Big Data – Security with Privacy

Response to RFI for National Privacy Research Strategy

From

The Participants of the NSF Big Data Security and Privacy Workshop (September 16-17, 2014)

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

Track Chairs: Elisa Bertino and Murat Kantarcioglu

Workshop Chair: Bhavani Thuraisingham

DRAFT October 16, 2014

ABSTRACT

This document provides a summary of the discussions on big data security and privacy at the NSF Workshop on this topic held at the University of Texas at Dallas on September 16 and 17, 2014. This document will continue to evolve over the next few months as we get feedback from the workshops participants and the final workshop report will be completed by February 2015. This document also includes an Appendix that consists of comments received from other communities that will be included in the final workshop report.

1. INTRODUCTION

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, data sets are exchanged across different organizations, resulting in these data sets 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 Security and Privacy Subgroup of the NIST Big Data Public Working Group. Inputs from different communities will be integrated into the final workshop report.

2. 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)[1] series, and journals, such as Transactions on Data Privacy[2]. However many such techniques either do not scale to very large data sets 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 data sets. 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 data sets and cooperate to learn the global data mining results without revealing the data in their own individual data sets. Fundamental work in this area includes: (i) techniques allowing two parties to build a decision tree without learning anything about each other data sets 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 issue by using a combination of perceptual

[1] https://petsymposium.org/2014/

[2] http://www.tdp.cat/

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.

3. 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 fit 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 is 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, involves 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 attacker may know everything is too pessimistic. Therefore, coming up with privacy risk definitions under realistic attacker scenarios would be needed.

4. 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 data sets 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 for 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 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 user’s 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 data sets. 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 data sets can sell privacy-preserving data analytic services based on these data sets. 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 result 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 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 mechanism that automatically block devices from recording/acquiring data when in certain location (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.

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

APPENDIX: Additional Comments

The following is the submission of the Security and Privacy subgroup of the NIST Big Data Public Working Group. This document will be integrated into the NSF Workshop Report.

The NIST Big Data Public Working Group (NBD-PWG) has established a subgroup aimed at Security and Privacy aspects of Big Data systems.

Introduction to the Working Group

NIST identified the goals of this Group on a public web site.

NIST Big Data

NIST Big Data Program

Welcome to NIST Big Data Public Working Group (NBD-PWG)!

NIST is leading the development of a Big Data Technology Roadmap. This roadmap will define and prioritize requirements for ***interoperability****,* ***portability****,* ***reusability****, and* ***extendibility*** *for big data analytic techniques and technology infrastructure in order to support secure and effective adoption of Big Data. To help develop the ideas in the Big Data Technology Roadmap, NIST is creating the Public Working Group for Big Data.*

■ ***Scope:*** *The focus of the NBD-PWG is to form a community of interest from industry, academia, and government, with the goal of developing a consensus* ***definitions****,* ***taxonomies****,* ***secure reference architectures****, and* ***technology roadmap****. The aim is to create vendor-neutral, technology and infrastructure agnostic deliverables to enable Big Data stakeholders to pick-and-choose best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added from Big Data service providers and flow of data between the stakeholders in a cohesive and secure manner.*

■ ***Deliverables:***

1. *Develop Big Data Definitions*
2. *Develop Big Data Taxonomies*
3. *Develop Big Data Requirements*
4. *Develop Big Data Security and Privacy Requirements*
5. *Develop Big Data Security and Privacy Reference Architectures*
6. *Develop Big Data Reference Architectures*
7. *Develop Big Data Technology Roadmap*

<End snip>

Security and Privacy Subgroup

The Security and Privacy subgroup is co-chaired by Arnab Roy (Fujitsu) and Mark Underwood (Krypton Brothers).

The larger Group has undertaken a multi-phase effort. Phase 1 work has been largely completed and will be published in coming months. As a result of work completed in Phase 1, we were able to contribute material to the US National Body contribution to the November 2014 ISO JTC1 Plenary on Big Data. This included several additions regarding security, privacy and information assurance (provenance and veracity in Big Data frameworks) related to PII.

As was the case with Phase 1, NBD-PWG Phase 2 is ongoing under NIST guidance provided by Wo Chang.

A distinctive feature of the NBD-PWG is its attempt to develop an overall taxonomy and reference architecture to guide builders and maintainers of Big Data systems. By tying security and privacy considerations to a reference architecture (RA), the team believes that it will be able to provide framework-level guidance. Target consumers for that guidance include:

- Enterprise architects
- GRC specialists, especially auditors
- Cybersecurity specialists
- Users of Big Data analytics platforms, including visualization
- Big Data software users, especially for, but not limited to open source tools such as Hadoop, Storm, Spark, Mahout, etc.
- Big Data forensics
- Privacy policy analysts

In addition to these activities, the NBD-PWG will host workshop panels at the IEEE Big Data 2014 Conference (Oct 2014). One of these panels works to foster collaboration between teams working on domain-specific security / privacy standards with counterpart practitioners in those fields actively developing privacy policies.

This approach – fostering collaborations between technology teams and practitioner within Communities of Interest -- could prove important in the larger NSF undertaking, as privacy considerations move in both broad, cross-domain and domain-specific directions. Current practice in data collection for NIH grant recipients may be different from practitioners working in educational research initiatives.

Privacy in the NBD-PWG Reference Architecture

Contributors to the Security and Privacy subgroup attempted to highlight connections between privacy use cases in existing Big Data implementations and the NBD-PWG RA. These use cases are being mapped to specific elements of the RA, using a readily understood, lightly-technical “crosswalk.”

Anticipated benefits from the subgroup’s work in both Phase 1 and Phase 2 include the following general areas. (We will flesh out these claims with more detail in a later version of the document.)

- Notional implementation of a Privacy fabric across components of the Big Data RA
- Proposed changes in software design patterns for privacy in Big Data systems
- New privacy design patterns in light of Big Data variety, especially for PII data injection / inference / infusion
- More systematic approach to Big Data provenance, extensible to device registration and emerging standards for the Internet of Things
- Provide useful connections between Big Data operational systems (e.g., system health) and privacy preservation
- Mapping of Oasis Privacy Management Reference Model (PMRM) to the Big Data RA (limited to publicly available documentation)
- Mapping of Privacy by Design recommendations to the Big Data RA (limited to publicly available documentation)
- Mapping of existing privacy policy-preserving protocols (e.g., Microsoft Active Directory) to Big Data settings
- Consideration of component-specific (e.g., data provider, GRC, audit, forensics and software test "points" within the Big Data software development life cycle (SDLC)
- More unified connection between privacy defense and threat mitigation / vulnerability assessment. Approach recommends using Big Data for privacy protection (logs, complex event fusion, etc.)
- Consideration of visualization dimensions for privacy and security

Use Case-Rich Approach

The subgroup has taken pains to limit its scope, where possible, to privacy considerations that are typically characteristic of Big Data projects.

A goal of the documents produced in Phase 2, and to a lesser extent in Phase 1, is to provide system implementers and auditors with use cases from which their own scenarios can be considered. As the RA security fabric is refined and socialized, the subgroup anticipates adding to the use case collection. These additions will provide specialized communities of interest with approaches taken by others, and mapped against vendor-neutral, current Big Data technologies seen through the organizing principles of the RA.

We learn from the work of others. In a December 2013 letter from the American Hospital Association to NIST regarding the latter's Preliminary Cybersecurity framework, the association pointed out the importance of **sector-specific work**. They pointed out 18 diverse cybersecurity sectors solely within the domain of hospital information systems. Identifying and responding to the needs of a much greater number of specific communities is likely to be key in the NSF endeavor as well.