Birds of a Feather Leak Together: The Set Bias Privacy Problem

Abhishek Tiwari

Citation: *A. Tiwari*, "Birds of a Feather Leak Together: The Set Bias Privacy Problem", Abhishek Tiwari, 2024. doi:10.59350/rvhrr-f1704

Published on: October 24, 2024

Secure multi-party computation (SMPC) enables organisations to collaborate on sensitive data analysis without directly sharing raw information. However, seemingly harmless aggregate outputs, particularly private set intersection (PSI), can leak individual-level information when analysed strategically over time. This post is based on research presented by [1] at the 31st USENIX Security Symposium and examines how privacy systems can be vulnerable to sophisticated inference attacks that exploit natural dataset biases. This post is also a nice continuation from previous post where we covered privacy-preserving multi-touch attribution case study from TikTok(see [2]) and briefly discussed "differential attacks" exploiting gaps in PSI protocols.

Understanding Output Privacy

While much attention in privacy-preserving computation focuses on protecting input data through encryption and secure protocols, output privacy presents a distinct challenge. Output privacy concerns the information that can be inferred from computation results, even when those results appear safely aggregated or anonymized.

Traditional approaches to output privacy rely on simple principles like k-anonymity or minimum threshold reporting. However, these approaches fail to account for the cumulative effect of multiple outputs over time. Each piece of output provides information about the underlying data – that's what makes it useful. Yet this same information, when combined with other outputs and external knowledge, can enable privacy attacks.

Moreover, sequential composition becomes critical as each output reveals information that can be combined with previous outputs. External knowledge, such as demographic patterns or natural data biases, can dramatically amplify privacy risks. Dynamic data adds complexity, as changes can either help privacy by making old inferences obsolete or hurt it by revealing information about the changes themselves.

Private Set Intersection (PSI)

PSI protocols form the foundation of many privacy-preserving computations, allowing two parties to identify common elements in their sets without revealing anything else. PSI protocol is a black box that receives a set X from one party and a set Y from the other. It internally computes the intersection size $|X \cap Y|$ and ends this value to one of the two parties.

PSI-CA (Cardinality) is a variant that reveals only the size of the intersection, while other variants like PSI-SUM enable computation of aggregates (like sums or averages) over values associated with intersecting elements. Private-ID assigns consistent random tags to elements through a shared random function.

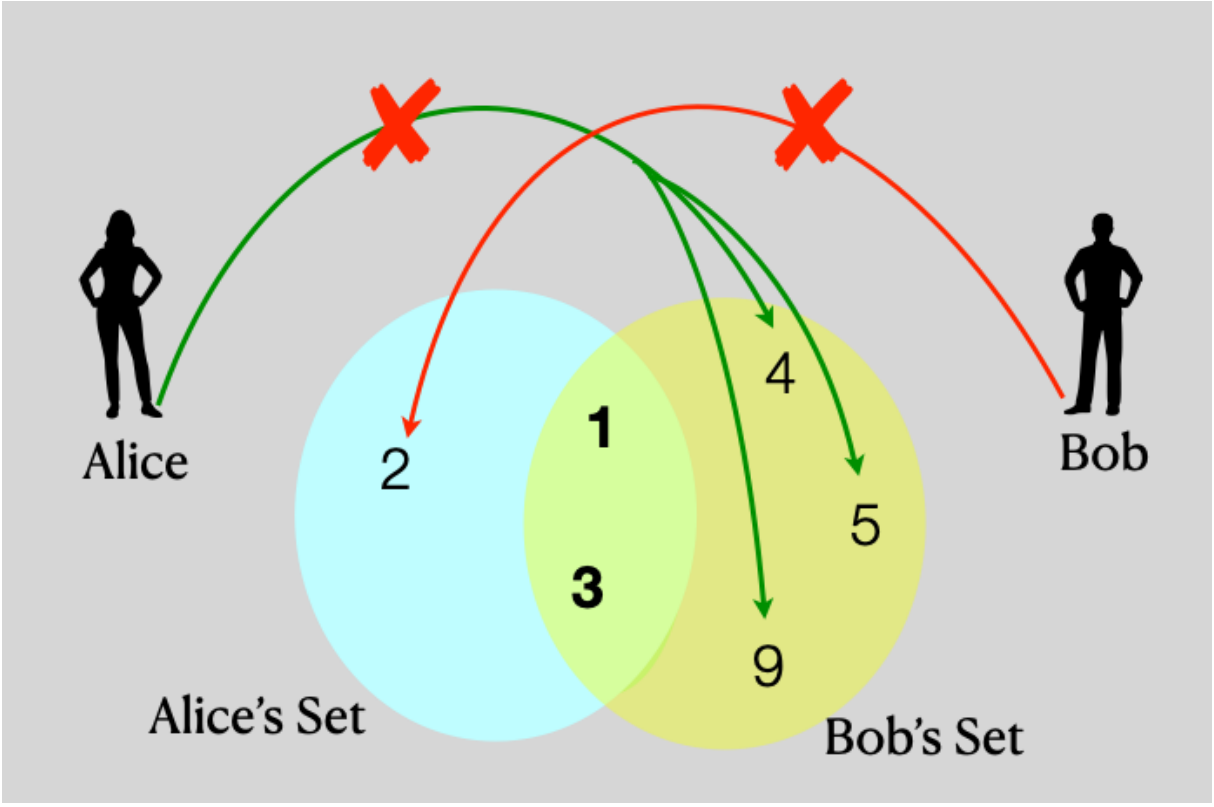

Figure 1: PSI determine the intersection of sets held by multiple parties, without revealing the non-intersecting components.. may be used in many practical applications including password checkup, DNA private matching, measuring ads efficiency, COVID-19 contact tracing etc. Image Credits Wikipedia.

Many real-world privacy-preserving systems rely on PCI-CA protocols. COVID-19 contact tracing apps use PSI-CA to let users check potential exposure while preserving patient privacy. Online advertising platforms employ protocols like PSI-SUM and Private-ID to measure ad performance metrics without exposing individual user data.

Vulnerability

While PSI protocols protect against direct data exposure, subtle privacy leaks can occur. For example repeatedly running PSI with strategically chosen inputs can reveal information about non-intersecting elements. Similarly, when sets evolve over time, changes in intersection sizes may leak membership information. Finally, implementation details may leak information through timing or memory access patterns i.e. side-channel attacks.

Attack Strategies

The Baseline Attack

The baseline attack employs a binary-search-like approach for efficient membership inference. Starting with a complete set of target elements, the attacker organizes them into a binary tree where each non-leaf node's elements are split into two subsets. Through depth-first search and strategic querying of intersection sizes, the attacker can identify or eliminate multiple elements with far fewer queries than testing each individually.

The Feature-Aware Attack

A more sophisticated approach incorporates known characteristics of target elements. Instead of random splitting, the attacker clusters elements based on features that correlate with membership probability. For example, in COVID-19 contact tracing, clustering based on symptoms like fever and cough makes it more likely to group infected patients together. This feature-aware strategy significantly accelerates the identification process.

Simulated Attacks

COVID-19 Contact Tracing Attack

In token-based contact tracing systems, users automatically exchange anonymous tokens via Bluetooth when in close proximity. The attack begins with strategically placed mobile phones collecting tokens from passing individuals. Through external observation and symptom screening data, the attacker associates symptoms with specific tokens.

Through 135 protocol queries, the feature-aware attack identified an average of 25.9 infected patients' tokens. This high success rate stems from strong correlation between symptoms and COVID-19 status, making the feature-aware approach 5.2-9.0 times more effective than traditional methods.

Ad Conversion Revenue Measurement Attack

Google's Private Join and Compute system lets advertisers calculate revenue from ad clicks using PSI-SUM. The attacker obtains potential `personal_ids` through first-party customer data and third-party marketplaces, already linked to demographic features.

With just 15-60 protocol invocations, the baseline attack found 2.0-7.5 set members from sets of 512-2048 target elements. Perfect accuracy was achieved due to the growing nature of the victim's set, as new ad clicks accumulate over time.

Ad Conversion Lift Measurement Attack

Facebook's Private Lift system uses Private-ID for measuring incremental conversions i.e. conversion lift. Despite being limited to one query per day, 30 daily protocol invocations achieved high-precision identification of viewer segments without false positives. The attacks consistently found 4-8 set members from various target set sizes, maintaining effectiveness despite weak demographic correlations.

Privacy Implications

The implications extend beyond technical vulnerabilities. In contact tracing, identifying specific infected individuals enables harassment and discrimination. For advertising platforms, revealing specific ad clickers exposes detailed information about individual interests and behaviors - especially problematic for sensitive topics.

The attack's effectiveness at exploiting natural dataset bias is particularly troubling. Features that make data useful for legitimate analysis become powerful tools for attackers. The very patterns that enable valuable insights also create vulnerabilities.

Defensive Approaches

The research explored several potential defensive strategies against set membership inference attacks, each with its own tradeoffs and implementation challenges.

Query Rate Limiting

The most straightforward defense involves restricting the number of intersection size queries allowed within a given time period. By limiting the attacker's ability to make multiple queries, this approach directly constrains their ability to gather enough information for successful inference attacks.

However, implementing effective query rate limiting proves challenging in practice. The fundamental difficulty lies in determining appropriate thresholds. Too restrictive limits hamper legitimate functionality, while too permissive ones may not provide adequate protection. Furthermore, the appropriate

threshold often depends on factors unknown to the defender, such as the attacker's target set size and background knowledge.

Real-world systems must also consider how query limits affect different use cases. For instance, contact tracing applications need frequent updates to remain effective, while advertising measurement might accommodate less frequent queries. Organisations must carefully balance security requirements against operational needs when setting these limits.

Pattern Detection and Monitoring

A more sophisticated approach involves monitoring query patterns to detect potential attacks. This defense strategy analyzes sequences of queries to identify suspicious patterns that might indicate an ongoing inference attack. For example, the binary-search pattern used in the baseline attack creates a distinctive sequence of intersection sizes.

Pattern detection systems can incorporate various signals: query frequency, set size distributions, overlap between consecutive queries, and changes in intersection sizes. Machine learning models might help identify subtle attack patterns that simple rule-based systems miss.

However, sophisticated attackers can modify their approach to avoid detection. They might introduce random variations in their query patterns, interleave attack queries with legitimate-looking ones, or distribute their queries across multiple accounts or time periods. This creates an ongoing cat-and-mouse game between attackers and defenders.

Differential Privacy Integration

By adding carefully calibrated noise to reported intersection sizes, systems can provide provable privacy guarantees while maintaining approximate utility.

The implementation requires several key decisions. The privacy parameter ε must be chosen to balance privacy and utility - smaller values provide stronger privacy but make results less accurate. The privacy budget must be allocated across multiple queries, considering how privacy guarantees degrade with repeated queries.

Advanced composition approaches can help to analyse cumulative privacy loss across multiple queries. Adaptive mechanisms might adjust noise levels based on query patterns or remaining privacy budget.

Data Perturbation Strategies

Beyond pure differential privacy, systems might employ various data perturbation strategies. Rounding intersection sizes to predetermined intervals reduces the precision of information revealed. Adding random elements to sets before computing intersections introduces uncertainty in results. Sampling from the underlying sets before computing intersections reduces the confidence of inferences.

These approaches offer practical alternatives to formal differential privacy, potentially providing better utility for specific applications. However, their privacy guarantees are generally weaker and harder to analyze rigorously.

Access Control and Segmentation

Organizational defenses might include stricter access control and data segmentation. Different users or applications might receive different levels of access to intersection size queries. High-precision results might require additional authorization or auditing. Large-scale queries might face additional restrictions compared to smaller ones.

Systems can also segment data temporally or by category, limiting the scope of any single query. This reduces the amount of information available through any particular access point, though it may complicate legitimate uses that need comprehensive data access.

Architectural Approaches

Fundamental architectural changes might provide stronger protections. For instance, systems might move away from providing exact intersection sizes, instead offering only approximate ranges or binary thresholds. Alternative protocols might reveal only derivative statistics rather than raw intersection sizes.

Some applications might support completely restructured approaches that avoid revealing intersection sizes altogether. For example, private contact tracing systems might use different cryptographic primitives or decentralised architectures that provide equivalent functionality without the same vulnerabilities.

Looking Forward

This research highlights a fundamental challenge in privacy system design: seemingly safe aggregate statistics can leak sensitive individual information when analyzed strategically over multiple observations.

References

[1] X. Guo, Y. Han, Z. Liu, D. Wang, Y. Jia, and J. Li, “Birds of a Feather Flock Together: How Set Bias Helps to Deanonymize You via Revealed Intersection Sizes,” in *USENIX Security Symposium*, 2022. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/guo

[2] A. Tiwari, “Privacy-Preserving Multi-Touch Attribution at TikTok,” 2024, *Abhishek Tiwari*. doi: 10.59350/rhrwf-tbs74.