Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 18 (2013) 1979 – 1988

Procedia Computer Science

International Conference on Computational Science, ICCS 2013

PREDICT: Privacy and Security Enhancing Dynamic Information Collection and Monitoring

Li Xiong*, Vaidy Sunderam, Liyue Fan, Slawomir Goryczka, Layla Pournajaf

Department of Mathematics and Computer Science, Emory University, 400 Dowman Dr., Atlanta, GA 30322, USA

Abstract

In this paper, we present an overview of our ongoing project PREDICT (Privacy and secuRity Enhancing Dynamic Information Collection and moniToring). The overall aim of the project is to develop a framework with algorithms and mechanisms for privacy and security enhanced dynamic data collection, aggregation, and analysis with feedback loops. We discuss each of our research thrusts with research challenges and potential solutions, and report some preliminary results.

Keywords: Differential Privacy; Secure Multi-Party Computation; Filtering; Monitoring; Feedback Control

1. Introduction

As new technological tools are being developed that facilitate the continuous collection and analysis of information in novel and sophisticated ways, participatory sensing and data surveillance [5, 21] are gradually integrated into an inseparable part of our society. At the same time, the Dynamic Data Driven Applications Systems (DDDAS) paradigm [9, 3] established in the last decade offers the promise of augmenting the effectiveness of such data collection and analysis. DDDAS entails a synergistic feedback loop between application simulations and data collection, in which data are dynamically integrated into an executing simulation to augment or complement the application model, and, conversely the executing simulation steers the data collection processes of the application system [9, 3]. The DDDAS concept is crucial to the *big data* problem in the surveillance systems in order to collect data in targeted ways, adapting dynamically to application needs, rather than ubiquitously. In addition, important *data privacy* issues frequently and increasingly arise in such data surveillance systems [33, 37, 3, 27]. Many of the complex and streaming data are personal and highly sensitive to privacy concerns as well as volume issues, and will benefit greatly from privacy enhanced DDDAS capabilities.

Driving Applications. Dynamic information monitoring has numerous applications in a wide range of domains including syndromic surveillance, intelligence data collection, traffic monitoring, emergency response, and web surveillance.

- *Syndromic Surveillance*. The terrorist attacks in 2001 and various disease outbreaks such as the 2009 outbreak of H1N1 Flu [1] and the recent outbreak in Germany of E.coli [2] have prompted much attention

*Corresponding author. Tel.: +1-404-727-0758
E-mail address: lxiong@emory.edu.

Selection and peer review under responsibility of the organizers of the 2013 International Conference on Computational Science
doi:10.1016/j.procs.2013.05.367

in syndromic surveillance systems [38, 6, 40]. Such systems seek to use health data in real time for early detection of large-scale disease outbreaks and bioterrorism attacks. Traditional syndromic surveillance systems mainly rely on health data from clinical and emergency room encounters. The recent proliferation of wireless and mobile technologies provides the opportunity for individuals to produce continuous streams of data about themselves (*self surveillance* [27]). A vast amount of data can be captured, such as detailed information about individuals' physical activities, locations (e.g., through text messages), and physiological responses (e.g., through small sensors). In the existing syndromic surveillance systems and research conducted to date, issues of privacy and confidentiality of the individuals (*data subjects*) as well as the sheer volume of the data have been known to hamper researchers' efforts. With dynamic feedback loops coupled with privacy protection, data can be anonymized to preserve individual privacy and then injected to real time simulations using diffusion models to simulate and predict the outbreak patterns. The predicted patterns in turn can be used to steer further data collection (e.g., from regions with increased risks) as well as for prevention and intervention purposes.

- *Intelligence Data Collection*. As recent events demonstrate, numerous situations exist where intelligence gathering is performed in crowd settings both non-deliberately by the general public and by principals who are anonymously embedded in the crowds. A canonical example is an uprising in a major city under hostile governmental control – the general public uses smart devices to report on various field data (*third party surveillance* [27]), but there may also be agents among the crowd, reporting similar data using similar media (e.g., Twitter) to avoid identification. In such situations, central agencies (or the distributed agent network) desire to dynamically steer the data collection through feedback loops (e.g. directing agents to specific data collection locations or requesting finer-grained data). This feedback loop may also take place on open media and it is important to protect the identity and location of the agents (*data contributors*).

Most existing surveillance systems have focused on analytical and modeling methods, with little attention to dynamic feedback loops and privacy requirements. In parallel, typical privacy protection techniques today [13, 20, 14] deal with static and persistent data, but are not sufficient in the surveillance systems where high-volume, complex data are acquired dynamically. New mechanisms are needed urgently to support privacy enhancing dynamic data monitoring with feedback loops while maintaining provable and quantifiable *privacy* guarantees and *data integrity* guarantees.

Contributions. In this paper, we present an overview of our ongoing project PREDICT (Privacy and secuRity Enhancing Dynamic Information Collection and moniToring). The overall aim of the project is to develop a framework with algorithms and mechanisms for privacy and security enhanced dynamic data collection, aggregation, and analysis with feedback loops, which will be valuable in situations such as the ones outlined above. We discuss each of our research thrusts with research challenges and potential solutions, and report some preliminary results.

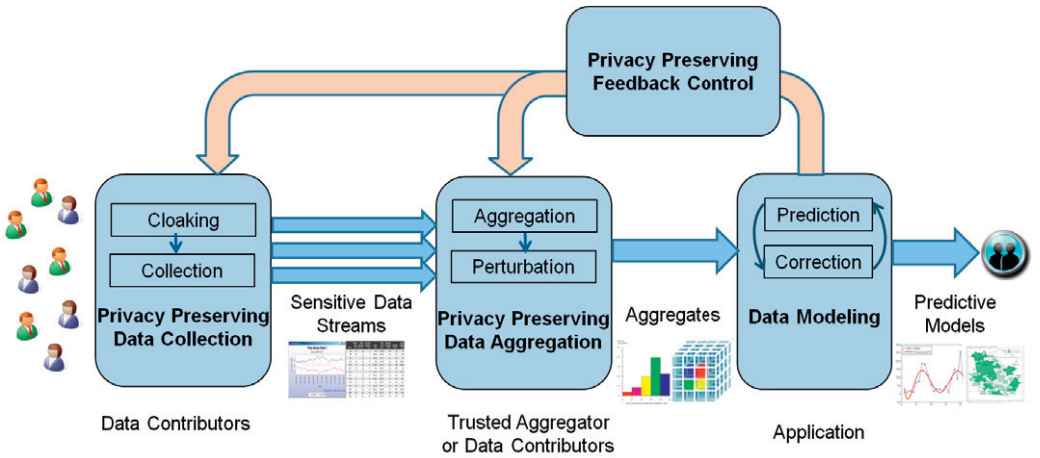

Fig. 1. PREDICT Overview

Figure 1 depicts an architectural viewpoint of PREDICT. The key innovation of PREDICT is the privacy enhanced feedback loops between data collection, data aggregation, and data modeling. The dynamic approach that leverages real time predictive data models and feedback loops to steer further data collection and privacy mechanisms is crucial to both enhance privacy and address the big data problem in real time surveillance systems. Implementing these feedback loops presents several unique challenges: (1) how to design the feedback control mechanisms for the privacy preserving data collection and aggregation while minimizing the privacy risk of data subjects and maximizing data integrity, (2) how to model and guarantee data integrity in the presence of perturbations introduced by the privacy mechanisms in addition to measurement uncertainty, and (3) how to guarantee the privacy of data contributors in the feedback loops when there is no trusted aggregator as the data contributors are mutually untrusted. Our project consists of the following research thrusts:

- **Task I: Privacy Preserving Data Collection and Aggregation with Feedback Control**. Sensitive data streams are collected, aggregated, and perturbed at selected time points to formally guarantee the state-of-the-art differential privacy for data subjects. We are designing a number of feedback loops to control the collection, aggregation, and perturbation process, including collection assignment (*how* to collect), sampling (*when* to aggregate), grouping (*how* to aggregate), and perturbation (*how* to perturb), based on feedback from previously observed aggregates and predictions as well as the privacy and integrity requirements from executing applications.
- **Task II: Dynamic Data Modeling and Uncertainty Quantification**. Aggregated and perturbed data streams are injected into predictive data models, which in turn correct the predictive data model. Data integrity is investigated in the presence of data perturbation introduced by the privacy protection mechanisms.
- **Task III: Secure Data Aggregation and Feedback Control without Trusted Aggregator**. While the privacy preserving data collection and aggregation in Task I can be implemented both by a centralized trusted aggregator or a decentralized group of data contributors, the decentralized case introduces additional privacy concerns of the data contributors (in addition to the data subjects). Secure decentralized mechanisms are developed to allow data contributors to securely aggregate their data with perturbations and receive feedback from applications without disclosing additional information to other data contributors.

By proactively building privacy into the design of a DDDAS system, privacy protections are integrated directly into the DDDAS loop. The effect is to minimize the unnecessary collection and uses of personal data by the system and guarantee the anonymized participation of individuals in the system. The outcome of the proposed research would be a suite of algorithms and mechanisms that enhance privacy for DDDAS, and have significant impact in enabling and promoting public confidence and trust in surveillance systems for critical applications in national security and public health.

2. Problem Setup

2.1. System Model

We consider a dynamic set of *data contributors* who are participating and contributing their own data (self surveillance) or other data (third party surveillance) in a surveillance system. We use *data subjects* to refer to the individuals represented in the collected data, which are the same as data contributors in the self surveillance case. We consider two system models: centralized model and decentralized model, depending on whether there is a trusted aggregator. In both models, there is an untrusted application or application run by an untrusted party for analysis and modeling (e.g. disease outbreak detection or intelligence analysis). Hence Task I and II are applicable to both models while Task III is only applicable to the decentralized model.

- *Centralized model with a trusted aggregator (Task I and II).* In the centralized model, the trusted aggregator (e.g., CDC offices in the syndromic surveillance scenario) collects the data, aggregates them, performs appropriate data perturbation, and outputs perturbed aggregates with privacy guarantee, which can be in turn used for modeling and predictive studies. In the feedback loops, the trusted aggregator receives the control from the running application for further data collection, aggregation, and perturbation.

- *Decentralized model without a trusted aggregator (Task I, II, and III).* In some scenarios, a trusted aggregator is not available (e.g., in the intelligence collection scenario). The data contributors need to perform aggregations and perturbations among themselves if needed and submit the aggregated result to the untrusted aggregator or the application directly. In the feedback loops, the control is sent to individual contributors as well.

2.2. Data Model

Sensitive data can be collected via explicit data collection (such as reported symptoms in clinical visits) or as contributed data streams through various sensors or devices (such as temperature sensors installed at airport gates or reported symptoms via mobile devices). Each individual may contribute a discrete set of values over time or a continuous data stream. At each time point k, the collective data of all individuals can be represented as a relational database table D, in which each tuple consists of all the attributes associated with one individual or subject. Figure 2(a) shows an example data set collected for flu surveillance at a single time point.

Age	State	Temperature (F)	Symptoms
22	GA	98	Coughing, Sore Throat
25	GA	104	Coughing, Headache
30	TX	102	Headache, Vomiting
35	TX	98	Headache, Sore Throat

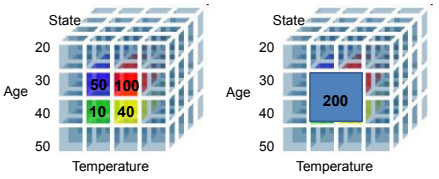

Fig. 2. Data model: (a) collected raw data at time k; (b) aggregated data cubes at time k.

For both data reduction and privacy protection, data is aggregated at selected time points. To model the aggregated data, we use a d-dimensional data cube, where d is the number of attributes in the collected data, to represent the aggregates. It is also called a cuboid in the data warehousing literature [31, 10]. Each dimension of the cube corresponds to a data attribute. Each cell of the cube represents an aggregated measure (e.g., the count of the individuals or data points), corresponding to the multidimensional coordinates of the cell. Figure 2(b) shows an example data cube representing the aggregated population count with respect to different combinations of age, state, and temperatures from the data in Figure 2(a). Note that at each time point, aggregates can be computed along different dimensions with different grouping or granularity depending on the feedback from the applications. Figure 2(b) shows another cube aggregated from the same data but with different grouping.

2.3. Privacy Model

Privacy of Data Subjects. In both centralized and decentralized models, we need to protect the privacy of data subjects represented in the collected data. We assume the end application and end users are untrusted - they may passively observe information to infer sensitive values of the data subjects. Further the analysis results may be shared with other untrusted parties. So our goal is to provide a provable privacy guarantee such that the end application will not learn anything about participating users in the system and whether they participated in the data collection.

Traditional approaches such as removing identifying attributes, generalizing or perturbing individual attribute values, have been shown to be susceptible to various attacks [20]. We use the state-of-the-art differential privacy [13, 14, 28] as our privacy model, which gives a strong and provable privacy guarantee. Differential privacy requires that the output of an aggregation or computation should not change significantly even if a single data subject had opted out of the data collection. Therefore, this assures an individual that any privacy breach will not be a result of presence of her record in the collected data. Formally, differential privacy is defined as follows.

Definition 2.1 (α-Differential privacy [12, 28])**.** *A randomized mechanism $\mathcal{A}$ satisfies unbounded α-differential privacy if for any neighboring databases D_1 and D_2 where D_1 can be obtained from D_2 by either adding or removing one tuple, and any possible output set S, $Pr[\mathcal{A}(D_1) \in S] \leq e^{\alpha} Pr[\mathcal{A}(D_2) \in S]$.*

A common mechanism to achieve differential privacy for a single aggregated value is the Laplace perturbation (LPA) [13], which adds systematically calibrated Laplace noise to the aggregates. Given an aggregate query Q, the global sensitivity [15] of Q, denoted by Δ_Q, measures the sensitivity of the query result $Q(D)$ if a data subject had opted out. In order to achieve α-differential privacy, the LPA mechanism returns $Q(D) + N$ in place of the original result $Q(D)$, where N is a random noise of Laplace distribution $Lap(\Delta_Q/\alpha)$ with a probability density function $Pr(x) = \frac{\alpha}{2\Delta_Q} e^{-|x|\alpha/\Delta_Q}$ [15].

Any sequence of aggregations from the same set of data subjects that each provides differential privacy in isolation also provides differential privacy in sequence (with accumulated privacy cost), known as *sequential composition* [32]. If a sequence of aggregations is conducted on *disjoint* data subjects, the privacy cost does not accumulate, but depends only on the worst guarantee of all aggregations, known as *parallel composition*.

Privacy of Data Contributors in Decentralized Model. In our decentralized system model with no trusted aggregator, the privacy of the data contributors need to be protected from the aggregator and other data contributors. We assume data contributors are either semi-honest - follows the protocol correctly but may passively observe information to infer sensitive information of other data contributors or data subjects, or malicious - can lie about the values being reported, but otherwise follows the protocol correctly. As in general cryptographic solutions, we make an assumption that at least a fraction of data contributors (e.g., a majority) are semi-honest. Remaining ones and the aggregator can be arbitrarily malicious.

In the feedback phase, our goal is to ensure that each contributor learns only whether and what data is being requested or collected from her and no additional information about whether and what data is collected from other contributors. In the data aggregation phase, our goal is to ensure the data aggregator or participating data contributors can only learn the aggregates and no additional information about the private data contributed by other data contributors, in addition to ensuring differential privacy of the aggregates for the data subjects. We use the secure multi-party computation (SMC) notion for this purpose [11, 22, 29]. In a multi-party computation protocol, a set of parties wish to jointly compute a function of their private data inputs. The protocol is *secure* if the parties learn only the result of the function but nothing else.

3. PREDICT Framework

3.1. Privacy Preserving Data Collection and Data Aggregation with Feedback Control

Given our system model, in order to monitor the data streams from the data contributors and to guarantee differential privacy for the input data streams, a simple yet infeasible approach can be achieved by issuing an aggregate query for each combination of the attribute values followed by a Laplace perturbation mechanism. Recall that the standard Laplace mechanism (LPA) adds systematically calibrated Laplace noise to the aggregates. Essentially, we can construct a data cube at each single time point consisting of perturbed aggregate measures for each cell. This approach suffers from two main drawbacks. First, it requires a total of m queries, where m is the number of cells in the data cube, giving rise to a scalability problem when the number of dimensions and domain sizes are large. Second, due to the potential correlations or overlap of the data subjects at different time points and the sequential composition of differential privacy, it will incur a high privacy cost quickly, or render the aggregated values useless due to a high level of noise being added. Few works [16, 7] that studied differential privacy under continual observation only consider a single counter (or in general a set of pre-fixed counters). More importantly, they defined *event-level* privacy to protect an event, i.e. one user's presence at a particular time point, rather than the presence of that user in the overall time range. The central challenge of applying differential privacy for continuous data monitoring is how to minimize the privacy cost by avoiding unnecessary aggregations and perturbations. Ideally, we should only compute the aggregates for desired regions at desired time points to preserve privacy cost, and hence preserve the data integrity.

The key idea in PREDICT is to use feedback loops from estimates and predictions based on previously observed aggregates and predictive models to dynamically control the collection, aggregation, and perturbation process. We briefly describe the feedback control, including the sampling rate (when to aggregate), aggregation grouping (how to aggregate), perturbation level (how much to perturb), and data collection control (how to collect). In this subsection, we outline the research tasks for the control mechanisms assuming the feedback is available. We assume at time step k, the *a priori* state estimate $\hat{x}_k^-$ is made based on a predictive model in the

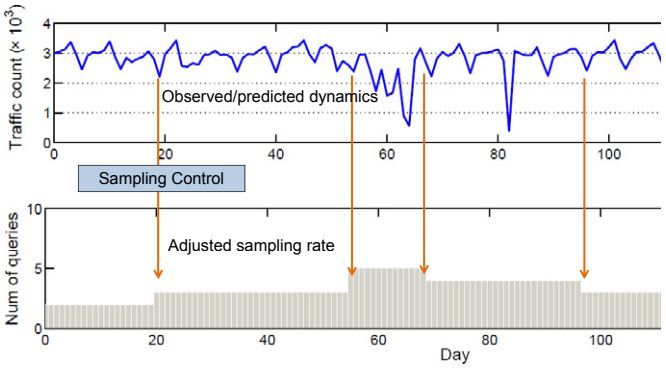

Fig. 3. Sampling Control using Feedback

application. The *a posteriori* state estimate $\hat{x}_k$ is based on a correction of *a priori* state estimate $\hat{x}_k^-$. In next subsection, we explain how the *a priori* and *a posteriori* are computed which in turn generates feedback for the aggregation and perturbation at the next time point.

Task I.1 Sampling Control using Error Feedback. We first study how to dynamically monitor a pre-defined aggregate (such as the percentage of clinical visits that are for "influenza-like illness", i.e. temperature > 100F AND (cough = TRUE OR sore throat = TRUE)) with sampling control. The key question is when to compute the aggregate (when to sample) such that the data dynamics can be accurately captured while maintaining the accumulated privacy cost below a given privacy bound α. Figure 3 illustrates the idea of sampling control using feedback. It shows an original aggregate data stream, *traffic count*, with different dynamics at different time periods. It also shows a desired aggregation strategy with different query rate or sampling rate at different time periods. We observe that the number of queries per time unit (sampling rate) increases at day 55, adapting to the significant fluctuations exhibited by the traffic count, and it drops beyond day 100, when there is little variation among the aggregated values. Ideally, the sampling control mechanism will achieve such dynamic and adaptive behavior. While this is intuitive, it is not a trivial task due to the inherent uncertainty of input data streams. The underlying dynamics of the aggregated data can not be observed directly and need to be carefully modeled. In addition, the raw aggregated data is not accessible to the application or feedback control mechanism due to privacy protections.

In our preliminary studies, we built an adaptive sampling controller which adjusts the sampling rate based on the feedback. The feedback is defined as the relative error between the *a posteriori* estimate and the *a priori* estimate at a particular time step. Note that the posterior estimate is only available when a noisy observation is sampled from the input stream at time step k_n. Thus no error is defined at non-sampling points. The model error measures how well the internal predictive model describes current data dynamics, supposing the *a posteriori* estimate $\hat{x}_{k_n}$ is close to the true value. We may infer that data is going through rapid changes if the error E_{k_n} increases with time. In response, the controller in our system will detect the errors and adjust the sampling rate accordingly. Given the model error as feedback, we adopt a PID (Proportional-Integral-Derivative) controller for the sampling control. For details, please refer to our recent papers [18, 17].

Task I.2 Aggregation Control using Prediction Feedback. The LPA based privacy mechanism is sensitive to the density of the data, e.g. an aggregated statistic from a sparse region will incur a higher relative error compared with a dense region with the same amount of perturbation. On the other hand, if we group multiple cells and compute the aggregate for a partition, the count for each cell has to be estimated assuming certain distribution of the data points in the partition. The dominant approach in histogram literature is making the *uniform distribution assumption*, where the frequencies of records in the bucket are assumed to be the same and equal to the average of the actual frequencies [25]. This introduces an *approximation error*. Ideally, when the density reaches certain level, we may wish to have a finer grained statistic for the sub-regions. For example, we may start with a monitor "number of flu occurrences". When the observed counter exceeds certain threshold, we may request more fine grained counters such as "number of flu occurrences in different regions". In other scenarios, we may wish to have a finer grained statistic when there are variances or uneven distributions in the region. In our intelligence

collection scenario, once the crowd center is identified for an uprising, it may be desired to collect more fine-grained data around the borders of the crowd to monitor how the crowd moves. Therefore in PREDICT project, we use aggregation control in the multidimensional data space based on predicted data values to jointly minimize the privacy cost and the noise introduced by the perturbation and the approximation.

Not surprisingly, finding the optimal multi-dimensional grouping or partitioning even without the privacy constraints is a challenging problem and optimal partitioning even in two dimensions is NP-hard [34]. In our prior work [39], a kd-tree (k-dimensional tree) based partitioning strategy based on a perturbed cell data cube has been shown to be feasible for exploiting the underlying data characteristics. For this project, we can use the predictions or estimates from the applications and apply various partitioning strategies to control the aggregation in the current time point. We are currently exploring multi-dimensional partitioning such as Binary Space Partitioning (BSP) and Quad-tree techniques to dynamically partition the multi-dimensional data space based on the predicted state such that data will be aggregated from similar sub-cubes.

Task I.3 Perturbation Control using Error Feedback. We also attempt to dynamically determine the level of perturbation for each partition using the feedback of the model error and the uncertainty requirement of the application. If the model error is high, which suggests significant data dynamics, the perturbation control mechanism can adjust the perturbation level so that a more precise aggregate with less noise can be obtained at next sampling point. Similar controller mechanisms like a PID controller can be explored.

In addition, applications may impose an uncertainty bound or requirement for a perturbed aggregate. In such case, a perturbation with minimum privacy budget required is invoked in order to satisfy the uncertainty requirement while minimizing the overall privacy cost. In general, when there is no specified integrity requirement, we can consider the overall privacy bound as a resource or budget and model the perturbation control problem as an online resource allocation problem, which we plan to explore in the future.

Task I.4 Dynamic data collection assignment. A final feedback loop is to control how to collect data. When multiple individual data contributors are available for collecting data (third party surveillance), the aggregator can coordinate the data collection process such that data collection coverage is maximized and data collection cost is minimized. Since we need to protect the identity as well as location privacy for individual data contributors, they can query the central aggregator for data collection tasks anonymously using cloaked locations.

We are currently designing stochastic optimization algorithms for coordinated data collection assignment. The goal is to optimally assign individual data contributors for data collection tasks (i.e. points of interests) based on feedback and data integrity requirement from applications as well as the cloaked (uncertain) locations of individual data contributors. The optimization goals may include maximizing data collection coverage, maximizing data integrity guarantee, and minimizing data collection cost (e.g. distance traveled by individual data contributors).

3.2. Task II. Dynamic Data Modeling with Uncertainty Quantification

An essential component that enables the feedback loop is the data modeling that provides real-time model-based prediction and correction based on the sampled or observed aggregates. The key challenge is how to model the data in the presence of perturbation error injected by the LPA privacy mechanism. Thus, the focus of Task II is to explore robust data assimilation and spatial interpolation techniques for estimating the current state of the system using sampled aggregates with uncertainty quantification in real time.

Task II.1 Data Modeling in Time Domain. In order to model the perturbed aggregates in time domain, we have applied several filtering (or data assimilation) techniques. Data assimilation [26, 4] is a general approach in which observations (or perturbations in our context) of the current (and possibly, past) state of a system are combined with the results from a prediction model (the forecast) to produce an analysis, which is considered as 'the best' estimate of the current state of the system. The model is then advanced in time and its result becomes the forecast in the next analysis cycle.

In our initial design, the prediction, i.e. prior estimate, is released at a non-sampling point, while the correction, i.e. posterior estimate based on the noisy observation and prediction, is released at a sampling point. We adopted a constant process model for a single pre-defined aggregate which is given by $x_{k+1} = x_k + \omega$ where k is the discrete time index and ω is a white Gaussian noise $p(\omega) \sim N(0, Q)$ with variance Q. The observed aggregate is perturbed by the Laplace mechanism and can be modeled by $z_k = x_k + \nu$ where ν is a Laplacian noise which follows

$p(v) \sim Lap(0, \lambda)$ with λ being the magnitude parameter determined by differential privacy mechanism. Since the measurement noise is non-Gaussian, the posterior density cannot be analytically determined without the Gaussian assumption about the measurement noise. We adopted two solutions to the posterior estimation challenge. One is to approximate the Laplace noise with a Gaussian noise, which can be then solved by the classic Kalman Filter [26]. The other is to simulate the posterior density function with Monte Carlo methods based on the Sampling-Importance-Resampling (SIR) particle filter. The details can be found in our recent paper [17]. We are currently extending the work to non-linear models for the time-series.

Preliminary Results. We have completed the design of a framework with Filtering and Adaptive Sampling for monitoring single time-series addressing the challenges in Task I.1 and Task II.1. We performed a set of experiments using the Kalman Filter and Particle Filter in combination with PID based sampling control on synthetic datasets as well as real traffic monitoring and flu datasets. Our approaches consistently outperform the baseline Laplace perturbation algorithm and the state-of-the-art Discrete Fourier Transform (DFT) based algorithm [35], which can be only applied in batch processing setting instead of real-time setting. For more detailed results and a demonstration description, please refer to [18, 19, 17]. We are currently extending the work to multi-dimensional time-series with spatial partitioning techniques and other feedback controls (Task I.2 and Task II.2).

Task II.2 Data Modeling in Multi-Dimensional Data Space. The spatial dependencies or homogeneities of neighbourhood characteristics legitimate the use of spatial interpolation methods to predict values for specified spatial locations using a limited number of sample data aggregates at nearby locations. In general, this is also applicable in the multi-dimensional data space. We are currently exploring both deterministic and stochastic methods, and in particular, two commonly adapted interpolators, Inverse Distance Weighting (IDW) and kriging. In addition, diffusion models will be explored as we build system prototypes for syndromic surveillance systems.

A widely used deterministic interpolation method is Inverse Distance Weighting (IDW). It is a local exact interpolator that interpolates values based only on the surrounding measured values of the interpolating location and functions of the inverse distances between the interpolating location and locations of the surrounding sample. On the other hand, stochastic methods, such as kriging [36], interpolate values not only based on the surrounding data values, but also based on the overall autocorrelation calculated by applying statistical models to all the known data points. Because of this, not only do stochastic methods have the capability of producing a prediction surface, but they also provide some measure of the certainty or accuracy of the predictions. This is important for allowing the perturbation control proposed in Task I.3. We plan to study detailed interpolation algorithms based on Kriging with uncertainty quantifications [8].

3.3. Secure Data Aggregation and Feedback Control without Trusted Aggregator

In our decentralized system model with no trusted aggregator, the privacy of the data contributors need to be protected from the aggregator and other data contributors. Such protection needs to be maintained in the entire feedback loop, i.e. both the feedback control phase and the data aggregation phase.

Task III.1 Secure Feedback Control. In the feedback phase, our goal is to ensure that each contributor learns only whether and what data is being requested or collected from her and no additional information about whether and what data is collected at other contributors. Depending on the current data model, applications may need to send feedback to individual data contributors, e.g. to collect more crowd data in certain geographic locations in the intelligence collection for city uprising scenario, without disclosing the control command to other data contributors or entities in the network.

A simple idea we plan to implement is based on the public key encryption. The data contributors can send their public keys to the application as they contribute their data. The application can in turn encrypt individual feedback control commands using corresponding public keys such that the contributor who is intended to receive the control commands can decrypt the message and follow the command for future data collection. We will explore other potential crypto or secure computation mechanisms and study the performance impact carefully during the project.

Task III.2 Secure Data Aggregation. In the data aggregation phase, our goal is to ensure the data aggregator or participating data contributors can only learn the aggregates and no additional information about the private data contributed by other data contributors, in addition to ensuring differential privacy of the aggregates for the

data subjects. The problem can be formulated as a secure multiparty computation (SMC) or distributed privacy preserving data sharing problem [22, 29], in which a set of parties jointly computes a function of their private data inputs such that the parties learn only the result of the function but nothing else.

In addition to leveraging existing SMC protocols, one particular challenge for secure data aggregation is that when the collected data involves personal data, the aggregates need to be perturbed (as in Task I) to protect the privacy of data subjects. Suppose a set of n semi-honest data collectors need to compute a perturbed sum to satisfy α-differential privacy. If we have one data contributor generate a Laplace noise and add it to the secure sum result, the sum will be disclosed to this contributor. Thus, the perturbation needs to be distributed as well. Our goal is to minimize the total noise added to the result, and ensure that each data contributor generates a noise such that the summation of the noise is sufficient to achieve α-differential privacy.

Preliminary Results. We have designed several SMC protocols for various privacy preserving aggregation and analytical tasks [30, 23]. We also did a comprehensive comparative study for the secure sum problem with differential privacy [24]. We studied several secure multiparty computation schemes: Shamir's secret sharing, perturbation-based, and various encryption schemes. Differential privacy of the final result is achieved by distributed Laplace perturbation mechanism (DLPA). Partial random noise is generated by all participants, which draw random variables from Gamma or Gaussian distributions, such that the aggregated noise follows Laplace distribution to satisfy differential privacy. We also introduced a new efficient distributed noise generation scheme with partial noise drawn from Laplace distributions. We compared the protocols with different privacy mechanisms and security schemes in terms of their complexity, security characteristics, and scalability both analytically and experimentally in a real distributed environment [24]. We are currently extending the approaches to dynamic aggregations with the feedback control.

4. Conclusion

We described our ongoing project PREDICT for privacy and security enhanced dynamic information collection and monitoring with feedback loops. Our key insights are: (1) privacy preserving data aggregation with perturbations can simultaneously achieve condensed data representation and privacy protection, (2) a dynamic approach that leverages real time predictive data models and feedback loops to drive further data collection and privacy mechanisms is crucial to both enhance privacy and address the big data problem in real time surveillance systems. While we have made some initial progress in building a framework for monitoring single time-series with adaptive sampling and filtering as well as secure aggregations with differential privacy for data collection when there is no trusted aggregator, many research tasks and challenges remain. We look forward to carrying out the research and hearing the feedback from the DDDAS community.

Acknowledgement

This research is supported by the Air Force Office of Scientific Research (AFOSR) DDDAS program under grant FA9550-12-1-0240.

References

[1] 2009 h1n1 flu. http://www.cdc.gov/h1n1flu/.
[2] Investigation update: Outbreak of shiga toxin-producing e. coli o104 (stec o104:h4) infections associated with travel to germany. http://www.cdc.gov/ecoli/2011/ecolio104/index.html.
[3] *Report of the August 2010 Multi-Agency Workshop on InfoSymbiotics/DDDAS, The Power of Dynamic Data Driven Applications Systems*. Workshop sponsored by: Air Force Office of Scientific Research and National Science Foundation.
[4] M. S. Arulampalam, S. Maskell, and N. Gordon. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–188, 2002.
[5] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. In *In: Workshop on World-Sensor-Web (WSW06): Mobile Device Centric Sensor Networks and Applications*, 2006.
[6] B. Cakici, K. Hebing, M. Grnewald, P. Saretok, and A. Hulth. Case: a framework for computer supported outbreak detection. *BMC Med Inform Decis Mak.*, 10(14), 2010.
[7] T.-H. H. Chan, E. Shi, and D. Song. Private and continual release of statistics. In *ICALP (2)*, pages 405–417, 2010.

[8] J. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley Series in Probability and Statistics. 2009.

[9] F. Darema. Dynamic data driven applications systems: A new paradigm for application simulations and measurements. In *Computational Science - ICCS 2004*, volume 3038 of *Lecture Notes in Computer Science*, pages 662–669. Springer Berlin / Heidelberg, 2004.

[10] B. Ding, M. Winslett, J. Han, and Z. Li. Differentially private data cubes: optimizing noise sources and consistency. In *SIGMOD*, 2011.

[11] W. Du and M. J. Atallah. Secure multi-party computation problems and their applications: a review and open problems. In *NSPW '01: Proceedings of the 2001 workshop on New security paradigms*, pages 13–22, New York, NY, USA, 2001. ACM.

[12] C. Dwork. Differential privacy. *Automata, Languages and Programming, Pt 2*, 4052, 2006.

[13] C. Dwork. Differential privacy: A survey of results. In M. Agrawal, D.-Z. Du, Z. Duan, and A. Li, editors, *TAMC*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2008.

[14] C. Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54, January 2011.

[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *3rd Theory of Cryptography Conference*, 2006.

[16] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. Differential privacy under continual observation. In *STOC*, pages 715–724, 2010.

[17] L. Fan and L. Xiong. Adaptively sharing time-series with differential privacy. *CoRR*, abs/1202.3461, 2012.

[18] L. Fan and L. Xiong. Real-time aggregate monitoring with differential privacy. In *CIKM*, pages 2169–2173, 2012.

[19] L. Fan, L. Xiong, and V. Sunderam. Fast: Differentially private real-time aggregate monitor with filtering and adaptive sampling (demonstration track). In *ACM SIGMOD*, 2013.

[20] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Surveys*, 42(4), 2010.

[21] S. L. Garfinkel and M. D. Smith. Guest editors' introduction: Data surveillance. *IEEE Security & Privacy*, 4(6), 2006.

[22] O. Goldreich. *Foundations of Cryptography: Volume 2, Basic Applications*. Cambridge University Press, New York, NY, USA, 2004.

[23] S. Goryczka, L. Xiong, and B. Fung. Secure distributed data anonymizatoin and aggregation with m-privacy. *IEEE Transactions on Data and Knowledge Engineering (TKDE)*, 2013.

[24] S. Goryczka, L. Xiong, and V. Sunderam. Secure multiparty aggregation with differential privacy: A comparative study. In *6th International Workshop on Privacy and Anonymity in the Information Society (PAIS)*, 2013.

[25] Y. Ioannidis. The history of histograms (abridged). In *Proc. of VLDB Conference*, 2003.

[26] R. E. Kalman. A new approach to linear filtering and prediction problems. 1960.

[27] J. Kang, K. Shilton, D. Estrin, J. Burke, and M. Hansen. Self-surveillance privacy. *Iowa Law Review*, 97, 2012.

[28] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 international conference on Management of data*, SIGMOD '11, 2011.

[29] Y. Lindell and B. Pinkas. Secure multiparty computation for privacy-preserving data mining. Cryptology ePrint Archive, Report 2008/197, 2008. http://eprint.iacr.org/.

[30] J. Liu, L. Xiong, J. Luo, and J. Z. Huang. Privacy preserving distributed dbscan clustering. *Transactions on Data Privacy*, 2013.

[31] K. M and H. J, W. *Data mining: concepts and techniques, Second Edition*. MorganKaufman, 2006.

[32] F. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *SIGMOD*, 2009.

[33] 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*, MobiSys, 2009.

[34] S. Muthukrishnan, V. Poosala, and T. Suel. On rectangular partitionings in two dimensions: Algorithms, complexity, and applications. In *ICDT*, pages 236–256, 1999.

[35] V. Rastogi and S. Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *SIGMOD*, 2010.

[36] D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, ACM '68, pages 517–524, 1968.

[37] K. Shilton. Four billion little brothers?: privacy, mobile phones, and ubiquitous data collection. *Commun. ACM*, 52:48–53, November 2009.

[38] M. M. Wagner, A. W. Moore, and R. M. Aryel, editors. *Handbook of biosurveillance*. 2006.

[39] Y. Xiao, L. Xiong, and C. Yuan. Differentially private data release through multidimensional partitioning. In *Secure Data Management*, pages 150–168, 2010.

[40] W. Yih, S. Deshpande, C. Fuller, D. Heisey-Grove, J. Hsu, B. Kruskal, M. Kulldorff, M. Leach, J. Nordin, J. Patton-Levine, E. Puga, E. Sherwood, I. Shui, and R. Platt. Evaluating real-time syndromic surveillance signals from ambulatory care data in four states. *Public Health Rep.*, 125(1), 2010.