Advancements in Private Set Union: Protocols, Security Frameworks, and Applications

Zihan Lin 1, Jingyan Wang [2, *] and Shuaiqi Xu [3]

[1] Maynooth College, Fuzhou University, Fuzhou, 350108, China

[2] Cyber Science and Engineering, Sichuan University, Chengdu, 610207, China

[3] Computer Science and Engineering, Tianjin University of Technology, Tianjin, 300382, China

* Corresponding Author Email: 2021141530055@stu.scu.edu.cn

Abstract. Private Set Union (PSU), a subset of Private Set Intersection (PSI) methodologies, is swiftly marking its importance in an era that progressively values computational integrity without infringing data privacy. This necessity sprouted from the intricate balance between ensuring rigorous privacy protection and fostering beneficial data sharing. PSU enables multiple entities to establish a union of their individual datasets without revealing their private information to each other. The foundational security models that drive PSU are both intricate and robust, with roots tracing back to the Computational Diffie-Hellman problems. These problems address the challenges of ensuring secure computations in multi-party scenarios. Attack strategies targeting PSU have evolved over time, leading to an enhancement in the defense mechanisms. Techniques utilizing encryption, randomization, and other cryptographic methods have seen consistent improvement, making privacy-preserved inter-party data analytics both feasible and efficient. Recent breakthroughs in PSU protocols have seen rigorous evaluations based on efficiency, security, and performance metrics. These evaluations aim to determine how these protocols fare under real-world conditions and what challenges they might encounter. Additionally, PSU's adaptability and relevance have been showcased across various industries. Whether it's digital libraries collaborating for enriched content while preserving copyright restrictions, financial institutions aiming to analyze transaction patterns without compromising individual account details, risk assessment firms sharing data to provide better evaluations without revealing sensitive details, joint graph computations ensuring optimal routing without exposing network infrastructures, or the ever-growing Internet of Things (IoT) ecosystem maintaining device interoperability while upholding user privacy; PSU's prominence is undeniable.

Keywords: Private Set Union, Data Privacy, Computational Diffie-Hellman, Security Models.

1. Introduction

Private Set Union stems from the pressing need in myriad real-world scenarios where multiple entities seek to conduct data analysis and computations, yet remain wary of unveiling their original datasets. Conventional data processing approaches often necessitate amalgamating datasets to facilitate computations. Yet, when delving into datasets laden with sensitive information, a direct data exchange could pave the way for significant privacy infractions. Navigating this precarious balance between safeguarding privacy and fostering data sharing, PSU technology crystallized as a seminal adaptation of the Private Set Intersection specifically tailored for functional computation landscapes.

Central to its design, PSU empowers two distinct parties, each equipped with their exclusive datasets, to decipher their union, all the while ensuring no data, barring the outcome, sees the light of day. To paint a vivid picture, consider the intricate challenge of curating an IP blacklist. Here, two entities, vested in maintaining this blacklist, aspire to discern the union, yet are reticent about laying bare their exhaustive IP blacklists. Within such confines, divulging intersection data morphs into a potential hazard, positioning PSU as the linchpin that conceals intersection specifics yet uncovers union data.

While PSI protocols abound, a majority are ill-suited to directly tackle the PSU conundrum. Spanning from 2019 to date, the academic arena has witnessed a burgeoning interest, with scholars delving deep into the research and fine-tuning of PSU protocols. Harnessing sophisticated techniques

like encryption and randomization, PSU stands as a beacon, ensuring an impregnable data privacy shield, even as it facilitates a harmonious and insightful data computation rendezvous between participating entities.

2. Security Model

The Computational Diffie-Hellman (CDH) problem stands as a cornerstone in the cryptographic landscape. When presented with two elements, ga and gb, the CDH challenge revolves around the computation of gab, where g represents a generator, and a and b serve as private keys, with ga and gb being their corresponding public keys. The inherent complexity of the CDH problem arises from the predicament that, despite knowledge of ga and gb, deriving gab remains formidable unless either a or b is disclosed. Its counterpart, the Decisional Diffie-Hellman (DDH) problem, rooted in the CDH problem, tasks one with ascertaining if, given four elements g, ga, gb, and gc, the relation c = ab holds true. Essentially, the DDH challenge is discerning whether gc mirrors gab. The enigma here is that, even with g, ga, gb, and gc at one's disposal, deducing if c equals ab remains elusive without access to a or b.

Within the realm of protocol behavior, the Semi-honest Model operates under the premise that all participants adhere diligently to the protocol's dictates, yet they might endeavor to glean maximal information from the operational procedure. The Coercion-resistant Semi-honest Model mirrors the Semi-honest stance but further postulates scenarios where attackers might compel participants into unveiling their confidential data or altering their actions. Meanwhile, the Malicious Model concedes that participants might instigate any conceivable assault, ranging from protocol deviations to data manipulation, thereby jeopardizing the protocol's sanctity. Several attack paradigms emerge in this discourse. The Static Attack envisages assailants with preordained targets prior to the protocol's initiation. In juxtaposition, the Dynamic Attack paradigm grants assailants the latitude to recalibrate their targets contingent on the protocol's trajectory and the intelligence harvested. A specialized subset of this, the Adaptive Attack, enables assailants to cherry-pick the most potent assault strategy, informed by the protocol's nuances and accumulated data. The Collusion Attack, meanwhile, portrays a scenario where multiple rogue participants connive, pooling resources and strategizing collectively, thereby mounting assaults that would be beyond the reach of solitary attackers.

Delving into the intricacies of the Semi-honest Model, validating the security of PSU necessitates a series of steps: delineating security aspirations, postulating an adversary endowed with boundless computational prowess yet remains in the dark, asserting that such adversaries remain incapable of accessing confidential data, and lastly, corroborating the protocol's integrity.

3. Existing PSU Protocol Designs

Kolesnikov et al. unveiled a groundbreaking protocol for private set union computation within the OT hybrid model. Apart from leveraging the OT extension, this protocol predominantly hinged on symmetric key primitives [1]. The linchpin of this approach was the innovative Reverse Private Member Test (RPMT) protocol. Garimella et al. put forth a novel approach to compute the intersection of any function, under the assumption that revealing the intersection’s cardinality does not compromise security. This method seamlessly integrates with any conventional 2PC protocol. Certain intersection computations, by harnessing identifiers, achieve a more direct and efficient resolution, circumventing the intricate steps of secret sharing. Benchmarking against state-of-the-art intersection computation protocols, such as that of Pinkas et al. from Eurocrypt 2019, this novel protocol trims communication overheads by approximately 2.5-3 times and boasts a swifter runtime on comparatively slower networks (50Mbps) [2]. Its private id functionality, drawing from PSU, is both streamlined and significantly quicker than its antecedents. In 2022, Jia et al. championed a novel PSU protocol blueprint, integrating shuffle techniques for the first time [3]. By sidestepping burdensome operations prominent in earlier renditions, like additive homomorphic encryption and

recurrent operations on receiving sets, this design underscored both efficiency and security while eliminating superfluous data leakage. Performance enhancements were palpable, with improvements ranging from 4-5 times in both WAN and LAN configurations via a singular thread.

Tu et al. introduced an inventive construction for unbalanced PSU, amalgamating Fully Homomorphic Encryption (FHE) with a fresh protocol dubbed Permutation Matrix Private Equivalence Test [4]. This protocol's genius lay in its ability to render the communication complexity of the larger dataset logarithmic, thereby slashing communication durations. Especially in unbalanced configurations, it bested preceding protocols. A unique trait of this approach was its capability to excel in situations with stark size discrepancies between parties' input/select vectors. Communication intervals plummeted by up to 37 times across varied network contexts, and runtime surged forward, clocking accelerations between 10 and 35 times. Yet, it's worth noting that the computational complexity saw a significant uptick. Zhang et al. architected a versatile framework grounded in Oblivious Transfer (OT) tailored for PSU, integrating a freshly minted protocol known as the Multi-Query Reverse Private Membership Test (mq-RPMT) to tackle inherent computational and communication imbalances [5]. Two primary structures were outlined for mq-RPMT: one rooted in symmetric key encryption coupled with general 2PC methodologies and the other in re-randomizable public-key encryption (PKE). Both avenues realized PSU with linear computational and communicative complexities. When pitted against contemporary PSU protocols, experimental outcomes illustrated the exceptional prowess of their PKE-centric protocol. Depending on dataset dimensions, communication overheads dipped between 3.7 to 14.8 times. In terms of runtime, this novel PSU design clocked speeds from 1.2x to 12x faster than existing top-tier methods, influenced by the network milieu.

4. Experiments Design

Pursuant to the directions outlined for enhancing PSU efficiency, a structured experimental design is set to unfold. The primary choice for the programming language is likely to be C/C++, given its superior efficiency compared to Java. The emphasis will be on executing multi-thread experiments, particularly within the application purview of IP blacklists. The benchmark for evaluation will encompass an array of current methodologies such as KRTW, GMRSS, JSZDG, and PKEPSU schemes. The overarching aim is to conduct a meticulous comparison of these strategies, evaluating their merits and potential drawbacks. Results will be systematically presented in tabular formats, spotlighting communication costs and runtime metrics across diverse data volumes and network ranges. To offer a more visual representation, line plots are set to be crafted, focusing on network bandwidth. These plots will provide a lucid comparison of communication complexity versus computational complexity for the proposed schemes. For clarity and ease of reference, minimum values will be annotated on these charts, allowing for immediate insights into performance metrics. In sum, this experimental design not only delves deep into the nuances of PSU efficiency but also presents findings in an accessible and comprehensive manner, paving the way for further advancements in the domain.

5. Typical application analysis

In the realm of big data information transmission and cloud computing, privacy computing has unveiled new horizons. A prominent strategy, presented in [5], addresses the protection of digital library users' privacy. It adeptly navigates the challenges of privacy infringements during the construction and operational phases of digital libraries. Furthermore, a notable solution, detailed in [6], underscores the nuances of user information privacy within mobile social networks. This solution hinges on similarity matching, striving to find the ideal equilibrium between attribute item distinction, the potency of similarity measurement benchmarks, and the rapidity of matching execution.

Turning to the financial sector, referenced in [7], there's a compelling discourse on the potential of large-scale applications of privacy computing. The narrative pivots from ensuring the sanctity of data circulation to fostering a secure environment for the dissemination and sharing of financial data, which is intrinsically sensitive and invaluable. Delving further, [8] illuminates the persistent challenges of trust impediments and data silos in financial digitization. The paper advocates for fortifying supply chains with privacy computing, thereby laying the groundwork for viable solutions. As for the nexus of privacy computing and privacy intersection within risk assessment contexts: collaborative risk evaluations routinely necessitate amalgamated lists to bolster precision. PSU emerges as a potent tool, especially evident in contemporary joint IP blacklists. Such a method enhances the pinpointing of malevolent sources by pooling blacklists spanning various overseers and assault categories, as delineated in [9]. Additionally, the BLAG system, spotlighted in [10], employs the PSU protocol, serving as a vanguard for aggregated IP blacklists.

On the topic of Privacy Computing and Privacy Intersection in Joint Graph Computing: [11] elucidates a specific application of PSU in the realm of aggregate graph computations. Here, both entities possess a diagram, be it network topologies or sales channel maps, and they endeavor to run specific algorithms on a conjoined diagram. Achieving this in a privacy-conservative manner mandates an evolved PSU protocol. The Internet-of-Things (IoT) landscape, particularly within vehicular contexts, grapples with ensuring the seamless, secure dispatch of location-centric services. This domain is riddled with challenges, from privacy threats to security pitfalls. Most extant privacy shields are anchored to centralized location servers, which unfortunately are susceptible to singular point failures and inadvertent privacy leaks. While anonymity and cryptography have been heralded as remedies, they bring along intricate encryption intricacies and a demanding resource appetite. Crafting a distributed, privacy-resilient data shield for IoT is paramount, especially within the smart city blueprint. In Vehicle Ad Hoc Networks (VANETs), proximity testing is ascendant, especially when contextualized within location-centric services. Yet, prevailing proximity testing modalities, highlighted in [12], aren't without their challenges. From safeguarding user location confidentiality to maintaining equilibrium in interactions and curbing computational burdens, the hurdles are manifold, as further detailed in [13,14].

6. Conclusion

Drawing from a thorough review of existing literature, our current research endeavor involves crafting a comprehensive survey on research related to Private Set Union. This will encapsulate a summary and introduction of extant PSU protocols, diving deep into the results achieved thus far and shedding light on potential avenues for future research. Present findings indicate that the prevailing trend in this domain leans heavily towards privacy union computations between two entities. There's a conspicuous absence of in-depth exploration regarding multi-party privacy unions. Interestingly, the spotlight tends to be on enhancing computational and communication prowess, often sidelining the vital aspect of rigorous security study and proofing.

Looking ahead, there are promising realms to venture into. For instance, the design and development of PSU protocols that come with reduced computational and communication complexities would be groundbreaking. Researchers could also delve into amalgamating techniques like Oblivious Transfer, Hash functions, One-Time Password Reverse Functions, Oracle functions, Diffie-Hellman key exchange, Decisional Diffie-Hellman assumptions, Multi-Party Computation, and the one-time-pad encryption method. To provide specific examples: One could look into crafting an Oblivious Transfer-based PSU protocol tailored for unbalanced datasets. Another promising avenue would be developing a Hash function-centric unbalanced PSU protocol. A third could explore Multi-party secure PSU protocols grounded in Oblivious Transfer methodologies. Lastly, the introduction of a PSU protocol wherein Public Key Encryption stands as the cornerstone technique could be revolutionary. As we move forward, striking a balance between efficiency and security will be paramount in shaping a resilient digital landscape.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Vladimir Kolesnikov, Mike Rosulek, Ni Trieu, and Xiaorui Sun. Scalable private set union from symmetric-key techniques. In International Conference on the Theory and Application of Cryptology and Information Security, pages 636–666, Cham, 2019. Springer International Publishing.

[2] Gayathri Garimella, Payman Mohassel, Mike Rosulek, Saeed Sadeghian, and Jaspal Singh. Private set operations from oblivious switching. In IACR International Conference on Public-Key Cryptography, pages 591–617. Springer, 2021.

[3] Yanxue Jia, Shufang Sun, Hui Shen Zhou, , et al. Shuffle-based private set union: Faster and more secure. In 31st USENIX Security Symposium (USENIX Security 22), pages 2947–2964, 2022.

[4] Binbin Tu, Yanhong Chen, Qingju Liu, et al. Fast unbalanced private set union from fully homomorphic encryption. Cryptology ePrint Archive, 2022.

[5] Cong Zhang, Yu Chen, Weiran Liu, Min Zhang, and Dongdai Lin. Linear private set union from multi-query reverse private membership test. Cryptology ePrint Archive, 2022.

[6] Hui Pan. (2011). Research on user privacy of digital library and its implications for cloud computing services. Information Theory and Practice, 34(4), 4.

[7] Dan Yanghe. (2015). Research on privacy issues based on similarity matching in mobile social networks. (Doctoral dissertation, Shanghai Jiao Tong University).

[8] Wang Guosai, Li Yi, Chen Kun, Shi Shi, Yang Zuyan. (2022). Financial application of privacy computing technology. Financial Development Research (8), 31-37.

[9] Chen JW. (2021). Private Computing Empowers Supply chain Finance. China Finance (24), 3.

[10] Hogan, K., Luther, N., Schear, N., Shen, E., Stott, D., Yakoubov, S., & Yerukhimovich, A. (2016, November). Secure multiparty computation for cooperative cyber risk assessment. In 2016 IEEE Cybersecurity Development (SecDev) (pp. 75-76). IEEE.

[11] Ramanathan, S., Mirkovic, J., & Yu, M. (2020, January). Blag: Improving the accuracy of blacklists. In NDSS.

[12] Brickell, J., & Shmatikov, V. (2005). Privacy-preserving graph algorithms in the semi-honest model. In Advances in Cryptology-ASIACRYPT 2005: 11th International Conference on the Theory and Application of Cryptology and Information Security, Chennai, India, December 4-8, 2005. Proceedings 11 (pp. 236-252). Springer Berlin Heidelberg.

[13] Zhou, Q., Zeng, Z., Wang, K., & Chen, M. (2022). Privacy Protection Scheme for the Internet of Vehicles Based on Private Set Intersection. Cryptography, 6(4), 64.

[14] Zhang, L., Gao, W., Chen, S., Ren, W., Choo, K. K. R., & Xiong, N. N. (2021). A Privacy-Preserving Proximity Testing Using Private Set Intersection for Vehicular Ad-Hoc Networks. IEEE Transactions on Industrial Informatics, 18(10), 7373-7383.