Abstract—Participatory Sensing (PS) is a new data collection paradigm based on the voluntary participation of many cellular users equipped with smart applications, a large diversity of sensors, and Internet connectivity at all times. Although many PS-based applications can be foreseen to solve interesting and useful problems, many of them have not been fully implemented and used in practice because of privacy concerns. Compounding the problem, privacy-preserving mechanisms introduce additional issues. For example, one of the most important problems is that of the quality of the information provided by the PS system to the final users. The problem is that, in order to protect the privacy of the users, most privacy-preserving mechanisms modify their real locations, which makes the reported data as if it had been measured from a different location, introducing noise or false information in the system and to the final users. Another important problem is that of the energy consumption. Privacy-preserving mechanisms consume extra energy and users are not very willing to use PS applications if they drain their batteries considerably faster. This paper proposes a hybrid privacy-preserving mechanism that combines anonymization, data obfuscation, and encryption techniques to increase the quality of information and privacy protection without increasing the energy consumption in a significant manner. A new algorithm is proposed that dynamically changes the cell sizes of the grid of the area of interest according to the variability of the variable of interest being measured and chooses different privacy-preserving mechanisms depending on the size of the cell. In small cells, where users can be identified easier, the algorithm uses encryption techniques to protect the privacy of the users and increase the quality of the information, as the reported location is the real location. On the other hand, anonymization and data obfuscation techniques are used in bigger cells where the variability of the variable of interest is low and therefore it is more important to protect the real location (privacy) of the user. We evaluated our hybrid approach and other privacy-preserving mechanisms using a real PS system for air pollution monitoring. Our experiments show the better performance of the proposed hybrid mechanism and the existing trade-offs in terms of privacy, quality of information to the final user, and energy consumption.

I. INTRODUCTION

Participatory Sensing (PS) is a new data collection paradigm based on the availability of millions of cellular users equipped with smart phones and applications, Internet connectivity at all times, and a large diversity of sensors. Given the number of cellular users, PS systems can form really large networks made of millions of mobile sensing devices in a fast and cost-efficient manner. Thus, PS have the potential to address important and new problems and make a difference in our daily lives and society altogether. For instance, an application that requires users to sense the level of pollution as they travel to build accurate pollution maps could be used by governmental organizations to monitor and control the amount of gases placed in the atmosphere by some industries and by the community for many other different purposes [12].

As with any other new technology, PS also brings new challenges that need to be addressed if we want this new sensing paradigm to bring all the expected benefits. For example, data accuracy and validity are more important than ever. Since sensors are supposed to be cheap and they are in the hands of the users, the potential for less accurate measurements is higher than in current systems [11]. Therefore, new algorithms need to be devised before applying inference and visualization techniques to counteract these problems. In addition, many users might not be willing to spend their own resources, such as energy and data plans, without getting something in return for sending the sensing reports. Appropriate incentive mechanisms need to be in place to motivate users to participate in the PS system [7]. Another vital problem, the one that we are concerned about in this paper, is privacy. Since participants reveal their spatio-temporal information when they submit their sensing reports, they are making their private information available at the same time. Thus, many PS systems cannot be implemented in practice because users might not be willing to participate if their private information is not protected.

Consequently, several privacy-preserving mechanisms have been designed for PS systems thus far. They are commonly based on anonymization, obfuscation, or encryption techniques. Anonymization-based mechanisms aim to protect the association between a participant and her private data (e.g., location) transforming the original data into a generalized data that represent a set of participants [2]. Obfuscation-based techniques aim at protecting the same association but transforming the private information of the user (e.g., location) with no consideration of other participants [15]. Finally, encryption-based mechanisms aim to protect the privacy of the participants using cryptographic techniques for the transmission and storage of the reported data [6]. Privacy-preserving techniques introduce new challenges to PS systems as well. For example, anonymization and obfuscation techniques introduce noisy data to the system. Data transformation decreases the quality of information presented to the final users. On
the other hand, encryption-based techniques require additional processing and additional data transmissions that increase the energy consumption on mobile devices.

In this paper, using P-sense [12], a pollution monitoring system as a PS system example, we study different privacy-preserving mechanisms to find answers to the following questions:

- What is the probability of locating a participant? What privacy mechanisms provide more (or less) privacy?
- What is the quality of the information provided to the final users? Are privacy mechanisms actually useful to the final users? How much do these privacy mechanisms modify/alter the real data?
- How much energy does the privacy-preserving mechanism consume?
- What kind of trade-offs or choices can we make to obtain reasonable levels of privacy, quality of information, and energy consumption?

In addition, we propose a new privacy-preserving mechanism for environmental sensing that combines anonymization, data obfuscation, and encryption techniques to increase the privacy of the users while improving the quality of information and the energy consumption. The proposed hybrid mechanism includes an algorithm that dynamically changes the cell sizes of the grid of the area of interest according to the variability of the variable of interest being measured and chooses different privacy-preserving mechanisms depending on the size of the cell. The mechanism is built upon the Point of Interest-based mechanism presented in [14], for location obfuscation and data anonymization, in the case of big cells where the variability of the variable of interest is low and therefore it is more important to protect the real location of the user, and a double encryption approach only used in the case in which users are located inside of small cells (where users can be identified easier) to protect their privacy and increase the quality of the information, as the reported location is the real location.

Using P-Sense [12], we evaluate the proposed hybrid mechanism and other important privacy-preserving mechanisms in terms of privacy given by the probability to be located \(P_{loc}\), quality of information given by the coefficient of determination \(R^2\), and energy consumption \(E\). Our results indicate the superiority of the proposed hybrid mechanism and show the trade-offs available in terms of these three performance metrics.

The rest of the paper is organized as follows. Section II includes a brief description of the privacy-preserving and inference mechanisms available in the literature and used in this paper. Section III describes the system architecture, the data processing model, and the threat model. Section IV describes our proposed hybrid privacy-preserving mechanism. Sections V and VI present the setting and metrics utilized in the performance evaluation and the performance results, respectively. Finally, Section VII presents the most important conclusions and provides directions for additional research.

II. RELATED WORK

This section provides a brief literature review of the most important mechanisms available in the literature, as they relate to this work.

A. Privacy-preserving mechanisms

Privacy-preserving mechanisms can be classified as anonymization, obfuscation and encryption. At first, in anonymization-based mechanisms, the user’s sensitive data are generalized to a group of users in such a way that the user cannot be distinguishable from the group. One important technique in this category is Tessellation, which enlarges each location point to a region called a Tile using the location of \(k\) users [3]. Once the tiles are defined, participants report their data tagged to the center of the tile instead of their actual locations. The main drawbacks of this technique are the need for a central entity knowing the location of the participants in order to define the tiles, which could be compromised by internal adversaries, and the lack of accuracy for certain applications due to the possibly large tile size.

Obfuscation techniques assume that the identity of the participants are or could be known. Differently from anonymization techniques, the key idea is to modify the real location of the participants without considering the location of other participants. As a result, most obfuscation techniques are run in the mobile devices. In this category, Random Perturbation and Points of Interest are widely-known techniques. Random perturbation, proposed in [1], modifies the original data set with random noise drawn from a known statistical distribution. For instance, the actual user location can be modified adding Gaussian noise [16]. On the other hand, the Points of Interest mechanism, initially designed for Location Based Services [4], defines a set of points of interest that are sensitive to the system. Users select the closest point of interest and query the system as if they were located in that point. The work presented in [14] makes an adaptation of this technique to PS applications. The obfuscation part of the system divides the area of interest into a set of regions called cells, each one centered on a point of interest according to the application requirements. Data sensed within the area of a cell are reported as sensed in the center of that cell.

Encryption-based techniques rely on cryptographic methods to guarantee the privacy of the participants. One of the most important mechanisms is the double encryption mechanism presented in [6]. The system uses two servers: an identification proxy and an application server. The mobile application encrypts the sensed data and uses the public key of the ID proxy to encrypt its identification data. The ID proxy, upon receiving the reports, validates the signature of the sender to guarantee the integrity of the data; however, this server does not have the capability to decrypt the sensed data because it is encrypted using the public key of the application server. Using this scheme, the system has a double protection layer against an adversary and therefore a very low probability of being hacked. However, this approach has a high energy consumption.
B. Inference mechanisms

Inference techniques are utilized in PS systems to estimate the variables of interest in those places where data are not available. In this area, Kriging is one of the most widely used techniques in geostatistics (a branch of statistics that focuses on spatio-temporal datasets). Kriging is a BLUE (Best Linear Unbiased Estimator) interpolator, i.e., it is a linear estimator that matches the correct expected value of the population (unbiased) while minimizing the variance of the observations (Best) [5], [13].

III. System Model

In this section we describe the system architecture, the data processing model, and the threat model used in the paper.

A. System architecture

Figure 1 shows the system architecture presented in [14] and used here to implement and evaluate the different privacy-preserving mechanisms studied in this work. The PS system consists of the following four independent entities:

1) **Mobile nodes**: set of mobile devices with enough computing power and communication capabilities to perform the sensing task, run the privacy mechanism, and transmit the data to the application server.

2) **Data broker**: responsible for sending the list of points of interest to mobile nodes, receiving the sensing reports, validating the mobiles’ signature, performing the initial data aggregation, and forwarding the anonymized data to the application server.

3) **Application server**: responsible for performing the inference process and providing the information to the final users.

4) **Points of interest generator server**: responsible for dividing the target area into cells and assigning locations to points of interest. This process is crucial in order to guarantee data privacy and a good quality of information.

B. Data processing model

Our data processing model consists of four modules, as shown in Figure 2. The first module consists of the sensed data acquired and reported by the users. These data are fed directly to the second module, which is in charge of applying the privacy mechanism and producing the protected data. The next module is the data verification module. Data verification, in the context of PS, consists on the process of detecting and removing spatial outliers to properly reconstruct the variables of interest. This module receives the protected data and verifies the validity of the data. Wrong measurements due to malfunctioning sensors and/or tampered data due to malicious users are removed by this module. Consider a pollution monitoring system where certain users would like to modify the sensed data in order to report a lower level of pollution in a specific area, and therefore, avoid penalty fees, and users reporting wrong data because of the low quality of the sensors or malfunctioning sensors. For this module, we rely on our previous work that implemented a hybrid neighborhood-aware algorithm for outlier detection that considers the uneven spatial density of the users, the number of malicious users, the level of conspiracy, and the lack of accuracy and malfunctioning sensors [9], [10]. In those publications, we showed that our algorithm performed as good as the best estimator, effectively removing the wrong and tampered data while considerably reducing the execution time. Figure 3 shows an example of how the data verification module reconstructs the spatial interpolation of the carbon monoxide variable after removing the tampered data. The final module is the inference engine, which receives all (good) measured data and uses Kriging to estimate the variable of interest throughout the entire area. By comparing real data from real locations, real data from modified locations, and estimated data using Kriging, we can determine the quality of information reported to the final users.

In addition, our model defines time slots for performing the sensing task called **Rounds**. Each round corresponds to a time interval in which participants sense and report data to the system. When the current round \( R_t \) is ended, the inference mechanism estimates the value of the variable of interest in each point of the area of interest, and a new round \( R_{t+1} \) starts.

C. Threat model

In order to design a robust mechanism to protect the privacy of the participants, we consider internal and external adversaries. Therefore, we assume that the adversary could compromise any component of the system, including eavesdropping the communication channel. Thus, we assume that
the adversary may have access to the data broker, application server, or both.

IV. THE PROPOSED HYBRID MECHANISM

The proposed privacy mechanism is a combination of the Points of Interest (POI) mechanism and the Double-encryption technique described in [14] and [6], respectively. At first, the POI mechanism is designed to obfuscate the actual location of the participants using a set of pre-defined locations called points of interest. Each Voronoi space defined for each POI is called a cell. Thus, before reporting to the data broker, the mobile device changes the actual location on the record for the location of the closest POI. The main advantage of the POI mechanism is the fact that there is no need to have a centralized entity knowing the actual locations of the participants. Further, when the data broker receives the obfuscated data, it computes the average of the reported data from all the participants in the same cell and reports only one record per cell to the application server. Thus, this mechanism anonymizes the obfuscated data, adding an extra layer of privacy to the model. The main problem of this mechanism is the noise introduced in the data as a consequence of changing the real locations of the users.

The double encryption scheme is designed to prevent the association between the participants and their records in the system without modifying the actual data, therefore improving the quality of information provided to the final user. In this scheme, the mobile device encrypts the sensed data \(D_i\) using the public key of the application server producing \(\text{Enc}_{\text{AS}}(D_i)\). Keep in mind that \(\text{Enc}_{\text{DB}}(\text{Enc}_{\text{AS}}(D_i))\) is transmitted to the data broker. Since the data broker and the application servers are independently managed, the double encryption mechanism makes it more difficult for an adversary to obtain the necessary data to identify the users. When the data broker receives the reported data, it de-encrypts those records obtaining \(\text{Enc}_{\text{AS}}(D_i)\) only. Therefore, an adversary hacking the data broker knows who sent the data but has no information about the data itself. Finally, the data broker sends the encrypted record \(\text{Enc}_{\text{AS}}(D_i)\) to the application server, where the record is de-encrypted to obtain the actual data \(D_i\). In this case, an adversary compromising the application server only, will know the sense data but will not know who originated it.

The question that arises here is: when should the mobile device use the POI or the double encryption mechanism? In order to answer this question, we need to refer the reader to the Inference stage of the system model. In such stage, the application server applies the Kriging technique, discussed in Section II-B, using the data reported by the participants. The result of applying Kriging to the reported data is a Map of estimated values \((M_{R_i})\) for round \(R_i\), for each point in the area of interest. This map presents to the final user the values of the variables of interest over the entire area -the pollution maps in our example. Then, based on the current \(M_{R_i}\), map, the Points of Interest generator server (Section III-A) runs Algorithm 1 to define the the new set of points of interest and Voronoi spaces (or cells) for the next round \(R_{i+1}\).

Algorithm 1 is the most important piece of the new hybrid mechanism. The main idea to calculate the variability of the variable of interest (e.g., pollution, temperature) after each round and change the size of the cells according to it. If the measurements present very low variability (measurements are within a small range of variation) inside a cell, the cell will very likely remain of the same size and the mechanism will choose to obfuscate the data, i.e., the users will send the reports using the location of the POI of their respective cells. On the other hand, if the variable of interest presents high variability (e.g., very different pollution or temperature values in different zones of the same cell), the algorithm will find those zones with different values and create new cells. If these new cells are smaller than a minimum cell size \(S_{\text{min}}\), the mechanism will choose to encrypt the data because otherwise the user might be recognized considerably easier (the user will be confined to a smaller area). In this manner, when the mechanism obfuscates the data, it protects the privacy of the user more and saves energy, and when it encrypts the data, it provides more accurate information about the variable of interest but it spends more energy. These trade-offs will be explored later in the paper in more detail.

Algorithm 1 Definition of Points of Interest for \(R_{i+1}\)

**Require:** \(M_{R_t}, k, n, S_{\text{min}}\)

1. \(c_{xy} \leftarrow \text{Kmeans}(M_{R_t}, k)\)
2. for each \(c_{xy} \in c_{xy}\) do
3. \(V_i \leftarrow \text{Voronoi}(c_{xy})\)
4. \(g_i \leftarrow \nabla M_{R_t}(V_i)\)
5. end for
6. \(\text{POI}_{xy} \leftarrow c_{xy}\)
7. \(g_{\text{min}} \leftarrow \text{minimum}(g)\)
8. \(g_{\text{max}} \leftarrow \text{maximum}(g)\)
9. divide \([g_{\text{min}}, g_{\text{max}}]\) into \(n\) subinterval \(I_j\)
10. for \(i = 1\) to \(k\) do
11. classify \(g_i\) in the appropriate \(I_j\)
12. \(\Phi_{xy} \leftarrow \text{Kmeans}(M_{R_t}(V_i), j)\)
13. \(\text{POI}_{xy} \leftarrow \text{POI}_{xy} \cup \Phi_{xy}\)
14. end for
15. for each \(\text{POI}_{xy} \in \text{POI}_{xy}\) do
16. if \(\text{Size}(\text{Voronoi}(\text{POI}_{xy})) < S_{\text{min}}\) then
17. \(\text{Mark POI}_{xy}\) for encryption
18. end if
19. end for
20. return \(\text{POI}_{xy}\)

Algorithm 1 requires the Map of estimated values \((M_{R_t})\), the number \(k\) of initial POIs, the maximum number of subcells \((n)\), and the minimum cell size \((S_{\text{min}})\) as input parameters. The first step seeks to find \(k\) cluster centroid locations \((c_{xy})\) in \(M_{R_t}\). Next, for each centroid \(c_{xy}\), the algorithm computes the Voronoi space \((V_i)\) in \(M_{R_t}\) and the gradient of change \((g_i)\) for the variable of interest in the Voronoi space. The minimum and the maximum gradient of change define an interval \(I\), which is subdivided into \(n\) subintervals \(I_1, I_2, ..., I_n\). In the next step,
each gradient of change $g_i$ is classified in the appropriate subinterval $I_j$ and the Voronoi space is subdivided finding $j$ clusters centroid locations ($\Phi_{xy}$). Note that the new list of $POI_{xy}$ for round $R_{t+1}$ is the union between the set $c_{xy}$ and each subset $\Phi_{xy}$. Finally, the size of the cell centered on each $POI_{xy}$ is compared to the minimum cell size ($S_{min}$) in order to determine which cells are too small to guarantee a good privacy level. Thus, when located in those cells, the mobile device selects the double encryption technique; otherwise, the mobile device selects the Points of Interest mechanism.

Figure 4 shows how the hybrid algorithm changes the shape and size of the grid cells over time according to the sensing conditions on each round to improve the quality of the information. The left figure shows the initial, completely squared grid that in the first round produces a quality of information of $R^2 = 0.719$. Five rounds later, the center figure shows how the shape and number of cells are changed according to the variability of the variable being measured and now the $R^2$ improves to 0.831. Finally, the right figure shows a yet different grid and cell sizes that after 10 rounds improves the quality of information metric to $R^2 = 0.900$.

V. PERFORMANCE EVALUATION SETTING

This section describes the performance metrics and experimental setting utilized to evaluate the performance of the privacy-preserving mechanisms considered in this work.

A. Performance metrics

In order to assess the performance of the system, we use the following performance metrics:

1) Probability to be located ($P_{loc}$): In order to measure the privacy level offered by our hybrid mechanism as well as the other privacy-preserving mechanisms, we calculate the Probability to be located ($P_{loc}$). In this work we only consider the snapshot case, i.e., we only calculate the probability to locate a user given the information contained in one round. The trajectory case, i.e., the probability of identifying a user given that the adversary has information from various rounds, is part of our current investigation and will be reported in another paper later. To calculate ($P_{loc}$), we first need to define the following:

- **Data record ($D_i$):** it is the $i$ tuple of the form $(L_{xy}, R_i, B_i)$ reported by a participant, where $L_{xy}$ is the reported location, $R_i$ is the current round, and $B_i$ is a vector containing the sensed values. Thus, $D_i$ does not contain any data that identifies the participant.
- **Map of estimated values ($M_{R_i}$):** it is a new matrix where $M_{xy}$ corresponds to the values estimated by the Inference Engine for the variables of interest at location $(x, y)$ in round $R_i$.
- **Cell location function ($Cell(D_i)$):** given a reported record $D_i$, this function returns the cell (i.e., Voronoi space ($V_j$) centered at $L_{xy}$) in which the record was anonymized/obfuscated from, and the size of the cell $S_{V_j}$.

$$Cell(D_i) \Rightarrow (V_j, S_{V_j})$$ (1)

- **Location estimator function ($Loc(M_{R_i}, D_i, Perc)$):** given a map of estimated values ($M_{R_i}$), a percentile $Perc$, and the $i$ data record $D_i$ in round $R_t$, this function returns an area $A$ of size $S_A$ in which the adversary concludes, with certainty $P_A$, that the issuer was located at when she generated record $D_i$.

$$Loc(M_{R_i}, D_i, Perc) \Rightarrow (A, S_A, P_A)$$ (2)

In order to produce $(A, S_A, P_A)$, this function generates a distribution of the differences between $D_i$ and each estimated value in the cell in which $D_i$ is located. Then, the function selects the $Perc$ percentile in such distribution.

- **Function to estimate the average area ($Loc(M_{R_t}, D_i)$):** given a map of estimated values ($M_{R_t}$) and the $i$ data record $D_i$ in round $R_t$, this function returns the average size of the estimated areas ($S_A$) in which the adversary concluded that the issuer of record $i$ was located at when she generated $D_i$. In order to compute $S_A$, Equation (2) is evaluated using $D_i$ and different values for $Perc$ (e.g., 5%, 10%, ..., 100%) generating a set of estimated areas: $(A_1, S_{A_1}, P_{A_1}), (A_2, S_{A_2}, P_{A_2}), ..., (A_k, S_{A_k}, P_{A_k})$. Then,

$$Loc(M_{R_t}, D_i) = S_A = \sum_{j=1}^{k} (S_{A_j} * P_{A_j})$$ (3)

- **Probability to be located ($P_{loc}$):** given an estimate of the average area where the issuer of record $i$ was located, this
function computes the probability of finding the actual location of the issuer of the data record $D_i$ in round $R_t$. 

$$P_{loc}(Loc(M_{R_t}, D_i)) = 1 - \frac{S_A}{S_T} \quad (4)$$

where $S_T$ is the size of the total area of interest. Notice that $P_{loc}$ increases as $S_A$ decreases because a larger $S_A$ means a larger area in which $D_i$ was located by the adversary (i.e., a lower probability of finding the actual location). Thus, a lower $S_A$ increases the probability of finding the actual location in which $D_i$ was generated.

2) Quality of information ($R^2$): The quality of information is given by the coefficient of determination ($R^2$), which is defined as:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (5)$$

where $y_i$ are the interpolated values generated by using the non-modified locations, $f_i$ are the interpolated values generated by using the values modified by the privacy mechanism, and $\bar{y}$ is the mean of the known data. Therefore, the quality of information $R^2$ provides a measure of how different the real data and the estimations are. By calculating the coefficient of determination, we can measure the effect of the privacy mechanisms in the quality of the final interpolations.

3) Energy consumption ($E$): In order to calculate this last performance metric, measured in Joules, we define three processes that take place in the mobile device. Assuming that sensed data are stored in the mobile device, the first process creates a packet with one or multiple records (samples) that may have been stored in memory. Once the packet is created, the mobile device applies the privacy-preserving mechanism and modifies the data to protect the privacy of the user. Finally, the packet is transmitted to the data broker.

Our model considers a constant voltage ($V$) and senses the actual current ($A_i$) provided to the device every $\delta t_i$ seconds. Then, Equation 6 was used to calculate the total energy used by the three processes described before.

$$E = \sum_i (V \ast A_i \ast \delta t_i) \quad (6)$$

B. The mechanisms and experimental setting

In order to collect real environmental data using a PS approach, we utilized the P-Sense system described in [12]. P-Sense is a PS system for air pollution monitoring that measures CO (ppm), CO$_2$ (ppm), combustible gases (ppm), air quality (4 discrete levels), temperature (F), and relative humidity (%). Figure 5 shows the measurements of four environmental variables after applying a spatial interpolation using a multivariate approach with the combination of the Kriging technique [11]. The figure shows the fine granularity of the measurements provided by P-Sense, definitively not attainable by static environmental stations located miles away from the campus.

The campus of the University of South Florida, in Tampa, Florida was the location for the data collection. Data was collected during three months, 3 times a day for one hour, 3 to 4 days a week. We used the data collected during one hour on a randomly chosen day as our dataset. We assume that the variable of interest does not change fast from one round to the next in each location. The area of the campus is approximately 3.9 km$^2$ which was represented by a grid of 108x108 square units in this project. Thus, each square unit represents 334 m$^2$. This mapping allows us to scale the system to cover larger areas, if needed. For instance, an area of interest of 100 Km$^2$ could be represented using the same grid with 108x108 square units but now each square unit would represent 8574 m$^2$.

For comparison, in addition to our proposed hybrid mechanism, we implemented the most widely known mechanisms in each of the three categories presented in Section II. We used the parameters found in [15] for the best Quality of Information for each mechanism as follows:

1) Tessellation: This mechanism was implemented with $k = 3$. Additionally, tiles were constructed using a grid of $x=5$ and the locations are ordered based on $x$-axis location and the tiles are as big as necessary to contain $k$ participants depending on its value.

2) Points of Interest: This mechanism was implemented dividing the area of interest using a grid of $5 \times 5$ cells.

3) Random Perturbation: This mechanism was implemented using a uniform distribution $\{1,10\}$ for the displacement of the actual locations in each axis.

4) Hybrid: This mechanism was implemented dividing the area of interest of 3.9 Km$^2$ using a grid of $5 \times 5$ cells for the first round. The parameters $k$ and $S_{min}$ for Algorithm 1 were set to 25 clusters following conclusions from [14] and 200 square units of 66800 m$^2$ each according to the results obtained in Table II, respectively.

VI. PERFORMANCE EVALUATION RESULTS

This section presents the evaluation of the privacy-preserving mechanisms according to the performance metrics presented in Section V-A1.

A. Privacy results

According to the system architecture presented in Figure 1, an adversary may have access to three different types of data sets: the data set available in the data broker, the data set available in the application server, or both. As a result, our privacy analysis will be divided in three parts, each one considering that the adversary has access to one of these three data sets. In doing this, we also have to be aware that not all privacy-preserving mechanisms use the same components presented in Figure 1. For example, in the case of anonymization...
schemes, the anonymizer has the actual data reported from the participants, which is equivalent to the data set available in the data broker.

1) Results considering that the adversary has access to the data broker: This scenario is equivalent to an adversary compromising the anonymizing infrastructure in the case of k-anonymity techniques. Thus, an adversary compromising these components have access to the participant’s actual data (e.g., actual location) in the case of tessellation. On the other hand, in the case of Random Perturbation and the Points-of-Interest techniques, the adversary has access to the obfuscated data only. Finally, in the case of our hybrid mechanism, the adversary gains access to a portion of the reported data, which has been obfuscated in the mobile device because the other portion has been encrypted and it is not accessible to the adversary in this server.

The first row in Table I ($P_{loc}$ at DB) shows the Probability to be located ($P_{loc}$) in this scenario based on Equation 4. In the case of Tessellation, $P_{loc}$ is 1 because an adversary compromising the anonymization infrastructure has access to the actual data containing the actual location of the participants. In the case of Random Perturbation, $P_{loc}$ is 0.979 because of the low average displacement needed in order to guarantee a good quality of information [15]. On the other hand, the Points of Interest mechanism has a $P_{loc}$ equal to 0.983 because the data broker does not have the obfuscated records not the actual ones.

Finally, the hybrid mechanism has a $P_{loc}$ of 0.985 for the data records available at the data broker; however, these records only represent 70.2% of all records because the other ones were encrypted and therefore not available in the data broker. As a result, $P_{loc} = 0.692$ for the hybrid algorithm.

2) Results considering that the adversary has access to the application server: In this scenario, the adversary compromises the application server, which is the component where the inference process is performed. In the case of Tessellation, the adversary does not have access to the actual data but the anonymized data. In the case of Random Perturbation, the adversary has access to the obfuscated data only. In the case of Points of Interest and our Hybrid mechanisms, the adversary has access to the anonymized data by the data broker as well as the actual data reported through the encryption path. Please recall that the amount of encrypted data in the application server depends on the size of the cells; mobile users who are located in cells smaller than $S_{min}$ are the only ones encrypting their records.

The second row in Table I ($P_{loc}$ at AS) presents $P_{loc}$ for each privacy mechanism. In the case of Random Perturbation, $P_{loc}$ is the same as the one presented on the first line ($P_{loc}$ at DB) because the data broker and application server have both the same records. In the case of Tessellation with $k = 3$ and sixty users, as we used in our experiments, the number of anonymized records is $\frac{1}{3}$ of the actual ones. Therefore, the average tile size is $\frac{1}{30}$ of the total area meaning that $P_{loc} = 0.95$ for the anonymized records sent by the data broker. Since each anonymized record represents 3 participant records, $P_{loc}$ for each actual record is $\frac{1}{3}$ of 0.95 which is 0.317. In the case of the Points of Interest mechanism, the application server has access to the anonymized records sent by the data broker. With a 5x5 grid, $P_{loc}$ for each anonymized record is $1 - \frac{1}{25} = 0.96$. Our experiments achieve a $k = 2.4$ (i.e., number of actual records represented by each anonymized records), producing a $P_{loc} = \frac{0.96}{2.4} = 0.4$ for each reported record from mobile devices. Finally, in the case of the hybrid mechanism, an adversary compromising the application server will have access to the encrypted records plus the anonymized records sent by the data broker. Thus, $P_{loc}$ is the sum of $P_{loc}$ for the encrypted records plus $P_{loc}$ for the anonymized records sent by the data broker. In the case of the encrypted records, $P_{loc} = 1$ for each one of the records because they contain the exact location of the users; however, these records only represent 29.8% of the total number of records. On the other hand, $P_{loc}$ for the anonymized records is 0.961, which is applied to the remaining 70.2% of the records. However, since each anonymized record represents 2.04 of the original records (2.04 is the average number of obfuscated/anonymous users in each cell), $P_{loc} = 0.298 + 0.961 \times 0.702/2.04 = 0.628$ as shown in Table I.

3) Results considering that the adversary has access to the data broker and application server: This is the worst case scenario because it assumes that the adversary has compromised both, the data broker and the application server. The third row in Table I ($P_{loc}$ at DB + AS) presents the $P_{loc}$ in the case of an adversary compromising both, the data broker and application server. In the case of Tessellation, $P_{loc} = 1$ because the adversary has access to the actual locations from the data broker. In the case of Random Perturbation, $P_{loc}$ is the same presented above because the data in the data broker and application server are the same. Therefore, in this case, it does not make any difference if an adversary compromises the data broker, the application server, or both. In the case of Points of Interest, $P_{loc}$ is the same as in the case where the data broker is compromised because it is the source of more detailed data compared to the application server. Finally, in the case of the hybrid mechanism, $P_{loc}$ is the sum of $P_{loc}$ at the data broker $P_{loc} = 0.692$ and $P_{loc}$ of the encrypted records 0.298 at the application server because an adversary compromising both components will have access to all records in the system; therefore the final $P_{loc}$ is 0.989.

B. Quality of information results

In [15], we concluded that the quality of information ($R^2$) is inversely proportional to the average displacement of the user with respect to their real locations regardless of the privacy-preserving technique applied. Therefore, the challenge here is
to improve $R^2$ without affecting the privacy of the participants.

Figure 6 shows the quality of information achieved by the different mechanisms for the Relative Humidity (RH) variable measured by P-Sense at USF. In the case of Random Perturbation, the quality of information is better because of the lower average displacement compared to the POI mechanism. On the other hand, Tessellation has the worst quality of information because of the larger average displacement. Finally, in the case of the hybrid mechanism, we obtained a significant improvement for $R^2$ because the average displacement is decreased. In the first round the hybrid technique uses the 5x5 grid but for the following rounds, the hybrid mechanism is able to detect regions with high variability of the variable of interest and break those cells into smaller cells, thus reducing the average displacement in those regions. Not only that, in those cells smaller than $S_{\text{min}}$, the mechanism makes the mobile user to encrypt the data, which contains the real location of the user, which improves the quality of information even further.

C. Energy-consumption results

The experiments to measure the energy consumption of the different mechanisms consisted of the transmission of 100 P-Sense records (i.e., location, timestamp, CO$_2$, CO, air quality, temperature, RH factor, and combustion gases concentration) using Wi-Fi and 3G Networks. Each record is anonymized, obfuscated, or encrypted depending on the selected privacy-preserving mechanism. We used a Hewlett Packard E3631A 0-6V, 5A/0-(+-) 25 V, 1A Triple Output DC Power Supply to provide power to the cell phone, a Samsung SPH-D720 Nexus S with an Android OS Version 4.0.3.

Figure 7 shows the comparison between the different privacy-preserving mechanisms using Wi-Fi. Obfuscation-based techniques, such as the POI and Random Perturbation, present the lowest energy consumption, as they do not require additional transmissions. Tessellation [8] has a higher energy consumption because it needs to establish a connection with the anonymization infrastructure in order to obtain the tile information before sending the report to the data broker. Finally, the double encryption mechanism is shown to be the worst in energy consumption because it needs to establish a SSL channel, which requires the transmission of the key, the double-encryption process, and the transmission of the data.

Figure 7 does not show the hybrid mechanism because the hybrid mechanism has different levels of energy consumption depending on the percentage of records encrypted. As explained before, a record will be encrypted or obfuscated depending on the cell size. So, in Figure 8 we show the energy consumption of the hybrid mechanism for different percentages of data encrypted using Wi-Fi and 3G networks. A 0% means that all records were obfuscated while a 100% means that all records were encrypted. Notice that the larger the percentage of records encrypted, the larger is the difference in energy consumption between Wi-Fi and 3G networks. Comparing Figures 7 and 8, it can be seen how the energy consumption of the hybrid mechanism is lower than Tessellation when the number of records encrypted is equal to 0 and equal to the Encryption mechanism when the percentage of records encrypted is 100% (case WiFi). For other values, the energy consumption of the hybrid mechanism increases linearly with the percentage of records encrypted.

D. Trade-offs

Table II presents the percentage of encrypted records (% Encrypted), the percentage of obfuscated records (% POI), the probability to be located at the data broker ($P_{\text{loc}}$ at DB), application server ($P_{\text{loc}}$ at AS) and both ($P_{\text{loc}}$ at DB + AS), as well as the energy consumption for transmitting 100 P-sense records using Wi-Fi (Wi-Fi) and 3G (3G) networks as we vary the minimum cell size parameter $S_{\text{min}}$. In addition, Figure 9 shows the trend of the quality of information depending on the minimum cell size ($S_{\text{min}}$). Notice that lower
values of $S_{\text{min}}$ mean lower percentages of encrypted records, which is translated into lower energy consumption as well as lower quality of information and higher $P_{\text{loc}}$ at the data broker. On the other hand, higher values of $S_{\text{min}}$ mean higher percentages of encrypted records, which translate into higher energy consumption, higher quality of information, and a higher $P_{\text{loc}}$ at the application server. In conclusion, it is clear that $S_{\text{min}}$ has an important effect on the privacy, quality of information, and energy consumption of the users. Therefore, this parameter can be changed by network administrators to better suit their needs and the needs of the users.

VII. CONCLUSIONS AND FUTURE RESEARCH

This paper presents a hybrid privacy-preserving mechanism that combines anonymization, data obfuscation, and encryption techniques to increase the privacy of the users while improving the quality of information and the energy consumption. The proposed hybrid mechanism includes an algorithm that dynamically changes the cell sizes of the grid of the area of interest according to the variability of the variable of interest being measured and chooses different privacy-preserving mechanisms depending on the size of the cell. Encryption techniques are used in small cells (where users can be identified easier) to protect the privacy of the users while increasing the quality of the information, as the reported location is the real location. Anonymization and data obfuscation techniques are otherwise used in bigger cells where the variability of the variable of interest is low and therefore it is more important to protect the real location of the user. We evaluated our hybrid approach and other privacy-preserving mechanisms using a real PS system for air pollution monitoring. Our experiments show the better performance of the proposed hybrid mechanism and the existing trade-offs in terms of privacy, quality of information to the final user, and energy consumption.

This paper points to two clear directions for further research. The first one relates to the privacy analysis of trajectories, i.e., calculating the privacy of each mechanism given that data is available from various rounds. The second area relates to incentive mechanisms. With our proposed hybrid mechanism, it should be easier to integrate an incentive mechanism and determine the number of samples needed in each cell to provide a good level of privacy, quality of information, and energy consumption.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the work of Yarelis Ares and Omar Aviles, REU students from Universidad del Turabo, Puerto Rico, for implementing the different privacy mechanisms and running the experiments to determine their energy consumption. Their work was supported by the National Science Foundation under grant No. 1062160, an REU Site on Ubiquitous Sensing.

REFERENCES