is not known, for example, if –in the long run- the intermittent application of each would lead to unintended disclosures. This is a well-known problem for any hybrid model, and needs to be studied in this case.

A less theoretical, but equally important issue is to be able to establish if the data that a statistical model returns to the data consumer would be useful at all, that is, would reduce the negative impact of this privacy protection. Acceptance from reasonable data consumers is important for the dissemination of these models.

On a different level, for the test case we made many assumptions, so it is necessary to experimentally explore what would be the effect of different queries on the threshold. For example, what percentage of the time the prototype returns a *k*-anonymized da-

taset instead of a DP summary. This requires a careful tuning of the k and the threshold to reflect more realistic values.

7 Acknowledgements

This work was partially supported by the Programa de Posgrado en Computación e Informática (PCI), the Escuela de Ciencias de la Computación e Informática (ECCI), the Centro de Investigaciones en Tecnologías de la Información y Comunicación (CITIC), and the Sistema de Estudios de Posgrado (SEP) all at the Universidad de Costa Rica (UCR), the Ministerio de Ciencia, Tecnología y Telecomunicaciones (MICITT), and by the Consejo Nacional para Investigaciones Científicas y Tecnológicas (CONICIT) of the Government of Costa Rica.

References

1. Cisco: Cisco Visual Networking Index : Global Mobile Data Traffic Forecast Update, 2015 – 2020. White Paper, Cisco (2016)
2. Ukil, A., Bandyopadhyay, S., Joseph, J., Banahatti, V., Lodha S.: Negotiation-based privacy preservation scheme in internet of things platform. In: Proceedings of the First International Conference on Security of Internet of Things, pp. 75-84. ACM New York (2012)
3. Yusuf, S.: Survey of Publish Subscribe Communication System. Technical Report, Department of Computer Science, Kent State University (2004)
4. Zyskind, G., Nathan, O., Pentland A.: Decentralizing Privacy: Using Blockchain to Protect Personal Data. In: Binnig C., Dageville, B. (eds.) Proceedings of the 2015 IEEE Security and Privacy Workshops (SPW '15), pp 180-184. IEEE Computer Society Washington, DC (2015)
5. Cormode, G., Srivastava, D.: Anonymized data. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (SIGMOD '09), pp. 1015-1018. ACM, New York (2009)
6. Angiuli, O., Blitzstein, J. , Waldo, J.: How to de-identify your data. Queue, vol. 13, no. 8. ACM New York (2015)
7. Clifton, C., Tassa, T.: On syntactic anonymity and differential privacy. Transactions on Data Privacy, vol. 6, no. 2, 161–183. IIIA-CSIC Catalonia (2013)
8. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03), pp. 2012-210. ACM New York (2003)
9. Dwork, C.: Differential privacy. In: In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
10. Sweeney, L.: k-anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, 557–570. World Scientific Publishing Co., Inc. River Edge, NJ (2002)
11. Iyengar, V.: Transforming data to satisfy privacy constraints. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02), pp. 279-288. ACM New York (2002)
12. Lodha, S., Thomas, D.: Probabilistic anonymity. In: Francesco Bonchi, Elena Ferrari, Bradley Malin, and Yücel Saygin (eds.) Privacy, Security, and Trust in KDD. LNCS, vol. 4890, pp. 56-79. Springer, Heidelberg (2008)

13. McSherry, F.: Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, pp. 19-30. ACM New York (2009).
14. Greveler, U., Glösekötterz, P., Justusy, B., Loehr, D.: Multimedia content identification through smart meter power usage profiles. In: Proceedings of the International Conference on Information and Knowledge Engineering (IKE), pp. 1-8. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2012)
15. Taylor, H.: Most people are "privacy pragmatists" who, while concerned about privacy, will sometimes trade it off for other benefits. Harris Poll. 17, pp. 1-6 (2003)