Smart IoT Data Collection

Nikos Fotiou, Vasilios A. Siris, Alexandros Mertzianis, George C. Polyzos
Mobile Multimedia Laboratory, Department of Informatics
School of Information Sciences & Technology
Athens University of Economics and Business, Greece
{fotiou, vsiris}@aueb.gr

***Abstract*—We present and experimentally evaluate procedures for efficient IoT data collection while achieving target requirements in terms of data accuracy and privacy protection. The procedures adjust the time period between consecutive measurements following an additive increase and multiplicative decrease (AIMD) scheme based on a target data accuracy and add noise to measurements using differential privacy techniques. The experimental evaluation involves real temperature and humidity measurements obtained from two testbeds through the FIESTA-IoT platform. Our results show that the AIMD adaptation of the measurement period is robust to different types of measurements from different testbeds, without having any tuning parameters, and the addition of noise to the sensor measurements using differential privacy has a negligible effect on the aggregate statistics.**

***Index Terms*—data accuracy, adaptive data collection, differential privacy, testbed experiments**

I. Introduction

The Internet of Things will involve a huge number of sensors. Periodically collecting data from all IoT sensors will waste a significant amount of communication and storage resources, in addition to a significant amount of energy, which impedes the scalability of IoT systems. Moreover, the collected data, or even correlating the collected measurements, may reveal sensitive information related to end-users' activities. Naive approaches to (big) IoT data collection - reflecting the "first collect everything then (try to) analyse everything" paradigm - can also impede the integration of IoT and Cloud systems and of IoT and big data analysis, since the immense amount of data has implications on the amount of network resources and the amount of computation resources.

Different IoT applications have different requirements in terms of accuracy, latency, and energy consumption. For example, an environmental monitoring application can require monitoring of real world phenomena with some degree of accuracy, while being tolerant to delays in receiving data updates from the IoT sensors. On the other hand, time-critical applications, such as security and critical infrastructure monitoring that involve both monitoring and actuation, can have strict requirements in terms of the delay for IoT sensor nodes to transmit their data to the applications. Finally, different IoT sensors can have different constraints in terms of battery consumption hence applications can require a different balance between data accuracy or timeliness and energy consumption. Motivated by the above, the goal of our work in this paper is to develop and experiment with procedures for efficiently collecting IoT data while achieving target requirements in terms of data accuracy, timeliness, energy efficiency, and privacy protection.

In summary, the contributions of the paper are the following:

- We define procedures for efficient data collection that satisfy target requirements in terms of data accuracy and privacy protection. The data accuracy-driven procedure adapts the period between measurements using an additive increase and multiplicative decrease (AIMD) scheme and has no tuning parameters. The privacy-driven procedure is based on differential privacy techniques.
- We evaluate the procedures with experiments involving real temperature and humidity measurements, that are obtained from two testbeds over the FIESTA-IoT platform[1].

The rest of the paper is structured as follows: In Section II we present the overall architecture of our measurement framework, focusing specifically on the data accuracy and privacy-driven strategies. In Section III we present and discuss our experimental results. In Section IV we present a brief summary of related work, identifying how the work contained in this paper differs. Finally, in Section V we conclude the paper identifying directions of ongoing and future research.

II. IoT data collection strategies

Our data collection framework BeSmart implements the following four strategies:

- Data accuracy-driven: This strategy considers the tradeoff between the data accuracy and the frequency of measurement requests. Specifically, the frequency of measurement requests is adapted (temporal adaptation) while maintaining a target data accuracy. Additionally, when there are many IoT sensors located in the same geographic area, the collector can exploit the spatial correlation of sensor measurements to adapt the subset of the sensors from which measurements are requested (spatial adaptation).
- Time-driven: This strategy ensures that the elapsed time since the timestamp of the last measurement is below some maximum delay; this elapsed time corresponds to the timeliness of data measurements. To select the appropriate time to request measurements in order to ensure the maximum elapsed time target, the delay from the time a request is sent by the collector until the time the

[1] The FIESTA-IoT platform provides uniform access to IoT data from heterogeneous testbeds. For more information see http://fiesta-iot.eu/

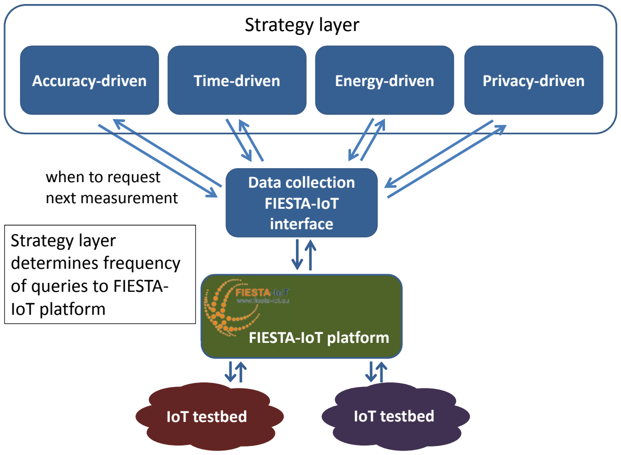


Fig. 1. BeSmart data collection architecture.

response (measurement) is received must be considered; the latter delay can involve the network delay, the delay of the IoT measurement middleware, and the delay at the particular IoT testbed where the sensor is located.

- Energy-driven: This strategy seeks to maximize the net benefit, which is the gain for a given data accuracy or timeliness minus the corresponding power consumption. The gain for a specific data accuracy or timeliness can be expressed using a utility function, while the power consumption depends on the frequency of measurements.
- Privacy-driven: This strategy targets to preserve end-user privacy by altering the accuracy of the individually retrieved results, without significantly altering their statistics. In particular, this strategy will (a) try to minimize the amount of times a specific sensor is requested for data (using also approaches developed for the data-driven and time-driven strategies), and (b) add "noise" to the retrieved results using differential privacy techniques.

The above strategies are implemented solely at the receiver-side, i.e. at the data collector (see Figure 1). This is the only option possible if the sensors cannot implement specific collection strategies, as is the case with the FIESTA-IoT platform which we consider in our experiments. Another advantage of the receiver-driven approach is that because sensor nodes typically have small processing and storage capabilities, the range of strategies they can implement can be limited. This is not the case with the processing capabilities at the data collector side.

The architecture of our BeSmart framework is shown in Figure 1. The strategy layer, which implements the aforementioned four strategies, determines the time period at which measurements are requested from the FIESTA-IoT platform. The FIESTA-IoT platform manages IoT data from heterogeneous systems and environments and their entity resources (such as smart devices, sensors, and actuators), and was developed by the EU-funded Federated Interoperable Semantic IoT/cloud Testbeds and Applications project. FIESTA-IoT enables experimenters to use a single Application Program Interface (API) for executing experiments over multiple IoT testbeds that are federated in a testbed agnostic way.

Next we discuss in more detail two of the strategies implemented in our framework: accuracy-driven and privacy-driven data collection. These two strategies will be evaluated with measurements obtained from the FIESTA-IoT platform in Section III.

A. Accuracy-driven data collection

The motivation for the accuracy-driven data collection strategy is that many applications require a particular data accuracy, and providing a higher accuracy offers no advantages. Hence, the goal of this strategy is not to select the period between measurement requests such that the time series of data measurements have the smallest deviation from the actual sensor values, which indeed can be period at which the IoT sensor obtains measurements for a particular phenomena. Rather, the strategy seeks to reduce the frequency of the measurements, hence reduce the amount of resources (processing, communication, and storage) necessary for data collection, while achieving a target data accuracy.

The data accuracy-driven strategy seeks to achieve a target accuracy according to which the last measurement obtained differs from the current sensor value by at most a target accuracy; this target accuracy can be expressed as a simple percentage, e.g. the last measurement differs from the current sensor value by at most 10%. Of course, only an Oracle with knowledge of all the future sensor values can achieve the above goal 100% of the time. Figure 2 illustrates when measurements are requested by the Oracle, with knowledge of all future measurement values. A target data accuracy defines an accuracy interval corresponding to the last measurement that was obtained. A new measurement is requested whenever the current sensor value is outside the accuracy interval of the last measurement. Observe from Figure 2 that measurements are requested more frequently, i.e. the period between consecutive measurement requests is smaller, when the measured values change at a higher rate. We will use the Oracle as a benchmark to compare the performance of the proposed adaptation procedure that we describe next.

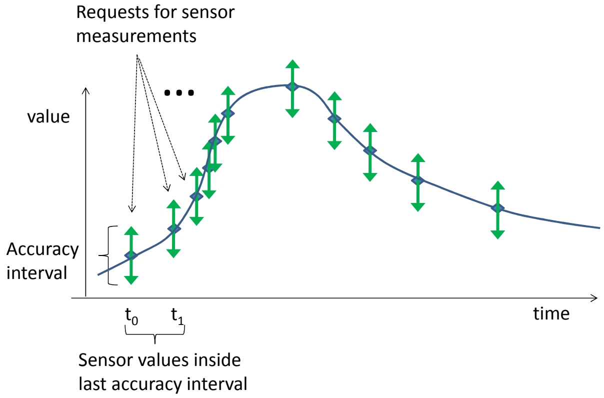


Fig. 2. Given a data accuracy target, which defines an accuracy interval, the goal is to request measurements such that at any time the current sensor value falls within the accuracy interval of the lasted measurement. Only an oracle can achieve this 100% of the time.

Since the value of future measurements is not known, the Oracle cannot be applied in practise. Instead, we propose that the time period between consecutive measurement requests is initially equal to some small value, which can be the interval in which the sensor produces (or obtains) its measurements, which we will refer to as unit interval. The time period T between consecutive measurements is increased by one unit interval, i.e. it increases linearly, as long as the current measurement is inside the accuracy interval based on the measured value when the interval T *was last adapted*. If the current measurement is outside this accuracy interval, then T is reduced to half its value, i.e. it decreases multiplicatively. The above additive increase and multiplicative decrease (AIMD) procedure resembles the AIMD adaptation of TCP's congestion window. Indeed, the motivation for the additive increase is that the procedure slowly tests larger periods between consecutive measurements, while the reduction to half is conservative to avoid requesting measurements to far apart, hence have measurements that deviate from the actual sensor values. A key feature of the proposed AIMD adaptation of the measurement period is that it is not based on some data model and involves no tuning parameters.

B. *Privacy-driven data collection*

This strategy targets to preserve end-user privacy by altering the accuracy of the retrieved results. This is achieved by adding "noise" to the retrieved results using *differential privacy* [1] techniques.

The goal of differential privacy is to allow the extraction of statistics for a population of users without revealing any information about particular individuals. This is achieved with the addition of some "noise" to the statistics extraction process. This noise guarantees that the contribution of a single individual to the calculated statistics is not "significant." Hence, differential privacy guarantees that the output of the statistics calculation process is only slightly impacted by the contributions of a single individual. However, extracting information about a group of users in a privacy preserving way provides only partial security: a provider should still be trusted to collect and maintain user-provided sensitive information. This is not necessary if noise is added at the source during the data collection phase and not during the data processing phase. A trivial approach for implementing this functionality is the so-called "survey based on random responses" [2]. In a nutshell, with this approach, if a user is asked a question that can be answered with a "Yes" or "No," (for example, "Is your access network congested?"), then she flips a fair coin, in secret, and answers the truth if it comes up tails. Otherwise she flips another coin in secret and answers "Yes" if it comes up tails or "No" otherwise. This approach allows users to retain very strong deniability, while at the same time the real ratio of "Yes" answers can be accurately estimated using $2{*}(Y-0.25)$, where Y is the portion of "Yes" responses.

For our privacy-driven strategy we use the RAPPOR protocol. The RAPPOR protocol [3] is an extension of the random responses approach, developed by Google and used in Google Chrome for collecting user statistics. This protocol allows questions with a richer dataset of possible answers (nevertheless, this dataset has to be pre-defined) and protects users' privacy even if the same question is asked many times. The RAPPOR data collection process is extremely lightweight, hence it can be implemented even in constrained devices. The basic idea behind RAPPOR is that whenever a user is asked a question, she is provided with a set of options. For each option the user plays the "random response" game. Eventually the user responds with a bit vector of size equal to the options' set: a bit in the vector set to 1 means that the corresponding option is selected by the user as one of the answers to the question. Due to the randomness of the response process, a user may select multiple options or none of the options, and an option may or may not correspond to the user's real answer.

III. Experiments

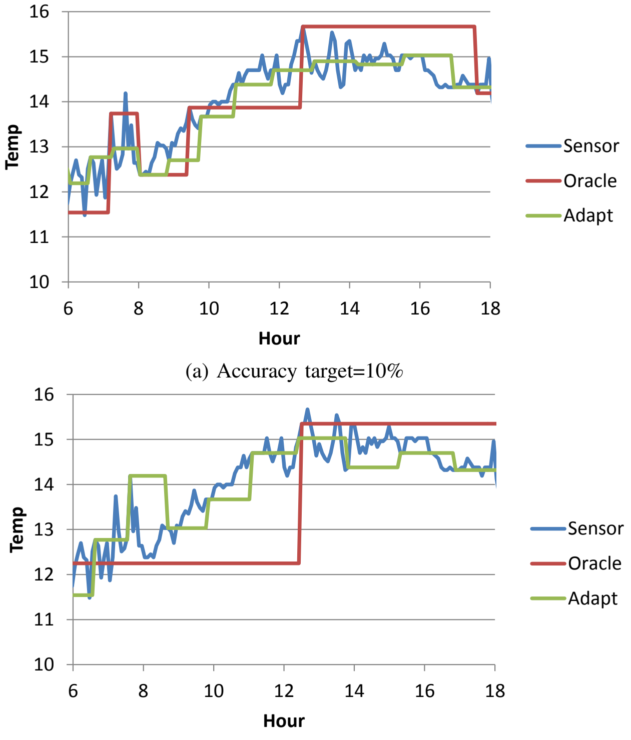
In this section we present experiments for the data accuracy-driven and the privacy-driven data collection strategies. Our experiments consider measurements of different data types (phenomena), namely temperature and humidity, obtained from the SmartSantander and FINE testbeds through the FIESTA-IoT platform. The SmartSantander testbed is located in Santander, northern Spain. The FINE testbed is located on the island of Crete, Greece.

A. *Accuracy-driven data collection*

Figure 3 shows the temperature measurements for the Oracle, which has knowledge of the future temperature values, and the AIMD adaptation procedure, for two data accuracy targets: 10% and 20%. At time periods where the graph for the Oracle and the AIMD adaptation procedure is horizontal, measurements are requested only at the beginning of the corresponding period. Hence, measurements are requested when the graph increases or decreases. Observe in Figure 3 that the measurements for the Oracle, compared to those for the AIMD adaptive scheme, are farther from the actual sensor values but still within the target accuracy interval; this is expected since the Oracle tries to reduce the number of measurements while allowing the maximum deviation of the measurements from the actual sensor values to be equal to the target data accuracy. Comparing Figure 3(a), which was obtained for target data accuracy 10%, with Figure 3(b), which was obtained for target data accuracy 20%, we observe that a larger data accuracy results in fewer measurements for both the Oracle and the AIMD adaptation procedure.

TABLE I
Temperature, SmartSantander, 96 hour time window, total # of sensor values: 1.147

Target accuracy	Oracle	AIMD adaptation
10%	measurements: 68 (5,9%)	measurements: 86 (7,5%) deviations: 56 (4,9%)
20%	measurements: 9 (0,8%)	measurements: 48 (4,2%) deviations: 2 (0,05%)



(a) Accuracy target=10%

(b) Accuracy target=20%

Fig. 3. Measurements with the Oracle and the AIMD adaptation procedure. Comparison of the two graphs illustrates the tradeoff between the target data accuracy and the number of measurements.

Table I quantifies the gains in terms of the reduced number of measurements of the Oracle and the AIMD adaptation procedure, for a 96 hour time window that includes a total of 1.147 temperature values with an interval of approximately 5 minutes between consecutive values. Both the Oracle and the AIMD adaptation procedure reduce the number of measurements by over 90%. In particular, for 10% target data accuracy, with the AIMD measurement period adaptation procedure only 86 out of the total 1.147 temperature values are requested (7,5%), which is very close to the number of measurements requested with the Oracle, which is 68 (5,9%). Moreover, the gains increase when the target data accuracy increases from 10% to 20%, i.e. the data accuracy becomes less stringent. The percentage of measurements which are outside the target data accuracy interval for the AIMD adaptation procedure is 4,9% and 0,05% of the total number of sensor values (1.147), for a target data accuracy 10% and 20%, respectively; this means that, for a 10% target data accuracy, only 56 out of the 1.147 temperature values (4,9%) were outside the 10% accuracy interval of the last measured value. For a 20% target data accuracy, only 2 out of the 1.147 (0,05%) temperature values where outside the 20% accuracy interval. On the other hand, since it has knowledge of future temperature values, which of course in practise is not possible, the Oracle can request measurements such that they all differ from the actual sensor values by at most the target data accuracy.

TABLE II
Humidity, SmartSantander, 96 hour time window, total # of sensor values: 977

Target accuracy	Oracle	AIMD adaptation
10%	measurements: 34 (3,5%)	measurements: 134 (13,7%) deviations: 51 (5,2%)
20%	measurements: 11 (1,1%)	measurements: 56 (5,7%) deviations: 52 (5,3%)

Table II quantifies the gains in terms of the reduced number of humidity measurements. The AIMD procedure again achieves significant gains by reducing the number of measurements requested by more than 85%. The results in Table II also illustrate the tradeoff between data accuracy and number of measurements, for both the Oracle and the AIMD adaptive procedure. Observe that the difference in performance of the Oracle and the AIMD procedure is larger for the humidity measurements than for the temperature measurements, Table I; this is because the humidity changes more abruptly compared to the temperature, for the time window in which the experiments were performed.

Table III quantifies the gains in terms of the reduced number of measurement requests for temperature measurements obtained from the second testbed, FINE, whose temperature sensors provided measurements every 10 minutes. Compared to the results in Table I, the number of measurements for both the Oracle and AIMD adaptation procedures is higher, but the gains are still significant: the AIMD procedure achieves over 80% reduction of the total number of measurements for a 10% target data accuracy, with only 5% deviations.

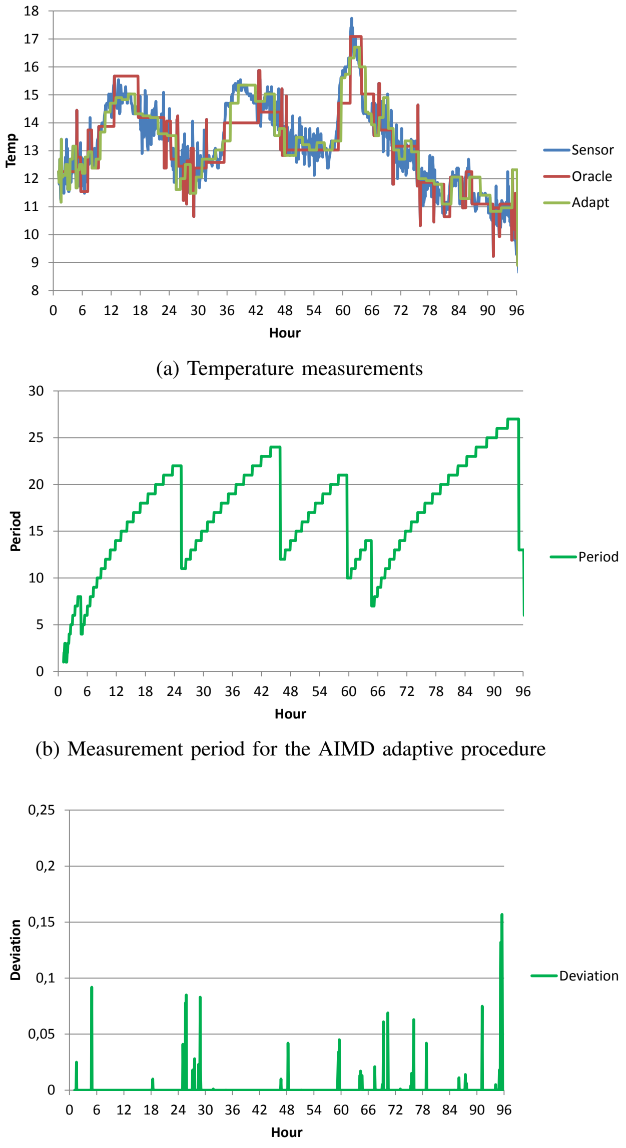
Figure 4(a) shows that the AIMD adaptation procedure can follow the varying trend of the temperature quite well, even though the values of the temperature exhibit periodicity and non-stationarity. Figure 4(b) shows the adaptation of the measurement period of the AIMD procedure, illustrating its additive increase and multiplicative decrease behavior. Finally, Figure 4(c) shows that the magnitude of the deviations for the AIMD procedure, i.e. the relative difference of the measured values compared to the sensor value when the measured values are outside the target accuracy, is typically less than 10%

B. *Privacy-driven strategy*

For the evaluation of the privacy-driven strategy we implemented the *basic one-time RAPPOR* algorithm, described in [3], and applied it for calculating averages of temperature measurements performed by sensors in the Smart Santander

TABLE III
Temperature, FINE, 96 hour time window, total # of sensor values: 628

Target accuracy	Oracle	AIMD adaptation
10%	measurements: 48 (7,6%)	measurements: 109 (17,4%) deviations: 32 (5,1%)
20%	measurements: 22 (3,5%)	measurements: 66 (10,5%) deviations: 29 (4,6%)



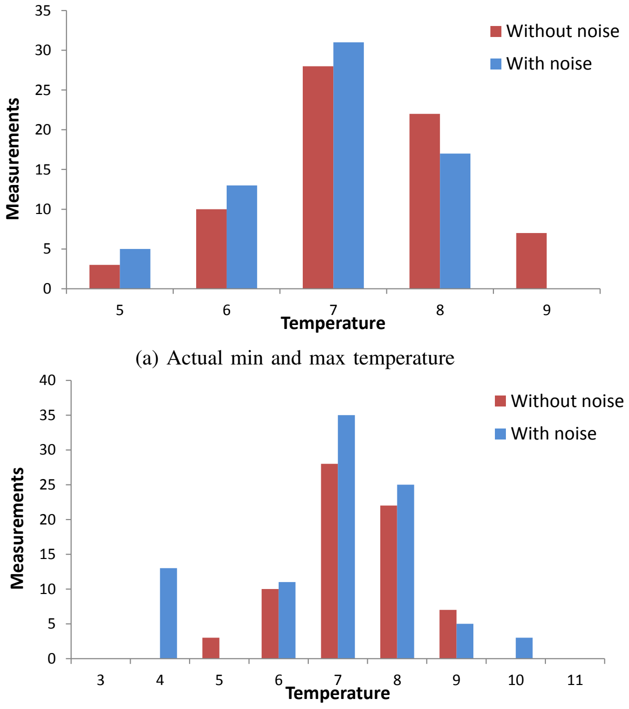
(a) Temperature measurements

(b) Measurement period for the AIMD adaptive procedure

(b) Deviations for adaptive scheme

Fig. 4. Performance of the AIMD adaptation procedure in a 96 hour time window of temperature measurements from SmartSantander.

testbed. In particular, we applied the algorithm over the measurements of 70 sensors in order to calculate hourly average temperatures. Each sensor was provided with a set of possible temperatures and for each element in the set the random responses game was executed. Hence, each sensor could respond "Yes" to some/all/none temperatures and its responses may or may not include the real measurement. Figure 5 illustrates[2] the distribution of measurements when the set of possible temperatures is confined by the actual minimum and maximum temperature, Figure 5(a), and when its boundaries are broader, Figure 5(b).



(a) Actual min and max temperature

(b) Expanded min and max temperature

Fig. 5. Distribution of measurements with and without noise addition with the limits of the possible temperatures (a) equal to the actual min and max temperature, (b) expanded.

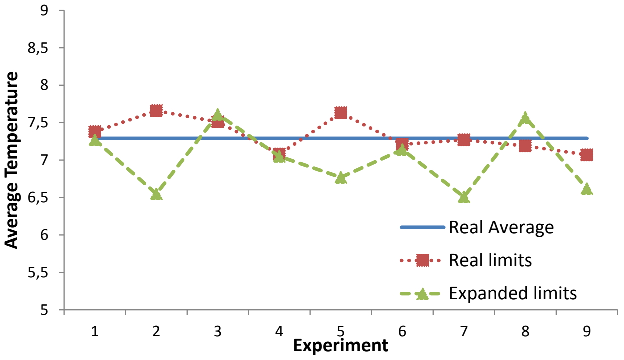


Fig. 6. Calculated average temperature per experiment.

Since the sensor responses are randomized, every time an experiment is repeated the sensor responses (and hence the calculated average temperature) differ. Figure 6 illustrates the calculated average temperature for a specific hour[3] compared to the real average temperature, when the set of possible temperatures is confined by the actual minimum and maximum temperature and when its boundaries are broader. The graph shows that the calculated averages are very close to the real averages, despite the addition of noise.

[2] In this paper we include a subset of our results. Interested users are encouraged to use our live demo located at https://mm.aueb.gr/fiesta/privacy.php

[3] The graphs for all other hours follow a similar pattern.

IV. Related work

Prior work on data collection in wireless sensor networks has proposed schemes that trade the quality of the data collected for energy efficiency. This work includes schemes based on models at the data collector that capture the correlation of data to reduce the number of data queries to sensors [4], [5], [6]. Some proposals combine pull-based (located at the data collector) and push-based (located at the sensors) mechanisms for reducing the number of messages [7], [8], [6] or consider purely push-based update strategies [9]. The data collection strategies investigated in this paper differ from the above in that they do not rely on models, as in [5], [4], [6], but rather adapt the period between consecutive measurements based on whether the measurements lie inside a target data accuracy interval. Moreover, the proposed data collection procedures are implemented solely at the data collector side (pull-based), hence do not require mechanisms at the sensor side, which is not possible when the sensor testbeds are not directly accessible or are under different administrative control, as is the case of the testbeds federated under the FIESTA-IoT platform.

In addition to the temporal correlation of data measurements, the spatial correlation of measurements from sensors located in the same area can be exploited, as in [10], [11], [12], which investigate in-network mechanisms for quality-driven sensor cluster construction and estimation of probabilistic models for capturing the temporal and spatial correlation. As noted above, our approach differs in that we consider pure pull-based procedures. The work in [13] proposes a utility-based model for optimally assigning data queries to sensors in a participatory sensing system, while [14] investigates a quality-driven function to select a fixed number of sensors for accomplishing a sensing task.

The trade-off between privacy and accuracy is a well studied problem which still attracts researchers' attention; see for example [15], [16], [17]. These works try to add as much noise as possible to the individual sensor measurements, while influencing as little as possible the aggregate statistics. The contribution of this paper is to evaluate the efficiency of these techniques using real IoT measurements.

V. Conclusions and future work

The proposed accuracy-driven data collection procedure employs an additive increase / multiplicative decrease (AIMD) adaptation of the time period between consecutive measurements. Experiments have shown that such an AIMD procedure is highly robust for measurements of different data types (phenomena), namely temperature and humidity, and for measurements from two different testbeds. The AIMD adaptation procedure is able to track the trends of the values measured in the presence of periodicity and non-stationarity. Moreover, unlike other proposed data collection schemes, it has no tuning parameters. The experiments for the privacy-driven strategy have shown that, despite the small number of sensors, the addition of noise to the sensor measurements using differential privacy has a negligible effect on the aggregate statistics. Ongoing work is investigating the tradeoff between data accuracy and power consumption. Additionally, we are combining the adaptation of the measurement period (temporal adaptation) with the dynamic selection of the subset of sensors to request measurements from (spatial adaptation); such an approach can yield higher energy efficiency, without sacrificing data accuracy.

Acknowledgements

This work was supported in part by the EU FIESTA-IoT project (Grant Agreement number: 643943 / H2020-ICT-2014-1) through the Open Call 3 Experiment "BeSmart: Smart IoT Data Collection".

References

[1] C. Dwork, F. McSherry, and A. Nissim, Kobbiand Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Springer Berlin Heidelberg, 2006, pp. 265–284.

[2] S. L. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[3] U. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," in *Proc. of ACM SIGSAC Conference on Computer and Communications Security*, 2014.

[4] D. Chu, A. Deshpande, J. Hellerstein, and W. Hong, "Approximate data collection in sensor networks using probabilistic models," in *Proc. of IEEE ICDE*, 2006.

[5] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong, "Model-driven data acquisition in sensor networks," in *Proc. of VLDB*, 2004.

[6] A. Jain, E. Chang, and Y.-F. Wang, "Adaptive stream resource management using Kalman Filters," in *Proc. of ACM SIGMOD*, 2004.

[7] Q. Han, S. Mehrotra, and N. Venkatasubramanian, "Energy efficient data collection in distributed sensor environments," in *Proc. of IEEE ICDCS*, 2004.

[8] Q. Han, D. Hakkarinen, P. Boonma, and J. Suzuki, "Quality-aware sensor data collection," *International Journal of Sensor Networks*, vol. 7, no. 3, pp. 127–140, May 2010.

[9] X. Tang and J. Xu, "Adaptive Data Collection Strategies for Lifetime-Constrained Wireless Sensor Networks," *IEEE Trans. on Parallel and Distributed Systems*, vol. 19, no. 6, pp. 721–734, June 2008.

[10] B. Gedik, L. Liu, and P. S. Yu, "ASAP: An Adaptive Sampling Approach to Data Collection in Sensor Networks," *IEEE Trans. on Parallel and Distributed Systems*, vol. 18, no. 12, pp. 1766–1783, December 2007.

[11] C. Liu, K. Wu, and J. Pei, "An Energy-Efficient Data Collection Framework for Wireless Sensor Networks by Exploiting Spatiotemporal Correlation," *IEEE Trans. on Parallel and Distributed Systems*, vol. 18, no. 7, pp. 1010–1023, 2007.

[12] C. Wang, H. Ma, Y. He, and S. Xiong, "Adaptive Approximate Data Collection for Wireless Sensor Networks," *IEEE Trans. on Parallel and Distributed Systems*, vol. 23, no. 6, pp. 1004–1016, June 2012.

[13] M. Riahi, T. Papaioannou, I. Trummer, and K. Aberer, "Utility-driven data acquisition in participatory sensing," in *Proc. of Int. Conf. on Extending Database Techn. (EDBT)*, 2013.

[14] M. Marjanovic, L. Skorin-Kapov, K. Pripuzic, A. Antonic, and I. Zarko, "Energy-aware and quality-driven sensor management for green mobile crowd sensing," *Journal of Network and Computer Applications*, vol. 59, pp. 95–108, January 2016.

[15] M. Andres, N. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proc. of ACM SIGSAC Conference on Computer and Communications Security*, 2013.

[16] G. A. Fink, "Differentially private distributed sensing," in *Proc. of 3rd IEEE World Forum on Internet of Things (WF-IoT)*, 2016.

[17] J. Chen, M. Huadong, and Z. Dong, "Private data aggregation with integrity assurance and fault tolerance for mobile crowd-sensing," *Wireless Networks*, vol. 23, no. 1, pp. 131–144, 2017.