An Elastic, Privacy-preserving Participatory Sensing Platform System and its Health Care Applications

Guoxing Zhan and Weisong Shi, *Senior Member, IEEE*

***Abstract*—The abundance of daily network-enabled computing devices and smart sensors are enabling participatory sensing applications in various areas including health care. While participatory sensing can greatly benefit the society and individuals, it encounters the obstacle of privacy concern. Considering the potential privacy leakage, the existing participatory sensing systems tend to limit the collected data to its internal use only. Additionally, the existing privacy research often do not take third-party applications into consideration. To conquer the challenge, we propose Woodward, an elastic, privacy-preserving participatory sensing system, using health care applications as an example. Woodward protects the user privacy and facilitates the data sharing with the third-party applications. It adopts an innovative anonymization process that allows high-precision query and impedes privacy attacks by great cost. We implemented Woodward with a health care application and evaluated the query precision and privacy protection quantitatively.**

I. Introduction

An increasing number of network-enabled computing devices permeate our daily lives. Some typical network-enabled consumer devices include smartphones, PDAs, and in-vehicle infotainment systems. In addition to their network capabilities such as WiFi, GPRS and Bluetooth, these devices are often equipped with sensors such as cameras, GPS, accelerometers and environmental monitoring units. Most of the network-enabled devices are "smart" in a sense that they can now run general-purpose application software. The network-enabled computing hardware is approaching a moment in which powerful, low-cost commodity devices will enable a new generation of applications [1]. This trend have laid the foundation for participatory sensing, in which daily network-enabled devices, such as cellular phones, are used to "form interactive, participatory sensor networks that enable public and professional users to gather, analyze and share local knowledge" [2], [3]. A participatory sensing system may be utilized to track commodity price, infer the seat availability in a coffee shop, or collect pollution and traffic readings [4], [5], [6]. Another important category of participatory sensing applications is towards the self-monitoring and self-management of patient health [7], [8]. With the availability of wireless biomedical sensors, a participatory sensing system will be able to collect biophysical data such as heart rate from the patient and deliver feedback accordingly back to the patient [9]. The data collection and the feedback delivery are performed through the computer networks. Such applications can lower the medical cost and make remote diagnoses possible [10].

While participatory sensing can bring great benefit in areas such as health care, there is a rise of concern over privacy leakage [11], [12], [13], [14]. When a user participates in a participatory sensing task, the sensing application could leak his personal information to an adversary. That would discourage the user's involvement in participatory sensing. Unfortunately, much existing work focuses on how to build the infrastructure to enable applications [4], [15], [16], [17] and generally does not take privacy into consideration. Meanwhile, certain participatory sensing platforms [18], [19], [20], [21] tend to limit the use of collected data to internally developed applications only and thus reduces the risk of privacy leakage. The restriction of the internal use of data prevents third-party applications from exploring the data and eliminates the benefit of data sharing. Further, the existing privacy research mainly concerns itself about the mechanisms to identify and prevent privacy issues [22], [23], [24], [25], [26], [27], [28], [29], [30] and tends to not support third-party applications.

To conquer the challenge, we propose Woodward, an elastic, privacy-preserving participatory sensing system. Woodward protects the user privacy and facilitates the data sharing with the third-party applications. It adopts a innovative anonymization process that allows high-precision query and impedes privacy attacks by great cost. We implemented Woodward with a health care application and evaluated the query precision and privacy protection quantitatively.

The rest of the paper is organized as follows: we give an overview of the Woodward system in Section II; the design of the Woodward server is described in Section III; the implementation of Woodward and its health care application is given in Section IV; the empirical evaluation is in Section V; the related work, future work and the conclusion are presented in Section VI and Section VII.

II. System Overview

Woodward is an elastic, privacy-preserving system to facilitate participatory sensing on network-enabled computing devices. This system allows arbitrary third-party applications to perform various query from third-party applications. Possible applications with Woodward may include health care, traffic analysis, events report, environmental monitoring, image search, and crowd analysis. We use self-monitoring and self-management of patient health as an exemplary type of applications to illustrate the system. As shown in Fig. 1, in

G. Zhan is a Ph.D. candidate in the Department of Computer Science, Wayne State University, Detroit, MI, 48202.
E-mail: gxzhan@wayne.edu

W. Shi is an Associate Professor of Computer Science at Wayne State University.
E-mail: weisong@wayne.edu

this Woodward system, a user utilizes his network-enabled handheld device (e.g., smartphone) and a few sensors including bio-medical sensors to collect healthcare-related data and send it to a central server - the Woodward server. The sensors are either integrated into the handheld device or connected wirelessly (e.g., via Bluetooth). The Woodward server stores the data, validates the data, anonymizes the data for privacy protection, and interact with the users and arbitrary third-party applications. The third-party applications can only access the anonymized data on the Woodward server and can submit health status feedback for a record accessed to the Woodward server. The Woodward server then delivers the feedback to the designated user. For another type of applications other than health care, the information flow is still the same; only the sensors and the applications are replaced accordingly.

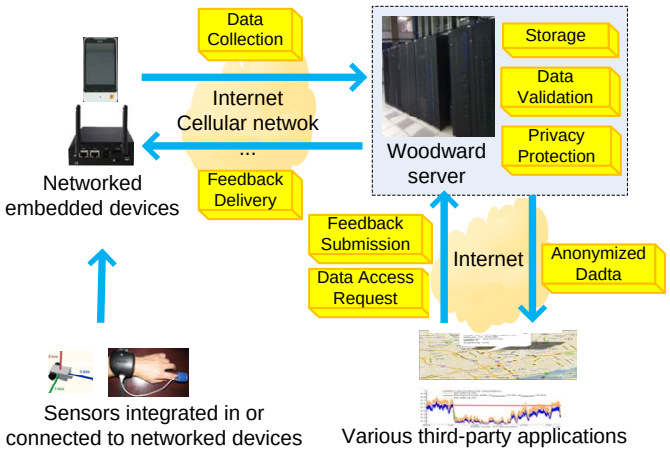

Fig. 1. The design of Woodward.

Thus, the Woodward system consists of three components (Fig. 1): the users submitting the data with network-enabled devices and sensors; arbitrary third-party applications; and the core component - the Woodward server. The third-party applications do not retrieve the data from the user directly; instead, all the data are sent to the Woodward server and the applications are only allowed to access the anonymized data from the Woodward server. The data flow is illustrated by Fig. 2. This requirement is crucial for the protection of the user's privacy and the reuse of data. If a third-party application directly accesses a user's original data, the user privacy is hardly guaranteed. On the other direction, the feedback flow is shown in Fig. 3. a third-party application does not directly deliver its generated feedback to a user because the application is not supposed to know the user's contact information due to the privacy requirement. Instead, the application submits the feedback for an anonymized record to the Woodward server first; the Woodward server then internally maps that anonymized subject of that feedback onto its true identity and delivers the feedback to the user.

The application should be aware that the data have gone through the anonymization process that adds noise to the original data. The anonymization process guarantees that any statistical query, including percentile query of any single value, will be highly precise. A statistical query concerns the

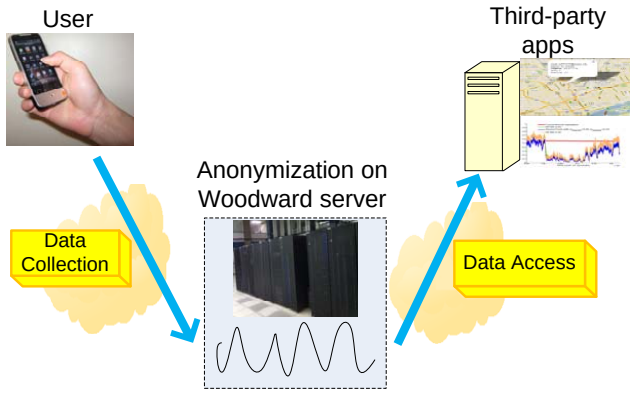

Fig. 2. The flow of user data.

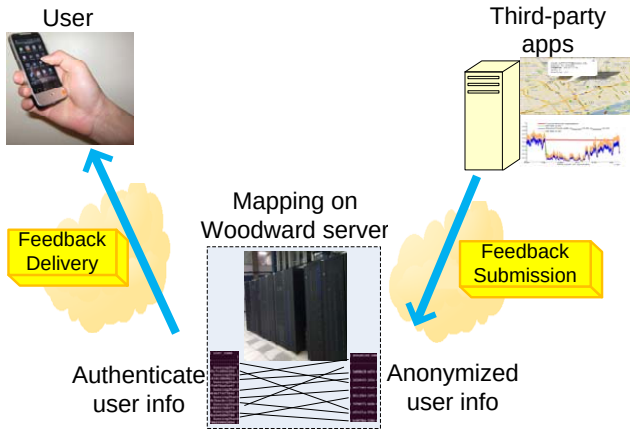

Fig. 3. The feedback flow.

statistical features that are based on the probability distribution of the data. Additionally, and importantly, for common values that occur frequently, the noises are small; for values that occur rarely, the noise can be large. Note that on the one hand, rare values, if exposed with only small noise, have a good chance of being traced by malicious attackers. On the other hand, common values are safe to be exposed with small noise added. The density of a neighborhood of a value can be decided either from a published result from the system or simply from performing a query; the range of the possible noise can also be determined similarly. Generally, most values fall into moderately densely populated areas and the anonymized data closely resemble the original data.

III. The Design of the Woodward Server

The Woodward server is designed to store received data, perform data validation, provide elastic application query interface and user feedback, and protect the user privacy. The design also aims to provide high-precision query answers and moderate performance. Before considering any other issues, we first want to present our approach to protect the user privacy as it closely relates to the rest of the system.

A. Privacy Protection at the Woodward Server

To protect the user privacy, we first state the privacy threat model. The Woodward server does not present the true

identity or contact information of a user to an application and an attacker will not be able to take advantage of any identity or contact information. However, based on certain prior knowledge about a particular user, the attacker may attempt to identify a certain anonymized record owned by that user. For example, the attacker might happen to that a user named Alice has an unusually high heart rate, 190 bpm. Then the attacker might search through all the anonymized data exposed by the Woodward server. If the data anonymization is not performed properly, the attacker can be lucky to find out some anonymized record with a heart rate reading of 192 bmp and sees no other anonymized record could better fit Alice. Thus, this attacker just identified an anonymized record belong to Alice. If that same record also contains other information (e.g., age) the attacker is interested in, the attacker could access that information and start harmful activities. Therefore, under this privacy threat model, with prior knowledge of an attribute (or multiple attributes) about a user, an attacker attempts to link at least one anonymized record to that user and exploit that record for other private information of the user. The current design of the anonymization process mainly assumes that an attacker only has the prior knowledge of a single attribute of a user though dealing with prior knowledge of multiple users can be achieved with a similar but more complicated scheme.

The Woodward server performs the anonymization process on the received data; and the application query request is executed against the anonymized data. The anonymization uses different schemes according to the types of the data. Our major interest here is the numeric biophysical data of the user (e.g., heart rate). Before that, we will first describe the anonymization for other types of data. For identity information such as the name and email address, the server maintains a secret one-to-one mapping that maps each identity into a unique meaningless symbol (e.g., a byte string). The mapping is maintained in such a way that it is impossible for a third-party to reverse the map to find out the original identity corresponding to an anonymized symbol. The reason that a symbol is still needed is that in a relation-entity database query often needs to know if two records are associated with the same user or not. For discrete attributes with only finite possible values, the data value is generally directly exposed to the applications unless the user indicates that the data should not be exposed; in the latter case, an "unknown" value will replace the original value for the application query. Generally, for each value out of the finite set, there can be a large number of users having that value; thus, that information is usually not sensitive. For text or binary data, the data is either completely hidden from the query or exposed to the application query, as specified the user. Regarding location data, the Woodward server anonymizes the exact location to a city-magnitude area. Though it is possible to exploit the existing approaches for location anonymization [31], [32], [33], [34], for our purpose with health care information, we are satisfied with this simple scheme.

We categorize the data as follows: 1. identity information such as the name and email address; 2. data attributes that can take values from a small finite set, e.g., the gender attribute with three possible values - "male", "female", and "unknown"; 3. text or binary information such as memo, images, and audio; 4. data of numerical value or data that can be transformed to continuous numeric value without losing information, such as heart rate, and location represented in latitude and longitude. This last category of data is usually most interesting to the applications. For the anonymization purpose the user address information is regarded as being of this category instead of the category of the user identity. The address is represented internally as a latitude-longitude pair.

For the first category - identity information, the server maintains a secret one-to-one mapping that maps each identity into a unique meaningless symbol (e.g., a byte string). The mapping is maintained in such a way that it is impossible for a third-party to reverse the map to find out the original identity corresponding to an anonymized symbol. The reason that a symbol is still needed is that query often needs to know if two records are associated with the same user or not. For the second category - discrete attributes with only finite possible values, the data value is generally directly exposed to the applications unless the user indicates that the data should not be exposed. If the user prefers that this data value should be kept secret, then an "unknown" value will replace the original value for the application query. For a database with a great amount of anonymous records (with identity anonymized), exposing the data attribute of this category is normally not a concern: a considerable portion of anonymous users usually share the same value from the finite set of all possible values. This is comparable to exposing exposing your gender on an anonymous questionnaire: there is little chance of privacy leakage since it is filled out anonymously. For the third category - text or binary information, the data is either completely hidden from the query or exposed to the application query. The users who upload text or binary information such as images for their own memory only will have that information completely unknown to the third-party application. However, when a user is willing to share a particular image and its title with everybody else, the Woodward server will have that information completely exposed to the application query. For the numeric biophysical data of the user, we need to take extra care to perform anonymization. These data are greatly valued by many applications. But exposing them directly (even without any explicit identity information) can be exploited by attackers with certain prior knowledge, as described in our privacy threat model.

Generally, if given the direct access to the original data of numerical values, an attacker can exploit the known rare-feature data or the known rare combination of multiple-feature data about a particular user and submit query request to obtain other unknown features of the user that interest the attacker. The attack may work because an anonymized symbol representing the user is associated with each user's record due to the need to identify whether multiple records are from the same user. The Woodward server maintains that anonymized symbol because it is important for a number of applications to associate multiple records with the same anonymous user. As an example of such applications, an application is to examine the hypothesis that a driver of higher heart rate tend to have higher accident rate. Suppose that the heart rate

and the accident rate are stored in two separate tables of a database. Then this application wouldn't be able to proceed unless given a clue as to whether a heart rate and a accident rate are from the same anonymous user. With a variety of applications, the best bet is to provide at least an anonymous symbol represent the user associated with a record. The goal of the anonymization process for the numeric biophysical data is to add maximal noise to the numeric data while maintaining high precision for the query. Observe that a statistical query regarding a numeric attribute usually can be broken down to arithmetic operations involving the following aspects: 1. relative frequency of values satisfying certain numeric inequality conditions; 2. numeric order of two attribute values; 3. numeric magnitude of the attribute value. As long as these aspects are kept "close" enough to their authenticate values, the high precision is maintained for the statistical query. Thus, we establish the following principles for anonymizing the numeric attribute. Let x be the random variable representing the numeric attribute; $\widetilde{x}$ be the random variable representing the anonymized version of x. Then the anonymized data should satisfy the following conditions: (a). Given any real number a, the difference of the probability (relative frequency) $|P(x < a) - P(\widetilde{x} < a)|$ must always not exceed a pre-set threshold; given any value d, $|P(x < d) - P(\widetilde{x} < \widetilde{d})|$ should not exceed the same threshold either. (b). For any values x_1, x_2, and their anonymized values $\widetilde{x}_1$, $\widetilde{x}_2$, the same order relation between x_1 and x_2 should very likely hold between $\widetilde{x}_1$ and $\widetilde{x}_2$ with a relaxed order relation. An anonymization-aware query should keep in mind that only a relaxed order but not a strict order is maintained, in a probabilistic sense. (c). Though $\widetilde{x}$ may occasionally deviate a lot from x, that should not occur so often that it affect certain statistical features such as variance. The condition (b) helps maintain the precision of a query involving the condition of two random variable comparison such as "heart_rate1 $>$ heart_rate2". It is crucial that a query involving comparison of numeric values takes anonymization into consideration and relax the strict inequality condition. Meanwhile, the conditions (a) and (c) together help maintain the precision of a general statistical value of the form $\int_S g(x) \cdot f(x) dx$, where $f(x)$ is the probability density function of the random variable x, $g(x)$ is the random numeric function the application is interested in, and S is the measurable set on which the integration is applied. The condition (a) indicates that the probability distribution of the anonymized data will be very close to that of the authenticate data. It also indicates that an observation of a user's anonymized data relative to others reflect almost the same relative situation as the authentic data. That means a third-party application can take advantage of a user's relative comparison to customize a feedback or recommendation for the user. That feedback or recommendation can be delivered through the Woodward server without the application's awareness of the user's contact. The condition (c) reflects that the the change of $g(x)$ is well limited after anonymization. Theoretically, the well-known Schwarz's Inequality $\int_S g(x) \cdot f(x) dx$ helps establish a loose upper bound on the deviation due to anonymization:

$$|\int_S \psi_1(x) \cdot \psi_2(x) dx| <= \sqrt{\int_S \psi_1(x)^2 dx} \cdot \sqrt{\int_S \psi_2(x)^2 dx}$$

To see this, let $\widetilde{f}(x)$ be the probability density function for the anonymized data. Then the deviation of the statistic is

$$\begin{aligned} & |\int_S g(x) \cdot \widetilde{f}(x) dx - \int_S g(x) \cdot f(x) dx| \\ = \; & |\int_S g(x) \cdot (\widetilde{f}(x) - f(x)) dx| \\ <= \; & \sqrt{\int_S g(x)^2 dx} \cdot \sqrt{\int_S (\widetilde{f}(x) - f(x))^2 dx} \end{aligned}$$

Before stating our algorithm of anonymizing the fourth category of numeric data, we would like to briefly describe the logic behind the algorithm. The algorithm achieves anonymization through adding certain noises to the data and it essentially depends on how to wisely decide the magnitude of the noise according to the specific data values. If there are K such candidates, then the attack would not succeed as long as K is large enough. This shares certain common features with the K-anonymity mechanism; however, differently, our scheme applies a random noise to each data value and ensure that there is a chance that $K - 1$ candidates exist to possibly be anonymized to a value retrieved by a malicious query. On the other, generally, the random noise should not be too large to cause an intolerable error to a query concerning the data magnitude. Considering the trade-off, for the sparse neighborhood, we select the random noise in a way that its magnitude will be reasonably small with a large likelihood but can be as high as the size of the sparse neighborhood with a small likelihood.

Fig. 16 and Fig. 5 visualize the idea. 10000 random values (original data) are generated between 50 and 220, based on the normal distribution. The anonymized process is applied to get the anonymized data. Fig. 16 shows the histograms of the the original data and the anonymized data. As indicated by

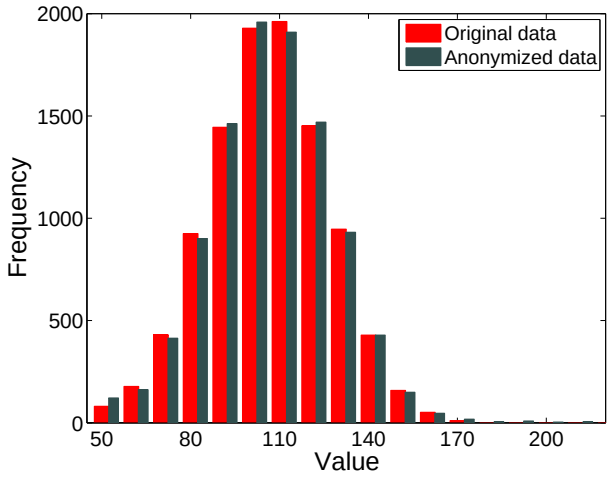

Fig. 4. Histograms of original and anonymized data with normal distribution.

Fig. 16, the original and the anonymized data show similar frequency distribution. Fig. 5 compares the original data and the anonymized data more closely. For each pair of (original value, anonymized value), a point is plotted. For data falling into the intense interval [80, 140] (Fig. 5), the anonymized values show very limited deviation from their original values. By contrast, for the sparse data out of that interval, the deviation between the original and the anonymized values can be as large as 60 (Fig. 5).

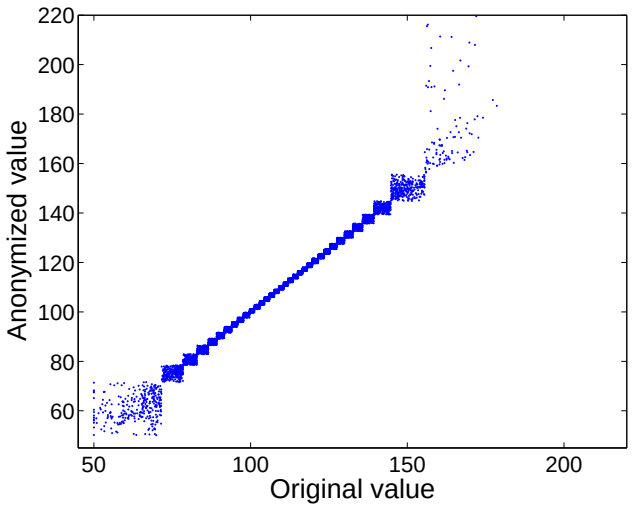

Fig. 5. Comparison of anonymized data and original data with normal distribution.

B. Anonymization on Numeric Biophysical Data

Now, we describe our anonymization algorithm of the numeric biophysical data as follows. Assume the size of the database has been large enough; otherwise, a query against a database with a few records will not be permitted due to privacy concern. The algorithm first divides the range of the numeric data into a series of contiguous neighborhood in the form of open or half-open intervals $(-\infty, I_0), [I_0, I_1), [I_1, I_2), [I_2, I_3), ..., [I_n, +\infty)$. Each interval contains a similar number of data values occurrence, with a value of multiple occurrence counted multiple times; the exceptions happen around those values that occur more than a few times. we denote that common relative frequency of almost all the neighborhoods (intervals) as RFNBH, i.e., the ratio of the data occurrence in that neighborhood to the total data occurrence. Thus, a neighborhood of a small size is denser than a neihgborhood of a greater size. The larger a neighborhood is, the sparser it is. After the neighborhood division, the anonymization process is applied to each existing data and incoming data according to the neighborhood they fall in. Each anonymized data value should still fall in the same neighborhood as its original neighborhood; it indicates that the difference of cumulative relative frequency between the original and the anonymized data should not exceed RFNBH - the common relative frequency of a neighborhood. In other words, the difference between the statistical distribution of the original and the anonymized data should not exceed RFNBH. The magnitude of the random noise applied is decided in a way that is almost proportional to the neighborhood size in a probabilistic sense. In other words, the sparser the neighborhood is, the higher noise is likely to be applied to the data values there. Importantly, the randomness of the noise applied discourages an attacker by presenting a whole neighborhood of candidate values to a malicious query targeting a particular user. There is a trade-off on the value of RFNBH: the smaller RFNBH is, the better query precision the anonymization maintains and the more privacy risk there is. A specific algorithm is given in Algorithm 1, with subprocedures in Function 2 and Function 3. Table I summarizes the notations and samples parameters used.

TABLE I
MAJOR NOTATIONS USED AND PARAMETERS WITH SAMPLE VALUES IN PARENTHESES.

Symbol	Meaning
X	An attribute.
x	An existing or incoming data value of attribute X.
$\widetilde{x}$	An anonymized value of x.
LBOUND(50)	the lower bound of the attribute value.
UBOUND(220)	the upper bound of the attribute value.
NbhList	List of neighborhoods covering the range of attribute X.
TOTAL	Total frequency of data occurrences.
RFNBH(0.03)	The common relative frequency of the neighborhoods
NZRT(0.5)	Ratio of noise magnitude over neighborhood size.
NZCF(0.7)	The confidence that noise falls into a major interval.
NZTH(10)	Maximal noise threshold.

Algorithm 1 For a numeric-valued attribute X in a database, anonymize its values.

procedure ANONYMIZE(attribute X)
  NbhList=NEIGHBORHOODDIVISION(X);
  for each existing/incoming value x of attribute X **do**
    identify its neighborhood $[I_{left}, I_{right})$ in *NbhList*;
    store $\widetilde{x}$=ANONYMIZE(x, $[I_{left}, I_{right})$);
  end for
end procedure

To defend against tough privacy attackers exploiting the fixed neighborhood, the neighborhood (interval) division can be run periodically with the subsequent noise-based anonymization. We suggest using a moderately long period so as not to overwhelm the computing resource.

C. Other Design Aspects

The Woodward server stores the anonymized data and allows an authorized third-party application to request database query on the anonymized database. The way how an application can send its query request mimics the way a user accesses a regular remote database: besides typical database privilege authorization, no other restriction applies. This gives the application the maximal freedom.

Regarding the data storage, the Woodward server stores the original and the anonymized data onto separate storage units. For any new data, an online anonymization is performed; the original and the anonymized data are then stored separately.

Function 2 Divide the attribute range into a series of contiguous neighborhood and return the neighborhood list.

function NEIGHBORHOODDIVISION(attribute X)
  NbhList={} ▷ List of neighborhoods
  $I_{right} = I_{left} =$ *LBOUND*;
  while $I_{left} <$ *UBOUND* **do**
    NumOfData $= 0$;
    while *NumOfData*$<$ *RFNBH*$*$*TOTAL* AND $I_{right} <$ *UBOUND* **do**
      NumOfData += occurrence frequency of I_{right};
      $I_{right} = min\{$*UBOUND*$, x | x > I_{right}\}$;
    end while
    if $I_{right} <$ *UBOUND* **then**
      Add $[I_{left}, I_{right})$ onto *NbhList*;
    else
      Add $[I_{left}, I_{right}]$ onto *NbhList*;
    end if
    $I_{left} = I_{right}$;
  end while
  return *NbhList*;
end function

Function 3 Return the anonymized value of x within its neighborhood $[I_{left}, I_{right})$.

function ANONYMIZE(Data x, $[I_{left}, I_{right})$)
  ALeft $= max(x -$ *NZTH*$, x -$ *NZRT* $* (x - I_{left}))$;
  ARight $= min(x +$ *NZTH*$, x +$ *NZRT* $* (I_{right} - x))$;
  XFR = occurrence relative frequency of x;
  ALeft $= x - (x -$ *ALeft*$)/(100 *$ *XFR* $+ 1)$;
  ARight $= x + ($*ARight* $- x)/(100 *$ *XFR* $+ 1)$;
  P=*NZCF* in the interval [*ALeft*, *ARight*);
  1-*NZCF* in $[I_{left},$ *ALeft*$)$, [*ARight*, $I_{right})$;
  Randomly get a value as $\widetilde{x}$ according the the above probabilistic distribution.
  return $\widetilde{x}$;
end function

For large data, we may use RAID or even a cluster database with shared-nothing structure.

For data submitted from an anonymous user, a validation process is performed to protect the database from pollution by erroneous data. To detect abnormal data like heart rate that is too high, the value is checked against the statistical distribution of the existing data. Further, we can take advantage of the source reputation for data validation. The specific valid approach can be found in our previous work [35].

IV. IMPLEMENTATION

We developed a user client program on a HTC Legend Android phone for data collection and feedback retrieval. As illustrated by Fig. 6, the Android client program the heart rate is wirelessly read from a Nonin 's Bluetooth-enabled sensor Avant 4100 worn around the user's wrist. In addition to the heart rate, the Android client program also collect location data with the internal GPS and get user input for a questionnaire about the user information such as age and email. All the collected data is sent to the Woodward server via WiFi. The same Android client program also displays the feedback generated by applications once it is available. The client exchanges messages with the Woodward server using XML. For the sake of security, all the network communication is protected using Secure Socket Layer (SSL).

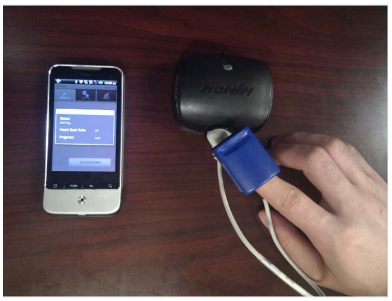

Fig. 6. The client program on an Android phone.

A. The Woodward Server

The Woodward server we developed stores data sent from the user into a MySQL database. An authenticate copy and an anonymized copy are stored in separate databases. The anonymization on a numeric attribute is performed according to Algorithm 1. When the server starts, it first prepares for the anonymization by performing the neighborhood division. Then, whenever a new records arrives, the server performs the online anonymization based on the outcome of the neighborhood division, with a small overhead. The server stores the user's identity information and his data. For security, only the hash of the user's password is stored in the authentic copy. The server maintains a one-to-one map between all true user names and their anonymized names. The anonymized names are generated from a secure random string generator that guarantees universal uniqueness. A sample name mapping is illustrated by Fig. 7. The server allows any third-party

user_name	anonymized_name
GuoxingZhan1	2d1b8425-2d3a-43c5-9234-b318c38b0623
GuoxingZhan8	42f1409a-1814-474a-8359-f98d3909289d
GuoxingZhan	5a098e28-edf4-4c0b-a568-d5cfcd95e395
GuoxingZhan4	601c35b4-2df6-4cc9-8e0f-463bff4ff642
GuoxingZhan2	b83f7081-01d9-4eb2-92bd-2fe0f0ad1e47
GuoxingZhan7	bc0078ba-538d-4ac3-84cb-dba5e6d8df88
GuoxingZhan6	e97e37ca-9fbe-44c6-b823-d409227b84f5
GuoxingZhan5	f4f66f71-bd3b-4a74-b3b9-8b7bac8cc332

Fig. 7. The one-to-one name mapping between true user names and anonymized user names.

application to perform read-only access to the anonymized copy with SQL. It also accepts the feedback an application generates towards a user and delivers the feedback to the user. The third-party application is not allowed to directly access the authentic copy. For the feedback, the application specifies the anonymized name of the user and the server maps that to the true user.

We created a sample third-party application that informs certain users of potential cardiovascular disease according to their heart rate readings. Specially, whenever a user is found to have a heart rate of at least 97% percentile among the group of user similar to his age [36], the application submits such feedback to the Woodward server targeting that user: "Your heart rate appears to be considerably higher than your peers. That reveals a certain risk of cardiovascular diseases. If you are interested in more details or need subscription service, please contact our eHealth group at xxx-xxx-xxxx."

V. Evaluation of Privacy Protection and Query Accuracy

With the data anonymization process, the woodward server presents the anonymized data to the third-party applications. For the anonymization process applied to a numeric attribute, we will evaluate the effectiveness of privacy protection and query accuracy. We generated a series of random values between 50 and 220 and applied the anonymization process according to Algorithm 1 and the parameters from Table I. The original random data were generated based on one of the following seven statistical distributions: uniform distribution, binomial distribution, normal distribution, Poisson distribution, chi-Squared distribution, Weibull distribution, and exponential distribution. For each distribution, 10,000 random values were generated as a complete set. Table II summarizes the statistical characteristics of the original random data, including the value range, the average, and the standard deviation. Table III summarizes the same statistics for the corresponding anonymized data. According to these statistics, the average and the standard deviation has very limited differences between the original and the anonymized data. On the other hand, except for the uniform distribution, the anonymization tends to enlarge the range of the data by various sizes. To explain it, note that the data are sparsely distributed at either the left or the right end of the original range, except for the uniform distribution. According to the anonymization process, the sparse areas tend to get larger noises. The larger noises at either end of the original range result in the enlarged range of the anonymized data. That effect is visualized in Fig. 5 and again in Fig. 8. Similar to Fig. 5, Fig. 8 plots each pair of (original data, anonymized data) as a point. The further a point is away from the diagonal line in the figure, the larger deviation there is.

TABLE II
Original random data.

Random data	Range	Average	Std Dev
Uniform distribution	50.02—220.00	135.26	49.12
Binomial distribution	82.00—136.00	110.10	7.39
Normal distribution	50.00—178.56	109.91	20.02
Poisson distribution	85.00—139.00	110.01	7.73
Chi-Squared distribution	51.26—210.28	85.04	18.89
Weibull distribution	55.52—154.27	104.44	15.06
Exponential distribution	50.00—180.93	62.10	12.00

Despite the differences in the range, overall, the empirical distributions of the original data and the anonymized data show very limited differences. We have seen the small differences of the frequency histograms for the normal distribution in Fig. 16.

TABLE III
Anonymized data.

Anonymized data	Range	Average	Std Dev
Uniform distribution	50.00—219.95	135.30	49.15
Binomial distribution	50.08—219.12	110.29	9.89
Normal distribution	50.15—219.56	109.96	20.58
Poisson distribution	50.18—215.59	109.93	9.36
Chi-Squared distribution	50.05—219.82	85.19	19.71
Weibull distribution	50.10—219.98	104.45	16.16
Exponential distribution	50.00—213.00	62.35	13.36

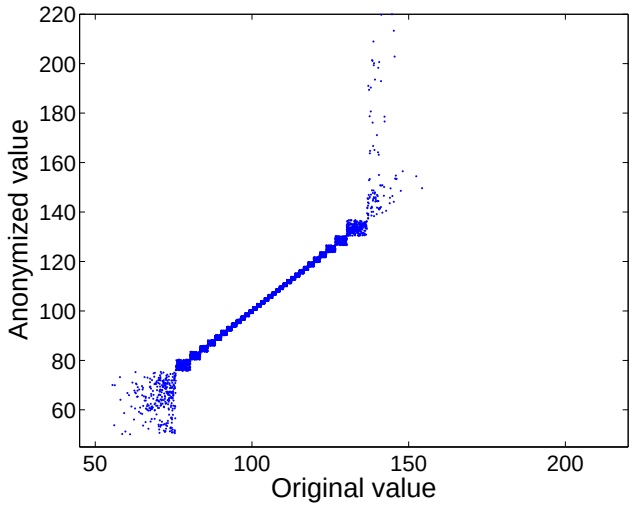

Fig. 8. Comparison of anonymized data and original data with Weibull distribution.

Again, Fig. 19 shows the very small deviation in the frequency histograms for the Weibull distribution. Additionally, their empirical cumulative distribution displays an almost perfect match (Fig. 10).

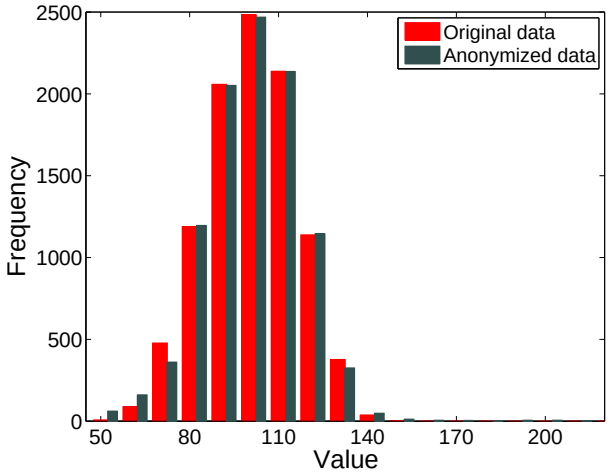

Fig. 9. Histograms of original and anonymized data with Weibull distribution.

The noise (i.e., the difference between an original value and its anonymized value) can vary, from a small scale to a very large scale. But on average, the noise tends to be small. Table IV summarizes the magnitude of the noise. Though the noise can range from 0 to 108, the average noise is no more

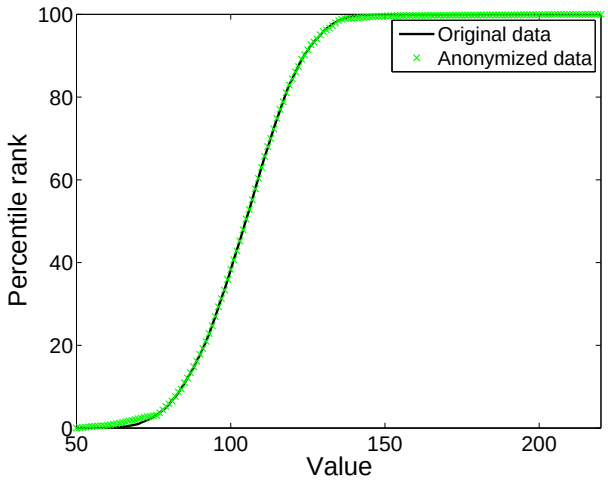

Fig. 10. Comparison of empirical CDF between anonymized data and original data with Weibull distribution.

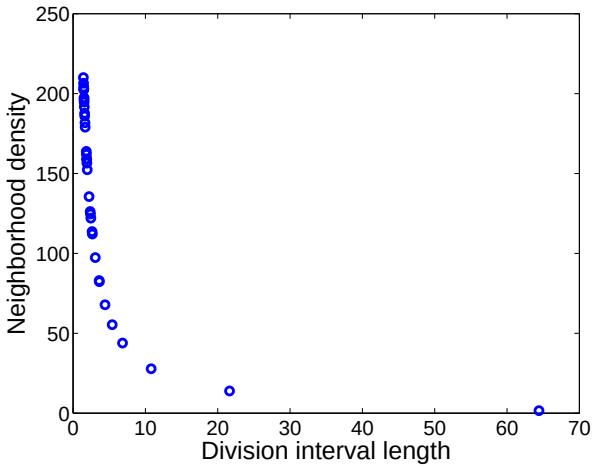

Fig. 11. Neighborhood density against division interval size for data with normal distribution.

than 1.4, with its standard deviation less than 5.

TABLE IV
NOISE MAGNITUDE

Noise magnitude	Range	Average	Std Dev
Uniform distribution	0.00—5.54	1.35	1.10
Binomial distribution	0.00—90.66	0.83	4.75
Normal distribution	0.00—59.63	0.95	2.33
Poisson distribution	0.00—79.57	0.71	3.50
Chi-Squared distribution	0.00—66.17	0.90	2.77
Weibull distribution	0.00—78.49	0.88	3.18
Exponential distribution	0.00—108.13	0.63	3.56

The noise is closely related the size of the division interval that the original value falls in. Roughly, the noise tends to be small for short intervals and greater for longer intervals. Table V summarizes the length of the division intervals. The length can vary from 0.3 to 116, corresponding to the various densities. As illustrated by Fig. 11, the longer the division interval is, the sparse the neighborhood is. The noise, on average, is roughly proportional to the length the division interval. Table VI summarizes the ratio of noise magnitude to division interval length. Though that ratio can vary from 0 to 1, generally, its average is from 0.18 to 0.27.

TABLE V
LENGTH OF DIVISION INTERVALS USED FOR ANONYMIZATION

Interval length	Range	Average	Std Dev
Uniform distribution	1.66—5.84	5.00	0.68
Binomial distribution	1.00—93.00	7.08	20.54
Normal distribution	1.43—64.42	5.00	11.15
Poisson distribution	1.00—88.00	7.39	19.99
Chi-Squared distribution	1.11—76.56	5.00	13.10
Weibull distribution	1.09—83.28	5.00	14.47
Exponential distribution	0.34—115.97	5.00	19.80

The percentile query shows that it is moderately accurate to use the anonymized data for estimating the percentile of an original value. Roughly, the percentile difference should have 100**RFNBH* as its upper threshold, where *RFNBH* is the common relative frequency of the neighborhoods. That corresponds to how the anonymization process divides the intervals

TABLE VI
RATIO OF NOISE MAGNITUDE TO DIVISION INTERVAL LENGTH

Noise/division interval	Range	Average	Std Dev
Uniform distribution	0.0000125—0.99	0.27	0.22
Binomial distribution	0.0000038—1.00	0.19	0.26
Normal distribution	0.0000141—0.99	0.26	0.21
Poisson distribution	0.0000036—1.00	0.18	0.25
Chi-Squared distribution	0.0000666—1.00	0.26	0.21
Weibull distribution	0.0000171—0.99	0.27	0.21
Exponential distribution	0.0000066—0.99	0.26	0.21

and adds noise. In the meantime, for discrete-valued numeric data (i.e., integers), the percentile difference may exceed 100**RFNBH* because of the biased noise introduced by the discreteness. If it is allowed to use non-discrete anonymized values, we may well control the percentile difference under that upper threshold with the following anonaymization process: first apply uniform tiny noise to the original data, then apply the original anonymization process to the data with tiny noise. The reason of applying tiny noise first is to break the clustering of discrete values and facilitate the splitting of the domain into division intervals (discrete values tend to cluster onto a few values). Table VII lists the percentile rank difference between the original data and the anonymized data for each distribution. Except for the two discrete distribution (binomial distribution and Poisson distribution), the data of all other distribution has a percentile rank difference between -3 and 3, which matches 100**RFNBH* (*RFNBH*=0.03). Additionally, the latter has a 0 differen on average.

TABLE VII
PERCENTILE QUERY ACCURACY

Percentile rank difference	Range	Average	Std Dev
Uniform distribution	-3—3	0.00	1.06
Binomial distribution	-6—4	-1.90	1.49
Normal distribution	-3—3	-0.00	1.05
Poisson distribution	-7—4	-1.83	1.53
Chi-Squared distribution	-3—3	-0.00	1.06
Weibull distribution	-3—3	-0.00	1.07
Exponential distribution	-3—3	-0.00	1.06

Finally, our anonymization process highly protects the user

privacy and discourages an attacker by the anonymized data. To quantify the efforts that the attacker needs to maliciously identify a user, for each numeric record, we define the attack cost as the number of records that falls between the original value and the anonymized value. Intuitively, the attack cost reflects the minimum number of records to check starting with the original value and before coming across the anonymized value. The attack knowing the original value from a certain user would have to examine through at least all those records before find out the corresponding anonymized record. This attack cost is the minimum cost that impedes the privacy attack and thus a very conservative estimation. The actual cost can be much high since an attacker can never be sure of the exact number of records falling between the original value and the anonymized value. The higher the attack cost is, the better our anonymization process protects the privacy. The attack cost for each original value can vary and is independent of the division interval it falls into. The attack cost roughly reflects the frequency of data falling into the corresponding division interval. Fig. 12 illustrates the attack cost for each value from a normal-distributed data set with 10,000 records. Visually, the attack cost is independent of where a trues lies. A value around the left end around 50 can have as a high attack cost as a value in the middle. Table VIII summarizes statistics of the attack cost for each distribution. Not only does the attack cost vary a lot, it also has a high value (86–197) on average. Fig. 13 illustrates the empirical cumulative distribution of the attack cost for the normal distribution data. The figure reveals: with a likelihood of 60%, the attack cost is at least 50; with a likelihood of 33%, the attack cost is at least 100; with a likelihood of 8%, the attack cost is at least 200.

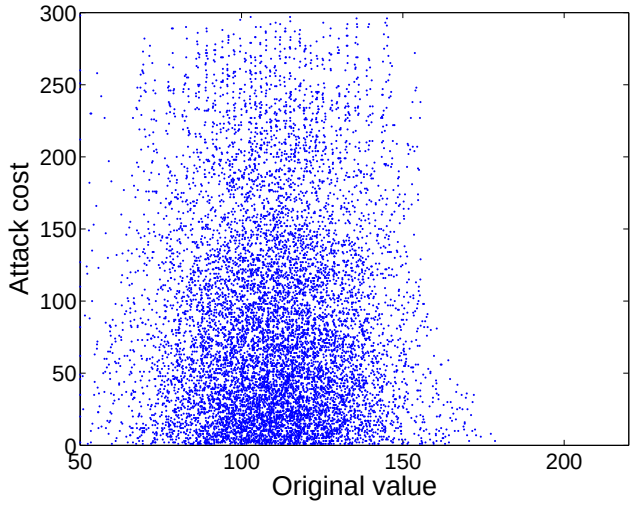

Fig. 12. Attack cost of identifying record from anonymized normal-distributed data with known original value.

VI. Related Work

The privacy leakage has arisen as one major concern involved in participatory sensing [11], [12], [13], [14]. The privacy attack comes in various forms. Besides the direct data theft, an attacker my attempt to identify a user or his activity either explicitly or implicitly by the user's usage of the computing hardware or software, such as IP/MAC addresses, usage pattern and device fingerprinting [37], [38], [39], [40]. The attacker may also attempt to analyze the data pattern [41], [42], [43], infer the user context [44], [45]

The existing research has explored the privacy protection with diverse approaches. Generally, these approaches fall into one of the following categories [41]: regulatory rules, privacy policies, anonymity, and obfuscation. The regulatory

TABLE VIII
Attack cost of identifying record from anonymized data with known original value

Attack cost	Range	Average	Std Dev
Uniform distribution	1—300	87.31	70.80
Binomial distribution	1—530	196.69	140.62
Normal distribution	1—298	86.45	70.32
Poisson distribution	1—671	189.58	141.51
Chi-Squared distribution	1—300	87.67	70.22
Weibull distribution	1—299	88.95	71.37
Exponential distribution	1—299	87.26	70.99

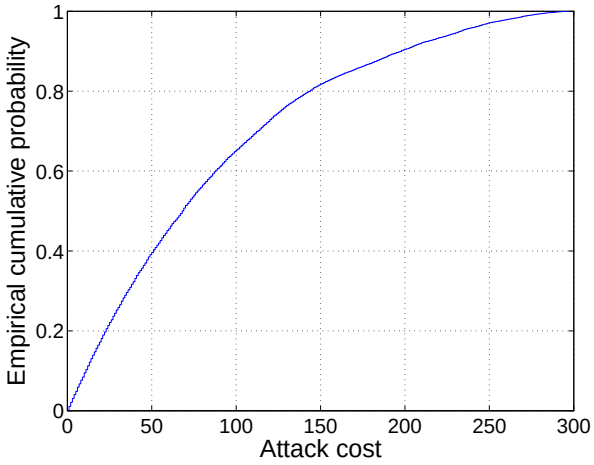

Fig. 13. Empirical CDF of attack cost of identifying record from normal-distributed anonymized data with known original value.

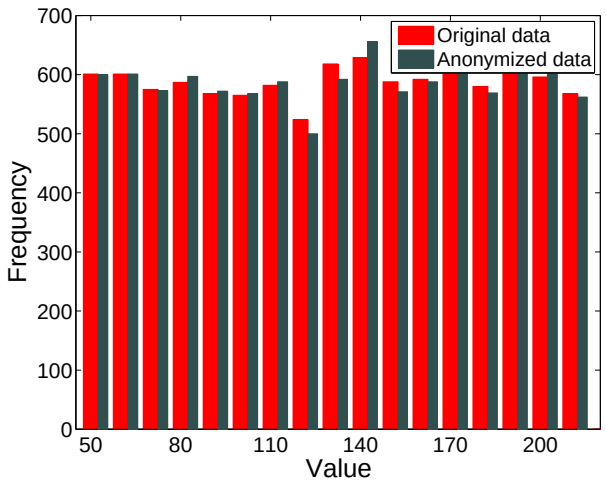

Fig. 14. Histograms of original and anonymized data with uniform distribution.

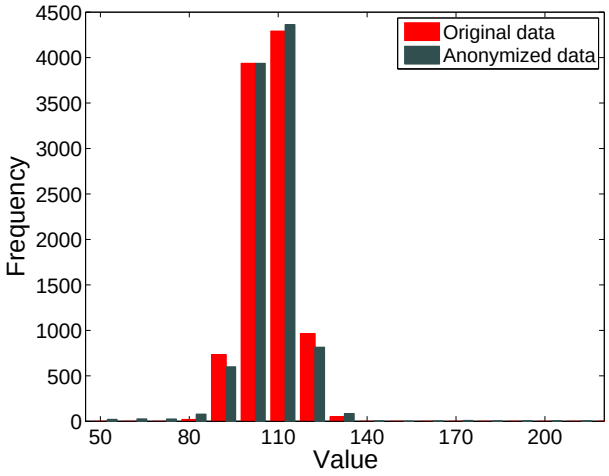

Fig. 15. Histograms of original and anonymized data with binomial distribution.

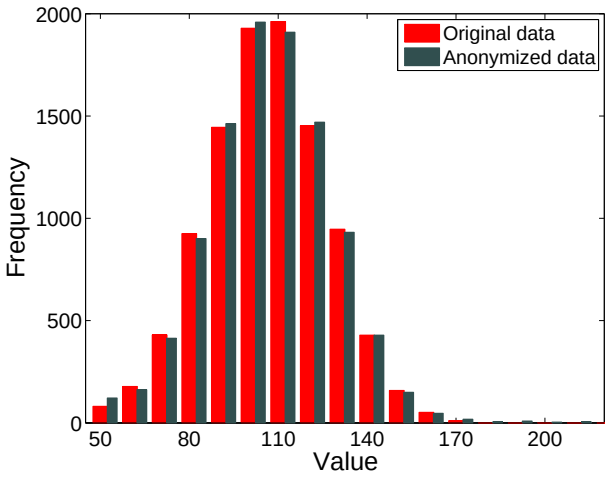

Fig. 16. Histograms of original and anonymized data with normal distribution.

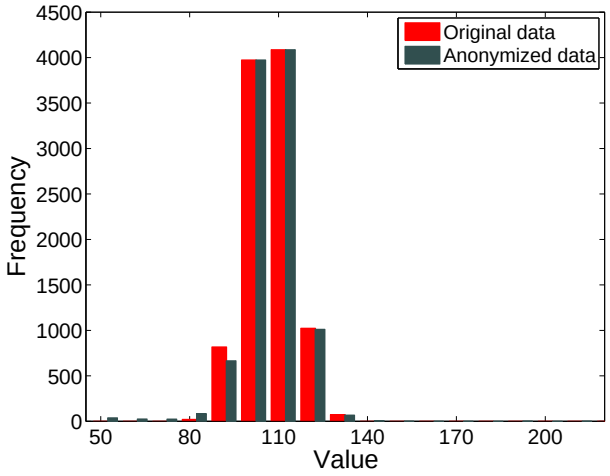

Fig. 17. Histograms of original and anonymized data with Poisson distribution.

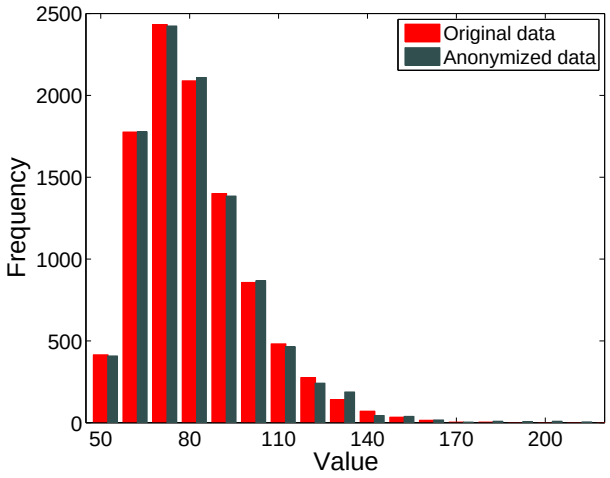

Fig. 18. Histograms of original and anonymized data with chi-squared distribution.

rules and privacy policies rely on administrative regulation and trust relationships. The anonymity-based approaches use pseudonym and group users to generate ambiguity [46], [47], [48]. Many such approaches are based on the concept of k-anonymity or its variants [49], [50], where privacy is obtained when it is unable to distinguish one entity from k-1 other entities. Typical examples occur in location-based services [31], [32], [33], [34]. Some of these approaches are known as ID rotating [51] and mix networksshmatikov:06,benjamin:06. The obfuscation-based approaches protect privacy by reducing the data quality [52], [53], [54], initially introduced for location-based services [55]. This category of approaches are also referred to as "cloaking" in a few research projects [26], [24], [56]. To quantify the privacy, the researcher have created different metrics. The k-anonymity-based approaches use the size of ambiguity set (k) as the level of privacy [49]. The obfuscation-based approaches may define privacy as the expected magnitude of the noise added onto the data or the duration to be able to track the user [55], [24].

Most of the existing participatory sensing platforms [18], [19], [20], [21] use the data to serve only internal applications and thus do not concern themselves with privacy protection. A recent project, *AnonySense* [22], built a participatory sensing platform to allow any third-party application to collect data from mobile users. *AnonySense* protects the user privacy by a mix network. The mix network allows users to send messages anonymously and mixes enough messages before reporting to applications. It mainly intends to unlink multiple data records from the same user. However, unlike our Woodward system, *AnonySense* does not the privacy attack based on prior knowledge of a certain user's record.

VII. Conclusions and Future Direction

We propose Woodward, an elastic, privacy-preserving participatory sensing system, using health care applications as an example. Unlike the existing participatory sensing platforms, Woodward protects the user privacy while supplying the anonymized data to arbitrary third-party applications. The innovative anonymization process adopted by Woodward greatly add high cost to privacy attacks; it also allows third-party

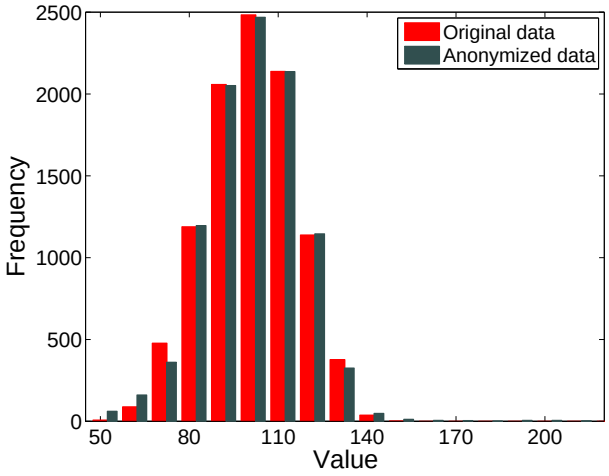

Fig. 19. Histograms of original and anonymized data with Weibull distribution.

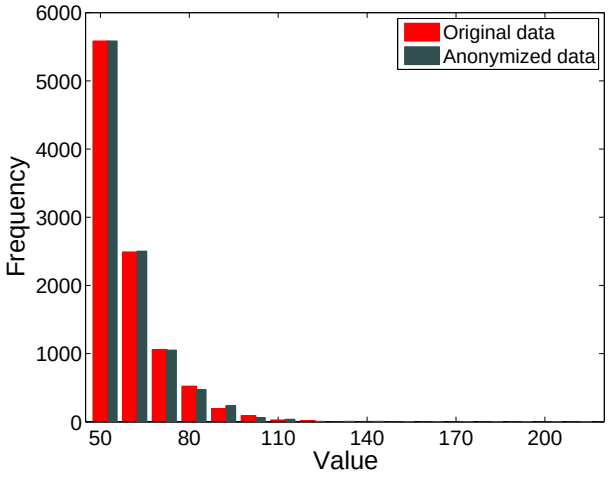

Fig. 20. Histograms of original and anonymized data with exponential distribution.

applications to perform statistical query with small under-threshold error. These features are not achievable by the existing privacy protection schemes. We implemented Woodward with a health care application and evaluated the query precision and privacy protection quantitatively. In the future, we plan to generalize the anonymization process to multi-dimensional data so as to further protect the privacy attacks based on a combination of known prior records. Additionally, we will develop versatile applications based on Woodward.

REFERENCES

[1] S. Gaonkar, J. Li, R. Choudhury, L. Cox, and A. Schmidt, "Micro-blog: sharing and querying content through mobile phones and social participation," in *MobiSys '08: Proceeding of the 6th international conference on Mobile systems, applications, and services*. New York, NY, USA: ACM, 2008, pp. 174–186.

[2] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. Srivastava, "Participatory sensing," *WSW06 at SenSys 06*, 2006.

[3] L. Deng and L. P. Cox, "Livecompare: grocery bargain hunting through participatory sensing," in *HotMobile '09: Proceedings of the 10th workshop on Mobile Computing Systems and Applications*. New York, NY, USA: ACM, 2009, pp. 1–6.

[4] A. Campbell, S. Eisenman, N. Lane, E. Miluzzo, and R. Peterson, "People-centric urban sensing," in *WICON '06: Proceedings of the 2nd annual international workshop on Wireless internet*, 2006.

[5] B. Hull *et al.*, "Cartel: A distributed mobile sensor computing system," in *Proc. of ACM SenSys 2006*, Nov. 2006.

[6] "Cens urban sensing project, 2007," http://research.cens.ucla.edu/projects/2006/Systems/Urban Sensing/.

[7] A. Lymberis, "Smart wearable systems for personalised health management: current r d and future challenges," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, 2003.

[8] J. Lee, "Smart health: Concepts and status of ubiquitous health with smartphone," in *ICT Convergence (ICTC), 2011 International Conference on*, sept. 2011, pp. 388 –389.

[9] R. Rednic, J. Kemp, E. Gaura, and J. Brusey, "Networked body sensing: Enabling real-time decisions in health and defence applications," in *Advanced Computer Science and Information System (ICACSIS), 2011 International Conference on*, dec. 2011, pp. 17 –24.

[10] I. Gondal, M. Iqbal, M. Woods, and S. Sehgal, "Integrated sensing and diagnosis – the next step in real time patient health care," in *Computer and Information Science, 2007. ICIS 2007. 6th IEEE/ACIS International Conference on*, july 2007, pp. 581 –586.

[11] T. Saponas, J. Lester, C. Hartung, S. Agarwal, and T. Kohno, "Devices that tell on you: privacy trends in consumer ubiquitous computing," in *SS'07: Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, 2007.

[12] N. Hopper, E. Vasserman, and E. Chan-Tin, "How much anonymity does network latency leak?" in *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security*. New York, NY, USA: ACM, 2007, pp. 82–91.

[13] T. Xu and Y. Cai, "Feeling-based location privacy protection for location-based services," in *CCS '09: Proceedings of the 16th ACM conference on Computer and communications security*, 2009.

[14] L. Cox, A. Dalton, and V. Marupadi, "Smokescreen: flexible privacy controls for presence-sharing," in *MobiSys '07: Proceedings of the 5th international conference on Mobile systems, applications and services*. New York, NY, USA: ACM, 2007, pp. 233–245.

[15] A. Parker, S. Reddy, T. Schmid, K. Chang, S. Ganeriwal, M. B. Srivastava, M. Hansen, J. Burke, D. Estrin, M. Allman, and V. Paxson, "Network system challenges in selective sharing and verification for personal, social, and urban-scale sensing applications," in *the Fifth Workshop on Hot Topics in Networks (HotNets-V)*, 2006.

[16] A. Alahmadi and B. Soh, "A smart approach towards a mobile e-health monitoring system architecture," in *Research and Innovation in Information Systems (ICRIIS), 2011 International Conference on*, nov. 2011, pp. 1 –5.

[17] S. Wang, W. Shi, B. Arnetz, and C. Wiholm, "Spartan: A framework for smart phone assisted real-time health care network design," in *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2010 6th International Conference on*, oct. 2010, pp. 1 –10.

[18] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell, "The jigsaw continuous sensing engine for mobile phone applications," in *the 8th ACM Conference on Embedded Networked Sensor Systems*, 2010.

[19] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda, "Peir, the personal environmental impact report, as a platform for participatory sensing systems research," in *Proceedings of the 7th international conference on Mobile systems, applications, and services*, 2009.

[20] C. Zhu, K. Li, Q. Lv, L. Shang, and R. Dick, "iscope: personalized multi-modality image search for mobile devices," in *Proceedings of the 7th international conference on Mobile systems, applications, and services*, ser. MobiSys '09. New York, NY, USA: ACM, 2009, pp. 277–290. [Online]. Available: http://doi.acm.org/10.1145/1555816.1555845

[21] T. Das, P. Mohan, V. Padmanabhan, R. Ramjee, and A. Sharma, "Prism: platform for remote sensing using smartphones," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, ser. MobiSys '10. New York, NY, USA: ACM, 2010, pp. 63–76. [Online]. Available: http://doi.acm.org/10.1145/1814433.1814442

[22] C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos, "Anonysense: privacy-aware people-centric sensing," in *MobiSys '08: Proceeding of the 6th international conference on Mobile systems, applications, and services*, 2008.

[23] B. Gedik and L. Liu, "Energy-aware data collection in sensor networks: A localized selective sampling approach," Georgia Institute of Technology, Tech. Rep., 2005.

[24] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in gps traces via uncertainty-aware path cloaking," in *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security*, 2007.

[25] G. Iachello, I. Smith, S. Consolvo, M. Chen, and G. Abowd, "Developing privacy guidelines for social location disclosure applications and services," in *SOUPS '05: Proceedings of the 2005 symposium on Usable privacy and security*. New York, NY, USA: ACM, 2005, pp. 65–76.

[26] M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in *Proc. of ACM MobiSys'03*, May 2003.

[27] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[28] R. Popa, H. Balakrishnan, and A. Blumberg, "Vpriv: Protecting privacy in location-based vehicular services," in *18th USENIX Security Symposium*, Montreal, Canada, August 2009.

[29] T. Ristenpart, G. Maganis, A. Krishnamurthy, and T. Kohno, "Privacy-preserving location tracking of lost or stolen devices: cryptographic techniques and replacing trusted third parties with dhts," in *SS'08: Proceedings of the 17th conference on Security symposium*. Berkeley, CA, USA: USENIX Association, 2008, pp. 275–290.

[30] J. Manweiler, R. Scudellari, and L. P. Cox, "SMILE: Encounter-based trust for mobile social services," in *Proceedings of ACM CCS 2009*, November 2009. [Online]. Available: http://www.cs.duke.edu/~lpcox/ccs196-manweiler.pdf

[31] G. Zhong and U. Hengartner, "Toward a distributed k-anonymity protocol for location privacy," in *Proceedings of the 7th ACM workshop on Privacy in the electronic society*, ser. WPES '08, 2008.

[32] Z. Gong, G.-Z. Sun, and X. Xie, "Protecting privacy in location-based services using k-anonymity without cloaked region," in *Proceedings of the 2010 Eleventh International Conference on Mobile Data Management*, ser. MDM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 366–371. [Online]. Available: http://dx.doi.org/10.1109/MDM.2010.33

[33] A. Gkoulalas-Divanis, P. Kalnis, and V. S. Verykios, "Providing k-anonymity in location based services," *SIGKDD Explor. Newsl.*, vol. 12, pp. 3–10, November 2010. [Online]. Available: http://doi.acm.org/10.1145/1882471.1882473

[34] S. Mascetti, C. Bettini, X. S. Wang, D. Freni, and S. Jajodia, "Providenthider: An algorithm to preserve historical k-anonymity in lbs," in *Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, ser. MDM '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 172–181. [Online]. Available: http://dx.doi.org/10.1109/MDM.2009.28

[35] G. Zhan, W. Shi, and J. Deng, "Sensortrust: A resilient trust model for wireless sensing systems," *Pervasive and Mobile Computing*, 2011.

[36] W. B. Kannel, C. Kannel, R. S. P. Jr., and L. Cupples, "Heart rate and cardiovascular mortality: The framingham study," *American Heart Journal*, 1987.

[37] J. Franklin, D. McCoy, P. Tabriz, V. Neagoe, J. Van Randwyk, and D. Sicker, "Passive data link layer 802.11 wireless device driver fingerprinting," in *Proceedings of the 15th conference on USENIX Security Symposium - Volume 15*. Berkeley, CA, USA: USENIX Association, 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1267336.1267348

[38] T. Kohno, A. Broido, and K. Claffy, "Remote physical device fingerprinting," in *Security and Privacy, 2005 IEEE Symposium on*, may 2005, pp. 211 – 225.

[39] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall, "802.11 user fingerprinting," in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, ser. MobiCom '07. New York, NY, USA: ACM, 2007, pp. 99–110. [Online]. Available: http://doi.acm.org/10.1145/1287853.1287866

[40] X. Gong, N. Kiyavash, and N. Borisov, "Fingerprinting websites using remote traffic analysis," in *Proceedings of the 17th ACM conference on Computer and communications security*, ser. CCS '10. New York, NY, USA: ACM, 2010, pp. 684–686. [Online]. Available: http://doi.acm.org/10.1145/1866307.1866397

[41] J. Krumm, "A survey of computational location privacy," *Personal Ubiquitous Comput.*, vol. 13, pp. 391–399, August 2009. [Online]. Available: http://dx.doi.org/10.1007/s00779-008-0212-5

[42] D. Ashbrook and T. Starner, "Using gps to learn significant locations and predict movement across multiple users," *Personal Ubiquitous Comput.*, vol. 7, pp. 275–286, October 2003. [Online]. Available: http://dx.doi.org/10.1007/s00779-003-0240-0

[43] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting places from traces of locations," in *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, ser. WMASH '04. New York, NY, USA: ACM, 2004, pp. 110–118. [Online]. Available: http://doi.acm.org/10.1145/1024733.1024748

[44] W. M. Newman, M. A. Eldridge, and M. G. Lamming, "Pepys: generating autobiographies by automatic tracking," in *Proceedings of the second conference on European Conference on Computer-Supported Cooperative Work*. Norwell, MA, USA: Kluwer Academic Publishers, 1991, pp. 175–188. [Online]. Available: http://dl.acm.org/citation.cfm?id=1241910.1241923

[45] Y. Matsuo, N. Okazaki, K. Izumi, Y. Nakamura, T. Nishimura, K. Hasida, and H. Nakashima, "Inferring long-term user properties based on users' location history," in *Proceedings of the 20th international joint conference on Artifical intelligence*, ser. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 2159–2165. [Online]. Available: http://dl.acm.org/citation.cfm?id=1625275.1625624

[46] M. Edman and B. Yener, "On anonymity in an electronic society: A survey of anonymous communication systems," *ACM Comput. Surv.*, vol. 42, pp. 5:1–5:35, December 2009. [Online]. Available: http://doi.acm.org/10.1145/1592451.1592456

[47] D. J. Kelly, R. A. Raines, M. R. Grimaila, R. O. Baldwin, and B. E. Mullins, "A survey of state-of-the-art in anonymity metrics," in *Proceedings of the 1st ACM workshop on Network data anonymization*, ser. NDA '08. New York, NY, USA: ACM, 2008, pp. 31–40. [Online]. Available: http://doi.acm.org/10.1145/1456441.1456453

[48] V. Shmatikov and M.-H. Wang, "Measuring relationship anonymity in mix networks," in *Proceedings of the 5th ACM workshop on Privacy in electronic society*, ser. WPES '06. New York, NY, USA: ACM, 2006, pp. 59–62. [Online]. Available: http://doi.acm.org/10.1145/1179601.1179611

[49] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, pp. 571–588, October 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=774544.774553

[50] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 754–759. [Online]. Available: http://doi.acm.org/10.1145/1150402.1150499

[51] M. Lei, X. Hong, and S. Vrbsky, "Protecting location privacy with dynamic mac address exchanging in wireless networks," in *Global Telecommunications Conference, 2007. GLOBECOM '07. IEEE*, nov. 2007, pp. 49 –53.

[52] S. T. Peddinti and N. Saxena, "On the limitations of query obfuscation techniques for location privacy," in *Proceedings of the 13th international conference on Ubiquitous computing*, ser. UbiComp '11. New York, NY, USA: ACM, 2011, pp. 187–196. [Online]. Available: http://doi.acm.org/10.1145/2030112.2030139

[53] A. B. Brush, J. Krumm, and J. Scott, "Exploring end user preferences for location obfuscation, location-based services, and the value of location," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, ser. Ubicomp '10. New York, NY, USA: ACM, 2010, pp. 95–104. [Online]. Available: http://doi.acm.org/10.1145/1864349.1864381

[54] A. RayChaudhuri, U. K. Chinthala, and A. Bhattacharya, "Obfuscating temporal context of sensor data by coalescing at source," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 11, pp. 41–42, April 2007. [Online]. Available: http://doi.acm.org/10.1145/1282221.1282226

[55] M. Duckham and L. Kulik, "A formal model of obfuscation and negotiation for location privacy," in *In Pervasive*, 2005, pp. 152–170.

[56] C.-Y. Chow, M. F. Mokbel, and X. Liu, "A peer-to-peer spatial cloaking algorithm for anonymous location-based service," in *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, ser. GIS '06. New York, NY, USA: ACM, 2006, pp. 171–178. [Online]. Available: http://doi.acm.org/10.1145/1183471.1183500