

# Research Statement

Ragib Hasan  
Email: rhasan7@jhu.edu

## 1 Research Accomplishments

My general research interests fall within the area of data security and trustworthiness in databases and file systems. More specifically, I am interested in data and storage security, with an emphasis on securing data history. While the digital nature of modern information has brought enormous benefits, it has also created new vulnerabilities. Unlike physical documents, digitally stored information can be rapidly copied, erased, or tampered with. The origin and derivation history of an object are also important in many applications. But in a distributed and potentially untrusted environment, both data and its history are subject to tampering, especially by insiders with financial or strategic motives. Access control alone cannot provide the necessary protection to data, when insiders with superuser access rights become adversaries. Therefore, there is an urgent need to develop schemes for ensuring the trustworthiness of data history.

In the early part of my PhD work, I developed a generic threat model for storage [HMLY05] and analyzed storage security breach incidents [HY06]. This initial exploration showed that data integrity is a critical problem in data processing. Many storage security issues in real life result from illicit modification of data in storage, and from not having trustworthy information about the data objects history. The focus of my dissertation research is, therefore, to develop tools for ensuring the integrity and confidentiality of data history. In particular, I have focused on areas where integrity of data history is important – e.g., data provenance in files, and trustworthy history preservation in regulatory compliant databases. For my dissertation, I have developed practical and efficient schemes to ensure protection of history from illicit tampering.

### 1.1 Data Provenance

Data provenance provides information about the history of a data object, i.e., the origin, lineage, and transformation of the object through its lifetime. Used extensively in scientific computing for many years, data provenance has recently been used in mainstream applications such as business, government, and health care. However, provenance information cannot be relied upon unless we can guarantee its integrity and trustworthiness. As data and its provenance flow between people and tasks in potentially untrusted environments, it becomes essential to provide integrity and confidentiality assurances for provenance. In my research, I have created schemes for providing integrity, confidentiality, and privacy guarantees for data provenance [HSW07, HSW09c, HSW09a]. As part of my research, I also developed proofs of correctness, as well as mechanisms for efficient storage and compaction of provenance records. Finally, I have developed the first ever implementation of a secure provenance system, which enabled applications to record secure provenance for files at a very low cost – 1%-3% for most real life file system workloads.

## 1.2 History Integrity in Regulatory Compliant Databases

Generic techniques developed for ensuring trustworthy history may be inefficient when we have to satisfy application specific requirements. So, in the second part of my dissertation, I focused on a particular application area – protecting the integrity of data history in database systems. Governments around the world have created legislation mandating the integrity of stored data. Regulations such as the Sarbanes-Oxley Act, HIPAA, and SEC Rule 17-a4 require data records to be term-immutable, i.e., to be retained in a tamper-evident manner for a multi-year retention period. This is especially important for the financial sector, where violations can bring strong monetary penalties and criminal prosecution. To address this problem, I have developed an efficient database architecture that provides audit-based compliance with these regulations [HWM09]. The new architecture is 10 times faster than the state of the art, and provides 100-fold faster audits. I have also developed the first ever formalization and implementation of litigation holds for database records [HW09].

## 1.3 Remembrance

Finally, my research looks into generalizing history, provenance and versioning of digital objects into the notion of remembrance, i.e., the ability of data to remember its past values and contextual information. I have proposed augmenting data objects in a system with the remembrance capability in order to provide a richer, more expressive data processing paradigm [HSW09b]. With remembrance as an intrinsic property of all data objects and system components, data objects no longer remain as dumb containers for information. Rather, we can perform informed computing based on not just the present, but the entire history or “memory”/remembrance of data.

## 1.4 Miscellaneous

Besides my dissertation research, I have developed a permit-based authorization system for mashups during my internship at Google’s security group. I was also actively involved in designing a multi-level security and authorization scheme for smart buildings. A prototype of our system is currently in the process of being deployed at Siebel Center for Computer Science at UIUC. Besides this, I have also collaborated with several research groups at NCSA, where I designed and implemented simulation tools for analyzing the security of the power grid.

# 2 Research Goals

In my future career, I want to change the world of data management by making it more trustworthy, through development of schemes to protect stored information against insider threats. In addition, I want to disseminate knowledge by developing teaching materials focusing on different aspects of data security and provenance. To achieve these goals, I want to pursue several directions in data security in files and databases, and continue my work towards making data trustworthy. Specifically, I want to focus on the following directions:

## 2.1 Securing provenance

Data provenance is an emerging research area. As provenance becomes widely used in the mainstream of computing, the need to strengthen different aspects of provenance security will become crucial, especially in high-stakes applications such as business and health-care. In my

dissertation, I have developed schemes for protecting integrity of provenance. However, there are still many open problems in provenance security. Among these problems, I want to pursue research on provenance security in three areas: formal analysis, confidentiality and access control, and support for privacy-preserving operations on provenance.

Researchers have recently started to look into the semantics of provenance security – emphasizing the fact that data and its provenance can have different confidentiality and privacy requirements. I want to explore the theoretical aspects of provenance security, and develop formal models for the semantics of provenance security. Next, I want to develop provably secure schemes for access-control, i.e., how to decide who should be able to access a given provenance record for a data object.

Privacy of provenance records is also a challenging problem – and in this area, I will design schemes to protect privacy of the data history while retaining the usefulness of provenance. In particular, performing privacy preserving queries on provenance data without leaking sensitive attributes is difficult, and I want to explore techniques to solve this problem.

Finally, I also want to work on the systems and deployment aspect of secure provenance. Making operating systems provenance-aware is challenging, as scalability and performance issues limit the use of many strong cryptographic techniques. To allow seamless deployment of security, we need to create an efficient, low-overhead, and generic tool. To this end, I want to develop a scalable, implementation-independent secure provenance system which can be used in conjunction with any provenance system or operating system. As this tool will be independent of the provenance formats as well as collection mechanisms, it can easily be used with many existing systems. This secure provenance tool will act as a middleware between the provenance collection layer and storage layers, and will have options for storage and cryptographic plug-ins – allowing usage of different storage and security schemes.

## **2.2 Application of secure provenance in databases**

Till now, most of the common uses of provenance in computing have been in the field of scientific computing. However, over 90% of today's business data is maintained electronically, and is therefore subject to tampering attacks. I want to develop architectures for data storage in enterprise class systems and databases with support for secure data provenance. This integration of provenance is not limited to storage alone – to have a complete picture of data history, we also need to explore techniques for securing the provenance of network communication as well as application execution provenance. Therefore, in future, I want to explore different techniques for securing the provenance in operating systems and network communications.

## **2.3 Application of secure provenance in social networks**

Today, social networking has become ubiquitous. However, it has become difficult to determine the trustworthiness of information propagating via social networks. I believe that secure history and provenance can provide a solution to this problem. For example, the provenance of a data item being propagated, exchanged, and disseminated via social networks can be used by a user to decide whether to trust it or not. Secure data provenance will ensure that an adversary did not tamper with the history for his/her own benefit.

## **2.4 Provenance in the clouds**

With the advent of Web 2.0 and cloud-computing, the whole paradigm of computing is at a crossroads. I envision that in the next few years, a large part of computing will shift to the

cloud-computing model. This switch will also bring new challenges related to the trustworthiness of data as well as computations performed in third-party clouds. More specifically, in a cloud-computing scenario, most of the data will reside in data clouds, while the applications will run in the clouds. In such a setup, clients require additional assurances regarding the confidentiality and integrity of their data and applications. I believe that provenance can play a crucial role here, and by augmenting cloud computing operations with provenance information, and ensuring the integrity of provenance, we can make cloud computing trustworthy.

## 2.5 Remembrance-based systems

In my CIDR 2009 paper, I proposed the concept of remembrance – i.e., augmenting all data objects with persistent memory, and considering history as an intrinsic property of data objects [HSW09b]. In this vision, everything in a computer, ranging from coarse-grained object such as files, to fine-grained data variables inside an application, will retain (part of) their previous states. Computations will be memory-aware – copying or other operations will carry over the memory of input objects to output objects. This will be very useful in distributed computing, especially in the cloud computing scenario, where a given piece of data item can originate from potentially untrusted principals or applications. Using such untrustworthy data as input in a computation will impact the output. If we retain the memory about the origin of data, we can later verify the entire derivation of an output by looking into the collection of memories.

Many past attempts at memory-retentive objects have focused only on piecemeal solutions to this problem. For example, provenance, versioning file systems, snapshots, transaction-time databases, etc., focus on retaining the memory of files, but this does not provide an end-to-end solution. When data is read by an application from a file with a versioned history, and then sent over the network, the original history is not retained, unless the application has built-in support for that. I would like to design systems that consider data and its history as an atomic container, and study the challenges, and performance issues in designing such a system. As an immediate goal, I would like to enhance the Linux operating system kernel with remembrance functionality for kernel data structures as well as integrated support for provenance and memory retention for files and network communication.

## 3 Broad Impact and Societal Benefits of My Research

Secure provenance and its application in different areas will have a significant impact on computing. First, it will allow us to look into the past of data objects. Based on the history, we can make informed decisions about the trustworthiness of the data item. As we move into the age of information, we also face challenges regarding the trustworthiness of information we receive from other and potentially untrusted people or systems. Securing data provenance will help ensure the trustworthiness we need for digital data. Next, my research in different applications of provenance will introduce trustworthy provenance in diverse areas such as cloud computing, network communication, and operating systems. To the best of my knowledge, there are no existing research projects that aim at using provenance in such systems in an end-to-end manner. Therefore, my proposed research will not only break new ground, but will also advance the state of the art.

In the bigger picture, research on data provenance security has significant societal benefits as well. If we can make digital data trustworthy with secure provenance, trust issues in many critical areas such as business and health-care can be greatly reduced. This is because the information in ordinary digital records can be tampered with, copied, or erased quickly and

silently, and in large quantities. Many of the corporate scandals of recent years involved illicit modification of accounting and financial data. If we can protect the integrity of data and its history, we can protect against such adversaries, leading to widespread adoption of trustworthy digital data processing. This will reduce costs, and benefit society as a whole.

## References

- [HMLY05] R. Hasan, S. Myagmar, A. Lee, and W. Yurcik. Toward a Threat Model for Storage Systems. In *ACM StorageSS*, 2005.
- [HSW07] R. Hasan, R. Sion, and M. Winslett. Introducing Secure Provenance: Problems and Challenges. In *ACM StorageSS*, 2007.
- [HSW09a] R. Hasan, R. Sion, and M. Winslett. Preventing History Forgery with Secure Provenance. *ACM Transactions on Storage (TOS)*, November 2009.
- [HSW09b] R. Hasan, R. Sion, and M. Winslett. Remembrance: The Unbearable Sentience of being Digital. In *CIDR*, 2009.
- [HSW09c] R. Hasan, R. Sion, and M. Winslett. The Case of the Fake Picasso: Preventing History Forgery with Secure Provenance. In *USENIX FAST*, 2009.
- [HW09] R. Hasan and M. Winslett. Trustworthy Vacuuming and Litigation Holds in Long-term, High-integrity Records Retention. UIUC-DCS Technical Report. (also in submission to a conference), 2009.
- [HWM09] R. Hasan, M. Winslett, and S. Mitra. Efficient Audit-based Compliance for Relational Data Retention. UIUC-DCS Technical Report, (also in submission to a conference), March 2009.
- [HY06] R. Hasan and W. Yurcik. A Statistical Analysis of Disclosed Storage Security Breaches. In *ACM StorageSS*, 2006.