AFRL-SR-AR-TR-10-0151

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-02-2010	Final	From 01-12-2006 To 30-11-2009

4. TITLE AND SUBTITLE
Systematic Control and Management of Data Integrity, Quality and and Provenance for Command and Control Applications

5a. CONTRACT NUMBER
FA9550-07-1-0041

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Elisa Bertino

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Purdue University West Lafayette IN 47907

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 N. Randolph St.
Arlington VA 22203
NL

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A - Approved for public release

20100427018

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The objective of this project is to design and develop a comprehensive approach to the problem of assuring high data integrity able to support data analysis and decision making. The project has achieved the several novel results: (1) Digital signatures techniques for graph and tree structured data; the techniques are both hiding and binding, that is, they assure integrity without releasing information (unlike the Merkle hash tree technique that leaks information). (2) Efficient privacy-preserving data matching protocols; these protocols use a combination of data sanitization techniques, like differential privacy, with secure multi-party computation techniques. (3) A model to assess the trustworthiness of data based also on provenance information; the model has been applied to sensor data and location data. (4) An assessment of the use of sanitized data for classification tasks; through extensive experiments, we have shown that classifiers obtained from sanitized data are usually good. (5) A system to enforce application-defined integrity policies and its implementation on top of the ORACLE DBMS.

15. SUBJECT TERMS
Data Integrity, Data Quality, Information Trustworthiness, Data Matching, Integrity Systems, Digital Signature Techniques, Sensor Data, Location Information

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Elisa Bertino
U	U	U	None		19b. TELEPHONE NUMBER *(Include area code)* 765-496-2399

Systematic Control and Management of Data Integrity, Quality and Provenance for Command and Control Applications

Final Report

PI: Elisa Bertino
Institution: Purdue University
West Lafayette, Indiana
Contract Number: FA9550-07-1-0041

CO-PI: Murat Kantarcioglu
Institution: The University of Texas at Dallas
Richardson, Texas

Submitted to: Air Force Office of Scientific Research
Submission Date: January 24, 2010
Period of Performance: December 1, 2006 – November 30, 2009
Subcontractor: The University of Texas at Dallas

Statement of Work

Organizations need to share data so that analysts can mine the data, extract knowledge and make effective decisions. While the problem of confidentiality while data sharing has been investigated, the problem of data integrity that has not received much attention. However, in order for analysts and war fighters to make effective decisions and take actions, data must be accurate and current. Without integrity, any information extracted from the available data cannot be trusted. It is important to observe that data integrity can be undermined not only by errors introduced by users and applications, but also by malicious subjects who may inject inaccurate data into a database with the goal of *deceiving* the users of the data. Therefore, it is critical that data integrity issues which include data quality and data provenance be investigated for organizational data sharing, situation assessment, multi-sensor data integration and numerous other functions to support the war fighter.

The objective of the proposed research is to design and develop a comprehensive approach to the problem of assuring high data integrity able to support data analysis and decision making. Relevant goals of the research are as follows: (a) to develop flexible high-assurance systems for data integrity, supporting all relevant integrity policy languages and able to exploit information on data provenance; (b) to develop a mathematical framework to perform risk assessment concerning data integrity; (c) to investigate their use for protection against deception through injection of incorrect data.

In order to provide a comprehensive approach to the problem of high integrity data, we need a multi-faceted solution. Key elements of the envisioned solutions are: (a) policy languages for integrity, allowing one to specify under which conditions and circumstances data can be "accepted" in the database, to specify which data need to be validated and when, which data can be used when making critical decision; (b) a comprehensive set of information concerning data provenance that, together with other available information, allows one to evaluate the integrity of data; (c) risk assessment functions, supporting the evaluation of the integrity risks, incurred when accepting certain data in the database, and of use risks, incurred when using certain data. In addition to developing theoretical foundations concerning risk assessments, the proposed research will result in the specification and implementation of policy languages, and in the implementation of several tools, such as tools for risk assessment and provenance information management.

Currently there is no comprehensive approach to the problem of high assurance data integrity. Research in areas such as information security and assurance, namely in the areas of integrity models, data quality, policy specification and architectures, as well as risk assessment have produced some techniques, tools and methodologies, which either address only a small aspect of the data integrity problem. The significant scientific merit of our proposed research is to provide a comprehensive approach to the problem of high assurance data integrity based on the use and enforcement of integrity policies and to the problem of integrity risk assessment. Our research has extensive broader impact as data integrity, quality and provenance are critical features not only for Defense and Intelligence applications but also for numerous other applications in Healthcare and Finance.

Results of the Research Effort

The research carried out during the project has focused on data integrity and trust, and on the use of data with difference confidence levels in queries. In addition, as part of the research, an integrity system has been designed and implemented for the management of integrity policies on top of commercially available DBMS. The problem of data integrity and trust has been tackled at three different levels, all required in comprehensive solutions for high assurance data integrity:

(i) Physical level by developing signature techniques secure against information leaks that are specialized for tree-structured data structures and graph-structured data structures.

(ii) Data provenance level by developing techniques to assign confidence levels to data, and trust levels to data providers. Such techniques allow the party receiving the data to assess how trusted data are based on the data sources and by comparing data facts received by different sources. A theoretical model has been developed and then applied to the problem of assessing the trustworthiness of locations of individuals and of sensor data. Also solutions have been proposed concerning the use of such data in queries according to their confidence level and data usage policies.

(iii) (iii) Data quality level by developing an approach to assess the quality of classification models extracted from anonymized data and protocols for privately and efficiently matching data from different files. The matching protocols developed in the project combine, through a two-step matching approach, the use of secure multi-party techniques with data sanitization techniques. Two different sanitization techniques are supported by our protocols: k-anonimization, and differential privacy.

Accomplishments and Findings

1. *Signature Techniques for Graph-structured Data and Tree-structured Data*

Summary: Data sharing among multiple parties with high integrity assurance is an important problem. An integrity assurance technique provides mechanisms using which a party can verify that the data has not been tampered with. Specific integrity assurance requirements and techniques depend on the structure according to which the data are organized. Because one of the most widely used data organization structures is the tree structure, the development of techniques specifically suited for data organized according to such tree structures is crucial. When addressing the problem of integrity for tree structures it is important to notice that each node typically contains some content and that the structural relationships between the nodes may establish some relationships between the contents in these nodes. Such relationships may be defined according to properties such as classification, indexing, temporal-orientation and sensitivity of the contents. Integrity of such relationships is referred to as *structural integrity*, whereas the integrity of the contents is referred to as *content integrity*. An integrity mechanism for tree structures must thus preserve both content and structural integrity. In many cases, such as military scenarios, an additional requirement is to maintain the confidentiality of the content and the structural information. By confidentiality we mean that: (i) a party receives only those nodes and the structural information that the party is allowed to access, according to the stated access control policies; (ii) a party should not receive nor should be able to infer any information about the content and presence (or absence) of nodes and structural information that the party is not allowed to access.

The Merkle hash technique is the most well known signature for tree structures and has been widely extended for use in high-assurance integrity content distribution systems. A drawback of such technique is that the integrity verification process does not preserve confidentiality. Merkle hash is binding (integrity) but not hiding (confidentiality); therefore it is vulnerable to inference attacks. The signature of a non-leaf node combines the signatures of its children nodes in a particular order. Further, in order to be able to compute the hash of a node during integrity verification, the signatures of a set of nodes/subtrees in the tree has to be revealed to the party, even if the party does not have access to these nodes/subtrees. By the mere fact that such signatures are received, the party may infer that a given node, to which it has access to, has a sibling/parent/child node, even though the party does not have access to it. Such an inference may lead to confidentiality breaches. More specifically, the integrity verification of a subtree S, which belongs to a tree T, by using the Merkle hash technique reveals (1) the signature (Merkle hash) of some nodes which are in T but not in S; (2) the structural relationship between a node x in S and some node y which is in T but not in S; and (3) the relative (structural) order between a node x which is in S and y, which is in T but not in S.

One way to avoid such information leakage is by pre-computing and storing a separate Merkle signature for every distinct subtree that may be the result of a query or a request to access the tree. However such a technique is impractical because the result of a query can be an arbitrary subtree and there can be an exponential number of such subtrees in a tree. The problem that our research addresses is as follows: The trusted owner Alice of a data item organized as a (rooted) tree T wants to digitally sign T once so that it can be queried or accessed many times. A user Bob should be able to verify the integrity of the content and structure of a subtree S (of T) that Bob is authorized to access. Any information about a node which is in T but not in S, its signature, its structural relationship with any other node in T should not be revealed to Bob. Obviously the Merkle hash technique cannot be effectively used for this purpose. We have thus proposed an integrity assurance technique for tree structures which is secure against the above information leakages and is also efficient. The distribution of data is often carried out through third parties, which are not completely trusted in the sense that the trusted owner (Alice) relies on the third party D for the distribution but does not wish to provide D the authority to sign on its behalf. This may not be due to a lack of

trust in *D* but merely a recognition of the fact that typical systems such as *D*'s are more vulnerable (to breakdowns, spy wares, insider misbehavior, or simply accidents) than the trusted owner. This model offers many advantages, but also offers a challenge: How does the third-party distributor *D*, which does not have the authority to sign (only Alice does), and prove to a user the integrity of the data provided to the user (Bob) without leading to leakage of information related to structure as well as content? The obvious answer is to sign the data (tree) once and store at *D* a number of integrity verification items that *D* can later on provide to any party which is legitimately requesting a subset of the data (a subtree) which it is authorized to access; these integrity verification items are cryptographic hashes or signatures that enable the user to verify the integrity of the subtree - both content and structure that it receives.

A recent technique proposed by us (see reference [7] in the publication list) known as "structural signature scheme" for trees overcomes the shortcomings of the Merkle hash technique. Even though the structural signature scheme is provably secure - binding and hiding, yet it leads to a negligible leakage. Moreover, such a scheme cannot be applied to graphs, primarily because structural signatures use post-order and pre-order numbers, which can be exploited to infer the existence of cross-edges and back-edges, thus leading to leakages. No prior such technique (which is both binding and hiding) exists for graph structured data.

A major result of our research is the definition of an efficient Zero-Leakage Integrity Assurance (ZLIA) scheme for tree-structured data that is provably binding and hiding. The approach is based on computing and storing integrity verification units (IVU), such as signed hashes, that can be sent by the data distributor to a user B as a proof of integrity along with a subtree/subgraph, without leaking. The approach also uses the notion of *secure-names* for nodes. The purpose of secure names is to convey the order of siblings (which node is to the left of which other node) without leaking anything else (e.g., whether they are adjacent siblings, how many other siblings are between them, etc). We have also extended such a scheme for application to DAG-structured graphs. The proposed ZLIA schemes not only guarantee strong security properties with semantic security (zero-leakage) but also have optimal efficiency in terms of signing, storage and integrity verification. The costs of signing and storing IVU at a distributor according to the proposed ZLIA scheme are $O(n)$ and $O(n)$, respectively, for trees, and $O(n^2)$ and $O(n^2)$, respectively, for DAGs, where n is the number of nodes. Only a constant number ($O(1)$) of IV Us for either scheme is sent to user B, for integrity verification of a subtree/subgraph with n nodes. Furthermore, integrity verification of paths and subsequences that are results of queries on paths (such as XPath queries) and sequences, respectively have also proposed as an application of the integrity assurance scheme for trees. Moreover, our proposed for trees incurs $O(d)$ cost for pinpointing d compromised data items in an ordered set of n items, which is an important problem in digital forensics, and previous solutions have higher cost (in $O(d^3 logn)$). Our signature scheme has been extended to support the signature of general graphs.

The salient features that the ZLIA scheme possesses over existing techniques are:

- It provides stronger security guarantees in terms of integrity and confidentiality; unlike the Merkle hash technique it is semantically secure.
- It facilitates precise and efficient detection of integrity violations in the sense that it precisely identifies the node or the structural relationship that has been compromised.
- While its signature generation time is comparable to that of the Merkle hash technique, the user-side integrity verification time that it requires is less than that of the Merkle hash technique.

Major findings and their significance to the field: The development of the first signature technique for tree-structured data and graph-structured data which is semantically secure; note that the very well known Merkle hash tree technique is binding but not hiding (thus is not semantically secure.

2. *A Model for Trust Evaluation based on Data Provenance and its Applications*

Summary: Today the need of sharing data within and across the organizations is more critical than ever. The availability of comprehensive data makes it possible to extract more accurate and complete knowledge and thus supports more informed decision making. However reliance on data for decision making processes requires data to be of good quality and trusted. We refer to such requirements as *high-assurance data integrity*. Without high-assurance integrity, information extracted from available data cannot be trusted. While there have been some efforts to ensure confidentiality when sharing data, the problem of high-assurance data integrity has not been widely investigated. Previous approaches have either addressed the problem of protection from data tampering, through the use of digital signature techniques, or the problem of semantic integrity, that is, making sure that the data is consistent with respect to some semantic assertions. However even though these techniques are important components of any comprehensive solution to high-assurance data integrity, they do not address the question on whether one can actually trust certain data. Those techniques, for example, do not protect against data deception, according to which a malicious party may provide on purpose some false data, or against the fact that a party is unable, for various reasons, to provide good data. Techniques, like those developed in the area of data quality, may help; however they often require the availability of good quality data sources against which one can compare the data at hand and correct them.

It is clear that in order to address the problem of high-assurance data integrity we need comprehensive solutions combining several different techniques. In particular, one important issue in determining data integrity is the trustworthiness of data provenance. For example, a malicious source provider may announce that a small company has successfully signed a big contract which is not true in reality. This information is then passed to a stock analysis agent, based on which the agent infers that the stock prize of that company will go up with high probability and send this information to end users. If the data users, based on this information, decide to acquire stocks of such company, they may end up with severe financial losses. In contrast, the data users will have a big chance to avoid such a loss if they know that the source provider is not very trustworthy.

Though a lot of research has been carried out for data provenance, they mainly focus on the collection and semantic analysis of provenance information. Little work has been done with respect to the trustworthiness of data provenance. Data provenance (also referred to as lineage) can be described in various terms depending on the application context. Buneman et al. define data provenance in the context of database systems as the description of the origins of data and the process by which it arrived at the database. Lanter refers to lineage of derived products in geographic information systems as information that describes materials and transformations applied to derive the data. Greenwood et al. expand Lanter's definition of provenance and view it as metadata recording the process of experiment workflows, annotations, and notes about experiments. In our work, we define data provenance to be the information that helps determine the derivation history of a data product starting from its original sources. Unlike previous definitions, our definition of data provenance is more general and covers all possible information that may influence the trustworthiness of the data.

To evaluate the trustworthiness of data provenance, we need to answer questions like "Where did the data come from? How trustworthy is the original data source? Who handled the data? Are the data managers trustworthy?" More specifically, for example, if data X is from source A, how do we determine the trustworthiness of source A?. If X arrives at D via B and C, how do we tell if X is accurate at D? Also if data X now from D and data Y coming from E are merged by source F, then how do we determine the trustworthiness of the resulting data?

To address these challenges, we have developed a data provenance trust model which estimates the level of trustworthiness of both data and data providers by assigning trust scores to them. Based on such trust scores, users can make more informed decisions

whether or not to use the data. To build such trust model, we take into account various aspects that may affect the trustworthiness of the data. These aspects are data similarity, data conflict, path similarity and data deduction. Similar data items are considered as supports to one another, while conflicting data items compromise trustworthiness of one another. Besides data similarity and data conflict, the way that the data was collected is also an important factor when determining the trustworthiness of the data. For example, if several independent sources provide the same data, such data is most likely to be true. Data deduction measures the effect of the data process (e.g. data mining) on the data. Usually, the trustworthiness of the resulting data depends on the trustworthiness of input data and the on the parties that process the data.

We also observe that a data is likely to be true if it is provided by trustworthy data providers, and a data provider is trustworthy if most data it provides are true. Due to such inter-dependency between data and data providers, we have developed an iterative procedure to compute the trust scores. To start the computation, each data provider is first assigned an initial trust score which can be obtained by querying available information about data providers. At each iteration, we compute the trustworthiness of the data based on the combined effects of the aforementioned four aspects, and re-compute the trustworthiness of the data provider by using the trust scores of the data it provides. When a stable stage is reached, that is, when the changes of trust scores are negligible, the trust computation process stops.

The trust model has then been instantiated to two different application domains:

- *Location information.* Trustworthiness of location information about particular individuals is of particular interest in the areas of forensic science and epidemic control. In many cases, location information is not precise and may include fraudulent information. With the growth of mobile computing and positioning systems, e.g., GPS and cell phones, it has become possible to trace the location of moving objects. Such systems provide us an opportunity to find out the true locations of individuals. We have applied our trustworthiness model to assess the trustworthiness of the location information of an individual based on different evidences from different sources. We have also identified a collusion attack that may bias the computation. Based on the analysis of the attack, we have developed algorithms to detect and reduce the effect of collusion attacks. Experimental results show the efficiency and effectiveness of our approach.
- *Sensor data.* Our approach for estimating the trustworthiness of sensor data uses the data item provenance as well as their provenance. We have introduced two types of data provenance: the physical provenance which represents the delivering history of each data item, and the logical provenance which describes the semantic meaning of each data item. The logical provenance is used for grouping data items into semantic events with the same meaning or purpose. By contrast, the tree-shaped physical provenance is used in computing trust scores, that is, quantitative measures of trustworthiness. To obtain trust scores, we have developed a cyclic framework which well reflects the inter-dependency property: the trust scores of data items affect the trust scores of network nodes, and vice versa. The trust scores of data items are computed from their value similarity and provenance similarity. The value similarity is based on the principle that "the more similar the values for the same event are, the higher trust scores are," and we compute it under the assumption of normal distribution. The provenance similarity is based on the principle that "the more different physical provenances with similar values are, the higher trust scores are," and we compute it using the tree similarity. Since new data items are continuously generated in sensor networks, we need to evolve (i.e., re-compute) the trust scores to reflect those new items. As evolution scheme, we have proposed a batch mode for computing scores (non)periodically along with an immediate mode. Experimental results show that our approach is very efficient.

The main novelties of our approach are:

- A new trust model based on data provenance which allows one to assess the trustworthiness of data and data providers.
- Algorithms to compute trust scores.
- Applications of the trust model to the problem of assessing the trustworthiness of location information and sensor data.

Major findings and their significance to the field: The first model for assigning trust to data based on data provenance and its usage in two important application domains.

3. *Data Usage based on Data Confidence Levels*

Summary: Today the need of sharing data within and across the organizations is more critical than ever. The availability of comprehensive data makes it possible to extract more accurate and complete knowledge and thus supports more informed decision making. However reliance on data for decision making processes requires data to be of good quality and trusted. We refer to such requirements as *high-assurance data integrity*. Without high-assurance integrity, information extracted from available data cannot be trusted. While there have been some efforts to ensure confidentiality when sharing data, the problem of high-assurance data integrity has not been widely investigated. Previous approaches have either addressed the problem of protection from data tampering, through the use of digital signature techniques, or the problem of semantic integrity, that is, making sure that the data is consistent with respect to some semantic assertions. However even though these techniques are important components of any comprehensive solution to high-assurance data integrity, they do not address the question on whether one can actually trust certain data. Those techniques, for example, do not protect against data deception, according to which a malicious party may provide on purpose some false data, or against the fact that a party is unable, for various reasons, to provide good data. Techniques, like those developed in the area of data quality, may help; however they often require the availability of good quality data sources against which one can compare the data at hand and correct them. Moreover, improving data quality may incur in additional, not negligible, costs. It is also important to notice that the required level of data quality depends on the purpose for which the data have to be used. For example, for tasks which are not critical to an organization, like computing a statistical summary, data with a low level of quality may be sufficient, whereas when an individual in an organization has to make a critical decision, very good data with high are required.

The problem is thus how to design a system that can provide data meeting the confidence level required for each data use. A first relevant question is thus how to specify which task requires high-confidence data. In situations where we do not have enough data with high-confidence level to allow a user to complete a task, a question is how can we improve the confidence of the data to the required level with minimum cost Yet another question could be: "There is a huge data volume. Which portion of the data should be selected for quality improvement?" When dealing with large data volumes, it is really hard for a human to quickly find out an optimal solution that meets the decision requirement with minimal cost. As we have proved (see reference [5] in the publication list), the problem is NP-hard.

To solve the above questions, we have developed a comprehensive framework based on four key elements. The first element is the association of confidence values with data in the database. A confidence value is a numeric value ranging from 0 to 1, which indicates the trustworthiness of the data. Confidence values can be obtained by using techniques like those developed by us in this project (see reference [8] in the publication list) which determine the confidence value of a data item based on various factors, such as the trustworthiness of data providers and the way in which the data has been collected. The second element is the computation of the confidence values of the query results based on the

confidence values of each data item and lineage propagation techniques. The third and fourth elements, which are the most novel contributions of this work, deal respectively with the notion of *confidence policy* and with strategies for incrementing the confidence of query results at query processing time. The notion of confidence policy is a key novel notion proposed as part of our approach. Such a policy specifies the minimum confidence level that is required for use of a given data item in a certain task by a certain subject. Such policies are declarative and therefore can be easily specified and modified. As a complement to the traditional access control mechanism that applies to base tuples in the database before any operation, the confidence policy restricts access to the query results based on the confidence level of the query results. Such an access control mechanism can be viewed as a natural extension to the Role-based Access Control (RBAC) which has been widely adopted in commercial database systems. Therefore, our approach can be easily integrated into existing database systems. Since some query results will be filtered out by the confidence policy, a user may not receive enough data to make a decision and he may want to improve the data quality. To meet the user's need, we have developed an approach for dynamically incrementing the data confidence level; such an approach is the fourth element of our solution. In particular, our approach selects an optimal strategy which determines which data should be selected and how much the confidence should be increased to satisfy the confidence level stated by the confidence policies. We assume that each data item in the database is associated with a cost function that indicates the cost for improving the confidence value of this data item. Such a cost function may be a function on various factors, like time and money. We have developed several algorithms to compute the minimum cost for such confidence increment.

It is important to compare our solution to the well-known Biba Integrity Model, which represents the reference integrity model in the context of computer security. The Biba model is based on associating an integrity level with each user and data item. The set of levels is a partially ordered set. Access to a data item by a user is permitted only if the integrity level of the data is "higher" in the partial order with respect to the integrity level of the user. Despite its theoretical interest, the Biba Integrity Model is rigid in that it does not distinguish among different tasks that are to be executed by users nor it addresses how integrity levels are assigned to users and data. Our solution has some major differences with respect to the Biba Integrity Model. First it replaces "integrity levels" with confidence values and provides an approach to determine those values. Second it provides policies by using which one can specify which is the confidence required for use of certain data in certain tasks. As such our solution supports fine-grained integrity tailored to specific data and tasks. Third, it provides an approach to dynamically adjust the data confidence level so to provide users with query replies that comply with the confidence policies. We have developed various heuristics to adjust such level and have carried out extensive performance studies which demonstrate the efficiency of our system.

The main novelties of our approach are:

- We propose the first systematic approach to data use based on confidence values of data items.
- We introduce the notion of *confidence policy* and *confidence policy compliant query evaluation*, based on which we have developed a framework for the query evaluation.
- We have developed algorithms to minimize the cost for adjusting confidence values of data in order to meet requirements.

Major findings and their significance to the field: The first model to support the notion of confidence policy and a query processing approach to evaluate the confidence of the query results and to dynamically adjust confidence values.

4. *Using Anonymized Data for Classification*

Summary: The verification of data integrity and quality may also require the use of machine learning techniques; therefore the learning of good learned models, like classification trees, is crucial. In general the availability of large data sets from which to learn the models is crucial. Very often however privacy may make not possible to have available data for machine learning tasks. For example, privacy sensitive information related to individuals, such as medical data, is today collected, stored and processed in a large variety of application domains. Such data is typically used to provide better quality services to individuals and its availability is crucial in many contexts. In the case of healthcare, for example, the availability of such data helps prevent medical errors and enhance patient care. Privacy sensitive data may have many important legitimate uses serving distinct purposes outside the specific domain in which it has initially been collected. For example, drug companies and researchers may be interested in patient records for drug development. Such additional uses of data are important and should certainly be supported. Yet, privacy of the individuals to whom the data is related should be assured as well. To address this challenge, we explored how to do classification over anonymized data. We propose a novel approach that models generalized attributes of anonymized data as uncertain information. In our approach, each generalized value of an anonymized record r is accompanied by statistics collected from records in the same anonymization class as r. This extra information, released with the anonymized data, supports the accurate computation of expected values of important functions for data analysis such as dot product and square distance. Consequently, we can seamlessly extend many classification algorithms to handle anonymized data.

In [6], we discuss in details how such an approach to compute the "expected square distance" and the "expected dot product" can be applied to various classification algorithms. Also in [1], based on our novel approach, we address issues related to different uses of anonymized data sets in the data mining process, specifically the following uses of anonymized data sets:

a. **Classifying Anonymized Data Using Data Mining Models Built on Anonymized Data:** This is the typical application of anonymized data sets for data mining. For example, consider a medical researcher who wants to build a classifier on some medical data set. Due to privacy concerns, the researcher is not provided direct access to the data. Instead, the classifier has to be built and tested over an anonymous version of the medical data set.
b. **Classifying Original Data Using Data Mining Models Built on Anonymized Data:** In some cases, data mining models built on anonymized data sets may need to be tested on the original data sets (i.e., the data sets with no modification). For example, several hospitals may collaborate to create one large anonymized data set which is then shared among all collaborating hospitals. For a researcher in a participating hospital who is interested in using data mining to classify the patients' of the hospital, there will be at least two options. Either, the researcher can build a classifier using the local data set and use it for new patients or she can build a classifier using the large anonymized data set that involves many more samples and use it for new patients. To see which of these two options is better, we need to be able to classify original data (e.g., the medical records of the new patient) using data mining models built on anonymized data. Notice that in the above scenario, the researcher could anonymize the set of new patients and convert the problem to classifying anonymized data with models built on anonymized data (which, as we explained above, is a much more simpler problem). Although a perfectly viable alternative, how the set of new patients is anonymized becomes crucially important. Inserting every new patient's data into the hospital's training data set and re-anonymizing the data would take too long. Adjusting anonymization parameters for the small set of new patient data set (to be anonymized independently) is not easy either. On the other hand, if the model built on anonymized data supported classifying original data, new patients could be classified quickly and accurately.
c. **Classifying Anonymized Data Using Data Mining Models Built on Original Data:** In some cases, we may need to classify anonymized data using data mining models

built on original data. For example, for homeland security applications, the Department of Homeland Security (DHS) can legally have access to enough original data to build data mining models (e.g., it may have a large enough number of legally retrieved phone records related to suspicious activities.). At the same time, due to privacy concerns, DHS may not be allowed to directly access individually identifiable data (e.g., all phone records stored in the repositories of a phone company) for classification purposes. Instead, DHS may be given access to anonymized data for such purposes. To handle such situations, we need to be able to classify anonymized data using data mining models built on original data.

d. **Distributed Data Mining Using Anonymized Data:** If individual privacy is the main concern, data sets that are anonymized locally could be distributed for building data mining models. For example each hospital locally anonymizes the patient discharge data and then shares the anonymized data with other hospitals. The obvious question is whether building classification models from the union of locally anonymized data sets results in good classification results.

Major findings and their significance to the field: This work is the first attempt to understand different aspects of building and using classifiers over anonymized data. More specifically,

- We developed a new approach for computing expected dot product and square distance on anonymized data by treating anonymized instances as uncertain information.
- We showed how this approach can be used for classification.
- We showed how our techniques can be directly used to handle cases where either the training or the test data is anonymized.
- We showed the effectiveness of using locally anonymized data sets for distributed data mining.
- Our method applies to any anonymization method based on generalization.

5. *Protocols for Private and Efficient Record Matching*

Summary. The process of identifying and linking different representations of the same real-world entity across multiple data sources is known as the *record linkage* problem. Since it is a key component of methodologies for data quality and integrity, record linkage has been investigated extensively. However especially after the introduction of powerful data mining techniques, privacy concerns related to sharing of individual information have pushed research towards the re-formulation of the record linkage problem and the development of new solutions.

In order to prevent privacy concerns from hampering sharing of private information, two main approaches have been developed. These are sanitization methods that perturb private information to obscure individual identity and cryptographic methods that rely on Secure Multi-party Computation (SMC) protocols. Sanitization techniques such as *k*-anonymization or differential privacy usually involve privacy metrics that measure the amount of privacy protection. Higher levels of protection typically translate into further deviation from the original data and consequently less accurate results. Therefore sanitization techniques involve trade-off between accuracy and privacy. Cryptographic techniques do not sacrifice accuracy to achieve privacy. The algorithms applied to private data are converted to series of functions with private inputs. Then, using SMC protocols, accurate results are obtained. Under reasonable assumptions regarding computational power of the adversary, SMC protocols guarantee that only the final result and any information that can be inferred from the final result is revealed. SMC protocols generally have some security parameters (e.g. encryption key sizes) that allow users to trade off between cost and privacy.

Both those approaches are not able to provide a comprehensive solution addressing all relevant application requirements with respect to privacy, cost, and accuracy. The goal of our work is to address the limitations of such approaches. We propose a novel method to

address the private record linkage problem that combines cryptographic techniques and sanitization techniques. Unlike existing methods, trade-off in our solution is along three dimensions: privacy, cost, and accuracy. To the best of our knowledge, ours is the first systematic approach in this direction.

We assume three participants in our method. These are two data holders, with the data sets to be linked, and the querying party, who provides the classifier which determines matching record pairs. The basic idea is to utilize sanitized data sets to accurately match or mismatch a large portion of record pairs so that the need for costly SMC protocols is minimized. We call this the *blocking* step. The *blocking* step provides cost savings proportional to the level of anonymity, set independently by each participant. Later on, the *blocking* step is followed by the *SMC* step, where unlabeled record pairs are labeled using cryptographic techniques. If the input data sets are too large, we may have to label significant amounts of record pairs using cryptographic techniques. In such cases, since cost of the private record linkage process is not known in advance, data holder parties might be unwilling to participate. That's why we also consider limiting the costs of cryptographic techniques. This also allows us to analyze the cost-accuracy and cost-privacy relationships. When the upper bound imposed on SMC costs is too low, some record pairs might remain unlabeled at the end of the *blocking* step. In order not to reveal irrelevant pairs, we label them non-matched. While this precaution ensures 100% precision, it degrades recall since some of those unlabeled record pairs might actually be matching. Fortunately, sanitized data sets can help reduce the effects. Based on generalizations of records, pairs that are more likely to match can be given priority in the *SMC* step. Two different versions of our matching protocols have been developed. The first version uses k-anonymization technique for the sanitization step, whereas the second uses differential privacy. Experiments conducted on real data sets have shown that our method has significantly lower costs than cryptographic techniques and yields much more accurate matching results compared to sanitization techniques, even when the data sets are perturbed extensively.

Our method has many advantages over existing methods. The main advantages can be summarized as follows:

- Costs are significant lower than, and at worst, equal to the costs of existing cryptographic techniques.
- Precision is always 100%, which implies that irrelevant record pairs are protected against disclosure and therefore privacy of these records is assured.
- Recall varies with the upper bound on SMC costs, imposed by participants.
- Our method applies to any anonymization method and any cryptographic technique. Participants can choose different anonymization methods, anonymity levels, quasi-identifier attribute sets.

Major findings and their significance to the field: The development of the first method for privacy-preserving record matching, supporting both equality and similarity, which is very efficient for very large data sets.

6. Systems for Data Integrity

Summary: We have designed an *integrity policy language*, supporting all relevant policies concerning data and an underlying system enforcing the policies expressed in this language.. Our approach consists of the following elements: (1) The notion of *integrity metadata template* by using which metadata information relevant for integrity can be specified. (2) A flexible *integrity control policy language* that is able to support various integrity requirements. The integrity metadata template describes data structures and values of information by which data integrity is determined. We have developed a metadata specification language for precisely describing such information which can be vary according to the application requirements. By using the language, we can specify a metadata template for a type of target data. Then, an instance of the metadata template is associated with each item of the target and the instance is used in integrity control policies whenever the associated data item is

accessed. The integrity control policy language describes the specifications and enforcement of application-dependent integrity policies. Our integrity policy language supports both Access Control Policy (ACP) and Data Validation Policy (DVP). ACP specifies what actions (e.g., read, insert, modify, or delete) on a data are permitted under which conditions to preserve data integrity when a data item is accessed. On the other hand, independently from data accesses, DVP specifies autonomous processes for monitoring and/or enhancing the data integrity.

To apply our high-assurance integrity solution to real applications, we have investigated how to integrate our approach into existing database management systems (DBMSs) such as Oracle. We have first carefully investigated two possible integration approaches (e.g., modifying DBMS code, exploiting a mediator between DBMS and applications to enforce integrity policies), and then found that Language Level Integration (LLI) is the most practical one. LLI is based on the approach that automatically translates integrity policies and integrity metadata templates into a DBMS specific language (e.g., SQL). LLI is an efficient and solid integration approach since it uses DBMS built-in facilities. The design has been developed specifically for the Oracle DBMS; however, we believe that can be easily adapted to other commercial DBMS.

For developing the LLI approach, we have designed in detail: 1) a database schema for describing metadata templates and 2) rules for translating integrity policy language into the DBMS understandable language. In our design, a metadata specification is translated into SQL DDL (Data Description Language) statements creating database tables whose attributes correspond with the data structure of metadata information. An instance of the metadata template is represented as a record in the table. Next, each integrity policy is translated onto a database trigger whose code automatically executed whenever a particular data item is accessed. The trigger executes the same procedure described in the corresponding integrity such as determining integrity by using metadata information (i.e., a record in a table), accepting/denying the access, and updating metadata information. In addition, for fine granularity access control, we use the Oracle VPD (Virtual Private Database) facility and the Oracle FGA (Fine Grained Auditing) functions. An implementation of the prototype has been completed and several performance experiments have been carried out.

The main novelties and key advantages of our approach are:

- The policy-based solution provides a comprehensive and solid framework for data integrity in information systems such as DBMSs.
- LLI is a practically applicable approach for deploying integrity policy languages since it does not require modifying the source code of the DBMS and nor the code of the applications running on top of the DBMS.
- LLI prevents unexpected security holes since it uses DBMS built-in functions that cannot be bypassed by applications.
- The detailed database schema and translation rules, even though designed for the Oracle DBMS, can be easily extended for other commercial DBMSs.

Major findings and their significance to the field: (1) The development of the flexible and simple to use policy-based data integrity framework able to support various integrity-related access control and data validation requirements. (2) The development of a practical and efficient approach to deploy integrity policies on top of existing commercial DBMSs.

Relevance to the AF's mission and potential civilian technology challenges

The developed techniques are relevant in protecting against deception and assuring that data are correct and can be trusted. Such techniques can be applied in any context in which crucial decision making takes place and when data need to be rapidly acquired by different sources. The use of our efficient privacy-preserving record linkage technique is crucial in situations in which different organizations need to determine if they have common data in their databases or digital archives without however sharing these databases and archives among each others. Relevant applications include surveillance and intelligence operations, homeland security, and security of financial organizations.

Personnel Supported by the Project or Associated with the Project

Purdue University:

Prof. Elisa Bertino, Department of Computer Science. Prof. Bertino has been in charge of directing the research activities. She has been involved in all research activities undertaken in the project.

Prof. Guy Lebanon, Department of Statistics. Prof. Lebanon has investigated issues in the area of statistical decision theory and its application to the problem of risk assessment for information security.

Dr. Dan Lin (postdoc), Department of Computer Science. Dr. Lin has been involved in the research on trust evaluation of data provenance.

Chenyun Dai (PhD student, supported by the project), Department of Computer Science. C. Dai has been involved in the research on trust evaluation of data provenance. He has developed an approach to assess the trustworthiness of information about locations of individuals. He has also implemented a prototype of the integrity system on top of the ORACLE DBMS.

Ashish Kundu (PhD student), Department of Computer Science. A. Kundu has been involved in the research on signature techniques for tree structured data. He has developed the theoretical research and carried out an extensive comparison of our technique with the Merkle Hash technique.

Hyo-Sang Lim, Department of Computer Science. Dr. Lim has been involved in the design of the integrity system. He has also developed an approach to assess the trustworthiness of sensor data.

Mohamed Nabeel (PhD student), Department of Computer Science. M. Nabeel has been involved in the research on signature techniques for XML data and delta publishing. He has developed the theoretical research and carried out implementation and experimental activities.

University of Texas at Dallas:

Prof. Murat Kantarcioglu, Department of Computer Science. Prof. Kantarcioglu has been in charge of directing the research activities at UT Dallas. He has been involved in all research activities undertaken in the project.

Prof. Bhavani Thuraisingham, Department of Computer Science. Prof. Thuraisingham has been involved in the research on the risk based access control. She has participated in the theoretical research and the design of the system.

Ali Inan (PhD Candidate), Department of Computer Science. Mr. Inan has been involved in the research on privacy preserving record linkage. He has developed the theoretical model and carried out the implementation and experimental studies.

Publication List

1. A. Kundu, E. Bertino, "How to Authenticate Graphs Without Leaking", *Proceedings of 13th International Conference on Extending Database Technology (EDBT 2010)*, March 22-16, 2010, Lausanne, Switzerland (to appear)
2. A. Inan, M. Kantarcioglu, G. Ghinita, E. Bertino, "Private Record Matching Using Differential Privacy", *Proceedings of 13th International Conference on Extending Database Technology (EDBT 2010)*, March 22-16, 2010, Lausanne, Switzerland (to appear)
3. C. Dai, H.S. Lim, E. Bertino, Y.S. Moon, "Assessing the trustworthiness of location data based on provenance", *Proceedings of the 17th ACM International Conference on Advances in Geographic Information Systems (ACM-GIS 2009)*, November 4-6, 2009, Seattle, Washington, ACM Press.
4. E. Bertino, C. Dai, M. Kantarcioglu, "The Challenge of Assuring Data Trustworthiness", Keynote paper, *Proceedings of 14th International Conference (DASFAA 2009)*, Brisbane, Australia, April 21-23, 2009, LNCS 5463, Springer.
5. C. Dai, L. Dan. M. Kantarcioglu, E. Bertino, E. Celikel, B. Thuraisingham, "Policy Complying Query Evaluation Based on Lineage Propagation", *Proceedings of Secure Data Management Workshop (SDM 2009)*, August 28, 2009, Lyon, France, LNCS 5776, Springer.
6. A. Inan, M. Kantarcioglu, E. Bertino, "Using Anonymized data for Classification", *Proceedings of 25th International Conference on Data Engineering (ICDE)*, IEEE, April 7-12, 2009, Shanghai, China.
7. A. Kundu, E. Bertino, "Structural Signatures for Tree Data Structures", *Proceedings of the 34th International Conference on Very Large Data Bases (VLDB 2008)*, August 23-28, 2008, Auckland, New Zealand.
8. C. Dai, L. Dan. E. Bertino, M. Kantarcioglu, "An Approach to Evaluate Data Thrustwortiness Based on Data Provenance", *Proceedings of Secure Data Management Workshop (SDM 2008)*, August 22, 2008, Auckland, New Zealand, LNCS 5159, Springer.
9. A. Inan, M. Kantarcioglu, E. Bertino, M. Scannapieco, "A Hybrid Approach to Private Record Linkage", *Proceedings of 24th International Conference on Data Engineering (ICDE)*, IEEE, April 7-12, 2008, Cancun, Mexico.
10. M. Nabeel, E. Bertino, "Secure Delta-Publishing of XML Content", Poster Paper, *Proceedings of 24th International Conference on Data Engineering (ICDE)*, IEEE, April 7-12, 2008, Cancun, Mexico.
11. E. Celikel, M. M. Kantarcioglu, B. Thuraisingham, E. Bertino, "Managing Risks in RBAC Employed Distributed Environments", *Proceedings of the Information Security (IS) 2007 International Symposium*, November 25-30, 2007, Vilamoura, Portugal, Lecture Notes in Computer Science 4804, Springer 2007.
12. M. Nabeel, E. Bertino, "A Structure Preserving Approach for Securing XML Documents", Invited Paper, *Proceedings of The Second International Workshop on Trusted Collaboration, 2007 (TrsutCol-2007)*, November 12, 2007, White Plains, New York, USA.
13. A. Kundu, M. J. Atallah, E. Bertino, "Zero-Leakage Integrity Assurance of Structured Data in an Environment of Untrusted Third Party Distributors", November 2009, submitted for publication.
14. H.S. Lim, C.Dai, E.Bertino, "*A Comprehensive Policy-based Approach for High-assurance Data Integrity in DBMSs*", November 2009, submitted for publication.
15. H.S. Lim, Y.S. Moon, E.Bertino "*Provenance-based Confidence Policy Management in Data Streams*" in preparation.
16. C.Dai, H.S.Lim, Y.S. Moon, E. Bertino, "*How Much Can you Trust Location Information? An Approach based on Physical and Logical Location Data*" in preparation.

Interactions and Transitions

a) Presentations at meetings:
   1. Presentation of the research results from this project to the kickoff meeting of the Northrop Grumman Cybersecurity Research Consortium (http://www.cerias.purdue.edu/site/news/view/cerias_partners_with_industry_academic_leaders_to_address_nations_cybersecu/)
   2. Participation of PI Bertino and student C. Dai to the *GIS 2009 Conference* to present a paper on the results of this project.
   3. Participation of PI Kantarcioglu to the *SDM 2009 Workshop* to present a paper on the results of this project.
   4. Participation of PI Bertino and Kantarcioglu to the AFOSR PI meeting in Washington DC, June 25, 2009
   5. Participation of PI Bertino to the *DASFAA 2009 Conference* to give a keynote talk focused on some of the results of this project.
   6. Participation of student A.Inan to the *ICDE 2009 Conference* to present a paper on the results of this project.
   7. Participation of PI Bertino and student A. Kundu to the *VLDB 208 Conference* and to the *SDM 2008 Workshop* to present papers reporting the results of this project.
   8. Participation of PI Kantarcioglu and student A. Inan to the *ICDE 2008 Conference* to present a paper on the results of this project.
   9. Participation of PI Bertino and Kantarcioglu to the AFOSR PI meeting in Washington DC, June 12, 2008.
   10. Participation of PI Bertino and student M. Nabeel to the *2007 Workshop on Trusted Collaboration* to present a paper on the results of the project.
   11. Participation of PI Kantarcioglu to the *2007 IS International Symposium* to present a paper on the results of this project.

b) New discoveries:
   1. A new signature technique for hierarchical data and graph-structured that is both binding and hiding and has semantic security.
   2. The first assessment about the use of anonymized data for classification tasks.
   3. The first efficient technique for privacy-preserving record matching.
   4. The first model for assessment of data trustworthiness based on provenance.
   5. The first system supporting a comprehensive policy language for integrity.