CERIAS Tech Report 2004-18

E-NOTEBOOK MIDDLEWARE FOR ACCOUNTABILITY AND REPUTATION BASED TRUST IN DISTRIBUTED DATA SHARING COMMUNITIES

by Paul Ruth, Dongyan Xu, Bharat Bhargava, Fred Regnier

Center for Education and Research in
Information Assurance and Security,
Purdue University, West Lafayette, IN 47907-2086

E-notebook Middleware for Accountability and Reputation Based Trust in Distributed Data Sharing Communities

Paul Ruth[⋆], Dongyan Xu[⋆†], Bharat Bhargava[⋆†], and Fred Regnier[‡]

Purdue University, West Lafayette, IN 47907. USA,
{ruth, dxu, bb}@cs.purdue.edu, fregnier@purdue.edu

Abstract. This paper presents the design of a new middleware which provides support for trust and accountability in distributed data sharing communities. One application is in the context of scientific collaborations. Multiple researchers share individually collected data, who in turn create new data sets by performing transformations on existing shared data sets. In data sharing communities building trust for the data obtained from others is crucial. However, the field of data provenance does not consider malicious or untrustworthy users. By adding accountability to the provenance of each data set, this middlware ensures data integrity insofar as any errors can be identified and corrected. The user is further protected from faulty data by a *trust view* created from past experiences and second-hand recommendations. A *trust view* is based on real world social interactions and reflects each user's own experiences within the community. By identifying the providers of faulty data and removing them from a *trust view*, the integrity of all data is enhanced

1 Introduction

In scientific research, scientists rely on experimental data to demonstrate their findings. The accuracy of the data is critical not only for the validity of the research results but also for the reputation of the scientist. Currently, a scientist's professional reputation is determined by peer review of papers submitted to conferences and journals for publication. Frequently, results obtained are based on complete data that does not accompany the paper. It is assumed that the integrity of the data has been maintained throughout.

To complicate matters even more, the recent growth in processing power and storage capacity along with the ease of communication through the Internet, has allowed scientists to create and process very large data sets based on locally derived data as well as data obtained from other scientists. Although a large data set can provide better results because of larger and more diverse sampling,

⋆ Department of Computer Science
† The Center for Education and Research in Information Assurance and Security (CERIAS)
‡ Department of Chemistry and The Bindley Bioscience Center at Purdue University

in order to be confident with the results, the origin of all data in the set must be known. In most scientific communities there is no standardized method for collecting and sharing data, which makes it difficult to achieve global data consistency, validity, and credibility. More specifically heterogeneity between labs may lay in the following:

- The condition and calibration of experimental instruments in different labs, and the condition and configuration of lab environments.
- The context of different experiments, such as the time, location, temperature of the experiments, and in the case of medical or social experiments, the age and ethnic group of human subjects.
- The protocol (and the strictness of its enforcement) of data generation, transformation, and derivation. For example, different labs may use different sampling rates, precision, and number of repetitions.
- The capacity, version, and configuration of computing platforms (both software and hardware) in different labs.
- Non-uniform data formats adopted by different labs, due to their formatting conventions and differences in software/hardware/instruments.

There is a need for a distributed environment that allows researchers to collaborate by sharing data while maintaining the complete history and source of all data sets. This by necessity would include those smaller sets which constitute the greater accumulation of data and the transformations from which they were combined. The field of data provenance is evolving out of this concern [2, 4, 6–8, 10–12, 14, 16, 17, 20, 25, 27]. Data provenance is the description of the origins of a piece of data and the process by which it arrived in a database [7]. Data provenance is often used to validate data or re-execute a derivation with different input parameters. Currently the field of data provenance is working on how to annotate large ad-hoc data sets in order to identify and correct erroneous data or rederive data sets based on new input. However, the existence of malicious and incompetent users has not been considered. To date, data provenance projects have considered all participants to be trustworthy and meta-data to be correct.

We have determined that data provenance schemes can also be used to store information regarding the validity of data sets. Similar to how the scientific community performs peer reviews on scientific research, shared data sets can be subjected to peer review before they are widely accepted. Users of the shared data set will be able to assess the set's integrity through a similar analytic process as that employed in the peer review process and malicious or incompetent users will be exposed.

In most fields of science, instruments for collecting data and the algorithms to operate on data are constantly advancing. Ideally, any system which expedites the communal sharing of data should record all of the context information related to the data's collection and transformation. By using our system, an individual scientist may investigate the history of a particular data set to determine if s/he disagrees with any collection techniques or transformation algorithms used to construct it. The scientist could then explore whether users with a previously

determined low reputation collected or derived any part of the data. Thus, by allowing the examiner to asses the product as a sum of its parts, s/he can produce a thorough peer review of the data.

It is important to investigate the source of collaborative data. Errors made at very low levels may never be seen once the data is integrated and replicated multiple times. One might consider the affect a misconfigured instrument may have on data obtained in any given data set. Unless the configuration/calibration of each instrument used to collect the data is recorded, it may never be possible to identify the problem at a later stage.

A more specific example is found in bioinfomatics. Here the functional annotation of proteins by genome sequencing projects is often inferred from similar, previously annotated proteins. The source of the annotation is often not recorded so annotation errors can propagate throughout much of the data base [5, 9, 13, 15, 21]. In extreme cases, data may be faked intentionally. In 2002, an external committee concluded that former Bell Labs researcher Hendrik Schön. manipulated and misrepresented data in 16 papers involving 20 co-authors [3]. These co-authors and an unknown number of scientists who have used parts of the 16 falsified papers all blindly trusted the integrity of Schön's data. It has been suggested, in the wake of this incident that the research community adopt a data auditing and validation system which can help verify the integrity of data and results independently.

We have designed a system that records the history of a data set similar to other data provenance systems which use a directed acyclic graph (DAG). However, our system establishes a cryptographic signature for each data set and its history. A user will then be accountable for the validity of each signed data set. If a data set contains material contributed by many other sources, the identity of those sources will be included in the history of the larger set. In this way, not only can the original faulty data set be found, but the researcher who made the mistake can be held accountable.

Once a system of accountability is in place, a trust management system based on real world notions of trust and reputation can be implemented. This will go a long way towards increasing the probability that an individual data set is valid and increase the integrity of the data in the entire community. Users will interact with each other and record their experiences. Each user will individually evaluate the probable integrity of each piece of data based on the unforgeable and irrefutable information contained in the signed histories, his or her personal experiences, and the recommendations of others.

The remainder of this paper is organized as follows: section 2 gives a brief summary of current related work, section 3 provides an overview of our project goals, section 4 shows the basic architecture of the system, section 5 provides the data history data structure, section 6 gives a detailed description of how *trust views* are determined and implemented, and the final sections talk about future work and our conclusions.

2 Related Work

Recently there has been increased activity in the area of data provenance. Research is focused on providing the correct annotations for recording the history of data. What follows is a very brief description of several data provenance projects.

The Chimera project [2, 10, 11] is is a system developed as part of the Grid Physics Network (GriPhyN) project. Chimera provides for ad-hoc sharing and creation of distributed data sets while recording the history of each data set. The purpose of Chimera is to play the role of a makefile for large systems comprised of many data sets distributed across a network. This distributed makefile allows for the recreation of large data sets when a smaller sub-data set is changed. However, Chimera neither provides accountability for the shared data, nor helps users determine which data sets are most likely to consist of valid information.

Earth System Science Workbench (ESSW) [4, 12] is for data centers studying earth sciences. The goal of the project is to allow participating data centers to search and obtain data while publishing their own data. This project does not consider malicious or incompetent users in the system.

The myGrid project [16, 25, 27] has more capabilities. myGrid is a complete e-Science system in which not only data will be shared but all electronic resources including: instruments, sensors, data, and computational methods. In essence, myGrid provides for an experiment to be done completely *in silico*. However the data in the system is assumed to be correct and of high integrity.

ESP2Net [20] is developing a Scientific Experiment Markup Language (SEML). SEML is based on XML and is a language which requires data provenance information be stored with all data. SEML is aimed at scientific data sharing.

The PENN Database Research Group led by Peter Bunemen [7] has done significant work at the lower levels of data provenance. Their work is focused on how to record data provenance within a database and does not consider the peer-to-peer relationships formed by the various data providers.

Audun Jøsang [18, 19, 24] has concentrated his research on the theoretical side of trust. Most of his work in the logic of trust relationships. More recently he has studied the trust relationship between agents in e-commerce.

Alfarez Abdul-Rahman [1] proposed a model of trust that mimics the real world trust each of us exhibits everyday when dealing with other people. His trust model allows for each participant to form their own opinion of other peers based on his or her own experiences with the system. Each user will independently form this opinion and the opinion with change as more experiences are created. We use Abdula-Rahman's trust model in our system.

RAID lab [26] has developed an approach to establish the trust of a principal, *Alice*, based on her history (i.e. a sequence of trust establishment events that involved *Alice*). They assume that *Alice* obtains a rating for each event that characterizes her behavior. Their approach is context-sensitive in that it considers the ratings and attributes associated with trust establishment events such as risk and event sequence patterns.

3 Overview

The goal of our research is to design and prototype an accountable, trust-aware, and data-centric e-notebook middleware. This e-notebook middleware is distributed, running on machines in individual research labs and possibly on larger servers (for example a campus-wide e-notebook that could be created at a university or large company). The e-notebook will record (1) the context in which raw data is generated (by communicating with on-board software) and (2) the history of curated data including data transformation, derivation, and validation. The individual, through his or her e-notebook, will digitally sign and be accountable for the result of every process performed. Based on the information recorded and experiences with others participating in the network, the distributed e-notebook will establish and maintain *trust views* for scientists sharing scientific data. We contend that these *trust views* and accountability for each data item will provide a measure of confidence in the shared data similar to the trust gained by the peer review process.

The e-notebook will change the way scientific data is compared and correlated. With the proposed e-notebook, a user will not only judge the value of a data set, the context in which the data was collected and the history (organized as a directed acyclic graph recording the steps of data collection and transformation from the very beginning) of how the data came to be can be used, improving the trustworthiness of scientific discoveries based on such comparison.

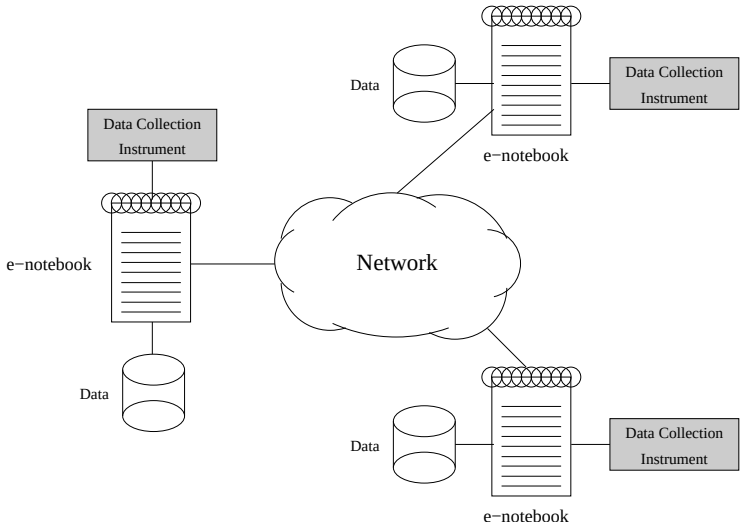

Fig. 1. High level view of the architecture. Users will participate in the community through an e-notebook. Each e-notebook will store some amount of data and participate in the querying for and sharing of data. In addition, each e-notebook will digitally sign and claim responsibility for data sets which it creates.

4 Architecture

The middleware architecture (figure 1) of the proposed system will be highly distributed and flexible. The key element of the system is the e-notebook. Each user will create his or her own notebook and through it collaborate with other users by querying for and retrieving data published on their e-notebooks. The access is also provided to instruments for collecting raw data. The e-notebook will do more than simply collect raw data from the instruments. It will also collect all contextual data (instrument settings, temperature, time, researcher's name, etc.) that the researcher might not think are important. Similar to other data provenance research projects the desired way to accomplish this is for the e-notebook to be connected directly to the instrument's on-board software. It will also be possible for a researcher to input data manually. It should be noted, however, that human error and the common desire to exclude seemingly irrelevant data demonstrates the benefit of automating this process. The e-notebook will also record all applications of transformations on a data set.

In addition to e-notebooks which belong to individual scientist, there may be e-notebooks that reside on servers for the purpose of sharing large amounts of data. An e-notebook of this type will be identical to a regular one and provide a sharing and storage facility for a group of users. Ideal sites for a server e-notebook may include universities and large companies. The only differences between a user e-notebook and a server e-notebook will be the size and the way that it is used. Server e-notebooks will have a larger storage capacity and higher bandwidth capabilities. A server e-notebook's intent is to provide a large repository for storing data that regular users might not want to store locally. The server e-notebook will query for and download any and all data which is to be shared. It may be desirable for the owner of a server e-notebook to allow other users to upload data to the server themselves.

5 Data History and Evidence

When data is collected, transformed, and combined in a distributed ad-hoc manner by different people with different agendas, the temporal history of the data is often lost. Data provenance is the recording of meta-data which describes the history of a data set. Our design of the data provenance system not only records the history of the data, but extends the current systems to include unforgeable and irrefutable evidence of what happened to the data and who performed those actions.

We use a data provenance scheme, similar to current data provenance systems [2, 4, 6–8, 10–12, 14, 16, 17, 20, 25, 27] in which a directed acyclic graph (DAG) is used to describe a data set's history (figure 2). In our design, a data set's DAG is a digitally signed claim of the history of a data set made by the user of the e-notebook from which it was created. Each node in the DAG contains a single data set and information describing how it was created. Some data sets are collected directly from instruments while others are created by performing transformations on one or more existing data sets. For each data set created through

transformations there will be a directed edge from the data set's node to each node used as input to the transformation. In figure 2, data sets 1-3 were collected directly from instruments while data sets 4-6 were the results of transformations.

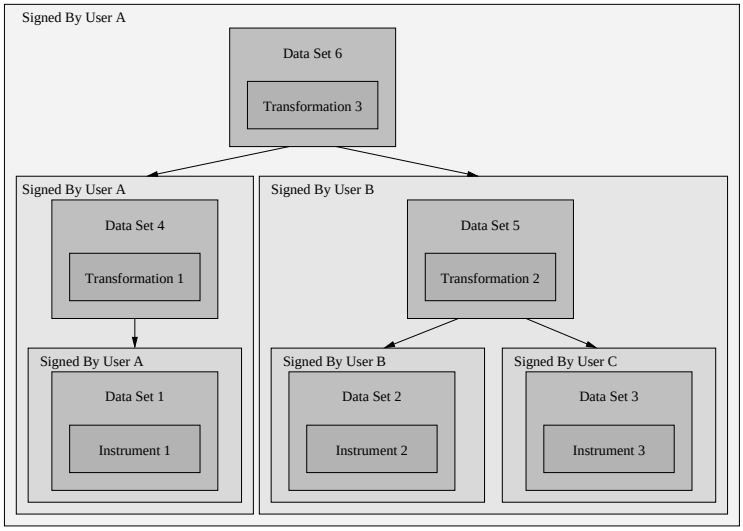

Fig. 2. Each data sets has a directed acyclic graph (DAG) that stores its history. Each DAG is created and digitally signed by a user. A user's digital signature is unforgeable and irrefutable evidence of how the data was created and who created it. Within each DAG a signed sub-DAGs can provide the histories of any data sets that contributed to the larger data set. Signed DAGs create accountability that can be used to form reputations.

When a user collects or derives a new data set, s/he creates a new node with directed edges to a copy of each node used to create the new one. The entire DAG is digitally signed by its creator. It is worth mentioning that within the signed DAG each sub-DAG remains digitally signed by and accountable to its original creator.

The purpose of digitally signing the DAG is to establish accountability. When a user signs a DAG, s/he claims creative responsibility and credit for the data. All sub-DAGs will remain signed by and accountable to their creators. The DAG is then published through the e-notebook for download and use by other users.

When a user downloads a data set, that user may wish to investigate its history by searching the DAG. In this manner, s/he can know all transformations which were applied, all users who were involved, and the context in which the data was collected. It can be known if any of the transformations, users, or contexts are inadequate for the intended use of the data set and if necessary, the material may be avoided.

In some cases, downloaded data sets may contain errors (intentional or otherwise). If an error is found, the digitally signed DAG is unforgeable and irrefutable evidence of what happened and who is responsible. At the very least, the evidence of errors in data sets should be used to adjust the reputation of the careless or malicious user, while at the most the evidence can be made public (possible in a court case) to prove malice. Although, the intent of the system is to increase the integrity of all data in the system by discouraging inappropriate use of the system, the evidence of carelessness and malice must be strong enough to present in court for this to be effective.

Figure 2 shows an example DAG for data set 6, which was created and signed by user A. With his signature A is claiming that he created data set 6 using transformation 3 and input data sets 4 and 5. Both data sets 4 and 5 are in turn signed by their creators. In this case data set 4 happens to have been created and signed by A who performed the transformation to create data set 6. The other input data set, 5 was created and signed by B. Because data set 5 is signed by B, A makes no claims to its validity. A only claims that he agreed to the use of data set 5. If data set 5, or any data which went into it, is discovered to be faulty, user A should disband the use of that data set and the creator of the first faulty data set is held accountable.

If any user were to obtain data set 2 and 3 along with transformation 2, s/he can validate user A's claim by recreating data set 5. If it is not possible to recreate data set 5 by applying transformation 2 to data sets 2 and 3, user A did not create data set 5 this way and incorrectly made the claim that s/he did. Once user A digitally signs data set 5's DAG and releases it to the community, user A can never assert s/he did not make this claim. If it can be shown that data set 5 was not created in the way user A claimed it was, the signed DAG is evidence that A released incorrect data to the community. Evidence of malice cannot be shown with the DAG and must be determined in some other way.

6 Trust Views

One novel technique used by our system is the formation of *trust views* resulting from the reputation of an e-notebook user. Using previous, first-hand experience and second-hand recommendations each user will decide how to trust other e-notebook users. As in real world situations involving trust, there is no universal value assigned to the integrity of each user. No person necessarily judges integrity in the same way as someone else. Each user may have his own algorithm for determining the integrity of others. We propose that using the signed history DAGs described in section 5 users have enough information to make value judgments. This will increase the probability of an individual obtaining valid data and raise the integrity of all the data in the system.

6.1 Trust Judgments

In order to create a *trust view*, each user must make judgments of how much and what kind of trust to assign other users. E-notebook users can make trust judg-

ments in any way they wish. At first, users might rely on off-line relationships. However, as experiences with the community increases, it becomes possible to use accountability information obtained from the signed data histories to make judgments about other's findings. There are endless possibilities in which to use signed histories to make trust judgments. Describing them all would be impossible. Listed below are a few properties which might lead to an increase in the level of trust assigned to a user:

- Consistently producing mistake free data sets.
- Quickly modifying data when mistakes are found in lower level data sets.
- Recommending users who provide quality data sets.

Alternatively, the next list of properties might lead to a reduction of a user's trust:

- Creating and signing a data set which is known to be intentionally fraudulent.
- Consistently making unintentional mistakes in the creation of new data sets.
- Using data which are known to be faulty in the creation of new data sets.
- Recommending users who provide faulty data sets.

In addition to personal experiences, trust judgments can be made using second hand recommendations. Building trust in recommendations can initially be done by accepting the positive assessments of other users who are known outside of the system. Once a base of trust has been established, one may trust the recommendation of users who are unknown outside the system.

Abdul-Rahman describes one social model for supporting trust in virtual communities [1]. In this research, agents trust each other by ranking all first-hand experiences into discrete categories (for example: very good, good, bad, very bad). If only first-hand experiences were considered, when deciding on the trust to award another agent the trust category with the most experiences in it is used. However, Abdul-Rahman provides for trusting through recommendations as well. Recommendations are made by sharing assessments based on first hand-experiences. However, an agent cannot use recommended experiences in the same way as first-hand experiences. The technique used is to calculate the semantic difference between recommendations received and first-hand experiences using those recommendations. Future recommendations can be modified by the semantic difference seen in the past to more accurately suggest amounts of trust to award. In other words, for each user who makes recommendations, the receiving users will calculate the typical difference between the recommendation and personally observed outcome. The typical difference can then be applied to adjust future recommendation from that user.

We have designed a similar model of social trust for users to determine the probability that a given data set is valid. In our system, agents are users and the categories are *very trustworthy*, *trustworthy*, *untrustworthy*, and *very untrustworthy*. It should be noted that any finite number of categories will work and we chose four categories to mirror Abdul-Rahman's work. Each user will record all first hand experiences and determine which category each experience should

belong to. At any given time, the trust level determined by first hand experience is the level associated with the category containing the most experiences. For example, if user A has 4 *very trustworthy* experiences and 5 *trustworthy* experiences with user B, then A applies the category *trustworthy* to B.

Recommendations are made by incorporating the experiences of others into one's rating. Each user has his or her own experiences and techniques for categorizing the experiences. For this reason, another user's recommendation must be adjusted to approximately fit his or her categorizations. To do this the past recommendations and the user's resulting experiences are used to find the semantic difference between the recommendations and his or her experiences. This is done as described in Abdul-Rahman's paper [1]. The semantic difference is then used to adjust each future recommendation. To complete the example, remember that user A determined that user B deserves the trust category of *trustworthy*. If user C has determined (from previous experiences) that when user A recommends *trustworthy*, C's personal experience has shown that a *untrustworthy* experience usually occurs. In this case the semantic difference says to reduce A's recommendation by one category. Therefore, C would adjust A's *trustworthy* recommendation to that of *untrustworthy*.

6.2 Trust Implementation

We propose a novel application of Role-based Trust-management language (RT_0) [22] to implement the social trust model described above. RT_0 uses *credentials* to delegate trust roles from one entity to another. Determining if an entity can have a particular role relies on finding a *credential chain* between the entity and the ultimate authority on that role. What follows is some background on RT_0 and credential chains.

Background on RT_0 and Credential Chains In RT_0 entities (users) declare *roles* of the form $U.r$, where U is a user and r is a role. Users can issue four types of *credentials*:

- Type 1: $U_1.r \longleftarrow U_2$
  Entity U_2 is a member of U_1's role $U_1.r$. U_1 and U_2 may be the same user.
- Type 2: $U_1.r_1 \longleftarrow U_2.r_2$
  All members of $U_2.r_2$ are to be included as members of $U_1.r_1$. U_1 and U_2 may be the same users. r_1 and r_2 may be the same roles.
- Type 3: $U_1.r_1 \longleftarrow U_1.r_2.r_3$
  Any member of $U_1.r_2$ (say U_2) is allowed to determine members of $U_1.r_1$ by adding a credential $U_2.r_3 \longleftarrow U_3$.
- Type 4: $U_1.r \longleftarrow f_1 \cap f_2 \cap ... \cap f_k$
  The intersection of any number of roles and users.

As an example we present a naive, but valid, strategy for the creation of *credential chains* for the purpose of recommending trust.

Each user i creates a role $U_i.trusted$. For each other user U_j that U_i trusts, U_i issues the credential:

$$U_i.trusted \longleftarrow U_j \tag{1}$$

In this simple case, determining if U_a trusts U_b is done be finding the *credential chain* (the number over the arrow refers back to the credential number as labeled in the paper):

$$Chain : U_a.trusted \xleftarrow{1} U_b$$

User U_c can be indirectly trusted by U_a by the appropriate users issuing the credentials as follows:

$$U_a.trusted \longleftarrow U_b \tag{2}$$
$$U_a.trusted \longleftarrow U_b.trusted \tag{3}$$
$$U_b \longleftarrow U_c \tag{4}$$

The *credential chain* that allows U_c to have the role $U_a.trusted$ is:

$$Chain : U_a.trusted \xleftarrow{3} U_b.trusted \xleftarrow{4} U_c$$

Although this set of *credentials* is useful it has a draw back. All users who are directly or indirectly trusted by U_b are trusted by U_a. Since U_a might trust U_c's data sets, but not trust U_c's recommendations, we need a more powerful set of *credentials*.

This example has shown the basic features of RT_0. Users in our system will be able to use any strategy they wish for creating *roles* and *credential*. The next section describes a better suggested strategy for creating *roles* and *credentials*.

Credential Chain Strategy We have created a strategy for creating roles and credential rules that allow for the implementation of the social trust model described in section 6.1. The trust model, as presented, provides four categories of trust: *very trustworthy*, *trustworthy*, *untrustworthy*, and *very untrustworthy*. Again, these categories were chosen because of their similarity to Abdul-Rahman's examples. However, any finite number of categories can be chosen.

Any user i subscribing to our strategy will first create four basic trust *roles*:

$U_i.vt$: very trustworthy
$U_i.t$: trustworthy
$U_i.ut$: untrustworthy
$U_i.vut$: very untrustworthy

A user j is awarded a certain amount of trust depending on which of four *roles* applies to that user. *Credentials* are needed to assign these *roles* to users. This set of *credentials* has to do with the first-hand experiences a user has had. These *credentials* require the creation of four additional *roles*.

$U_i.exp\_vt$: Users awarded *very trustworthy* by first-hand experiences
$U_i.exp\_t$: Users awarded *trustworthy* by first-hand experiences
$U_i.exp\_ut$: Users awarded *untrustworthy* by first-hand experiences
$U_i.exp\_vut$: Users awarded *very untrustworthy* by first-hand experiences

Because personal experience is always more important than recommendations the first-hand experience roles will directly linked to the basic roles by the *credentials*:

$$U_i.vt \longleftarrow U_i.exp\_vt \tag{5}$$
$$U_i.t \longleftarrow U_i.exp\_t \tag{6}$$
$$U_i.ut \longleftarrow U_i.exp\_ut \tag{7}$$
$$U_i.vut \longleftarrow U_i.exp\_vut \tag{8}$$

If most of U_i's first-hand experiences with U_j are good experiences, U_i will create a credential rule $U_i.exp\_t \longleftarrow U_j$. The role $U_i.t$ is given to U_j by the *credential chain*:

$$Chain : U_i.t \xleftarrow{5} U_i.exp\_t \longleftarrow U_j$$

Next, *credentials* need to be created to incorporate second hand recommendation of other users. If the other user subscribes to this strategy, s/he will record his or her first-hand experiences and create *credentials* according to these experiences.. A user will link to his or her first-hand experience roles in a manner consistent with the trust model. In the model, a user must record recommendations of other users and compare these recommendations with his or her own first-hand experiences. The difference between the recommended values and the observed values will be applied to all new recommendations as an adjustment. The effect on *credential* will be that a recommendation by U_j of role $U_j.t$ may be, in U_i's eyes, equivalent to $U_i.ut$. This will be the case when U_j rates others higher than U_i, possible because his or her standards are lower. U_i may adjust U_j's recommendations by submitting the *credentials*:

$$U_i.t \longleftarrow U_j.exp\_vt \tag{9}$$
$$U_i.ut \longleftarrow U_j.exp\_t \tag{10}$$
$$U_i.vut \longleftarrow U_j.exp\_ut \tag{11}$$
$$U_i.vut \longleftarrow U_j.exp\_vut \tag{12}$$

If U_j had first-hand experiences with U_k which produced the *credential* $U_j.exp\_t \longleftarrow U_k$, the *credential chain* from U_k to U_i would grant U_k the role $U_i.ut$ and would be:

$$Chain : U_i.ut \xleftarrow{10} U_j.exp\_t \longleftarrow U_k$$

In this case U_i has determined that U_j usually recommends at one level higher than U_i's personal experience shows to be true. All of the recommendations have been adjusted down by one level. Notice that U_i will not except any of U_j's recommendation to the *role* $U_i.vt$. In general, the transformation from U_j's recommendations to U_i's trust values does not have to adjust all levels in the same direction or by the same about. As an example. U_i's experience with U_k may l produce the *credentials*:

$$U_i.vt \longleftarrow U_k.exp\_ut \tag{13}$$
$$U_i.t \longleftarrow U_k.exp\_t \tag{14}$$
$$U_i.ut \longleftarrow U_k.exp\_vut \tag{15}$$
$$U_i.vut \longleftarrow U_k.exp\_vt \tag{16}$$

This situation probably would not happen, but is still acceptable.

If there are a significant number of users making recommendations, there may be conflicting results of the *credential chains* (more than one basic role may be applied to a single user). For this reason the final decision on the appropriate role to apply to the user is made by counting the number of times each *role* is applied. In a similar fashion to the model, the *role* that was applied most is chosen. A user may even weight recommendation to achieve a weighted sum.

For the trust model to work each user should follow this strategy for creating *credentials* based on the semantic differences between his or her own experiences and the recommendations of others. However, if any user accidentally or maliciously creates faulty *credential chains*, the semantic differences applied to that user will adjust the recommendations accordingly.

There are many other possible strategies using RT_0 and *credential chains*. We plan on developing more and studying how different strategies interact with each other.

7 Future Work

We have many ideas for increasing the capabilities of our system. First, we would like to look at how much of the credential creation can be automated. Currently, validation of data sets must be done manually and rating of first-hand experiences must be done by a human. We think that some decisions about experiences can be automated and the credential chains can be updated accordingly.

Second, we would like to look at different strategies that users may use in determining trust and creating credential chains. We expect to find that not all strategies work well together and would like to answer these questions: Which strategies do work together? Is there a best strategy? If so, what is the best strategy?

Li [22] has proposed algorithms for distributed credential chain discovery. We would like to extend Li's work by discovering not just credential chains, but also directed credential graphs. It may be that a user trusts data using several

different credential chains that form a directed graph. This graph could be used to find the chain that provides the greatest amount of trust.

We would also like to find credential chains or graphs with which we can find data that is trusted by some set of users. This could be used by the community to find the data sets which the community as a whole tends to trust. This data would be the best to use when drawing results to be presented to the community.

8 Conclusion

This paper proposes an e-notebook data sharing middleware for scientific collaboration. The aim of the system is to create a virtual community where scientists sharing files are accountable for the files they share. We would also like to encourage the formation of natural *trust views* among these scientists. Accountability for shared data and the repercussions of obtaining a negative reputation will not only help scientists identify valid data but raise the integrity of the data in the entire system. Future research will refine the trust model as well as the data history with the goal of creating distributed community file sharing systems with integrity similar to the professional peer review process in which malicious or incompetent users are exposed and there contributions are removed.

Acknowledgments

This work is supported in part by a grant from the e-Enterprise Center at Purdue University, a gift from Microsoft Research, and grants from the National Science Foundation IIS0209059 and IIS0242840.

References

1. Abdul-Rahman, A., Hailes, S.: Supporting trust in virtual communities. In: Proceedings of the 33rd Hawaii International Conference on Systems Sciences. (2000)
2. Annis, J., Zhao, Y., Vockler, J., Wilde, M., Kent, S., Foster, I.: Applying chimera virtual data concepts to cluster finding in the sloan sky survey. (2002)
3. Beasley, M., Datta, S., Kogelnik, H., Kroemer, H., Monroe, D.: Report of the investigation committee on the possibility of scientific misconduct in the work of Hendrik Schön and co-authors. Technical report (2002)
4. Bose, R.: A conceptual framework for composing and managing scientific data lineage. In: 4th International Conference on Scientific and Statistical Database Management. (2002) 15–19
5. Brenner, S.E.: Errors in genome annotation. Trends in Genetics **15** (1999) 132–133
6. Buneman, P., Khanna, S., Tajima, K.: Data archiving. In: Workshop on Data Derivation and Provenance. (2002)
7. Buneman, P., Khanna, S., Tan, W.C.: Why and where: A characterization of data provenance. In: International Conference on Database Theory (ICDT). (2001)
8. Cavanaugh, R., Graham, G.E.: Apples and apple-shaped oranges: Equivalence of data returned on subsequent queries with provenance information. In: Workshop on Data Derivation and Provenance. (2002)

9. Devos, D., Valencia, A.: Intrinsic errors in genome annotation. Trends in Genetics **17** (2001) 429–431
10. Foster, I., Vockler, J., Wilde, M., Zhao, Y.: Chimera: A virtual data system for representing, querying, and automating data derivation. In: Proceedings of the 14th International Conference on Scientific and Statistical Database Management. (2002)
11. Foster, I., Vockler, J., Wilde, M., Zhao, Y.: The virtual data grid: A new model and architecture for data-intensive collaboration. In: Proceedings of the 2003 CIDR Conference. (2003)
12. Frew, J., Bose, R.: Earth system science workbench: A data management infrastructure for earth science products. In: SSDBM 2001 Thirteenth International Conference on Scientific and Statistical Database Management. (2001) 180–189
13. Galperin, M., Koonin, E.: Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. Silico Biol. **1** (1998) 55–67
14. Gertz, M.: Data annotation in collaborative research environments. In: Workshop on Data Derivation and Provenance. (2002)
15. R.Gilks, W., Audit, B., Angelis, D.D., Tsoka, S., Ouzounis, C.A.: Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics **18** (2002) 1641–1649
16. Greenwood, M., Goble, C., Stevens, R., Zhao, J., Addis, M., Marvin, D., Moreau, L., Oinn, T.: Provenance of e-science experiments - experience from bioinformatics. In: Proceedings UK e-Science All Hands Meeting 2003 Editors. (2003)
17. Howe, B., Maier, D.: Modeling data product generation. In: Workshop on Data Derivation and Provenance. (2002)
18. Jøsang, A.J.: The right type of trust for distributed systems. Proceedings of the 1996 New Security Paradigms Workshop (1996)
19. Jøsang, A.J., Hird, S., Faccer, E.: Simulating the effect of reputation systems on e-markets. Proceedings of the First International Conference on Trust Management (2003)
20. Kaestle, G., Shek, E.C., Dao, S.K.: Sharing experiences from scientific experiments. In: Proceedings of the 11th International Conference on Scientific and Statistical Database Management (SSDBM). (1999)
21. Karp, P.D.: What we do not know about sequence analysis and sequence databases. Bioinformatics **14** (1998) 753–754
22. Li, N., Winsborough, W.H., Mitchell, J.C.: Distributed credential chain discovery in trust management: extended abstract. In: ACM Conference on Computer and Communications Security. (2001) 156–165
23. Mann, B.: Some data derivation and provenance issues in astronomy. In: Workshop on Data Derivation and Provenance. (2002)
24. Patton, M.A., Jøsang, A.: Technologies for trust in electronic commerce. Electronic Commerce Research Journal **4** (2004)
25. Stevens, R.D., Robinson, A.J., Goble, C.A.: ^{my}Grid: Personalized bioinformatics on the information grid. Bioinformatics **19** (2003) i302–i304
26. Zhong, Y., Lu, Y., Bhargava, B.: Dynamic trust production based on interaction sequence. Technical Report CSD TR 03-006, Department of Computer Sciences, Purdue University (2003)
27. Zhao, J., Goble, C., Greenwood, M., Wroe, C., Stevens, R.: Annotating, linking and browsing provenance logs for e-science. In: Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data. (2003)