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 modification to a privacy negotiation model for the Internet of Things (IoT) that uses de-identification by traditional methods only and incorporates Differential Privacy as an alternative to balance privacy protection and data utility. We then analyze the implementation of a Proof of Concept of this modification, through the implementation of a process that selects whether to deliver a de-identified dataset or aggregated statistical values, according to a decision flow using a electricity consumption database published in Australia, and perform a step-by-step description of this process. Finally we offer a conclusion and future work section.
Keywords: Privacy Negotiation– Internet of 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[1] per month at the end of 2015, 1.6 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 to deal with this problem is to allow data consumers and producers to negotiate privacy agreements in advance. However, most data negotiation models deal with either traditional de-identification techniques [3,4] or with statistical databases being consulted interactively [5]. The purpose of this paper is to explore the possibility of integrating two such models attempting to satisfy certain privacy requirements but at the same time provide with some important data that was censored because of conflicting privacy requirements. Specifically, we use as the base privacy-negotiation model the one proposed by Ukil et al in [2] and

[1] One exabyte is equivalent to one billion gigabytes, and one thousand petabytes.

incorporate Differential Privacy (DP) [6] to this model. We present the application of the modified model to a particular case in IoT and analyze the implications of the proposed changes for the privacy of the device user.

The rest of the paper is organized as follows: Section 2 describes how a privacy negotiation could be set up; Section 3 briefly deals with Differential Privacy and the implementation used in this work; in Section 4 a description of how DP was incorporated in the negotiation model is given; Section 5 proposes a proof-of-concept case, while Section 6 shows a small example step by step. Finally, in Section 7 the conclusions and future work are outlined.

2 Negotiating access to attributes

This Section deals with a series of concepts that are needed to properly explain the privacy-compliance modifications proposed using the base negotiation model [2]. We first situate the intervention within the ITU-T reference model, then define the stakeholders in the process, and finally explain how the negotiation model proposed by Ukil et al. [2] sets the privacy parameters.

Based on the ITU-T reference model [7], an IoT architecture can be referenced as multi-layered components with different functions. The device layer, thru sensors and devices gathers data from the users, data producers, and using different communication protocols in the network layer consolidates and aggregates data into the service and application support layer, that includes the event processing and analytics functions. Finally, the application layer exposes information to stakeholders and application users. This architecture model is depicted in Figure 1. Applications can also be mapped to different IoT initiatives like Smart Cities, Smart Transport, Smart Buildings, Smart Energy, Smart Industry, Smart Health, Smart Living among others. The specific example of IoT used for this paper is the electricity consumption analysis, where smart meters gathers information from households and provides information to investigators and data analysts to develop new technologies to reduce or improve electricity.

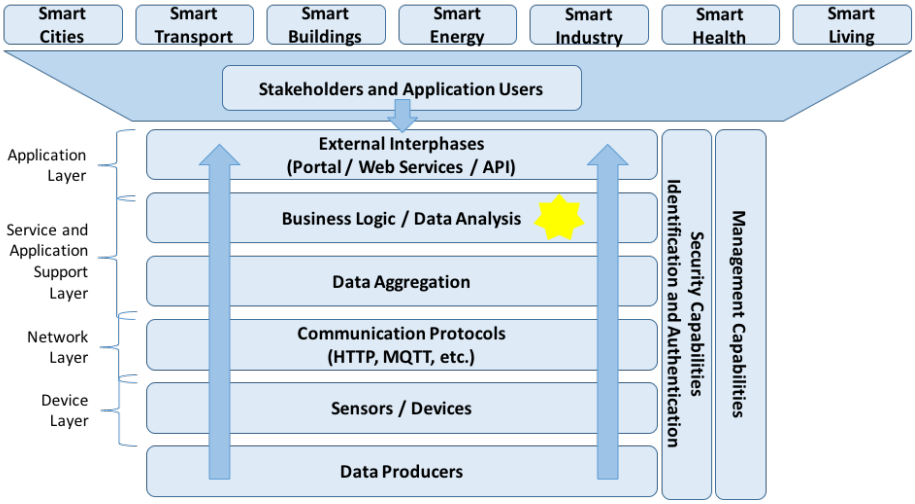


Figure 1. IoT Layered Architecture. Adapted from the ITU-T Reference Model [7]

The data that is gathered from different households carries more information that it is expected. According to Greveler et al. [3], energy consumption data transmitted to the applications located in the Event and Analytics layer in the IoT model referenced above, allow intrusive identification about devices located inside the households of data producers (e. g. 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 a electricity consumption survey, could reveal information that the house owner might not have wanted to share.

A privacy negotiation model 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.

Ukil et al. [2] faced the problem of utility-aware privacy preservation in IoT proposing a model where the data producer and the data consumer communicated through a module that integrated the data consumer requests for access, with the data producer

To define and negotiate privacy, there should be a discussion about the attributes (columns) in the database being accessed. According to the base model, the privacy negotiation module has rules incorporating local laws and basic de-identification procedures [2]. This means that all personally identifiable information (PII) is excluded from the agreement by default. With regard to quasi-identifiers and sensitive information, the base model applies the rules given in Lodha et al. [8] to define the type and degree of de-identification to be applied.

The negotiation process proceeds in two asynchronous steps. First the data producer interacts with the module to set an access matrix that relates all the relevant attributes to specific data consumers (Figure 2). The access given is either "as-is" (no masking) or as de-identified data (masking). The second step involves notifying the consumer that the data will be shared at the minimum disclosure policy that results from the producers included in any specific request. If the consumer agrees, then its privacy-preserving rule is included in the set.

Ukil et al [2] propose to use SafeMask [9] to implement the base model. SafeMask is an interactive de-identification tool, which intercepts data traffic and masks sensitive contents according to an established privacy policy, without changing the code of the application producing the data, working as a protection layer over the application itself.

Data Field \ Data Consumer	Federal Tax Agency	Local Government	Public University Research Department	Utilities and Services Company
Service Type	1	1	0	1
County	0	0	0	1
Total Occupants	1	0	1	1
Children Younger than 11	0	0	0	1
Children 11 to 17	0	1	0	1
Occupants older than 70	1	1	0	1
Household Income	0	0	0	1
Renting	1	0	1	1

Figure 2. Negotiation matrix of data producer. Adapted from Ukil et at. (2012).

Unfortunately, de-identification techniques, such as the ones used by SafeMask (e.g. k-anonymity) have serious problems (see Narayanan et al in [10], for example). For the volume of data generated by IoT devices, and given that consumers will be constantly aggregating data generated by the queries, de-identification may be rendered unfeasible, useless to defend the producer, or the data may be de-identified to such level that it will be useless to the consumer. This is why we propose a modification of the Negotiation Module based on Differential Privacy (DP) complementing de-identification, as a possible solution.

3 Differential Privacy (DP)

Differential Privacy (DP) was originally proposed by Dwork in [6]. It refers to a privacy goal requirement that must be satisfied by algorithms (or mechanisms) that describe a given data set using disturbed statistical values like an average or the count of elements in the data set. This goal is basically set by the epsilon (ε) value, that is the

difference between the probabilities of receiving the same result from a randomized algorithm against two different data sets that differ in just one record, so it can be guaranteed that a re-identification was not caused by the participation in a data set. A smaller value of ε represents stronger privacy, and values are usually set between 0 and 1, like 0.1 or ln(2), for instance.

McSherry in [11] described a practical implementation of differential privacy with C# integrated with the LINQ declarative language called PINQ. Elingsson et al. in [12] also introduced RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response) as a widely-applicable, practical mechanism that provides strong privacy guarantees combined with high utility, yet is not founded on the use of trusted third parties.

Because of its ease for integration into a .NET web application and specially to focus in the proposed model implementation instead of the learning curve of other Differential Privacy implementations, it was selected PINQ as the DP provider for this proof of concept, but other solutions could be considered in later implementations.

PINQ has basic implementations of four aggregations: a) Count, that returns the number of records in the dataset; b) Sum, that returns the addition of values in the dataset; c) Median that returns the 50^{th} percentile; and, d) Average, that returns the arithmetic mean of values in the dataset. Nevertheless, PINQ does not restrict programmers to that fixed set of computations or functions, it allows to add new functions if they are required. For the proof of concept analyzed in this research, we are using this set of aggregations, just to ensure feasibility on the analysis scope.

The noise that guarantees privacy in computations is added based in a randomized computation yielding nearly identical distributions obtained from nearly identical inputs. Added noise is symmetric exponential noise scaled by the parameter ε (epsilon parameter) used to denote the privacy level guarantee as well (ε-privacy). Count and Average uses Laplace scaled by ε as the symmetric exponential distribution, which has the property that the probability of any outcome decreases by a factor of e^{ε} with each unit step away from its mean. Mean is translated by one unit scales the probability of any output by a multiplicative factor of e^{ε} or less. Average and Median computations are implemented using the exponential mechanism of McSherry and Talwar [13]. Average's accuracy is calculated as $2/\varepsilon$ divided by the data set's size (record count).

4 Adding Differential Privacy to the model

The assumption in this process is that the negotiation process itself is a black box that results in a series of de-identification privacy-preserving rules and an access matrix for each data producer regarding possible data consumers. This work does not examine the methods to accomplish this. The modifications are done to the process that

generates the response to a particular query from the data producer (with some pre-negotiated rules already in place). A data loss threshold is added to determine whether to deliver (a) a de-identified dataset (as the base model proposes) or (b) a set of available statistical descriptors for the data obtained from a DP implementation.

The process followed by the modified privacy negotiation module is depicted in Figure 3. 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 (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 for this process is calculated (7) integrated with 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).

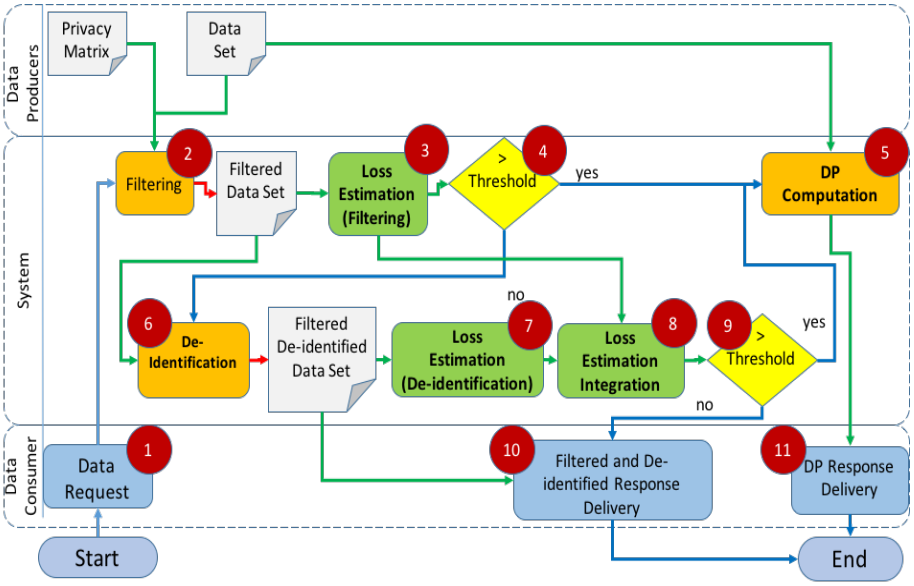


Figure 3. Protection method selection process, accordingly to the Query Analysis result.

The information loss computation that occurs in (3), (7), and (8), could be achieved through a variety of methods [14–22]). The one chosen for this implementation was the one by Nergiz et al. [22] because it works on generalization hierarchies, the technique used to de-identify quasi-identifiers in the data set.

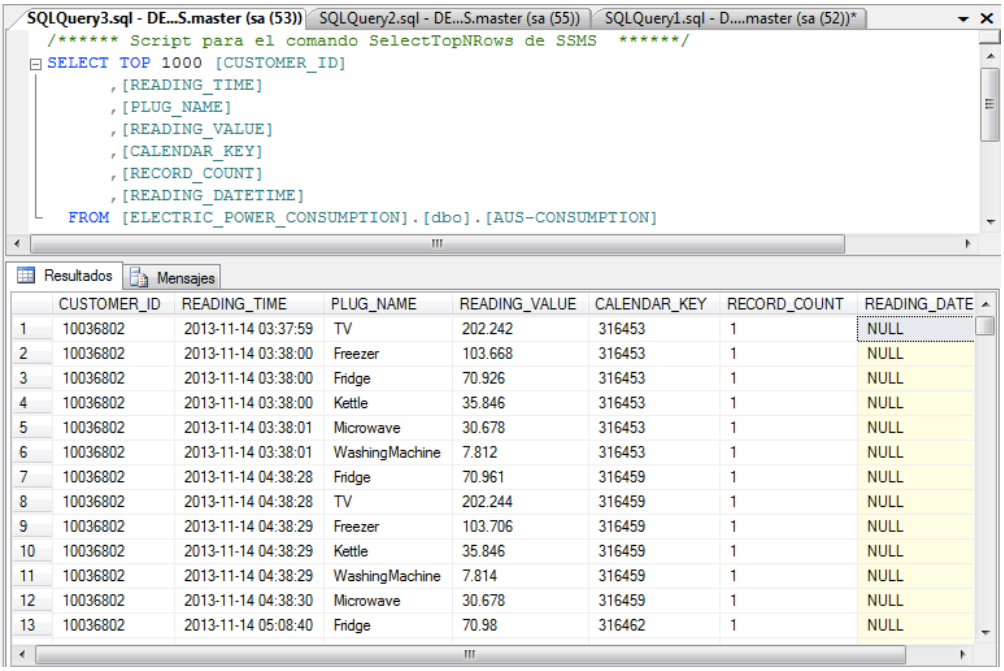
In this modified model, there is always a delivery of possibly useful information to the data consumer, even if it is in the form of aggregates, if the application of the rules for de-identification damaged the utility of the set beyond the agreed threshold. However, because of DP properties, the privacy guarantees to the data producers involved in the query are still maintained.

To briefly assess the soundness of the modified model, we will describe a case using electricity consumption in the next Section.

5 Application to a case

The application that will be used to set a proof of concept works with electricity consumption, where each sensor sends its data to servers over the Internet that 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. However, analysis of this sensed data will be used to recommend other users of the system. Furthermore, the application will need to process information gathered from different households on a particular region. Data will be very detailed and it could enable the profiling of people living in the house. 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 [23]. In most cases the data is not encrypted, and represents a risk for the producer of the data.

The data that was used for this case is 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. Figure 4 shows a screenshot of access to this collection.



	CUSTOMER_ID	READING_TIME	PLUG_NAME	READING_VALUE	CALENDAR_KEY	RECORD_COUNT	READING_DATE
1	10036802	2013-11-14 03:37:59	TV	202.242	316453	1	NULL
2	10036802	2013-11-14 03:38:00	Freezer	103.668	316453	1	NULL
3	10036802	2013-11-14 03:38:00	Fridge	70.926	316453	1	NULL
4	10036802	2013-11-14 03:38:00	Kettle	35.846	316453	1	NULL
5	10036802	2013-11-14 03:38:01	Microwave	30.678	316453	1	NULL
6	10036802	2013-11-14 03:38:01	WashingMachine	7.812	316453	1	NULL
7	10036802	2013-11-14 04:38:28	Fridge	70.961	316459	1	NULL
8	10036802	2013-11-14 04:38:28	TV	202.244	316459	1	NULL
9	10036802	2013-11-14 04:38:29	Freezer	103.706	316459	1	NULL
10	10036802	2013-11-14 04:38:29	Kettle	35.846	316459	1	NULL
11	10036802	2013-11-14 04:38:29	WashingMachine	7.814	316459	1	NULL
12	10036802	2013-11-14 04:38:30	Microwave	30.678	316459	1	NULL
13	10036802	2013-11-14 05:08:40	Fridge	70.98	316462	1	NULL

Figure 4. Partial view of electricity consumption database published by Department of Industry, Innovation and Science of Australian Government.

The collection includes electricity use (in kwh) measured each 30 minutes for a year approximately using a smart meter to collect this information. It also provides basic demographic information based on a survey, with information that could be used to infer customers. This information is published in CSV files, that were imported to a Microsoft SQL Server Express 2008 R2, and has 10,828,120 records that include customer Id, reading time, collection meter and date time stamp. 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.

The web application would be the one that the data consumer will be using. It has a user interface that allows the user to query thr database offering the four aggregate computations implemented in the DP solution (PINQ). For the limited purposes of this paper, the interface allows to set the epsilon parameter, the privacy budget and the data set size before loading data and showing calculations through an output window. The data loss threshold is set to 50% and a standard k value is set to 5 (for k-anonymity), but those parameters could be set as part of the configuration process for a usable implementation of this model. The user interface is depicted in Figure 5.

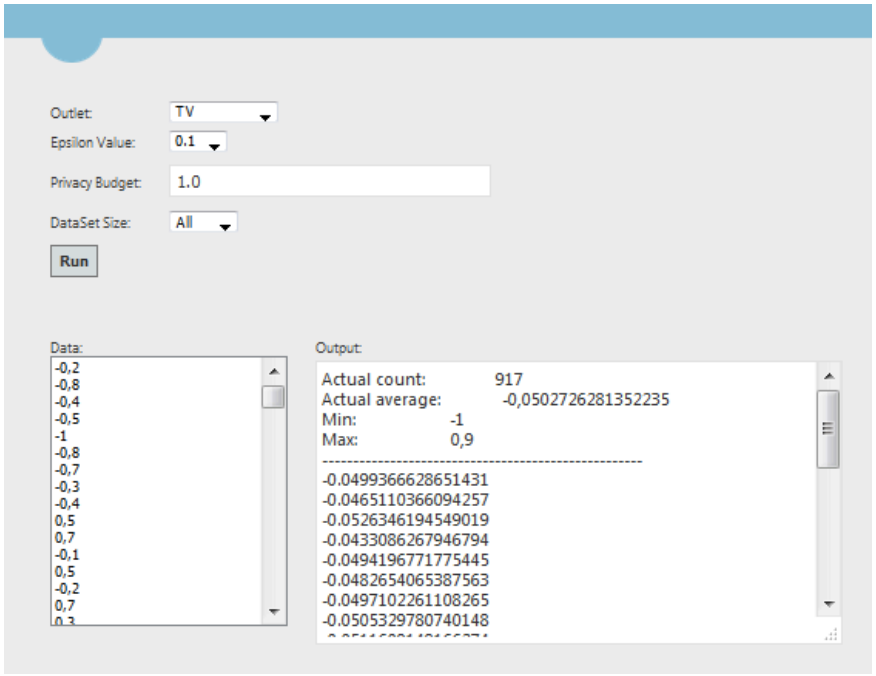


Figure 5. Modified model user interface.

This form is part of the presentation layer, implemented in ASP.NET presentation technology, using Microsoft Internet Information Services 7 on Microsoft Windows 7 Professional and running on an Oracle Virtual Box engine version 5.0.16. In the same ASP.NET application there is encapsulated the proposed Privacy Process Engine, using 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 Express data base management system. The logical architecture of the modified model is shown in Figure 6.

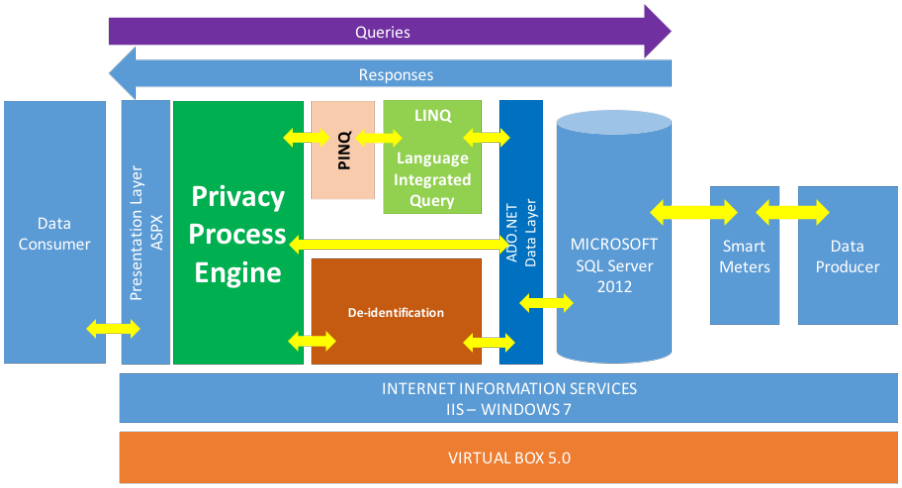


Figure 6. Logical Architecture for the modified model.

6 Running the Model

Using a small data set, just to be able to describe the full process, Table 1 shows the records for the full dataset that satisfies the data consumer query: "Provide County, Sex, Age, Reading_Value from all records that County equals {101, 102, 103} and Outlet='TV'" for a total of 18 records.

Table 1. Data set used to describe the full process.

Customer 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

When this record set is first filtered against the pre-set values from the privacy matrix, a total of 8 records must be removed from this data set to satisfy it since the related data producers will not allow to show information for at least one of the requested columns. Those records that must be removed are shown with a shadow on them. The privacy matrix is shown in Table 2.

Table 2. Privacy matrix for Data Consumer 2.

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 first information loss calculation is 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 it means that there is only chance to lose less than 0.056 in the de-identification process to provide the filtered and de-identified dataset. Otherwise, DP will be provided as response to the query. To choose DP at this time would reduce the processing effort of de-identifying the filtered data set. Therefore, the de-identification process has to proceed.

To provide k=2 k-anonymity the quasi-identifiers were categorized as follows: County: [100,101] and [102,103]; Sex: * (suppressed); Age (years): [less than 32[, [32-40[, [40,48[, [48,56[, [56-64[and [64 and more[. Any combination of these quasi-identifiers ranges provides 0 records, or 2 or more records, accomplishing the k=2 requirement. The information loss metric was calculated for this step using those ranges and was estimated for the 10 records in the filtered data set. Its value was 0.44. It means that 44% of information was lost using this de-identification requirement. Table 3 shows how filtered and de-identified data looks.

Table 3. Filtered and de-identified data set that guaranties k=2 k-anonymity. ID is a identifier value, QI a Quasi-identifier and Data is offered raw

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 a value of 0.95, exceeding the preset 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 the 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.

7 Conclusions and Future Work

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 still downloadable from Microsoft official web site and supported in Windows 10, their last operating system.

It would be very valuable if attribute matrix, which is actually a two-dimensional, 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.

The testing of this model in a formal experiment, could help to evaluate the feasibility and accuracy in both objectives: keeping privacy in different contexts for data producers; and keeping data utility on data consumer queries.

8 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 , 2010 – 2015, Growth Lakel. (2016). http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html.

[2] A. Ukil, S. Bandyopadhyay, J. Joseph, V. Banahatti, S. Lodha, Negotiation-based privacy preservation scheme in Internet of Things platform, Proc. First Int. Conf. Secur. Internet Things. (2012) 75–84. doi:10.1145/2490428.2490439.

[3] G. Cormode, D. Srivastava, Anonymized data, Proc. 35th SIGMOD Int. Conf.

Manag. Data - SIGMOD '09. (2009) 1015. doi:10.1145/1559845.1559968.
[4] O. Angiuli, J. Blitzstein, J. Waldo, How to de-identify your data, Queue. 13 (2015) 20–39. doi:10.1145/2838344.2838930.
[5] I. Dinur, K. Nissim, Revealing information while preserving privacy, Proc. Twenty-Second ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (2003) 202–210. doi:10.1145/773153.773173.
[6] C. Dwork, Differential privacy, Proc. 33rd Int. Colloq. Autom. Lang. Program. (2006) 1–12. doi:10.1007/11787006_1.
[7] International Telecommunication Union, Overview of the Internet of things, Ser. Y Glob. Inf. Infrastructure, Internet Protoc. Asp. next-Generation Networks - Fram. Funct. Archit. Model. (2012) 22.
[8] S. Lodha, D. Thomas, Probabilistic anonymity, PinKDD'07 Proc. 1st ACM SIGKDD Int. Conf. Privacy, Secur. Trust KDD. (2007) 56–79. doi:10.1007/11539452_16.
[9] S.L. V. Banahatti, Safemask, in: 5th World TCS Tech. Archit. Glob. Conf., Prune, India, 2009.
[10] A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets, Proc. - IEEE Symp. Secur. Priv. (2008) 111–125. doi:10.1109/SP.2008.33.
[11] F. McSherry, Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis, Commun. ACM. 53 (2009) 89–97. doi:10.1145/1810891.
[12] Ú. Erlingsson, V. Pihur, A. Korolova, RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, Proc. 2014 ACM SIGSAC Conf. Comput. Commun. Secur. - CCS '14. (2014) 1054–1067. doi:10.1145/2660267.2660348.
[13] F. McSherry, K. Talwar, Mechanism Design via Differential Privacy, 48th Annu. IEEE Symp. Found. Comput. Sci. (2007) 94–103. doi:10.1109/FOCS.2007.66.
[14] R.J. Bayardo, DP thru optimal k anonymization, (n.d.).
[15] V.S. Iyengar, Transforming data to satisfy privacy constraints, Proc. Eighth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '02. (2002) 279. doi:10.1145/775047.775089.
[16] J. Domingo-Ferrer, F. Sebé, A. Solanas, A polynomial-time approximation to optimal multivariate microaggregation, Comput. Math. with Appl. 55 (2008) 714–732. doi:10.1016/j.camwa.2007.04.034.
[17] J.-L. Lin, M.-C. Wei, An efficient clustering method for k-anonymization, Proc. 2008 Int. Work. Priv. Anonymity Inf. Soc. - PAIS '08. (2008) 46. doi:10.1145/1379287.1379297.
[18] M.E. Kabir, H. Wang, E. Bertino, Efficient systematic clustering method for k-anonymization, Acta Inform. 48 (2011) 51–66. doi:10.1007/s00236-010-0131-6.
[19] G. Ghinita, P. Karras, P. Kalnis, N. Mamoulis, Fast data anonymization with low information loss, Proc. 33rd Int. Conf. Very Large Data Bases. (2007) 758–769.

http://dl.acm.org/citation.cfm?id=1325938\nhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.138.3217.

[20] D. Su, J. Cao, N. Li, E. Bertino, H. Jin, Differentially Private K-Means Clustering, Codaspy. (2016) 26–37.

[21] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, L. Murphy, A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners, Trans. DATA Priv. 7 (2014) 337–370.

[22] E. Nergiz, C. Clifton, Thoughts on k-Anonymization, Data Knowl. Eng. 63 (2007) 622–645. https://static.aminer.org/pdf/PDF/000/300/785/thoughts_on_k_anonymization.pdf.

[23] U. Greveler, B. Justus, D. Loehr, Multimedia content identification through smart meter power usage profiles, Comput. Priv. Data Prot. (2012). http://www.nds.rub.de/media/nds/veroeffentlichungen/2012/07/24/ike2012.pdf.