

# Research Statement

Ragib Hasan

Email: rhasan7@jhu.edu

Web: <http://www.ragibhasan.com>

Since the dawn of the human civilization, data processing has been an integral part of our lives. But only in recent years, we have had to process massive volumes of data. With the advent of computing, our techniques for processing, storing, and exchanging data have changed. Today, our data is highly mobile, and is processed, modified, and transmitted through different principals throughout its lifetime. Recent progress in distributed computing has brought new paradigms such as cloud computing to the forefront. However, the distributed nature and mobility has made it harder to trust data.

The long term goal of my research is to develop mechanisms and tools that will lead to widespread use of digital data processing in our everyday lives. A fundamental prerequisite of that is to make data processing trustworthy in a distributed environment. In my dissertation research, I have focused on securing history and provenance of data in file systems and database management systems. I have designed, implemented, and evaluated provably secure, efficient, and practical systems for making data history trustworthy in these application domains. Today, our data processing paradigms are moving towards cloud computing and massive online social networks. In my future research, I want to enable widespread adoption of distributed data processing in the real world, by designing schemes and tools to make cloud computing and social networks trustworthy.

## 1 Research Accomplishments

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 object's history. The focus of my dissertation research was, therefore, to develop tools for ensuring the integrity and confidentiality of data history. In particular, I 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 [Has09], I developed practical and efficient schemes to ensure protection of history from illicit tampering. Finally, during the first year of my postdoctoral work, I have mainly focused on the location provenance problem for mobile devices.

### 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 [HW11]. 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 [HW10].

## 1.3 Secure Location Provenance

As mobile computing gets popular, location-based services have become widespread. Many of these services use the location of a user for access-control decisions, authentication, information sharing, and policy evaluations. However, malicious users can misreport their locations. A global system for tracking users is not scalable and can raise severe privacy concerns. So, we need a scheme in which a user can collect location proofs from the locations it visits and later present these proofs to a verifier. The challenge is to ensure that both individual location proofs and the sequence of location proofs, i.e., the user’s location provenance cannot be manipulated or forged by malicious users acting alone, or colluding with the proof-issuers and/or other users. To solve this, I have developed schemes for secure location provenance. My research showed that that none of the previous work in this area are secure against most of the collusion attacks. To tackle these shortcomings, I have designed a witness-attested model for creating endorsed location proofs [HB]. Under such a model, location proofs will need to be endorsed by witnesses that are physically and temporally co-located with the user at the same location. I have also designed a space, energy, and computationally efficient scheme that allows a user to prove the order of any arbitrary subset of her previous locations. I have implemented and evaluated the location provenance schemes on the Google Android Platform, and demonstrated that my schemes are practical and efficient in today’s mobile devices.

## 1.4 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.

## 1.5 Miscellaneous

Besides my dissertation research, I have developed a permit-based authorization system for mashups during my internship at Google’s security group [HCS<sup>+</sup>08]. 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 the University of Illinois. 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 research agenda, I want to focus on the fundamental question: *How can we process massive volumes of data in a trustworthy and scalable manner?* In particular, I want to explore several application domains where large volumes of data need to be processed: cloud computing and social networking. I am also deeply interested in the techniques for securing the history and provenance of data, location of physical and virtual objects and users, and the use of verified provenance information in security.

Today, cloud computing has emerged as the future of computation and data storage. At the same time, emergence of social networking has caused massive amounts of information to be disseminated through the Internet. Cloud computing and social networking open up new problems in trustworthy and efficient data processing. For widespread adoption of cloud computing, we need to bring accountability to the cloud, and provide guarantees on confidentiality, integrity, and availability of the data and computation performed inside a cloud. For information gathered and exchanged in a social network, we need techniques to evaluate the trustworthiness of data which has passed through many principals. Therefore, in the next five years of my research career, I want to develop technology that will lead to widespread adoption of cloud computing as well as trustworthy data dissemination in social networks.

### 2.1 Accountability in cloud computing

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. However, before that can happen, we have to solve some fundamental problems of cloud computing.

*Why hasn’t every organization moved to cloud computing rather than maintaining their own data processing systems?* The answer is: *trust*. Today’s data and compute clouds are essentially opaque systems, where the clients have no control over and limited information about what happens inside the cloud. This in turn leads to less trust on the computation done and data stored in clouds. Lack of accountability and concern over the integrity and confidentiality of data and computation limit widespread adoption of clouds in the mainstream of computing. Clients also do not have any trustworthy method of monitoring the data storage and progress of running jobs. These problems have caused many companies to resort to “private clouds” for their sensitive data and computations, which, unfortunately, is contrary to the main philosophy of cloud computing – to provide computing as a service in a flexible, on-demand manner.

*How can we make clouds accountable?* Accountability in clouds is needed for both customers and cloud providers. A large source of client concern is that, cloud users do not see what happens inside a cloud and how their data is handled. Clients have to fully trust the cloud providers to act honestly and not breach the confidentiality of data and computations. Cloud providers, on the other hand, do not want to disclose the cloud topology and operational details. We need to balance the opposing needs of the providers and clients.

In my future research, I want to make cloud computing more trustworthy and reliable, by bridging the accountability gap. I plan to approach this problem in two directions – by designing cryptographic constructs and mechanisms that will allow the cloud provider to prove the confidentiality and integrity of the data and computation, and by building distributed, efficient, and scalable systems for trustworthy monitoring of clouds without disclosure of sensitive cloud topology information.

One possible approach for overcoming this mutual distrust can be the application of secure data provenance in a cloud environment. My dissertation research on data provenance security lays the groundwork on which we can tackle the problem of data integrity in a cloud. If a client can receive a verifiable and non-forgable provenance of the data item’s journey through a cloud, then the client can determine the trustworthiness of the data item. At the same time, this would benefit cloud providers in a different way – by enabling them to determine the sources of misconfiguration, errors, and possible attacks on the cloud. An accountable cloud would be very attractive to customers worried about data and computational integrity. However, the problem is non-trivial, as the cloud providers control virtually every aspect of a cloud. I would like to extend the data provenance security schemes I have developed as part of my dissertation to take into account the asymmetric trust scenario in a cloud environment.

Monitoring of ongoing cloud computation is a more difficult problem. How can a client know that the cloud provider is accurately running the task? A possible solution is to make performance monitors an integral part of cloud computing systems. However, this is hard since the cloud provider virtually has complete control over the applications run in the cloud. A possibility is to use a combination of trusted hardware to perform remote attestation of the monitors. In my research, I want to design mechanisms to verify the authenticity of the performance monitor outputs.

In summary, the focus of my research in cloud computing would be to design new mechanisms and tools that will make clouds more transparent and accountable, while preserving the confidentiality of topology and other sensitive configuration information. I believe that this transparency will remove the barriers to widespread adoption of cloud computing in business and healthcare applications.

## **2.2 Trustworthiness of data in social networks**

Today, social networking has become ubiquitous. Massive volumes of information are exchanged in social networks everyday. Social networks such as Facebook and Twitter have become part of our media landscape. However, it is difficult to determine the trustworthiness of information propagating via social networks. How do we trust information that has changed many hands and has been aggregated, processed, and summarized through its lifetime? Also, at the scale of billions of users, verifying the identity of users is difficult. We need objective and scalable mechanisms to determine the trustworthiness of such information. 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. My research in this direction will focus on designing objective schemes, algorithms, and systems for determining trustworthiness of information by taking into account its life history.

## References

- [Has09] Ragib Hasan. *Trustworthy History and Provenance for Files and Databases*. PhD thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, October 2009.
- [HB] Ragib Hasan and Randal Burns. Where Have You Been? Secure Location Provenance for Mobile Devices. Under submission.
- [HCS<sup>+</sup>08] Ragib Hasan, Richard Conlan, Brian Slesinsky, Nandakumar Ramani, and Marianne Winslett. Please permit me: Stateless delegated authorization in mashups. In *Proceedings of the 24th Annual Computer Security Applications Conference (ACSAC)*, December 2008.
- [HMLY05] Ragib Hasan, Suvda Myagmar, Adam Lee, and William Yurcik. Toward a Threat Model for Storage Systems. In *ACM StorageSS*, 2005.
- [HSW07] Ragib Hasan, Radu Sion, and Marianne Winslett. Introducing Secure Provenance: Problems and Challenges. In *ACM StorageSS*, 2007.
- [HSW09a] Ragib Hasan, Radu Sion, and Marianne Winslett. Preventing History Forgery with Secure Provenance. *ACM Transactions on Storage (TOS)*, 5(4):1–43, December 2009.
- [HSW09b] Ragib Hasan, Radu Sion, and Marianne Winslett. Remembrance: The Unbearable Sentience of being Digital. In *Proceedings of CIDR*, 2009.
- [HSW09c] Ragib Hasan, Radu Sion, and Marianne Winslett. The Case of the Fake Picasso: Preventing History Forgery with Secure Provenance. In *Proceedings of USENIX FAST*, February 2009.
- [HW10] Ragib Hasan and Marianne Winslett. Trustworthy Vacuuming and Litigation Holds in Long-term, High-integrity Records Retention. In *Proceedings of the 13th International Conference on Extending Database Technology (EDBT)*, 2010.
- [HW11] Ragib Hasan and Marianne Winslett. Efficient Audit-based Compliance for Relational Data Retention. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, 2011.
- [HY06] Ragib Hasan and William Yurcik. A Statistical Analysis of Disclosed Storage Security Breaches. In *ACM StorageSS*, 2006.