WORKSHOP REPORT
BIG DATA SECURITY AND PRIVACY
Sponsored by the National Science Foundation
September 16-17, 2014
The University of Texas at Dallas

ABSTRACT

This report describes the issues surrounding big data security and privacy and provides a summary of the National Science Foundation sponsored workshop on this topic held in Dallas, Texas on September 16-17, 2014. Our goal is to build a community in big data security and privacy to explore the challenging research problems.

Acknowledgements

We thank the National Science Foundation for providing support for this workshop. We also thank the program directors Dr. Chris Clifton and Mr. Jeremy Epstein for their support and encouragement for this workshop. We thank the workshop participants and the organizers of this workshop.

This document was prepared by the following people:

Workshop Chair: Dr. Bhavani Thuraisingham

Workgroup Chairs: Dr. Elisa Bertino and Dr. Murat Kantarcioglu

Workshop Coordinator: Ms. Rhonda Walls

Contact:

Dr. Bhavani Thuraisingham
Louis Beecherl Jr. Distinguished Professor of Computer Science
Executive Director, Cyber Security Research and Education Institute
The University of Texas at Dallas
bhavani.thuraisingham@utdallas.edu
http://www.utdallas.edu/~bhavani.thuraisingham/

@CyberUTD

1. INTRODUCTION

Recently a few workshops and panels have been held on Big Data Security and Privacy. Examples include the ACM CCS workshop on Big Data Security, ACM SACMAT and IEEE Big Data Conference panels. These workshops and panels have been influenced by different communities of researchers. For example, the ACM CCS workshop series is focusing on Big Data for security applications while the IEEE Big Data Conference is focusing on cloud security issues. Furthermore, these workshops and panels mainly address a limited number of the technical issues surrounding big data security and privacy. For example, the ACM CCS workshop does not appear to address the privacy issues dealing with regulations or the security violations resulting from data analytics.

To address the above limitations, we organized a workshop on Big Data Security and Privacy on September 16-17, 2014 in Dallas, Texas sponsored by the National Science Foundation (NSF) [1]. The participants of this workshop consisted of interdisciplinary researchers in the fields of higher performance computing, systems, data management and analytics, cyber security, network science, healthcare, and social sciences who came together and determined the strategic direction for Big Data security and privacy. NSF has made substantial investments both in cyber security and big data. It is therefore critical that the two areas work together to determine the direction for big data security. We made a submission based on the workshop results to the National Privacy Research Strategy [2]. This document is the workshop report that describes the issues in Big Data security and privacy, presentations at the workshop and the discussions at the workshop. We hope that this effort will help toward building a community in Big Data security and privacy.

The organization of this report is as follows. Section 2 describes the issues surrounding Big Data security and privacy. The workshop participants were given these issues to build upon during the workshop discussions. A summary of the workshop presentations is provided in section 3. A summary of the discussions at the workshop is provided in section 4. We submitted a version of this summary to the National Privacy Research Strategy [2]. Next steps are discussed in section 5. We have supplemented this report with three appendices. The workshop agenda is provided in Appendix A. The list of workshop participants is provided in Appendix B. Links to the presentations given at this workshop as well as the position papers submitted are provided in Appendix C.

REFERENCES

[1] http://csi.utdallas.edu/events/NSF/NSF%20workshop%202014.htm

[2] https://www.nitrd.gov/cybersecurity/nprsrfi102014/BigData-SP.pdf

2. ISSUES IN BIG DATA SECURITY AND PRIVACY

2.1 INTRODUCTION

This section describes issues in Big Data security and privacy that were given to the workshop participants to motivate the discussions. These issues include both security and privacy for big data as well as Big Data management and analytics for cyber security. While Big Data has roots in many technologies, database management is at its heart. Therefore in this section we will discuss how data management has evolved and will then focus on the Big Data security and privacy issues.

Database systems technology has advanced a great deal during the past four decades from the legacy systems based on network and hierarchical models to relational and object database systems. Database systems can also now be accessed via the web and data management services have been implemented as web services. Due to the explosion of web-based services, unstructured data management and social media and mobile computing, the amount of data to be handled has increased from terabytes to petabytes and zetabytes in just two decades. Such vast amounts of complex data have come to be known as Big Data. Not only does big data have to be managed efficiently, such data also has to be analyzed to extract useful nuggets to enhance businesses as well as improve society. This has come to be known as Big Data Analytics.

Storage, management and analysis of large quantities of data also result in security and privacy violations. Often data has to be retained for various reasons including for regulatory compliance. The data retained may have sensitive information and could violate user privacy. Furthermore, manipulating such big data, such as combining sets of different types of data could result in security and privacy violations. For example, while the raw data removes personally identifiable information, the derived data may contain private and sensitive information. For example, the raw data about a person may be combined with the person's address which may be sufficient to identify the person.

Different communities are working on the Big Data challenge. For example, the systems community is developing technologies for massive storage of big data. The network community is developing solutions for managing very large networked data. The data community is developing solutions for efficiently managing and analyzing large sets of data. Big Data research and development is being carried out both in academia, industry and government research labs. However, little attention has been given to security and privacy considerations for Big Data. Security cuts across multiple areas including systems, data and networks. We need the multiple communities to come together to develop solutions for Big Data security and privacy. This section describes some of the issues in Big Data security and privacy. An overview of Big Data management and analytics is provided in section 2.2. Security and privacy issues are discussed in section 2.3. Big data management and analytics for cyber security are discussed in section 2.4. Our goal towards building a community is discussed in section 2.5.

2.2 BIG DATA MANAGEMENT AND ANALYTICS

Big Data management and analytics research is proceeding in three directions. They are:

(i) Building infrastructure and high performance computing techniques for the storage of big data;

(ii) Data management techniques such as integrating multiple data sources (both big and small) and indexing and querying big data;

(iii) Data analytics techniques that manipulate and analyze big data to extract nuggets.

We will briefly review the progress made in each of the areas. With respect to building infrastructures, technologies such as Hadoop and MapReduce as well as Storm are being developed for managing large amounts of data in the cloud. In addition, main memory data management techniques have advanced so that a few terabytes of data can be managed in main memory. Furthermore, systems such as HIVE and Cassandra as well as NoSQL databases have been developed for managing petabytes of data.

With respect to data management, traditional data management techniques such as query processing and optimization strategies are being examined for handling petabytes of data. Furthermore, graph data management techniques are being developed for the storage and management of very large networked data.

With respect to data analytics, the various data mining algorithms are being implemented on Hadoop and MapReduce based infrastructures. Additionally, data reduction techniques are being explored to reduce the massive amounts of data into manageable chunks while still maintaining the semantics of the data.

In summary, Big Data management and analytics techniques include extending current data management and mining techniques to handle massive amounts of data as well as developing new approaches including graph data management and mining techniques for maintaining and analyzing large networked data.

2.3 SECURITY AND PRIVACY

The collection, storage, manipulation and retention of massive amounts of data have resulted in serious security and privacy considerations. Various regulations are being proposed to handle Big Data so that the privacy of the individuals is not violated. For example, even if personally identifiable information is removed from the data, when data is combined with other data, an individual can be identified. This is essentially the inference and aggregation problem that data security researchers have been exploring for the past four decades. This problem is exacerbated with the management of Big Data as different sources of data now exist that are related to various individuals.

In some cases, regulations may cause privacy to be violated. For example, data that is collected (e.g., email data) has to be retained for a certain period of time (usually 5 years). As long as one keeps such data, there is a potential for privacy violations. Too many regulations can also stifle innovation. For example, if there is a regulation that raw data has to be kept as is and not manipulated or models cannot be built out of the data, then corporations cannot analyze the data in innovative ways to enhance their business. This way innovation may be stifled.

Therefore, one of the main challenges for ensuring security and privacy when dealing with big data is to come up with a balanced approach towards regulations and analytics. That is, how can an organization carry out useful analytics and still ensure the privacy of individuals? Numerous techniques for privacy-preserving data mining, privacy-preserving data integration and privacy-preserving information retrieval have been developed. The challenge is to extend these techniques for handling massive amounts of often networked data.

Another security challenge for Big Data management and analytics is to secure the infrastructures. Many of the technologies that have been developed including Hadoop, MapReduce, Hive, Cassandra, PigLatin, Mahout and Storm do not have adequate security protections. The question is, how can these technologies be secured and at the same time ensure high performance computing?

Next the Big Data management strategies such as access methods and indexing and query processing have to be secure. So the question is how can policies for different types of data such as structured, semi-structured, unstructured and graph data be integrated? Since Big Data may result from combining data from numerous sources, how can you ensure the quality of the data?

Finally, the entire area of security, privacy, integrity, data quality and trust policies has to be examined within the context of Big Data security. What are the appropriate policies for Big Data? How can these policies be handled without affecting performance? How can these policies be made consistent and complete?

This section has listed just some of the challenges with respect to security and privacy for big data. We need a comprehensive research program that will identify the challenges and develop solutions for big data security and privacy. Security cannot be an afterthought. That is, we cannot incorporate security into each and every big data technology that is being developed. We need to have a comprehensive strategy so that security can be incorporated while the technology is being developed. We also need to determine the appropriate types of policies and regulations to enforce before Big Data technologies are employed by an organization. This means researchers in multiple disciplines have to come together to determine what the

problems are and explore solutions. These disciplines include high performance computing, data management and analytics, network science, and policy management.

2.4 BIG DATA ANALYTICS FOR SECURITY APPLICATIONS

While the challenges discussed in section 2.3 deal with securing Big Data and ensuring the privacy of individuals, Big Data management, and analytics techniques can be used to solve security problems. For example, an organization can outsource activities such as identity management, email filtering and intrusion detection to the cloud. This is because massive amounts of data are being collected for such applications and this data has to be analyzed. Cloud data management is just one example of big data management. The question is, how can the developments in big data management and analytic techniques be used to solve security problems? These problems include malware detection, insider threat detection, intrusion detection, and spam filtering.

2.5 COMMUNITY BUILDING

The various issues surrounding Big Data security and privacy were discussed at the beginning of the workshop and five keynote presentations were given at the workshop that addressed many of these issues. In addition, several position papers were submitted by the workshop participants and subsequently presentations based on these papers were given at the workshop. These papers and presentations set the stage for the two breakout sessions held during the workshop. One of these sessions focused on the security and privacy issues while the other focused on the applications. The presentations and the discussions at the workshop are summarized in sections 3 and 4 of this report. Our goal is to build a community in Big Data security and privacy.

3. SUMMARY OF WORKSHOP PRESENTATIONS

This section summarizes the presentations at the workshop. These presentations and the position papers can be found at http://csi.utdallas.edu/events/NSF/NSF%20papers%202014.htm

Keynote Presentations: We had five keynote presentations to motivate the workshop participants. These keynote presentations discussed the various Big Data security and privacy initiatives at NIST, Honeywell and IBM as well as discussed some of the research challenges. The opening keynote given by **Wo Chang** from NIST discussed the initiatives at NIST on Big Data and provided an overview of the Big Data workgroup. Later **Arnab Roy** from Fujitsu provided some details of the work by the Big Data security and privacy subgroup of this workgroup. **Elisa Bertino** from Purdue discussed issues and challenges of providing security with privacy. **Raj Rajagopalan** from Honeywell discussed big data security and privacy challenges for Industrial Control Systems. **Sandeep Gopisetty** from IBM discussed the Big Data Enterprise efforts at IBM while **Murat Kantarcioglu** from UT Dallas provided an overview of the Big Data security and privacy initiatives at UT Dallas.

There were several presentations given by the workshop participants. Below we give a summary of these presentations.

Towards Privacy Aware Big Data Analytics: Barbara Carminati from the University of Insubria Italy described a framework for privacy aware big data analytics. This framework included layers for privacy policy specifications, a unified query model, fine grained enforcement and a dashboard. She went on to discuss the functions of each layer.

Formal Methods for Preserving Privacy While Loading Big data: Brian Blake from the University of Miami discussed how formal methods can be incorporated into approaches to handle privacy violations when multiple pieces of information are combined. In particular, he discussed the creation of a software life cycle and framework for big data testing.

Authenticity of Digital Images in Social Media: Balkirat Kaur from North Carolina A&T State University discussed novel solutions for detecting tampered images in social media. In particular, she discussed an approach for creating and capturing image signatures.

Business Intelligence meets Big Data: An Overview of Security and Privacy: Claudio Ardagna from the University of Milano in Crema discussed the notions of full data and zero latency analysis within the context of big data security and privacy.

Towards Risk-Aware Policy based Framework for Big Data Security and Privacy: James Joshi from the University of Pittsburgh described a framework for big data security and privacy that takes risk into consideration. He discussed how realizing such a framework involves the integration of policy engineering and risk management approaches.

Big Data Analytics: Privacy Protection using Semantic Web Technologies: Csilla Farkas from the University of South Carolina discussed the use of semantic web technologies for representing policies and data and subsequently reasoning about these policies to prevent security and privacy violations.

Securing Big Data in the Cloud: Towards a more focused and data driven approach: Ragib Hasan from the University of Alabama at Birmingham described the challenges in secure cloud computing and discussed a data driven approach to provide some solutions. In particular, he discussed the need to look at the data life cycle and ensure trustworthy computation and attribution. He stated that provenance should be a fundamental part of clouds.

Privacy in a World of Mobile Devices: Tim Finin from the University of Maryland, Baltimore County discussed approaches to providing privacy in a mobile computing environment. He states that our privacy is at risk due to the proliferation of mobile devices and discussed ways of ensuring privacy.

Access Control and Privacy Policy Challenges in Big Data: Ram Krishnan from the University of Texas at San Antonio stated that data is being used in unplanned ways that were unforeseen during the time of collection. He then discussed the challenges for access control and privacy policy specification and enforcement for big data applications.

Timely Health Indicators Using Remote Sensing and Innovation for the Validity of the Environment: David Lary from The University of Texas at Dallas who is a natural scientist by training discussed the big data challenges for remote sensing with applications in human health. This presentation provided an overview of an application that manages and analyzes big data and showed the need to handle data privacy.

Additional Presentations: The workshop also had additional presentations including the following. Big Noise in Big Data: Research Challenges and Opportunities in Heterogeneous Sensor Data Integration by Calton Pu from Georgia Tech. and Accelerating the Performance of Private Information Retrieval Protocols using Graphical Processing Units by Gabriel Ghinita from the University of Massachusetts in Boston. Calton showed us a demonstration of integrating heterogeneous sensor data and discussed the need for data security and privacy while Gabriel discussed approaches and challenges for private information retrieval. Presentations related to Big Data security and privacy were also given by Anna Squicciarini from Pennsylvania State University and Guofei Gu from Texas A&M University. Topics discussed included social media privacy and malware attacks. Finally, Andrew Greenhut from Raytheon said a few words about security and privacy needs for defense applications.

Final Thoughts on the Presentations: As can be seen, the presentations covered a wide range of topics including security and privacy issues as well as applications such as healthcare. In addition, various types of frameworks for Big Data security and privacy were also discussed. Technologies discussed included social media, image processing, mobile data, and sensor information management. These presentations set the stage for the workshop discussions that took place as part of the break-out sessions. The discussions are summarized in section 4.

4. SUMMARY OF THE WORKSHOP DISCUSSIONS

4.1 INTRODUCTION

This section provides a summary of the discussions on Big Data security and privacy at the NSF Workshop. The workshop consisted of keynote presentations, presentations by the participants and workgroup discussions. We organized two workgroups, one on Big Data security and privacy led by Dr. Elisa Bertino and the other on Big Data Analytics for Cyber Security led by Dr. Murat Kantarcioglu. While the major focus of the workshop was on privacy issues due to Big Data management and analytics, we also had some stimulating discussions on applying big data management analytics techniques for cyber security. Therefore, this section provides a summary of the discussions of both workgroups.

The organization of this section is as follows. The philosophy behind Big Data security and privacy is discussed in section 4.2. Privacy enhanced techniques are discussed in section 4.3. A framework for Big Data privacy is discussed in section 4.4. Research challenges and interdisciplinary approaches to Big Data privacy are discussed in section 4.5. An overview of Big Data management and analytics techniques for cyber security is provided in section 4.6. The section is concluded in section 4.7. References for this section are given in section 4.8.

4.2 PHILOSOPHY FOR BIG DATA SECURITY AND PRIVACY

As discussed by Bertino [1], technological advances and novel applications, such as sensors, cyber-physical systems, smart mobile devices, cloud systems, data analytics, and social networks are making possible to capture, and to quickly process and analyze huge amounts of data from which to extract information critical for security-related tasks. In the area of cyber security, such tasks include user authentication, access control, anomaly detection, user monitoring, and protection from insider threat [2]. By analyzing and integrating data collected on the Internet and Web, one can identify connections and relationships among individuals that may in turn help with homeland protection. By collecting and mining data concerning user travels and disease outbreaks, one can predict disease spreading across geographical areas. And those are just a few examples; there are certainly many other domains where data technologies can play a major role in enhancing security.

The use of data for security tasks is however raising major privacy concerns [3]. Collected data, even if anonymized by removing identifiers such as names or social security numbers, when linked with other data, may lead to re-identify the individuals to which specific data items are related to. Also, as organizations such as governmental agencies often need to collaborate on security tasks, datasets are exchanged across different organizations, resulting in these datasets being available to many different parties. Apart from the use of data for analytics, security tasks such as authentication and access control may require detailed information about users. An example is multi-factor authentication that may require, in addition to a password or a certificate, user biometrics. Recently proposed continuous authentication techniques extend user authentication to include information such as user keystroke dynamics to constantly verify the user identity. Another example is location-based access control [4] that requires users to provide to the access control system information about their current location. As a result, detailed user mobility information may be collected over time by the access control system. This information if misused or stolen can lead to privacy breaches.

It would then seem that in order to achieve security, we must give up privacy. However this may not be necessarily the case. Recent advances in cryptography are making possible to work on encrypted data – for example for performing analytics on encrypted data [5]. However, much more needs to be done as the specific data privacy techniques to use heavily depend on the specific use of data and the security tasks at hand. Also current techniques are not still able to meet the efficiency requirement for use with big data sets.

In this document, we first discuss a few examples of approaches that help with reconciling security with privacy. We then discuss some aspects of a framework for data privacy. Finally, we summarize research challenges and provide an overview of the multi-disciplinary research needed to address these challenges.

The Appendix includes the inputs from the NIST Big Data Security and Privacy Working Group. Inputs from different communities will be integrated into the final workshop report.

4.3 EXAMPLES OF PRIVACY-ENHANCING TECHNIQUES

Many privacy-enhancing techniques have been proposed over the last fifteen years, ranging from cryptographic techniques such as oblivious data structures [6] that hide data access patterns to data anonymization techniques that transform the data to make more difficult to link specific data records to specific individuals; and we refer the reader for further references to specialized conferences, such as the Privacy-Enhancing Symposium (PET) series (https://petsymposium.org/2014/) and journals, such as Transactions on Data Privacy (http://www.tdp.cat/). However, many such techniques either do not scale to very large datasets and/or do not specifically address the problem of reconciling security with privacy. At the same time, there are a few approaches that focus on efficiently reconciling security with privacy and we discuss them in what follows.

- Privacy-preserving data matching: Record matching is typical performed across different data sources with the aim of identifying common information shared among these sources. An example is matching a list of passengers on a flight with a list of suspicious individuals. However, matching records from different data sources is often in contrast with privacy requirements concerning the data owned by the sources. Cryptographic approaches, such as secure set intersection protocols, may alleviate such concerns. However, these techniques do not scale for large datasets. Recent approaches based on data transformation and mapping into vector spaces [7], and combination of secure multiparty computation (SMC) and data sanitization approaches such as differential privacy [8], and k-anonymity [9,10] have addressed scalability. However, work needs to be done concerning the development of privacy-preserving techniques suitable for complex matching techniques, based for example on semantic matching. Security models and definitions also need to be developed supporting security analysis and proofs for solutions combining different security techniques, such as SMC and differential privacy.

- Privacy-preserving collaborative data mining: Conventional data mining is typically performed on big centralized data warehouses collecting all the data of interest. However, centrally collecting all data poses several privacy and confidentiality concerns when data belongs to different organizations. An approach to address such concerns is based on distributed collaborative approaches by which the organizations retain their own datasets and cooperate to learn the global data mining results without revealing the data in their own individual datasets. Fundamental work in this area includes: (i) techniques allowing two parties to build a decision tree without learning anything about each other's datasets except for what can be learned by the final decision tree [11]; (ii) specialized collaborative privacy-preserving techniques for association rules, clustering, k-nearest neighbor classification [12]. These techniques are however still very inefficient. Novel approaches based on cloud computing and new cryptographic primitives should be investigated.

- Privacy-preserving biometric authentication: Conventional approaches to biometrics authentication require recording biometrics templates of enrolled users and then using these templates for matching with the templates provided by users at authentication time. Templates of user biometrics represent sensitive information that needs to be strongly protected. In distributed environments in which users have to interact with many different service providers, the protection of biometric templates becomes even more complex. A recent approach addresses such an issue by using a combination of perceptual hashing techniques, classification techniques, and zero-knowledge proof of knowledge (ZKPK) protocols [13]. Under such approach, the biometric template of a user is processed to extract from it a string of bits which is then further processed by classification and some other transformation. The resulting bit string is then used, together with a random number, to generate a cryptographic commitment. This commitment represents an identification token that does not reveal anything about the original input biometrics. The commitment is then used in the ZKPK protocol to authenticate the user. This approach has been engineered for secure use on mobile phones. Much work remains, however, to be done in order to reduce the false rejection rates. Also different approaches to authentication and identification techniques need to be investigated based on recent homomorphic encryption techniques.

4.4 MULTI-OBJECTIVE OPTIMIZATION FRAMEWORK FOR DATA PRIVACY

Although there are attempts at coming up with a privacy solution/definition that can address many different scenarios, we believe that there is no one size fits all solution for data privacy. Instead, multiple dimensions need to be tailored for different application domains to achieve practical solutions. First of all, different domains require different definitions of data utility. For example, if we want to build privacy-preserving classification models, 0/1 loss could be a good utility measure. On the other hand, for privacy-preserving record linkage, F1 score could be a better choice. Second, we need to understand the right definitions of privacy risk. For example, in data sharing scenarios, probability of re-identification given certain background knowledge could be the considered right measure of privacy risk. On the other hand, $\varepsilon=1$ could be considered appropriate risk for differentially private data mining models. Finally, the computational, storage and communication costs of given protocols need to be considered. These costs could be especially significant for privacy-preserving protocols that involve cryptography. Given these three dimensions, one can imagine a multi-objective framework where different dimensions could be emphasized:

- **Maximize utility, given risk and costs constraints:** This would be suited for scenarios where limiting certain privacy risks are paramount.
- **Minimize privacy risks, given the utility and cost constraints**: In some scenarios, (e.g., medical care), significant degradation of the utility may not be allowed. In this setting, the parameter values of the protocol are (e.g., ε in differential privacy) chosen in such way that we try to do our best in terms of privacy given our utility constraints. Please note that in some scenarios, there may not have any parameter settings that can satisfy all the constraints.
- **Minimize cost, given the utility and risk constraints:** In some cases, (e.g., cryptographic protocols), you may want to find the protocol parameter settings that may allow for the least expensive protocol that can satisfy all the utility and cost constraints.

To better illustrate these dimensions, consider the privacy-preserving record matching problem addressed in [9]. Existing solutions to this problem generally follow two approaches: sanitization techniques and cryptographic techniques. In [9], a hybrid technique that combines these two approaches and enables users to trade-off between privacy, accuracy, and cost similar to multi-objective optimization framework discussed here. These multi-objective optimizations are achieved by using of a blocking phase that operates over sanitized data to filter out in a privacy-preserving manner pairs of records that do not satisfy the matching condition. By disclosing more information (e.g., differentially private data statistics), the proposed method incurs considerably lower costs than cryptographic techniques. On the other hand, it yields significantly more accurate matching results compared to sanitization techniques, even when privacy requirements are high. Using different privacy-parameter values allow for different cost, risk and utility outcomes.

To enable the multi-objective optimization framework for data privacy, we believe that more research needs to be done to identify appropriate utility, risk and cost definitions for different application domains. Especially, defining right and realistic privacy risks is paramount. Many human actions ranging from oil extraction to airline travel, involve risks and benefits. In many cases, such as trying to develop an aircraft that may never malfunction, avoiding all risks are either too costly or impossible. Similarly, we believe that avoiding all privacy risks for all individuals would be too costly. In addition, assuming that an attacker may know everything is too pessimistic. Therefore, coming up with privacy risk definitions under realistic attacker scenarios would be needed.

4.5 RESEARCH CHALLENGES AND MULTIDISCIPLINARY APPROACHES

Comprehensive solutions to the problem of security with privacy for Big Data require addressing many research challenges and multidisciplinary approaches. We outline significant directions in what follows:

- Data Confidentiality: Several data confidentiality techniques and mechanisms exist – the most notable being access control systems and encryptions. Both techniques have been widely investigated. However for access control systems for Big Data we need approaches for:

- *Merging large numbers of access control policies*. In many cases, Big Data entails integrating data originating from multiple sources; these data may be associated with their own access control policies (referred to as "sticky policies") and these policies must be enforced even when the data is integrated with other data. Therefore policies need to be integrated and conflicts solved.
- *Automatically administering authorizations for big data and in particular for granting permissions*. If fine-grained access control is required, manual administration on large datasets is not feasible. We need techniques by which authorization can be automatically granted, possibly based on the user digital identity, profile, and context, and on the data contents and metadata.
- *Enforcing access control policies on heterogeneous multi-media data*. Content-based access control is an important type of access control by which authorizations are granted or denied based on the content of data. Content-based access control is critical when dealing with video surveillance applications which are important for security. As for privacy such videos have to be protected. Supporting content-based access control requires understanding the contents of protected data and this is very challenging when dealing with multimedia large data sources.
- *Enforcing access control policies in big data stores*. Some of the recent Big Data systems allow its users to submit arbitrary jobs using programming languages such as Java. For example, in Hadoop, users can submit arbitrary MapReduce jobs written in Java. This creates significant challenges to enforce fine-grained access control efficiently for different users. Although there is some existing work [14,15] that tries to inject access control policies into submitted jobs, more research needs to be done on how to efficiently enforce such policies in recently developed Big Data stores.
- *Automatically designing, evolving, and managing access control policies*. When dealing with dynamic environments where sources, users, and applications as well as the data usage are continuously changing, the ability to automatically design and evolve policies is critical to make sure that data is readily available for use while at the same time assuring data confidentiality. Environments and tools for managing policies are also crucial.

- Privacy-preserving data correlation techniques: a major issue arising from Big Data is that correlating many (big) data sets one can extract unanticipated information. Relevant issues and research directions that need to be investigated include:
  - *Techniques to control what is extracted and to check that what is extracted can be used and/or shared*.
  - *Support for both personal privacy and population privacy*. In the case of population privacy, it is important to understand what is extracted from the data as this may lead to discrimination. Also when dealing with security with privacy, it is important to understand the tradeoff of personal privacy and collective security.
  - *Efficient and scalable privacy-enhancing techniques*. Several such techniques have been developed over the years, including oblivious RAM, security multiparty computation, multi-input encryption, homomorphic encryption. However, they are not yet practically applicable to large datasets. We need to engineer these techniques, using for example parallelization, to fine tune their implementation and perhaps combine them with other techniques, such as differential privacy (like in the case of the record linkage protocols described in [7]). A possible further approach in this respect is to first use anonymized/sanitized data, and then depending on the specific situation to get specific non-anonymized data.
  - *Usability of data privacy policies*. Policies must be easily understood by users. We need tools for the average users and we need to understand user expectations in terms of privacy.
  - *Approaches for data services monetization*. Instead of selling data, organizations owning datasets can sell privacy-preserving data analytic services based on these datasets. The question to be addressed then is: how would the business model around data change if privacy-preserving data analytic tools were available? Also if data is considered as a good to be sold, are there regulations concerning contracts for buying/selling data? Can these contracts include privacy clauses be incorporated requiring for example that users to whom this data pertains to have been notified?

- *Data publication*. Perhaps we should abandon the idea of publishing data, given the privacy implications, and rather require the data user to use a controlled environment (perhaps located in a cloud) for using the data. In this way, it would be much easier to control the proper use of data. An issue would be the case of research data used in universities and the repeatability of data-based research.
- *Privacy implication on data quality*. Recent studies have shown that people lie especially in social networks because they are not sure that their privacy is preserved. This results in a decrease in data quality that then affects decisions and strategies based on these data.
- *Risk models.* Different types of relationship of risks with big data can be identified: (a) big data can increase privacy risks; (b) big data can reduce risks in many domains (e.g. national security). The development of models for these two types of risk is critical in order to identify suitable tradeoff and privacy-enhancing techniques to be used.
- *Data ownership.* The question about who is the owner of a piece of data is often a difficult question. It is perhaps better to replace this concept with the concept of stakeholder. Multiple stakeholders can be associated with each data item. The concept of stakeholder ties well with risks. Each stakeholder would have different (possibly conflicting) objectives and this can be modeled according to multi-objective optimization. In some cases, a stakeholder may not be aware of the others. For example a user to whom a data pertains (and thus a stakeholder for the data) may not be aware that a law enforcement agency is using this data. Technology solutions need to be investigated to eliminate conflicts.
- *Human factors*. All solutions proposed for privacy and for security with privacy need to be investigated in order to determine human involvement, e.g. how would the user interact with the data and his/her specific tasks concerning the use and/or protection of the data, in order to enhance usability.
- *Data lifecycle framework.* A comprehensive approach to privacy for big data needs to be based on a systematic data lifecycle approach. Phases in the lifecycle need to be identified and their privacy requirements and implications need to be identified. Relevant phases include:
  - *Data acquisition* – we need mechanisms and tools to prevent devices from acquiring data about other individuals (relevant when devices like Google glasses are used); for example can we come up with mechanisms that automatically block devices from recording/acquiring data when in certain locations (or notify a user that recording devices are around). We also need techniques by which each recorded subject may have a say about the use of the data.
  - *Data sharing* – users need to be informed about data sharing/transferred to other parties.

Addressing the above challenges require multidisciplinary research drawing from many different areas, including computer science and engineering, information systems, statistics, risk models, economics, social sciences, political sciences, human factors, psychology. We believe that all these perspectives are needed to achieve effective solutions to the problem of privacy in the era of Big Data and of how reconcile security with privacy.

4.6 BIG DATA ANALYTICS FOR CYBER SECURITY

To protect important digital assets, organizations are investing in new cyber security tools that need to analyze Big Data ranging from log files to e-mail attachments to prevent, detect and recover from cyber attacks [16]. As a part of this workshop, we explored the following topics:

- *What is different about Big Data analytics (BDA) for Cyber security?*: The break-out participants pointed out that BDA for cyber security needs to deal with adaptive, malicious adversary that can potentially launch attacks to avoid being detected (i.e., data poisoning attacks, denial of service, denial of information attacks etc.). In addition, BDA for cyber security need to operate in high volume (e.g., data coming from multiple intrusion detection systems and sensors) and high noise environments (i.e., constantly changing normal system usage data is mixed with stealth advanced persistent threat related data). One of the important points that came out of this discussion is that we need BDA tools that can integrate data from host, network, social networks, bug reports, mobile devices, and internet of things sensors to detect attacks.

- *What is the right BDA architecture for Cyber Security*?: We also discussed whether we need different types of BDA system architectures for cyber security. Based on the use cases discussed, participants felt that existing BDA system architectures can be adapted for cyber security needs. One issue pointed out was that real time data analysis must be supported by a successful BDA system for cyber security. For example, once a certain type of attack is known, the system needs to be updated to look for such attacks in real time including re-examining the past data to see whether attacks occurred in the past.
- *Data Sharing for BDA for Cyber Security?:* It emerged quickly during our discussions that cyber security data needs to be shared both inside the organization and among organizations. In addition to obvious privacy, security and incentive issues in sharing cyber security data, participants felt that we need common languages and infrastructure to capture and share such cyber security data. For example, we need to represent certain low level system information (e.g., memory, cpu states, etc.) so that it can be mapped to similar cyber security incidents.
- *BDA for Preventing Cyber Attacks?:* There was substantial discussion on how BDA tools could be used to prevent attacks. One idea that emerged is that BDA systems that can easily track sensitive data using the capture provenance information can potentially detect attacks before too much sensitive information is disclosed. Based on this observation, building provenance-aware BDA systems could be needed for cyberattack prevention. Also, BDA tools for cyber security can potentially mine useful attacker information such as their motivations, technical capabilities, modus operandi, etc. to prevent future attacks.
- *BDA for Digital Forensics?:* BDA techniques could be used for digital forensics by combining or linking different data sources. The main challenge emerged as identifying the right data sources for digital forensics. In addition, it was not immediately clear, what data to capture? What to filter out? (Big noise in Big Data) What to link? What to store and how long? How to deal with machine generated content and Internet of things?
- *BDA for Understanding the Users of the Cyber Systems:* Participants believe that BDA could be used to mine human behavior to learn how to improve the systems. For example, an organization may send phishing emails to its users and give a security re-training for those who are fooled by such a phishing attack. In addition, BDA techniques could be used to understand and build normal behavior models per user to find significant deviations from the normal.

Overall, during our workshop discussions, it became clear that all of the above topics have significant research challenges and more research needs to be done to address them.

4.7 CONCLUSION

Our workshop has explored the issues surrounding Big Data Security and Privacy as well as applying Big Data Management and Analytics Techniques for Cyber Security. As massive amounts of data are being collected, stored, manipulated, merged, analyzed, and expunged, security and privacy concerns will explode. We need to develop technologies to address security and privacy issues throughout the lifecycle of the data. However technologies alone will not be sufficient. We need to understand not only the societal impact of data collection, use and analysis, we also need to formulate appropriate laws and policies for such activities. Our workshop has explored the initial directions to address some of the major challenges we are faced with today. We need an interdisciplinary approach consisting of technologists, application specialists, social scientists, policy analysts and lawyers to work together to come up with viable and practical solutions.

4.8 REFERENCES

[1] E. Bertino, "Security with Privacy – Opportunities and Challenges" Panel Statement, *COMPSAC 2014.*

[2] E. Bertino, Data Protection from Insider Threats. Morgan&Claypool, 2012.

[3] B. Thuraisingham: Data Mining, National Security, Privacy and Civil Liberties. SIGKDD Explorations 4(2): 1-5 (2002)

[4] M. Damiani, E. Bertino, B. Catania, P. Perlasca, "GEO-RBAC: A Spatially Aware RBAC", *ACM Transactions on Information and System Security* 10(1), 2007.

[5] D. Liu, E. Bertino, X. Yi, "Privacy of Outsourced K-Means Clustering", *Proceedings of the 9th ACM Symposium on Information, Computer and Communication Security*, Kyoto (Japan), June 4-6, 2014.

[6] H. X. Wang, K. Nayak, C. Liu, E. Shi, E. Stefanov, Y. Huang, "Oblivious Data Structures", IACR Cryptology ePrint Archive 2014: 185.

[7] M. Scannapieco, I. Figotin, E. Bertino, A. Elmagarmid, "Privacy Preserving Schema and Data Matching", *Proceedings of 2007 ACM SIGMOD International Conference on Management of Data*.

[8] M. Kuzu et al. "Efficient Privacy-aware Record Integration", *Proceedings of Joint 2013 EDBT/ICDT Conferences, EDBT'13*, Genoa, Italy, March 18-22, 2013, ACM.

[9] A. Inan, M. Kantarcioglu, G. Ghinita, E. Bertino, "A Hybrid Approach to Private Record Matching", IEEE Trans. Dependable Sec. Comput. (TDSC) 9(5):684-698 (2012)

[10]A. Inan, M. Kantarcioglu, E. Bertino, M. Scannapieco, "A Hybrid Approach to Private Record Linkage", ICDE 2008:496-505

[11] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining", in *Advances in Cryptology*, Springer-Verlag, Aug. 20-24 2000.

[12] J. Vaidya, Y. Zhu, C. Clifton, "Privacy Preserving Data Mining", *Advances in Information Security* 19, Springer 2006, pp.1-121.

[13] H. Gunasinghe, E. Bertino, "Privacy Preserving Biometrics-Based and User Centric Authentication Protocol for Mobile Devices", *Proceedings of 2014 Network and System Security (NSS2014)*, Xi'an, China, October 15-16, 2014.

[14]S. Khan, K. Hamlen, M. Kantarcioglu, "Silver Lining: Enforcing Secure Information Flow at the Cloud Edge", IC2E 2014:37-46

[15]H. Ulusoy et al. "Vigiles: Fine-Grained Access Control for MapReduce Systems,", 2014 IEEE International Congress on Big Data (BigData Congress) pp.40-47

[16]S Kar, "Gartner Report: Big Data will Revolutionize Cyber Security in the Next Two Years", CloudTimes.Org, Feb. 12, 2014.

5. NEXT STEPS

This report has described the security and privacy issues for Big Data and has also provided summaries of the workshop presentations and discussions. We made a submission to the National Privacy Research Strategy on October 16, 2014 that was based on the workshop summary. We will participate in the National Privacy Research Strategy Conference in Washington DC February 18-20, 2015 and will be giving a presentation of the workshop summary at this event.

Our goal is to build a community in Big Data Security and Privacy. We plan to establish a social media presence on this topic so that the participants at this workshop as well as others can continue to exchange ideas on this very important and critical topic.

APPENDIX A: WORKSHOP AGENDA

September 16, 2014

7:30-8:30am - Registration and Breakfast

8:30-9:00am - Welcome and Introductions
Chris Clifton, NSF and Bhavani Thuraisingham, UT Dallas

9:00-9:40 - Keynote Address #1
Wo Chang, NIST

9:40-10:20 - Keynote Address #2: Security with Privacy
Elisa Bertino, Purdue University

10:20-10:30am – Break

10:30-11:10am - Keynote Address #3: Security and Privacy of Big Data: A NIST Working Group Perspective
Arnab Roy, Fujitsu

11:10-3:00pm - Lunch + Following 20 minute presentations each:

Barbara Carminati, Pietro Colombo and Elena Ferrari, University of Insubria
Towards Privacy aware Big Data analytics

M. Brian Blake and Iman Saleh, University of Miami
Formal Methods for Preserving Privacy for Big Data Extraction Software

Balkirat Kaur, Malcolm Blow and Justin Zhan, North Carolina A&T University
Authenticity of Digital Images in Social Media

Claudio Agostino Ardagna and Ernesto Damiani, University of Milan
Business Intelligence meets Big Data: An Overview on Security and Privacy

James Joshi, University of Pittsburgh
Towards Risk-aware Policy based framework for Big Data Security and Privacy

Csilla Farkas, University of South Carolina
Big Data Analytics: Privacy Protection Using Semantic Web Technologies

Ragib Hasan, The University of Alabama at Birmingham, and Anthony Skjellum, Auburn University

Securing Big Data in the Cloud: Towards a More Focused and Data Driven Approach

Calton Pu, Georgia Institute of Technology
Big Noise in Big Data: Research Challenges and Opportunities in Heterogeneous Sensor Data Integration

Tim Finin, University of Maryland, Baltimore County
Semantics for Big Data, Security and Privacy

3:00-5:30pm - Breakout sessions

5:30-6:30pm - Presentations from the breakout sessions

6:30-7:30pm - Shuttle to hotel

7:30-9:00pm - Dinner at hotel, Harmony I Ballroom

September 17, 2014

7:30-8:30am - Registration and Breakfast

8:30-9:10 - Keynote Address #5: Industrial Control Systems: The Next Frontier for Cybersecurity and Big Data
Raj Rajagopalan, Honeywell

9:10-9:50 - Keynote Address #6: Big Data Security and Privacy Initiatives at UT Dallas
Murat Kantarcioglu, The University of Texas at Dallas

9:50-10:30am - Keynote Address #7: Big Data Enterprise
Sandeep Gopisetty, IBM Corp.

10:30-10:40am – Break

10:40am-12:00pm noon - Following 20 minute presentations each:

Ram Krishnan, University of Texas at San Antonio
Access Control and Privacy Policy Challenges in Big Data

David Lary, The University of Texas at Dallas
Holistics 3.0: Multiple Big Datasets and Machine Learning

Gabriel Ghinita, University of Massachusetts, Boston
Accelerating the Performance of Private Information Retrieval (PIR) Protocols Using Graphical Processing Units (GPUs)

12:00 noon-2pm - Breakout sessions with lunch

2-3pm - Presentations of the Breakout sessions

3-3:30pm - Wrap-up and Next Steps
Chris Clifton and Bhavani Thuraisingham

APPENDIX B: WORKSHOP PARTICIPANTS

Chris Clifton, NSF
Bhavani Thuraisingham, UT Dallas (Workshop Chair)
Elisa Bertino, Purdue University (Workgroup Chair and Keynote)
Murat Kantarcioglu, The University of Texas at Dallas (Workgroup Chair and Keynote)
Wo Chang, NIST (Keynote)
Arnab Roy, Fujitsu (Keynote)
Raj Rajagopalan, Honeywell (Keynote)
Sandeep Gopisetty, IBM Corp. (Keynote)
Barbara Carminati, University of Insubria
M. Brian Blake, University of Miami
Balkirat Kaur, North Carolina A&T University
Justin Zhan, North Carolina A&T University
Claudio Agostino Ardagna University of Milan
James Joshi, University of Pittsburgh
Csilla Farkas, University of South Carolina
Ragib Hasan, The University of Alabama at Birmingham
Anthony Skjellum, Auburn University
Calton Pu, Georgia Institute of Technology
Tim Finin, University of Maryland, Baltimore County
Ram Krishnan, University of Texas at San Antonio
David Lary, The University of Texas at Dallas
Gabriel Ghinita, University of Massachusetts, Boston
Anna Squicciarini, Pennsylvania State University
Guofei Gu, Texas A&M University
Andrew Greenhut, Raytheon
Marina Blanton, Notre Dame University
Eunis Santos, University of Texas at El Paso

The following members from The University of Texas at Dallas attended the workshop
Latifur Khan
Alvaro Cardenas
Zhiqiang Lin

Local Arrangements
Rhonda Walls

Web Chair
Rhonda Walls
Nathan McDaniel

APPENDIX C: Links to the Workshop, Presentations, and Position Papers

Workshop Details
http://csi.utdallas.edu/events/NSF/NSF%20workshop%202014.htm

Presentations and Position Papers
http://csi.utdallas.edu/events/NSF/NSF%20papers%202014.htm

Submission to the National Privacy Research Strategy
https://www.nitrd.gov/cybersecurity/nprsrfi102014/BigData-SP.pdf