Application of Compression-Aware Algorithms to Improve the Performance of Multi-Party Computation Protocols
(Technical Topic)

An Analysis of the Decision Making Process in Inter-Organizational Data Sharing
(STS Topic)

A Thesis Project Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Yonathan Fisseha

Fall, 2019

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signature _____

Approved _____ Date _____
Kathryn A. Neeley, Associate Professor of STS, Department of Engineering and Society

Approved _____ Date _____
Nathan J. Brunelle, Assistant Professor of Computer Science, Department of Computer Science

1. Introduction

The past two decades of human history are perhaps best characterized by progress made in the computing frontier. One major consequence of this progress is the growth in the amount of data generated throughout the world. A report from GovLab shows that Internet data alone increased by 1696% from 2005 to 2012 (Govlab, 2014, n.p.). This growth is expected to continue into 2020; the global data, according to a report by CSC, is expected to grow to 35 zettabytes by the year 2020 (CSC, 2012, n.p.). Furthermore, the same report shows 80% of this data is stored and processed by organizations instead of individuals; thus actualizing the benefits of data to the fullest extent requires understanding how organizations manage this data. Data is insightful when it is interconnected and aggregated because it can give a more complete representation of the underlying information. For example, the California Franchise Tax Board, which handles personal and corporate income taxes for California, effectively used data sharing to identify people who fail to pay their taxes by aggregating data from federal, state, county, and local sources that indicate the existence of nonfilers and estimates of amount owed and then matching this data against accounts receivable data for the state (Fedorowicz, 2006, p.10). Without this sophisticated data aggregation, the board could not have a complete understanding of taxpayers in the state.

Although data sharing is a powerful technique when successful, it is socially and technically challenging to implement and maintain. As the Harvard Business Review Survey found out, “only 14% of the respondents [to the survey] claim that they effectively share data with external digital ecosystem partners” (Harvard Review, 2018, n.p.). These silos of data kept behind the digital and physical walls of organizations hinder organizations from fully utilizing

the power of the collective data. Currently, organizations share data by trusting one of the involved parties or an independent third-party, and this requirement of trust often discourages data sharing as a whole. Therefore, organizations need the technical tools to securely and efficiently share data with other organizations without trusting a third-party. Consequently, the technical topic is focused on providing a cryptographically secure and efficient protocol for data sharing that does not involve a trusted third-party. The STS topic is focused on understanding the social challenges organizations face when sharing data and aims to help organizations better navigate the decision making process of data sharing.

2. Technical Topic: Application of Compression-Aware Algorithms to Improve the Performance of Multi-Party Computation Protocols

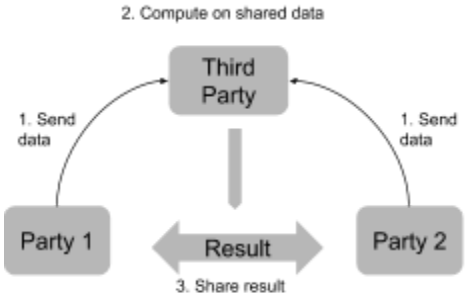
The state-of-the-art solutions for data sharing often outsource the data and computation to a computationally or technically more capable third-party. Involving a third-party, as shown in Figure 1.A, requires trusting the third-party with the aggregated data; this trust is often misplaced and has consequently led to data breaches, leaks, and abuse. For example, in 2018 internal threats and unauthorized access by the trusted party were the second most common forms of data breach, exposing 404 million consumer records (ITRC, 2015; Groot, 2019). Thus, organizations need a secure way to share and compute on aggregated data without trusting a third-party.

Multi-party computation (MPC) protocols allow parties to jointly perform some computation without disclosing their private input to each other, and collectively share the result of the computation on the aggregated data. Generally, the computation is defined as a function with multiple inputs to compute on the aggregated data, where each of the inputs to the function is privately provided by the involved parties. MPC was first introduced by Andrew Yao (Yao,

1982), where he detailed Garbled Circuits as a realization of MPC alongside the necessary security proofs. In Garbled Circuits, one of the parties becomes the sender and one of them becomes the receiver (Party-1 and Party-2 in Figure 1.B, respectively). Party-1 first enumerates the possible inputs into the function and their corresponding outputs into a table. It then encrypts each row of this table using a two-key symmetric encryption scheme, where the keys are the two possible inputs of the function in that row. Party-1 then sends the encrypted table, or garbled circuit, to the other party along with the key that corresponds to Party-1's input. Party-2 attempts to decrypt each row in the garbled circuit using Party-1's key and its own key. Since each row is uniquely encrypted, Party-2 can decrypt only one row and learns nothing else in the process. Party-2 can then send the decrypted result back to Party-1 for further post-decryption processing.

Recent projects have successfully used MPC in a number of cases to aggregate and compute on privacy sensitive data. For example, in 2008 beet farmers and processors in Denmark performed a nation-wide double auction to negotiate the price per unit of beets. Each processor specified how many beets they were willing to buy at a given price and, similarly, the farmers specified how many they were willing to sell at each price point. Using an MPC protocol, they were able to find the price that balances supply and demand, which would be impossible to compute if the data of either group was exposed, as it would corrupt the auction (Bogetoft, 2008, p.2). Despite its success in some cases, MPC has not achieved widespread adoption across different industries due to its high latency. Our project aims to improve the performance of MPC protocols in two ways. First, by efficiently compressing the input of each party, thus effectively reducing the size of the encrypted data the parties must send, receive, and process. While a naive compression technique can reduce the size of the encrypted data enroute

to the receiving party, it requires the receiving party to decompress the data so that it can compute the function on it. The decompression process can increase the total overhead of the protocol and possibility lose all performance gains from the compression step. Thus, our approach includes a second step of augmenting the desired function with compression-awareness, such that the computation proceeds on the compressed input without the decompression step. My work aims to demonstrate this principle on a specific algorithm called Longest Common Subsequence (LCM), which finds the longest common sequence of characters between two or more input strings, by augmenting it to process strings compressed using the Lemple-Ziv compression scheme (J. Ziv, 1978). The LCM algorithm design builds on the work of Crochemore et al. (2003), which solves the sequence alignment problem, a problem similar to LCM, by first compressing the strings using Lemple-Ziv.



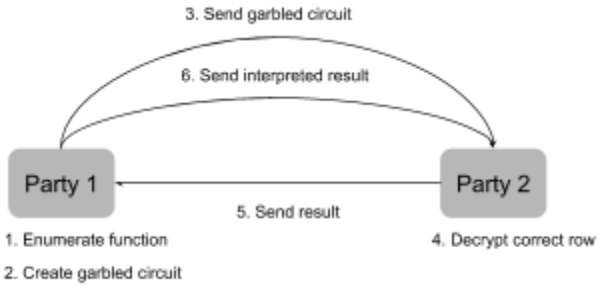


Figure 1.A: 3rd Party setup **Figure 1.B MPC protocol**

The involvement of a third-party requires the parties to trust it with their data as well as the aggregated data. The MPC protocol protects the privacy of each party's input thus parties don't need to trust each other. Created by Author.

3. STS Topic: An Analysis of the Decision Making Process in Inter-Organizational Data Sharing

Since inter-organizational data sharing is a sociotechnical challenge, organizations must solve social challenges beyond the technical domain. In a study on government agencies' data sharing, Yang et al. classify the challenges of inter-organizational sharing into three categories based on their literature review as shown in Figure 2. Consequently, when considering data sharing in the context of organizations, including private organizations, the analysis must include these factors as well as other influences, like the initial motivation of the organization to share data. These factors will vary significantly between industries; here, I consider data sharing in public health, private organizations, and academia.

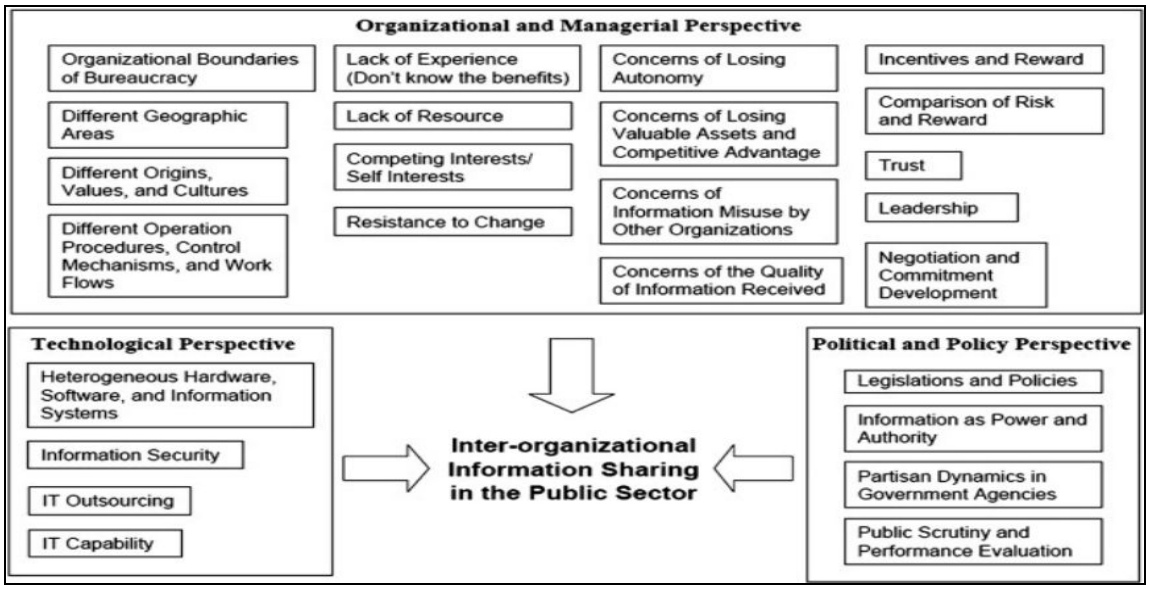


Figure 2: The three categories of barriers to data sharing in government agencies. Aside from the technical challenges of data sharing, most of the barriers actually come from the Organizational and Political categories. Adopted from Yang et al. (Yang, 2005, np)

Inter-organizational data sharing is not only useful but often necessary for public health organizations to fulfill their duties. In a 2014 literature review, van Panhuis et al. conclude that

lack of trust and privacy protection laws are two of the many barriers for data sharing in public health. For example, in 2016 two county health departments in Illinois collaborated to create a regional Community Health Improvement Plan, involving 3059 citizen surveys and 70 agencies (Tazewell, 2016, p.5). One of the committees was tasked with improving behavioral health in the justice sector and thus had to combine data from the criminal justice system with various health indicators. While the county jail cooperated, all other criminal justice stakeholders did not cooperate due to federal laws that limit data sharing and the lack of trust among them. The committee was able to reach a compromise where the data providers limited identifying information but still provided useful data (Schmit, 2019, n.p.). While policies and guidelines on inter-organizational data sharing and how to utilize the data "can facilitate relationship building, risk reduction, and trust development in inter-organizational information sharing projects" (Yang, 2011, n.p.), existing policies often fail to do so or have counterproductive effects due to extreme limitations.

Unlike the public health sector, the private sector doesn't often have the necessity to share data, since private organizations often are not designed with such interdependencies. Consequently, data sharing among such organizations depends on their willingness to share or other incentives. In a 2012 study on information sharing in supply chain, Timon et al. showed that a strong partnership is needed before a strong will to share data arises. A strong partnership is necessary because data sharing requires organization to release confidential financial and strategic information to partners who might be competitors. Willingness to share can be used in this decision making process as a proxy to evaluating the partnership strength and the level of trust (Timon, 2012, n.p.). More generally, the risk and reward comparison of data sharing is not

built into the business model of current day private organizations, and as a result, these organizations preemptively avoid sharing their data.

In contrast to the private sector which seem to lack the motivation to share data, academia is built on the principle of data sharing to avoid redundancy and increase reliability of research (Fetcher, 2015, p.9). However, here too data sharing is registered. These barriers include losing competitive advantage and degree of control on the individual level, and limitations put in place by funding agencies on the organization level. Younger researchers are especially reluctant to share research data with the community, while people over 50 are more likely to share. This is in part because data is not currently valued as much as publications, thus tenure evaluations similarly don't reward raw data as much as published papers. The problem is compounded by the need for control on how the data is used, since otherwise the receiver of the data could publish results first (the equivalent of first-to-market in the private sector). Similarly on the organizational level, researchers are either prohibited or not encouraged by their employers to share data.

As demonstrated above, although the general categories of Yang et al. are thematically useful, the decision making process on data sharing varies significantly across different industries. The STS project aims to apply techno-selectivity, which expresses the process of evaluating a given concept against the organization's values before accepting it, to help organizations navigate the data sharing decision making process. Techno-selectivity can give a vivid picture of the values that can support and coexist with data sharing in the data-centric world. The main challenge with this research is the lack of case studies and raw data from

various industries, and since I will not be able to collect the data myself, the project will most likely try to extrapolate based on the available data.

4. Conclusion

As our modern society increasingly becomes data dependent, the organizations that process and manage data should not continue to exist in isolation. Rather, they should aggregate their data and compute on the aggregated data to gain a more complete depiction of the information represented by the data. This project explores the challenge of inter-organizational data sharing both from a technical and social perspective. The technical topic aims to improve the efficiency of MPC protocols using compression-aware algorithms, which can provide a new layer of security and control for organizations interested in data sharing. The STS topic explores the social challenges related to inter-organizational data sharing through an analysis of three industries, and suggest techno-selectivity as a technique to navigate the decision making process. A successful completion, including implementation, of the technical project can provide an efficient and cryptographically secure means for data sharing and collaboration, which, as discussed previously, increases the value of the data and consequently, the capabilities of these organizations. Similarly, a successful completion of the STS project can provide a comprehensive understanding of data sharing beyond the technical difficulties paving the way to making data sharing a common business practice.

References

Abbott, Mike, and Rob Schimek. The Data Sharing Economy: Quantifying Tradeoffs that Power New Business Models. AIG.

Archer, D. W., Bogdanov, D., Lindell, Y., Kamm, L., Nielsen, K., Pagter, J. I., … Wright, R. N. (2018). From Keys to Databases—Real-World Applications of Secure Multi-Party Computation. The Computer Journal. doi: 10.1093/comjnl/bxy090

Harvard Business Review, (2018). An Inflection Point for the Data-Driven Enterprise. *Harvard Business Review*.

Bogetoft, P., Christensen, D. L., Damgård, I., Geisler, M., Jakobsen, T. P., Krøigaard, M., ... & Schwartzbach, M. I. (2008). Multiparty Computation Goes Live. *IACR Cryptology ePrint Archive*, *2008*, 68.

IBM Security, (2019). Cost of a Data Breach Study. [*IBM*]. Retrieved from https://www.ibm.com/security/data-breach

CSC (2012), "Big data universe beginning to explode." [A non-profit research report]

Crochemore, Maxime, et al. A Subquadratic Sequence Alignment Algorithm for Unrestricted Scoring Matrices. *SIAM Journal on Computing*, vol. 32, no. 6, 2003, pp. 1654–1673., doi:10.1137/s0097539702402007.

Fecher, Benedikt & Friesike, Sascha & Hebing, Marcel. (2015). What Drives Academic Data Sharing?. PLoS ONE. 10. 10.1371/journal.pone.0118053.

Fedorowicz, J., Gogan, J. L., & Williams, C. B. (2006). *The e-government collaboration challenge: Lessons from five case studies*. IBM Center for Business of Government.

Giannopoulos, Thanos & Mouris, Dimitris. (2018). Privacy Preserving Medical Data Analytics using Secure Multi Party Computation. An End-To-End Use Case.. 10.13140/RG.2.2.19303.70562.

Goldwasser, S. (1997, August). Multi party computations: past and present. In *Proceedings of the sixteenth annual ACM symposium on Principles of distributed computing* (pp. 1-6). ACM.

Govlab. The GovLab Index: The Data Universe. *The Governance Lab @ NYU*, 9 Jan. 2014, http://thegovlab.org/govlab-index-the-digital-universe/.

Groot, J. D. (2019, January 3). The History of Data Breaches.

ITRC . 2015. Data Breach Reports. Identity Theft Resource Center. [Annual report]

J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding. In *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530-536, September 1978. doi: 10.1109/TIT.1978.1055934

K. Smith, L. Seligman and V. Swarup, Everybody Share: The Challenge of Data-Sharing Systems. *Computer*, vol. 41, no. 9, pp. 54-61, Sept. 2008.doi: 10.1109/MC.2008.387

Schmit, C., Kelly, K., & Bernstein, J. (2019). Cross Sector Data Sharing: Necessity, Challenge, and Hope. *The Journal of Law, Medicine & Ethics*, *47*(2_suppl), 83-86.

Tazewell County Health Department (2016). Annual Report. Retrieved from https://www.tazewellhealth.org/DocumentCenter/View/406/2016-annual-report

Timon C. Du, Vincent S. Lai, Waiman Cheung, and Xiling Cui. 2012. Willingness to share information in a supply chain: A partnership-data-process perspective. *Inf. Manage.* 49, 2 (March 2012), 89-98. DOI=http://dx.doi.org/10.1016/j.im.2011.10.003

Yang, T. M., & Maxwell, T. A. (2011). Information-sharing in public organizations: A literature review of interpersonal, intra-organizational and inter-organizational success factors. *Government Information Quarterly*, *28*(2), 164-175.

Yao, Andrew C. Protocols for Secure Computations. *23rd Annual Symposium on Foundations of Computer Science (Sfcs 1982)*, 1982, doi:10.1109/sfcs.1982.38.

Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., ... & Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC public health*, *14*(1), 1144.

Verhulst, S. (2014, September 17). Mapping the Next Frontier of Open Data: Corporate Data Sharing. (Guest Blog Post). Retrieved from https://www.unglobalpulse.org/mapping-corporate-data-sharing.

Zaheer, A., McEvily, B., & Perrone, V. (1998). Does trust matter? Exploring the effects of interorganizational and interpersonal trust on performance. *Organization science*, *9*(2), 141-159.