Statement of Professional Goals and Objectives

Yuzhe Tang
Syracuse University, NY, USA, Email: ytang100@syr.edu

In the past decade, our society has started to see the birth and explosion of digitized personal data, brought about by technical advances in electronic health-care, Internet of Things (IoT), gene sequencing, mobile devices, etc. The use of this personal big-data in computerized systems, while bringing unprecedented convenience to human daily life and society, also causes significant concerns on cyber-security and privacy leakage, evident from controversies on the news released on a daily basis. As a result, our society and its safety have and will become increasingly dependent on the advances and deep understanding of computer science in general and cyber-security in particular. The dependency creates urgent needs for researchers and engineers in the related areas. My professional goals and objectives are to address these needs by conducting research on cyber-security, educating future computer-science engineers and scientists, and providing professional services that facilitate the research and education activities. This document presents my goals and the plan towards achieving them.

I. Research Objective

My research philosophy is to design elegant computer-science techniques and apply them to real, emerging applications with systems evaluation. I am particularly interested in and focus on secure protocol design for modeling new applications and practical protocol implementation with systems-level performance efficiency.

My long-term research goal is to empower average computer users with the capability of conducting secure computations of their personal data using third-party computing resources. Given the explosion of personal data (e.g. IoT device readings, sequenced genome data, electronic health records, etc.) and the popularity of outsourced computation to an untrusted third-party entity (e.g. Amazon-like cloud services), it can be expected that *computation on sensitive personal data using untrusted computing resources* will become the new computing norm in the age of big-data, and the techniques of secure computation [1], [2], [3], [4] hold the promise of ensuring data security and helping users take (back) control of their own personal data. Towards this end, my current research is two-fold: 1) Secure multi-party databases, where data owners in mutually untrusted domains jointly conduct database query processing yet without disclosing any privacy-sensitive information other than the query result, 2) Log-structured authenticated data storage, where a data owner outsourcing sensitive data to an untrusted cloud can assure security properties (e.g. freshness, data integrity, etc) yet with low overhead.

A. Secure Multi-Party Databases

I first came to work in the space of multi-party data systems in my PhD study from 2009. My dissertation research was about designing a privacy-preserving locator service, called ϵ-PPI [5], [6], [7], which serves as a federated search engine to facilitate searching on a distributed dataset. Such a service can be found useful in many emerging data-federation applications. A canonical application is locating patients' past medical records in the point of care in electronic health-care. ϵ-PPI enables the patient and her physician to quickly locate remote hospitals holding the patient's past records, in a way that does not leak privacy. In ϵ-PPI, secure multi-party computations [1], [2], [3], [4] were used in the service construction and were applied to special computations (e.g. set union). My current work on secure multi-party databases significantly expand the scope of my dissertation research in both target applications (from locator services to generic database applications) and computation model (from simple set computations to more expressive SQL queries).

In a data-federation scenario, multiple autonomous data owners holding sensitive personal data collectively conduct joint computation in the hope of extracting new knowledge from the federated big-data. The information sharing during the joint computation causes privacy leakage, presenting a major barrier to practical data-federation applications. The proposed secure multi-party database is a system that ensures strong data confidentiality for the joint SQL query processing. To the best of our knowledge, this is the first systems research work in the space of strongly secure multi-party database (strongly secure in the sense of ensuring semantically secure and oblivious database operators), and it addresses the key issue on privacy leakage in the federated big-data applications in Health-care information exchange [8], [9], [10], federated genome data networks [11], etc.

My research method in this project is to systematically revisit the design choices made in classic database systems (e.g. database optimizer, executor, etc) from the new perspective of strong security and multi-party applications. This new perspective leads to a series of technical design challenges. Our ongoing work, partially summarized in [12], tackles the challenges and consists of secure protocol design for multi-party database executions in the presence of various relational operators, redesign of database optimizer aware of the cost of multi-party computations, systematic implementation adaptable to various multi-party computation infrastructures (e.g. Yao's Garble circuits [1] and GMW protocols [2], etc). The project was partially supported by a seed research grant from Cyber Research Institute (CRI).

B. Log-structured Authenticated Data Storage

The second project is log-structured authenticated storage system. The technique can be found desirable when a cloud client outsources her sensitive data to (largely) untrusted cloud service. The research aim is to enable data authentication at very low cost. In designing systems, we observe there exists a tension between the requirement of data authentication and the goal of performance efficiency. On the one hand, performance efficiency in cloud storage entails log-structured design, as Log-Structured Merge trees (LSM trees) [13] are widely applied in the cloud storage space [14], [15], [16]. On the other hand, the problem of data authentication can be tackled by the protocols of authenticated data structures (ADS) [17], [18], [19], [20], [21], [22]. However, when designing an overall system, it becomes difficult to reconcile *the ADS at the protocol level and the LSM tree at the systems level* (i.e. "when LSM trees meet ADS"), as the two structures follow different (tree) design patterns. This structural "gap", if left untended, could lead to significant performance slowdown (due to inconsistent cross-layer systems design), as identified in our prior research [23].

To tackle the challenge and to bridge the gap, our proposed approach [24] is by leveraging the trusted execution environment (TEE) available in recent hardware, such as Intel SGX CPU [25]. We have proposed a log-structured data-authentication protocol and implemented it on Google's LSM store, LevelDB [26], in a novel fashion. Our current implementation ensures key-value data authentication, performance efficiency (no extra disk seeks), while minimizing the trust to cloud. My follow-up research in this project is to enable the seamless integration of authentication protocols with log-structured storage systems beyond just LSM trees. By planned technology transfer, my goal in this project is to provide open-source secure-storage systems and to facilitate the wider adoption of cloud outsourcing in security-oriented applications.

In addition to the above two projects, my future work aims at enabling systems-level efficiency and protocol-level security in techniques such as block-chain [27] and Tor networks [28]. My unique perspective is to identify concrete applications in domains of health-care, genome data sharing, etc.

In general, my research methodology can be summarized by a two-step process: 1) Identify new emerging security-oriented applications and advances in secure systems support, and then 2) (re)visit the technical designs (especially the classic ones that were tested by time) from the new perspectives of the applications. This broad and full-stack research enables me to identify unique research problems. In particular, the attention on fast evolving domains such as applications and systems support (e.g. architectural advances) can result in new, previously understudied scenarios. For instance, my research on log-structured authenticated storage is an example in point which marries the new computer hardware (Intel SGX CPU) with new applications (public-cloud data outsourcing). The second step is more technical which usually involves protocol-level design for formal security, systems-level design for performance efficiency, and prototype implementation/evaluation. This second step requires deep understanding of existing techniques in research literature and even in textbook; and my approach is to closely integrate this research requirement with teaching, as will be elaborated in the next section.

II. Teaching Objective

My teaching goal is to empower my students with necessary knowledge, skills, and learning methods in computer science that will help them find better employment opportunities. To achieve the goal, my teaching method in general is to motivate my students by sparking their curiosity in the class knowledge, and to guide them through the learning process by first showing the overview and outcome and then working out the details. I have applied these teaching methods on graduate-level courses and plan to exercise the same to the undergraduate teaching.

I have taught applied cryptography in a research seminar course (CIS700 Big Data and Cloud Security), and one lecture was about the concept of cryptographic hash functions. To grab my students' attention from the beginning, I started the lecture by playing an in-class game based on the birthday paradox; that is, to sample 7 or 8 students and ask them to reveal their birthdays (the day in a month). The purpose of the game is to give my students a real-world *context* where the concept of cryptographic hash can be explained clearly: Collision in cryptographic hash was explained in analogy to two students having the same birthday, and the collision resistance is about the small probability (or hardness) that a collision occurs. After this was done, it became natural to introduce more advanced concepts such as pre-image resilience, etc.

This applied cryptography course was taught from the perspective of cryptographic protocol users, instead of designers. Thus my teaching focus was on the protocol-level specification and security definition, and I deliberately tune down the protocol construction, a part requiring complex mathematics and that is not used as much in practice. This is consistent with my teaching goal of empowering students with the skills needed in their future workplace where cryptographic protocols are used but rarely redesigned.

I have taught CIS600/CSE655 Advanced Computer Architecture as a graduate-level core course. One of the challenges in teaching this course is its broad scope of seemingly unrelated topics. Instead of teaching every topic in the 500-page textbook, my teaching strategy is to selectively teach few topics in depth. For instance, hardware was treated as a blackbox, and the teaching emphasizes the architectural features that make difference to upper-level software design such as the data hazard that affects compiler design. To emphasize the use of hardware through instructions, I designed several projects and small programming tasks, such as false sharing in cache, through which students had hands-on and personal experience in observing the performance difference architectural features can bring about. One of my students said in the post-course survey that

"Professor Tang did an excellent job of presenting the information. I really like how he tied the technologies to the real world examples". The overall goal of my course design, development, and teaching is to empower my students with skills that they will find themselves competitive in the future employment market. Through the teaching experience, I honed my skills in handling challenging teaching situations, such as engaging students in a large class, incentivizing students' discussion in research seminar course, effectively managing TAs and graders to ensure students learning, etc.

There is a *synergy* between my teaching and research. On the one hand, my teaching facilitates the research activities. First, advising students for research requires the same teaching skills, as early-stage PhD students need to be closely guided. Second, teaching gives me another channel to identify potential PhD students. One of my PhD students was recruited from my research seminar course. Third, my research method of revisiting classic computer-science in the new context can be facilitated by the teaching activities on these classic topics. On the other hand, my research results in new teaching materials. My research on authenticated data storage results in a software emulator for Intel SGX instructions. This emulator has been successfully used as course projects in CIS/CSE600 Advanced Computer Architecture and CIS700 Big-data and Cloud Security.

My future courses will cover the topics on which I have expertise, including applied cryptography, database systems, and other computer-science areas. In addition to research seminars (CIS/CSE700), the future courses in my scope include graduate-level core courses (CIS/CSE600) or upper-level undergraduate electives (CIS/CSE400) on applied cryptography, database systems. I am also very interested in teaching introductory undergraduate courses (CIS/CSE 200/300) on algorithms, data structures, complexity theory, concrete mathematics, etc. I will prefer teach these courses in a top-down fashion (i.e. from real applications down to the more technical part) and from programmer's perspective (i.e. featuring programming tasks and labs). For instance, I am currently developing the applied-cryptography course at the graduate level (CIS600). My approach is to start from real security applications (e.g. password management) and then to present various cryptographic primitives necessary in these applications. My course will feature a series of programming assignments, such as using high-level cryptographic library (e.g. NaCL[1]) to construct a working protocol for the application of credential and password management.

III. Services

One of my career goals is to provide professional services to the research community and university. Towards this end, I have currently reviewed research manuscripts submitted for publication in journals and conferences. I have served as program committee members for computer-science conferences including ICDCS and IEEE Cloud, and served reviewers for journals, such as TOCS, TKDE, TSC, TWeb, TCSVT.

References

[1] A. C. Yao, "How to generate and exchange secrets (extended abstract)," in *27th Annual Symposium on Foundations of Computer Science, Toronto, Canada, 27-29 October 1986*, 1986, pp. 162–167. [Online]. Available: http://dx.doi.org/10.1109/SFCS.1986.25

[2] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game or A completeness theorem for protocols with honest majority," in *Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987, New York, New York, USA*, 1987, pp. 218–229. [Online]. Available: http://doi.acm.org/10.1145/28395.28420

[3] C. Liu, X. S. Wang, K. Nayak, Y. Huang, and E. Shi, "Oblivm: A programming framework for secure computation," in *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015*, 2015, pp. 359–376. [Online]. Available: http://dx.doi.org/10.1109/SP.2015.29

[4] O. Goldreich and R. Ostrovsky, "Software protection and simulation on oblivious rams," *J. ACM*, vol. 43, no. 3, pp. 431–473, 1996. [Online]. Available: http://doi.acm.org/10.1145/233551.233553

[5] Y. Tang, L. Liu, A. Iyengar, K. Lee, and Q. Zhang, "e-ppi: Locator service in information networks with personalized privacy preservation," in *IEEE 34th International Conference on Distributed Computing Systems, ICDCS 2014*. IEEE Computer Society, 2014, pp. 186–197. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6888815

[6] Y. T. L. Liu, "Privacy-preserving multi-keyword search in information networks," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2424–2437, 2015. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2015.2407330

[7] Y. Tang, T. Wang, and L. Liu, "Privacy preserving indexing for ehealth information networks," in *CIKM*, C. Macdonald, I. Ounis, and I. Ruthven, Eds. ACM, 2011, pp. 905–914.

[8] "Nwhin: http://www.hhs.gov/healthit/healthnetwork."

[9] "Shin-ny: http://www.health.ny.gov/technology/projects/."

[10] "Gahin: http://www.gahin.org/."

[11] "Beacon, https://genomicsandhealth.org/files/public/Beacon-FAQ.pdf."

[12] Y. Tang and W. Zhuang, "Towards building practical secure multi-party databases," in *IEEE SecDev*, 2016.

[13] P. E. O'Neil, E. Cheng, D. Gawlick, and E. J. O'Neil, "The log-structured merge-tree (lsm-tree)," *Acta Inf.*, vol. 33, no. 4, pp. 351–385, 1996.

[14] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. Gruber, "Bigtable: A distributed storage system for structured data (awarded best paper!)," in *OSDI*, 2006, pp. 205–218.

[15] "http://hbase.apache.org/."

[16] "http://cassandra.apache.org/."

[17] R. Tamassia, "Authenticated data structures," in *Algorithms - ESA 2003, 11th Annual European Symposium, Budapest, Hungary, September 16-19, 2003, Proceedings*, 2003, pp. 2–5. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-39658-1_2

[18] C. Martel, G. Nuckolls, P. Devanbu, M. Gertz, A. Kwong, and S. G. Stubblebine, "A general model for authenticated data structures," *Algorithmica*, vol. 39, no. 1, pp. 21–41, Jan. 2004. [Online]. Available: http://dx.doi.org/10.1007/s00453-003-1076-8

[19] C. Papamanthou, R. Tamassia, and N. Triandopoulos, "Authenticated hash tables," in *Proceedings of the 2008 ACM Conference on Computer and Communications Security, CCS 2008, Alexandria, Virginia, USA, October 27-31, 2008*, 2008, pp. 437–448. [Online]. Available: http://doi.acm.org/10.1145/1455770.1455826

[1] https://nacl.cr.yp.to/

[20] ——, "Authenticated hash tables based on cryptographic accumulators," *Algorithmica*, vol. 74, no. 2, pp. 664–712, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00453-014-9968-3
[21] Y. Zhang, J. Katz, and C. Papamanthou, "Integridb: Verifiable SQL for outsourced databases," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-6, 2015*, 2015, pp. 1480–1491. [Online]. Available: http://doi.acm.org/10.1145/2810103.2813711
[22] P. Devanbu, M. Gertz, C. Martel, and S. G. Stubblebine, "Authentic data publication over the internet," *Journal of Computer Security*, vol. 11, p. 2003, 2003.
[23] Y. Tang, T. Wang, L. Liu, X. Hu, and J. Jang, "Lightweight authentication of freshness in outsourced key-value stores," in *Proceedings of the 30th Annual Computer Security Applications Conference, ACSAC 2014, New Orleans, LA, USA, December 8-12, 2014*, C. N. P. Jr., A. Hahn, K. R. B. Butler, and M. Sherr, Eds. ACM, 2014, pp. 176–185. [Online]. Available: http://doi.acm.org/10.1145/2664243.2664244
[24] Y. Tang and J. Chen, "Log-structured authenticated data storage with minimal trust," 2016.
[25] "Intel corp. software guard extensions programming reference, 2014 no. 329298-002."
[26] "http://code.google.com/p/leveldb/."
[27] "Blockchain, https://en.wikipedia.org/wiki/Blockchain_(database)."
[28] "Tor project, https://donate.torproject.org/."