Split, Count, and Share: A Differentially Private Set Intersection Cardinality Estimation Protocol

Michael Purcell[1] **Yang Li**[1] **Kee Siong Ng**[1]

[1]School of Computing, Australian National University, Canberra, Australian Capital Territory, Australia

Abstract

We describe a simple two-party protocol in which each party contributes a set as input. The output of the protocol is an estimate of the cardinality of the intersection of the two input sets. We show that our protocol is efficient and secure. We show that the space complexity and communication complexity are constant, the time complexity for each party is proportional to the size of their input set, and that our protocol is differentially private. We also analyze the distribution of the output of the protocol, deriving both its asymptotic distribution and finite-sample bounds on its tail probabilities. These analyses show that, when the input sets are large, our protocol produces accurate set intersection cardinality estimates. We claim that our protocol is an attractive alternative to traditional private set intersection cardinality (PSI-CA) protocols when the input sets are large, exact precision is not required, and differential privacy on its own can provide sufficient protection to the underlying sensitive data.

1 INTRODUCTION

Secure multiparty computation (SMC) protocols allow multiple parties, each of whom holds some input to a given function, to jointly evaluate that function without sharing their inputs with one another [Goldreich, 1998]. Unfortunately, SMC protocols are expensive [Damgård et al., 2010] and do not provide any protection for sensitive information that may be revealed by their output.

For example, suppose that the nature of some business is such that its customers would prefer that their association with that business be kept confidential. Further, suppose that some other organization was allowed to use an SMC protocol to find out how many customers the two business have in common. Such a protocol is known as private set intersection cardinality (PSI-CA) protocol. While a PSI-CA protocol would protect any customers who were associated with only one of the two parties, it would necessarily compromise the privacy of anyone associated with both parties.

Preventing this kind of privacy breach requires a different notion of privacy. Traditional SMC protocols can be problematic because their outputs are exact. Attackers are able to use such protocols to make inferences about the data with extremely high confidence [Dinur and Nissim, 2003]. If the output of an SMC protocol was inexact, perhaps because the exact output was perturbed with some kind of noise, then an attacker would necessarily be less confident in any inferences that they made based on those outputs. This is the idea that underpins the concept of differential privacy (DP) [Dwork et al., 2006]. DP ensures that the output of a function will not reveal whether a record was used to compute that output and provides a degree of plausible deniability to entities that contribute sensitive data to a data set.

In this paper, we will describe the *Split, Count, and Share protocol*, a differentially private alternative to PSI-CA. Our protocol relies on differential privacy to protect all elements of both parties' sets. By doing so, we sacrifice precision; our protocol only produces accurate cardinality estimates when the input sets are large.

Also, our protocol provides only a differential privacy guarantee. It does not provide any security guarantee whatsoever in the sense of traditional SMC protocols [Cramer et al., 2015, Evans et al., 2018, Lindell, 2020]. In exchange, however, we are able to dramatically reduce the time, space, and communiation complexity of our protocol compared with PSI-CA protocols.

We will describe our protocol in detail, derive parameter values sufficient to guarantee (ϵ, δ)-differential privacy, and prove rigorous bounds on the difference between the estimates produced by our protocol and the actual cardinality of the intersection of the two input sets. We will conclude with a discussion that shows that our protocol is practically useful when both parties have large input sets.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

2 RELATED WORK

Traditionally, private set intersection (PSI) has been a topic of interest primarily in the field of secure multiparty computing. As such, much of the literature concerning PSI addresses questions of what can be achieved by mutually distrustful parties who are unwilling to reveal any information about their inputs to one another. More precisely, an SMC protocol is considered secure if the participants learn no more than what they could learn in an ideal world, where a trusted curator is present to ensure all participants' inputs are kept secret from each other and a correctly computed output is returned to the participants [Goldreich, 1998].

Despite the fact that Split, Count, and Share provides no security guarantee in the sense of traditional SMC protocols, it nevertheless occupies a similar niche to existing PSI and PSI-CA protocols. So, we will focus our survey of related work on existing SMC protocols and emphasize the fact that because we provide a different kind of security guarantee, it is difficult to make direct comparisons between the performance of our protocol and the performance of those mentioned below.

2.1 EXACT PSI

A wide variety of cryptographic primitives have been proposed as components on which private set intersection (PSI) protocols can be built. Among these, Diffie-Hellman (DH) key exchange and oblivious transfer (OT) extension are the most common. Generally, DH-based protocols have lower communication complexity but greater time complexity than OT-extension based protocols, which dominate the field for larger sets. But the unique characteristics of DH-based protocols seem to increase speed for small sets PSI as shown in a recent work by Rosulek and Trieu [2021].

The state-of-the-art OT-extension based PSI protocols in the semi-honest setting are the computation time optimized protocol by Kolesnikov et al. [2016], communication time optimized protocol by Pinkas et al. [2019a] and an efficient balanced protocol by Chase and Miao [2020]. Although beyond scope of this work, Pinkas et al. [2020], Rindal and Schoppmann [2021] and Garimella et al. [2021] describe some recent efficient PSI developments in the malicious model. Some of these protocols, when analyzed under the semi-honest model, are almost as efficient as those mentioned above. Furthermore, the authors have done some thorough theoretical and experimental comparisons with selected state-of-the-art protocols to demonstrate the efficiency of their protocols.

PSI-CA protocols compute only the size of the intersection of the input sets rather than the intersection itself. Many such protocols have been proposed, including those that work by modifying an underlying PSI protocol [Freedman et al., 2004, 2016] and those that work by post-processing the output of a circuit-based PSI protocol [Pinkas et al., 2019b]. In either case, the most efficient PSI-CA protocols to date [Cristofaro et al., 2012, Freedman et al., 2016, Ion et al., 2020, Debnath et al., 2021, Trieu et al., 2022] have linear computation and communication costs in the input set sizes. There are also some recently developed efficient application-driven PSI-CA protocols [Dittmer et al., 2020, Duong et al., 2020, Trieu et al., 2020] and efficient multiparty private set intersection protocols [Chandran et al., 2021].

2.2 APPROXIMATE PSI

Dramatic efficiency gains can be had by approximating the cardinality of the intersection of two input sets rather than computing it exactly. Doing so, however, inevitably comes at the cost of decreased precision. Some early works of PSI-CA approximation are Freedman et al. [2004] and Egert et al. [2015]. Recently, several works including Dong and Loukides [2018], Sparka et al. [2018], and Hu et al. [2021] have proposed using sketches to compute intersection cardinality estimates in less than linear time.

The most efficient PSI-CA approximation protocol that we are aware of [Dong and Loukides, 2018] realizes logarithmic computation and communication time complexity in the largest possible cardinality value. Moreover, the protocol's approximation error can be tuned to adjust the accuracy and efficiency tradeoff. To illustrate the superior efficiency of this protocol for large sets, the authors compared it with an efficient exact PSI-CA protocol [Cristofaro et al., 2012] and an early approximation protocol [Egert et al., 2015]. The experimental results show that for set size 10^6, Cristofaro et al. [2012] and Egert et al. [2015] (at 1% error rate) have computation time 3507.38 and 488.48 seconds respectively, whilst the FM sketch based protocol (at 1% error rate) has computation time only 2.97 seconds. For more experimental results, see Dong and Loukides [2018].

2.3 DP PSI

As a popular privacy-enhancing technique that addresses data contributors' membership privacy concerns, differential privacy (DP) [Dwork et al., 2006] has a long history of application in SMC protocols. The early work of Beimel et al. [2008] studied the feasibility of using DP to increase the efficiency of secure function evaluation protocols. More recent work by Groce et al. [2019] also demonstrated the effectiveness of DP in reducing standard PSI running costs. The use of DP in this work is to replace bin padding, which plays a significant role in hiding the actual sizes of the bins that contain hashed elements.

Another recent combination of DP and PSI is described in Kacsmar et al. [2020], which proposed differentially private mechanisms for both PSI and PSI-CA, for the imbal-

anced database setting, where the server holds a much larger database than the client. By using homomorphic encryption, the proposed mechanisms provide a stronger protection to the data, in the sense that the server learns nothing about the client's data due to encryption and the client learns DP guaranteed set intersection (cardinality). The communication complexity of the proposed mechanisms is at optimal $O(m)$ because of the help of homomorphic encryption, where m is the smaller database size. The computation time is $O(m+n)$, which does not take into account ciphertext size expansion due to encryption.

3 SPLIT, COUNT, AND SHARE

Suppose Alice holds the set $A \subset S$ and Bob holds the set $B \subset S$. Alice would like to estimate $|A \cap B|$, Bob is willing to work with Alice to help her do so, but neither party is willing to reveal the elements of their set to the other. To simplify the subsequent analysis, we will assume that Alice and Bob are willing to share the cardinality of their sets.

Also, suppose that for $1 \leq i \leq r$ Alice and Bob can randomly partition their sets into two subsets $A_{i,0}, A_{i,1} \subset A$ and $B_{i,0}, B_{i,1} \subset B$ in such a way that if $x \in A \cap B$ then Alice and Bob will put x into the same set; i.e. Alice puts x in $A_{i,j}$ if and only if Bob puts x in $B_{i,j}$. We will call each such partitioning a *round* and assume that the splitting decisions in each round are independent of all other rounds.

Finally, after each round both parties count how many of their set elements were put into each subset. That is, for all i Alice computes $V_i = |A_{i,1}|$ and Bob computes $W_i = |B_{i,1}|$. Recall that Alice and Bob make the same splitting decisions for all $x \in |A \cap B|$. If $|A \cap B|$ is large relative to $|A|$ and $|B|$ then V_i and W_i will be strongly correlated. Conversely, if $|A \cap B|$ is small relative to $|A|$ and $|B|$ then V_i and W_i will be weakly correlated. So, the sample correlation between V_i and W_i can be used to estimate $|A \cap B|$.

Notice that if Bob publishes his vector of counts, then Alice could use that information to make inferences about Bob's set. If Bob uses a differentially private release mechanism to perturb his vector of counts, he could publish the perturbed counts without revealing which elements comprise his set B. Alice could then compute the correlation between her vector of counts and Bob's perturbed vector of counts to derive an estimate of $|A \cap B|$.

In principle, Bob could use any differentially private mechanism to perturb his counts. The binomial mechanism (see Appendix A), however, is a particularly appealing choice for this application. Notice that Bob's unperturbed counts will be binomially distributed. If Bob uses the binomial mechanism to perturb his vector of counts, then his perturbed counts will be binomially distributed as well. Furthermore, Bob can generate his perturbed counts by simply augmenting his set with an appropriate number of dummy elements.

3.1 DESCRIPTION

A more precise description of the Split, Count, and Share protocol is as follows (see Section 3.5 for a derivation of the value of $n_{\epsilon,\delta}(r)$):

(Negotiate) Alice and Bob agree on:

1. a number of rounds to perform $r \in \mathbb{N}$,
2. differential privacy parameters (ϵ, δ),
3. r independent random oracles $\{\mathcal{E}_i\}_{i=1}^r$ where we have $\mathcal{E}_i : \mathbb{Z} \to \{0,1\}$ for all $1 \leq i \leq r$.

Split Alice and Bob use independent random oracles to partition their sets.

1. For $1 \leq i \leq r$ and $j \in \{0,1\}$, Alice computes
$$A_{i,j} = \{x \in A \colon \mathcal{E}_i(x) = j\}.$$
2. For $1 \leq i \leq r$ and $j \in \{0,1\}$, Bob computes
$$B_{i,j} = \{x \in B \colon \mathcal{E}_i(x) = j\}.$$

Count Alice and Bob count the number of elements in each of their split-sets.

1. For $1 \leq i \leq r$, Alice computes
$$V_i = |A_{i,1}|.$$
2. For $1 \leq i \leq r$, Bob computes
$$W_i = |B_{i,1}| + \text{Binomial}(n_{\epsilon,\delta}(r), 1/2).$$

Share Bob shares his (perturbed) counts with Alice.

1. Bob sends $\{W_i\}_{i=1}^r$ to Alice.

(Estimate) Alice estimates $|A \cap B|$.

1. Alice computes
$$\widehat{|A \cap B|} = \frac{4}{r} \sum_{i=1}^{r} (V_i - \mu_V)(W_i - \mu_W), \quad (1)$$
where $\mu_V = |A|/2$ and $\mu_W = (|B| + n_{\epsilon,\delta}(r))/2$.

Crucially, Bob receives no output from the protocol. Alice should not share $\widehat{|A \cap B|}$ with Bob. Because Bob knows the values of his perturbed counts that Alice used to compute $\widehat{|A \cap B|}$, he could use the value of Alice's estimate to make inferences about her set.

If Bob wants to estimate $|A \cap B|$, then Alice should instead use the differentially private mechanism to perturb her counts and send those perturbed counts to Bob. Bob can then use his (unperturbed) vector of counts and Alice's perturbed vector of counts to compute his estimate. This is conceptually equivalent to Alice and Bob running the protocol a second time, but with their roles reversed.

3.2 MOTIVATION

To motivate our choice of estimator as described by Equation (1), we first need to establish some notation. Notice that for $1 \leq i \leq r$ we have

$$V_i = X_i + Z_i$$
$$W_i = Y_i + Z_i + N_i$$

where X_i, Y_i, Z_i, and N_i are independent binomial random variables with $X_i \sim \mathcal{B}(|A| - |A \cap B|, 1/2)$, $Y_i \sim \mathcal{B}(|B| - |A \cap B|, 1/2)$, $Z_i \sim \mathcal{B}(|A \cap B|, 1/2)$, and $N_i \sim \mathcal{B}(n_{\epsilon,\delta}(r), 1/2)$.

So, if $\mu_V = \mathbf{E}[V_i]$ and $\mu_W = \mathbf{E}[W_i]$ then

$$\mu_V = |A|/2$$
$$\mu_W = (|B| + n_{\epsilon,\delta}(r))/2.$$

Similarly, if $\sigma_V^2 = \text{Var}(V_i)$ and $\sigma_W^2 = \text{Var}(W_i)$ then

$$\sigma_V^2 = |A|/4$$
$$\sigma_W^2 = (|B| + n_{\epsilon,\delta}(r))/4.$$

Notice that μ_V, μ_W, σ_V^2, and σ_W^2 are defined in terms of the known quantities $|A|$, $|B|$, and $n_{\epsilon,\delta}(r)$.

If we let $\sigma_{VW} = \text{Cov}(V_i, W_i)$, then we have

$$\sigma_{VW} = |A \cap B|/4. \tag{2}$$

Because σ_{VW} depends on the unknown quantity $|A \cap B|$, we cannot use it directly. We can, however, estimate σ_{VW} via the sample covariance $\hat{\sigma}_{VW}$ where

$$\hat{\sigma}_{VW} = \frac{1}{r}\sum_{i=1}^{r}(V_i - \mu_V)(W_i - \mu_W). \tag{3}$$

Together, (2) and (3) suggest that $4\hat{\sigma}_{VW}$ is a reasonable estimator of $|A \cap B|$.

3.3 ADDITIONAL NOTATION

In what follows, it will be convenient to work with the correlations rather than covariances. Observe that if we let $\rho_{\epsilon,\delta}(r)$ be the correlation between V_i and W_i, then we have

$$\rho_{\epsilon,\delta}(r) = \frac{\sigma_{VW}}{\sigma_V \sigma_W} = \frac{|A \cap B|}{\sqrt{|A|\,(|B| + n_{\epsilon,\delta}(r))}}.$$

Because $\rho_{\epsilon,\delta}(r)$ depends on the unknown quantity $|A \cap B|$, we cannot use it directly. We can, however, estimate $\rho_{\epsilon,\delta}(r)$ via the sample correlation $\hat{\rho}$. If we let $\tilde{V}_i = (V_i - \mu_V)/\sigma_V$ and $\tilde{W}_i = (W_i - \mu_W)/\sigma_W$ then we have

$$\mathbf{E}\left[\tilde{V}_i \tilde{W}_i\right] = \rho_{\epsilon,\delta}(r),$$
$$\text{Var}\left(\tilde{V}_i \tilde{W}_i\right) = 1 + \frac{|A \cap B|^2 - 2|A \cap B|}{|A|\,(|B| + n_{\epsilon,\delta}(r))}.$$

and

$$\hat{\rho} = \frac{\hat{\sigma}_{VW}}{\sigma_V \sigma_W} = \frac{1}{r}\sum_{i=1}^{r}\tilde{V}_i \tilde{W}_i. \tag{4}$$

3.4 COMPLEXITY

To carry out the Split, Count, and Share protocol, Bob must compute $\mathcal{E}_i(b)$ for all $1 \leq i \leq r$ and $b \in B$. He must then generate r binomial random variables to perturb each element in his count vector. So, the total time complexity of the protocol for Bob is $O(r|B|)$. Alice must compute $\mathcal{E}_i(a)$ for all $1 \leq i \leq r$ and $a \in A$. She must then compute the correlation between the two vectors of counts. So, the time complexity of the protocol for Alice is $O(r|A|)$.

Observe that the Split, Count, and Share protocol is a streaming protocol. That is, Alice and Bob do not need to store the outputs of $\mathcal{E}_i$. Indeed, they do not even need to store the elements of their sets. Instead, they can each maintain a set of r accumulators. Alice can take a single pass through her set, incrementing her ith accumulator whenever $\mathcal{E}_i(a) = 1$. After she does so, Alice's ith accumulator will contain the value of $|A_{i,1}|$. Similarly, Bob can take a single pass through his set, incrementing his ith accumulator whenever $\mathcal{E}_i(b) = 1$. After he does so, Bob's ith accumulator will contain the value of $|B_{i,1}|$.

As such, the space complexity of the protocol is determined by the space required to store the two vectors of counts. If we assume that both parties will use a sixty-four bit integer to store each count, then the space complexity of the protocol is $O(r)$. If A or B is small, then this complexity can be reduced by using fewer than sixty-four bits for each counter. In this case, the complexity of the protocol is $O(r \log_2(|A|) + r \log_2(|B|))$.

The communication complexity of the protocol is determined by the amount of data that Bob must send to Alice when he sends her his vector of perturbed counts. As such, the communication complexity of the protocol is $O(r)$.

3.5 SECURITY

The security of the Split, Count, and Share protocol is entirely dependent on the noise that Bob adds to his counts before sharing them with Alice. As mentioned above, Bob will use the binomial mechanism to perturb his vector of counts. The binomial mechanism is characterized by two parameters, n and p. We will restrict our attention to the case where $p = 1/2$. We will let $n_{\epsilon,\delta}(r)$ be the smallest value of n that provides (ϵ, δ)-differential privacy for the r-round version of the Split, Count, and Share protocol.

The privacy guarantees provided by many differentially private release mechanisms depend on the sensitivity of the input query. Precisely how this sensitivity is measured depends on the release mechanism. As discussed in Agarwal et al. [2018], the privacy guarantee provided by the binomial mechanism depends on three sensitivity parameters, Δ_1, Δ_2, and Δ_∞. The precise nature of this dependence is described by the following Lemma.

Lemma 1. *If f_r is the function that computes the vector of counts for the r-round Split, Count, and Share protocol, that is $f_r(B) = (|B_{1,1}|, |B_{2,1}|, \ldots, |B_{r,1}|)$, then $\Delta_1 f_r = r$, $\Delta_2 f_r = \sqrt{r}$, and $\Delta_\infty f_r = 1$.*

Proof. We have $f_r = (f_{r,1}, f_{r,2}, \ldots, f_{r,r})$ where the coordinate functions $\{f_{r,i}\}_{i=1}^r$ are independent counting queries. Because $f_{r,i}$ is real valued, $\Delta_p f_{r,i} = \Delta f_{r,i}$ for all p. Furthermore because $f_{r,i}$ is a counting query we have $\Delta f_{r,i} = 1$ for all $1 \leq i \leq r$. Therefore

$$\begin{aligned}
\Delta_1 f_r &= \sum_{i=1}^{r} |\Delta f_{r,i}| = r, \\
\Delta_2 f_r &= \left(\sum_{i=1}^{r} (\Delta f_{r,i})^2 \right)^{1/2} = \sqrt{r}, \\
\Delta_\infty f_r &= \max_{1 \leq i \leq r} |\Delta f_r| = 1.
\end{aligned}$$

□

Armed with the values of the relevant sensitivity parameters, we can use the following theorem to determine the value of $n_{\epsilon,\delta}(r)$ required to ensure that the binomial mechanism is (ϵ, δ)-differentially private.

Theorem 2. *Suppose that f_r is the function that computes counts for r rounds of the Split, Count, and Share protocol and that $\delta > 0$. Let*

$$\begin{aligned}
\phi_\delta(r) &= \sqrt{8r \log\left(\frac{1.25}{\delta}\right)} \\
\psi_{\delta,1}(r) &= \frac{4r}{3(1-\delta/10)} \\
\psi_{\delta,2}(r) &= \frac{10\sqrt{r \log(10/\delta)}}{(1-\delta/10)} \\
\psi_{\delta,\infty}(r) &= \frac{8}{3}\left(\log\left(\frac{1.25}{\delta}\right) + \log\left(\frac{20r}{\delta}\right)\log\left(\frac{10}{\delta}\right)\right).
\end{aligned}$$

Furthermore let $\psi_\delta(r) = \psi_{\delta,1}(r) + \psi_{\delta,2}(r) + \psi_{\delta,\infty}(r)$ and

$$n'_{\epsilon,\delta}(r) = \left(\frac{\phi_\delta(r) + \sqrt{\phi_\delta(r)^2 + 4\psi_\delta(r)\epsilon}}{2\epsilon} \right)^2.$$

If we have

$$n_{\epsilon,\delta}(r) \geq \max\left(n'_{\epsilon,\delta}(r),\ 92\log\left(\frac{10r}{\delta}\right),\ 8 \right) \tag{5}$$

then the mechanism used to compute Bob's counts for the r-round Split, Count, and Share protocol, that is $\mathcal{M}(B, f_r(\cdot); n_{\epsilon,\delta}(r))$, is (ϵ, δ)-differentially private.

Proof. If $\Delta_1 f_r$, $\Delta_2 f_r$, and $\Delta_\infty f_r$ are as in Lemma 1, then Corollary 12 (see Appendix A) implies the result. □

Observe that the value of $n_{\epsilon,\delta}(r)$ grows with r. That is, as the number of rounds r increases, so too does the amount of noise required to ensure a given level of differential privacy. Conceptually, this is because during each round, the noisy count that Bob shares with Alice reveals some information about his set. Increasing the amount of noise that Bob adds during each round ensures that the total amount of information that he reveals to Alice is limited. The following theorem describes the growth rate of $n_{\epsilon,\delta}(r)$.

Theorem 3. *If δ, $\phi_\delta(r)$, $\psi_\delta(r)$ are as in Theorem 2 and $n_{\epsilon,\delta}(r)$ is the smallest value that satisfies* (5)*, then for all $\epsilon > 0$ we have*

$$\lim_{r\to\infty} \frac{n_{\epsilon,\delta}(r)}{r} = C_{\epsilon,\delta}$$

where

$$C_{\epsilon,\delta} = \left(\frac{\phi_\delta(1) + \sqrt{\phi_\delta(1)^2 + \frac{16\epsilon}{3(1-\delta/10)}}}{2\epsilon} \right)^2.$$

Proof. Notice that we have

$$\begin{aligned}
\lim_{r\to\infty} \frac{\phi_\delta(r)}{\sqrt{r}} &= \sqrt{8\log\left(\frac{1.25}{\delta}\right)} \\
\lim_{r\to\infty} \frac{\psi_\delta(r)}{r} &= \frac{4}{3(1-\delta/10)}.
\end{aligned}$$

Therefore the result follows from Theorem 2. □

3.6 UTILITY

Having determined the amount of noise that Bob needs to introduce to ensure that the r-round Split, Count, and Share protocol is (ϵ, δ)-differentially private, it is natural to ask how accurately Alice can estimate $|A \cap B|$ using only her vector of counts and Bob's vector of perturbed counts. This will depend on a variety of factors including: the values of ϵ and δ, the number of rounds performed r, the size of Alice's set $|A|$, and the size of Bob's set $|B|$.

Broadly speaking, Alice's accuracy improves as r, ϵ, and δ increase. Crucially, while increasing $|A|$ and $|B|$ worsens Alice's absolute accuracy (i.e. doing so increases the absolute magnitude of her approximation errors), it improves her relative accuracy. So, if Alice is interested in the relative magnitude of her approximation errors, then the Split, Count, and Share protocol will provide better utility as the input sets A and B get larger.

At first glance, Equation (4) appears to suggests that we can simply invoke the Strong Law of Large Numbers to analyze the performance of $\hat{\rho}$ as an estimator of $\rho_{\epsilon,\delta}(r)$. Unfortunately, because $n_{\epsilon,\delta}(r) = O(r)$ (see Theorem 3), we have $\lim_{r\to\infty} \rho_{\epsilon,\delta}(r) = 0$. So, the situation is a bit more complicated and will require more careful analysis.

In what follows we will prove bounds on the probability that $|\widehat{A \cap B}|$ differs from $|A \cap B|$ by arbitrary threshold values. To do so, we will first characterize how well the sample correlation $\hat{\rho}$ approximates the true correlation $\rho_{\epsilon,\delta}(r)$. Because we have $|\widehat{A \cap B}| = \hat{\rho}\sqrt{|A|(|B| + n_{\epsilon,\delta}(r))}$ and $|A \cap B| = \rho\sqrt{|A|(|B| + n_{\epsilon,\delta}(r))}$, we can "lift" those bounds to describe how well $|\widehat{A \cap B}|$ approximates $|A \cap B|$. Our first result shows that the distribution of Alice's errors is approximately normally distributed with mean zero and variance $\nu_{\epsilon,\delta}(r)/r$.

Theorem 4. *Let Φ be the cumulative distribution function (CDF) of a standard normal random variable, that is $\Phi(x) = \mathbf{P}\{\mathcal{N}(0,1) \leq x\}$. For all $t \geq 0$ we have*

$$\lim_{r\to\infty} \mathbf{P}\left\{|\hat{\rho} - \rho_{\epsilon,\delta}(r)| \geq t\sqrt{\frac{\nu_{\epsilon,\delta}(r)}{r}}\right\} = 2\Phi(-t),$$

where

$$\nu_{\epsilon,\delta}(r) = 1 + \frac{|A \cap B|^2 - 2|A \cap B|}{|A|(|B| + n_{\epsilon,\delta}(r))}.$$

Proof. Notice that if $S_r = \sum_{i=1}^{r} \tilde{V}_i\tilde{W}_i$, then we have

$$\mathbf{P}\left\{|\hat{\rho} - \rho_{\epsilon,\delta}(r)| \geq t\sqrt{\frac{\nu_{\epsilon,\delta}(r)}{r}}\right\} = \mathbf{P}\left\{\left|\frac{S_r - r\rho_{\epsilon,\delta}(r)}{\sqrt{r\nu_{\epsilon,\delta}(r)}}\right| \geq t\right\}.$$

Recall, $\mathbf{E}\left[\tilde{V}_i\tilde{W}_i\right] = \rho_{\epsilon,\delta}(r)$ and $\text{Var}\left(\tilde{V}_i\tilde{W}_i\right) = \nu_{\epsilon,\delta}(r)$. Therefore, because $\tilde{V}_i$ and $\tilde{W}_i$ are bounded, we have $\mathbf{E}\left[\left|\tilde{V}_i\tilde{W}_i - \rho_{\epsilon,\delta}(r)\right|^3\right] < \infty$ and the result follows from the Berry-Esseen Theorem (Theorem 13). □

Notice that the deviations described by Theorem 4 are absolute errors rather than relative errors. This is significant because $\lim_{r\to\infty} \rho_{\epsilon,\delta}(r) = 0$. So, when we multiply by $\sqrt{|A|(|B| + n_{\epsilon,\delta}(r))}$, we will find that the accuracy of Alice's cardinality estimates will depend on $|A|$ and $|B|$.

Corollary 5. *Let $|\widehat{A \cap B}| = 4\hat{\sigma}_{VW}$. For all $t \geq 0$ we have*

$$\lim_{r\to\infty} \mathbf{P}\left\{\left||\widehat{A \cap B}| - |A \cap B|\right| \geq t\sqrt{|A|C_{\epsilon,\delta}}\right\} = 2\Phi(-t),$$

where $\Phi(x) = \mathbf{P}\{\mathcal{N}(0,1) \leq x\}$ is the cumulative distribution function of a standard normal random variable.

Proof. Observe that

$$|\hat{\rho} - \rho_{\epsilon,\delta}(r)| = \left|\frac{4\hat{\sigma}_{VW} - |A \cap B|}{4\sigma_V\sigma_W}\right|.$$

So, we have

$$\mathbf{P}\left\{|\hat{\rho} - \rho_{\epsilon,\delta}(r)| \geq t\sqrt{\frac{\nu_{\epsilon,\delta}(r)}{r}}\right\} = \mathbf{P}\left\{\left|\frac{|\widehat{A \cap B}| - |A \cap B|}{4\sigma_V\sigma_W}\right| \geq t\sqrt{\frac{\nu_{\epsilon,\delta}(r)}{r}}\right\}.$$

Furthermore, because $\lim_{r\to\infty} \nu_{\epsilon,\delta}(r) = 1$, Theorem 3 implies that

$$\lim_{r\to\infty} 4t\sigma_V\sigma_W\sqrt{\frac{\nu_{\epsilon,\delta}(r)}{r}} = t\sqrt{|A|C_{\epsilon,\delta}}.$$

Therefore, Theorem 4 implies the result. □

Corollary 5 shows that for large r, Alice's absolute errors will generally be on the order of the square root of the size of her set A. As such, the larger A is, the smaller relative errors will be. This fact is the basis for our claim that the Split, Count, and Share protocol is particularly well suited for settings where both Alice and Bob's input sets are large. Notice also that Corollary 5 implies that there is a law of diminishing returns as r increases. As we approach the asymptotic regime, the marginal cost for increasing r remains constant while the marginal utility gain for doing so steadily decays.

Our last result is a concentration inequality that shows that the sample correlation will be close to the true correlation with high probability. This result is a finite-sample bound. That is, it is a statement that applies for all values of r rather than a statement that applies only in the limit as r diverges.

Theorem 6. *For all $t \geq 0$ we have*

$$\mathbf{P}\left\{|\hat{\rho} - \rho_{\epsilon,\delta}(r)| \geq t\right\} \leq 2\exp\left(\frac{-rt^2}{6 + 4t}\right).$$

Proof. This follows from Bernstein's Inequality, Khintchine's Inequality, and the Legendre duplication formula. See Appendix B for details. In particular, this result is a direct consequence of Theorem 14. □

As with Theorem 4, we can "lift" Theorem 6 to make precise statements about the accuracy of Alice's cardinality estimates. Because this operation is so similar to that demonstrated in the preceding discussion, and because the formulae involved are significantly more complicated, we state the following result without proof.

Corollary 7. *For all $\gamma \geq 0$ we have*

$$\lim_{r\to\infty} \mathbf{P}\left\{\left||\widehat{A \cap B}| - |A \cap B|\right| \geq \gamma\sqrt{|A|}\right\} \leq U_{\epsilon,\delta}(\gamma),$$

where

$$U_{\epsilon,\delta}(\gamma) = 2\exp\left(\frac{-\gamma^2}{6C_{\epsilon,\delta}}\right).$$

4 DISCUSSION

To use the Split, Count, and Share protocol in practice, Alice and Bob must agree on a suitable set of parameter values. We can subdivide these parameters into two types. The first describe how Alice and Bob will make their splitting decisions in each round of the protocol. The second describe the privacy guarantees that the protocol will provide.

To reason about the first type of parameter, recall that Alice and Bob need to agree on a collection of r independent random oracles to use as splitting functions. Cryptographically secure hash functions are a natural class of functions to use to implement these random oracles in practice. Furthermore, if $\mathcal{E}$ is a hash function with a digest length of r bits, then Alice and Bob can compute all of the splitting decisions for a given input with a single function evaluation. So, setting $\mathcal{E} = \text{SHA3-512}$ and $r = 512$ may be reasonable defaults.

To reason about the second type of parameters, recall that only Alice will receive the output of the protocol. So, Bob is more concerned with privacy than utility. He needs the values of the privacy parameters (ϵ, δ) to be small enough to guarantee sufficient protection for his data. Alice is more concerned with utility than privacy. She needs the values of the privacy parameters (ϵ, δ) to be large enough to guarantee that her estimates will be sufficiently accurate.

As a general rule [McSherry, 2017], δ should be chosen to be negligible relative to $1/|B|$. As such, a cryptographically small value such as $\delta = 2^{-128}$ may be a reasonable default. It is less clear what a reasonable default value for ϵ might be. Appropriate values for ϵ depend on how much privacy loss Bob is willing to tolerate and how often he expects to participate in the protocol. Because the utility guarantees for the protocol are given in terms of the relative standard error of Alice's estimates, the choice of ϵ also depends on the size of her set.

Figure 1 depicts the relationship between the value of ϵ and the approximate standard error of Alice's cardinality estimates. Here, we let $r = 512$, $\delta = 2^{-128}$, $|A| = |B|$, and $|A \cap B| = |A|/2$. Each curve in the graph describes this relationship for a different value of $|A|$. Observe that the standard error decreases as ϵ and $|A|$ increase. Furthermore, notice that we have $\lim_{|A|\to\infty} 4\sigma_V\sigma_W/|A| = 1$ and $\lim_{|A|\to\infty} \nu_{\epsilon,\delta}(r) \approx 1.25$. So, as $|A|$ increases, the relative standard error converges to $\sqrt{1.25/512} \approx 0.05$.

So, we see that if both of their sets are large, then Alice and Bob can use the Split, Count, and Share protocol to compute fairly accurate estimates of how many elements their sets have in common. For relatively small values of ϵ, say $\epsilon = 0.05$, Alice can compute cardinality estimates that are accurate to within $0.1 \cdot |A|$ approximately 96% of the time. Doing so requires only that both parties hash their set elements and then update each of $r = 512$ counters after computing each hash.

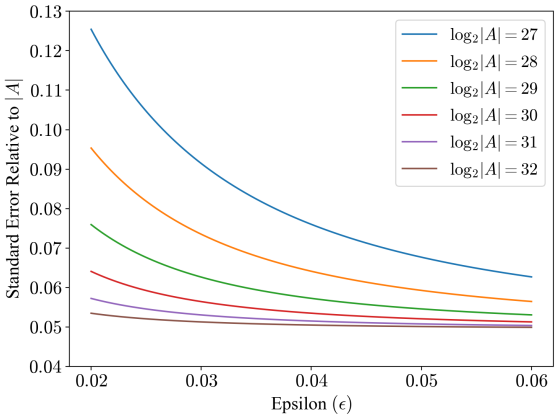

Figure 1: Standard Error Curves.

4.1 DISCRETE GAUSSIAN NOISE

One idiosyncrasy of the Split, Count, and Share protocol as described above is the distribution of the noise that Bob uses to ensure differential privacy. As described, Bob uses the binomial mechanism to perturb his counts. This allowed us to compute rigorous finite-sample tail bounds on the estimate errors that the protocol will produce. This, in turn, allows us to make strong statements about the kinds of utility guarantees the protocol can provide. There are, however, alternative noise distributions that Bob could use.

In particular, he could use the discrete Gaussian mechanism Canonne et al. [2020]. That is, Bob could use a release mechanism that perturbs his counts with noise drawn from a discrete Gaussian distribution. This is appealing because the discrete Gaussian mechanism has been shown to outperform the binomial mechanism. In this case, by "outperform" we mean that if parameters are chosen such that the two mechanisms provide equivalent privacy guarantees, then the variance of the noise produced by the discrete mechanism is smaller than that produced by the binomial mechanism.

Indeed, Figure 2 depicts the performance of two versions of the Split, Count, and Share mechanism, one using the discrete Gaussian mechanism and one using the binomial mechanism. As in Figure 1, we let $r = 512$, $\delta = 2^{-128}$, $|A| = |B|$, and $|A \cap B| = |A|/2$. Each curve in the graph depicts the absolute difference, averaged over one thousand experiments, between the set cardinality estimate produced by the protocol and the true value of $|A \cap B|$. For each value of $|A|$ we have a curve describing the performance of the binomial mechanism, depicted with a solid line, and a curve describing the performance of the discrete Gaussian mechanism, depicted with a dashed line. In all cases we see that the discrete Gaussian mechanism outperforms the binomial mechanism. This difference is fairly small, however, and both versions exhibit similar qualitative behavior.

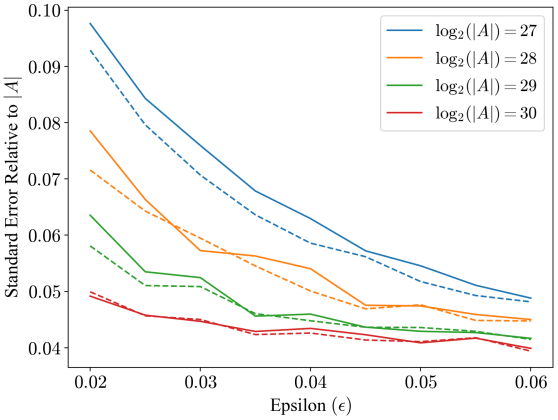

Figure 2: Binomial vs Discrete Gaussian Noise.

Unfortunately, the output distribution of Bob's perturbed counts when he uses the discrete Gaussian mechanism is complicated. As such, the utility guarantees that we provide above no longer apply. Given the empirical evidence presented here, however, it might be reasonable to assume that these utility guarantees are conservative. So, if Alice and Bob are interested in maximizing utility subject to some preexisting privacy constraints, then they may be best served by using the version of the protocol in which Bob uses the discrete Gaussian mechanism to perturb his counts.

5 FUTURE WORK

The results described above suggest several natural avenues for future research. Most obvious among these is to explore alternative protocols that could provide better utility than the Split, Count, and Share protocol. The naïve approach of combining a traditional SMC protocol to compute the exact set intersection cardinality and a simple differentially private release mechanism is one such alternative that is optimal with respect to the utility that it provides. Unfortunately, traditional PSI-CA protocols are expensive and thus impractical to use when Alice and Bob's sets are large. So, any alternative protocols must be efficient in terms of time and space complexity to accommodate large inputs.

Notice that the cosine of the angle between the characteristic vectors of the sets A and B is equal to the correlation between Alice and Bob's vectors of counts in the Split, Count, and Share protocol. Furthermore, in Charikar [2002], the author describes how SimHash can be used to estimate the size of the intersection of two sets by estimating the angle between their characteristic vectors. So, SimHash could be used as an explicit basis for other differentially private set intersection cardinality estimation protocols. Perhaps other set similarity estimation algorithms, e.g. MinHash [Broder, 1997], could be used in similar ways as well.

A DIFFERENTIAL PRIVACY

Definition 8 (Dwork et al. [2006])**.** A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ϵ, δ)-differentially private if for all $S \subset \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ with $\|x - y\|_1 \leq 1$ we have

$$\mathbf{P}\{\mathcal{M}(x) \in S\} \leq e^{\epsilon}\, \mathbf{P}\{\mathcal{M}(y) \in S\} + \delta.$$

If $\delta = 0$ we say that $\mathcal{M}$ is ϵ-differentially private.

Definition 9 (Dwork et al. [2006])**.** Let $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^d$ be an arbitrary d-dimensional function. The ℓ_p sensitivity of f is $\Delta_p f = \max\{\|f(x) - f(y)\|_p \colon x, y \in \mathbb{N}^{|\mathcal{X}|}, \|x - y\|_1\}$.

Definition 10 (Agarwal et al. [2018])**.** Suppose that we have $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{Z}^r$. The binomial mechanism, denoted $\mathcal{M}_B(x, f(\cdot); n)$, adds noise $N_i \sim \text{Binomial}(n, 1/2)$ to each of the r components of the output of f. That is,

$$\mathcal{M}_B(x, f(\cdot); n) = f(x) + (N_1, N_2, \ldots, N_r),$$

where $\{N_i\}_{i=1}^r$ are independent random variables.

Theorem 11 (Agarwal et al. [2018])**.** *Let $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{Z}^r$ and $\delta > 0$. Let $\Delta_i = \Delta_i f$ for $i \in \{1, 2, \infty\}$ and let*

$$\phi = \Delta_2 \cdot \sqrt{8 \log\left(\frac{1.25}{\delta}\right)}$$

$$\psi_1 = \Delta_1 \cdot \frac{4}{3(1 - \delta/10)}$$

$$\psi_2 = \Delta_2 \cdot \frac{10\sqrt{\log(10/\delta)}}{(1 - \delta/10)}$$

$$\psi_\infty = \Delta_\infty \cdot \frac{8}{3}\left(\log\left(\frac{1.25}{\delta}\right) + \log\left(\frac{20r}{\delta}\right)\log\left(\frac{10}{\delta}\right)\right).$$

Finally, let $\psi = \psi_1 + \psi_2 + \psi_\infty$.

If $n \geq \max(92\log(10r/\delta), 8\Delta_\infty)$ then the binomial mechanism $\mathcal{M}_B(x, f(\cdot); n)$ is (ϵ, δ)-differentially private for

$$\epsilon \geq \frac{\phi}{\sqrt{n}} + \frac{\psi}{n}.$$

Proof. See Appendix B of Agarwal et al. [2018]. □

Corollary 12 (Agarwal et al. [2018])**.** *Suppose that f, δ, ϕ, and ψ are as in Theorem 11 and $\epsilon > 0$. If*

$$n' = \left(\frac{\phi + \sqrt{\phi^2 + 4\psi\epsilon}}{2\epsilon}\right)^2$$

then for every

$$n \geq \max\left(n',\ 92\log\left(\frac{10r}{\delta}\right),\ 8\Delta_\infty(f)\right) \tag{6}$$

the binomial mechanism $\mathcal{M}_B(x, f(\cdot); n)$ is (ϵ, δ)-differentially private.

Proof. This follows from an application of the quadratic formula to determine the smallest value of n required to ensure that (6) holds. □

B CONCENTRATION INEQUALITIES

Theorem 13 (Berry-Esseen Theorem)**.** *Let $\{X_i\}_{i=1}^r$ be a sequence of independent and identically distributed random variables with $\mathbf{E}[X_i] = 0$, $Var(X_i) = 1$, and $\mathbf{E}[|X|^3] = \beta_3 < \infty$. Let $S_r = \sum_{i=1}^r X_i$ and let F_r denote the cumulative distribution function of $S_r/\sqrt{r}$. That is $F_r(x) = \mathbf{P}\{S_r \leq x\sqrt{r}\}$. Let Φ denote the cumulative distribution function of a standard normal random variable. That is, $\Phi(x) = \mathbf{P}\{\mathcal{N}(0,1) \leq x\}$. Then there exists a finite positive absolute constant C_0 such that*

$$\sup_{x\in\mathbb{R}} |F_r(x) - \Phi(x)| \leq \frac{C_0\beta_3}{\sqrt{r}}.$$

Proof. See Berry [1941], Shevtsova [2011]. □

Theorem 14. *Suppose that $\{(X_i, Y_i)\}_{i=1}^n$ are independent random vectors with $X_i \sim$ Binomial$(n_X, 1/2)$ and $Y_i \sim$ Binomial$(n_Y, 1/2)$ for all $1 \leq i \leq n$. If we let $\tilde{X}_i = (X_i - \mu_X)/\sigma_X$ and $\tilde{Y}_i = (Y_i - \mu_Y)/\sigma_Y$ then*

$$\mathbf{P}\left\{\frac{1}{n}\left|\sum_{i=1}^n \tilde{X}_i\tilde{Y}_i - \mathbf{E}[\tilde{X}_i\tilde{Y}_i]\right| \geq t\right\} \leq 2\exp\left(\frac{-nt^2}{6+4t}\right).$$

Proof. Bernstein's Inequality implies that it suffices to show that for all $p \geq 2$ we have

$$\mathbf{E}\left[\left|\tilde{X}_i\tilde{Y}_i\right|^p\right] \leq \frac{3}{2}p!2^{p-2}.$$

To that end, observe that for all $1 \leq i \leq n$, if $p \geq 2$ we have

$$\begin{aligned}
\mathbf{E}\left[\left|\tilde{X}_i\tilde{Y}_i\right|^p\right] &\leq \sqrt{\mathbf{E}\left[\tilde{X}_i^{2p}\right]}\sqrt{\mathbf{E}\left[\tilde{Y}_i^{2p}\right]} && (7)\\
&\leq \frac{2^p}{\sqrt{\pi}}\Gamma\left(p+1/2\right) && (8)\\
&= \frac{2^p}{\sqrt{\pi}}\left(\frac{\sqrt{\pi}\Gamma(2p+1)}{2^{2p}\Gamma(p+1)}\right) && (9)\\
&= \frac{(2p)!}{2^p p!},
\end{aligned}$$

where (7) is an application of the Cauchy-Schwartz inequality, (8) is an application of Khintchine's inequality, and (9) is an application of the Legendre duplication formula.

It remains to show that for all $p \geq 2$ we have

$$\frac{(2p)!}{2^p p!} \leq \frac{3}{2}p!2^{p-2}.$$

Let $f(p) = \frac{(2p)!}{2^p p!}$ and $g(p) = \frac{3}{2}p!2^{p-2}$ and observe that we have $f(p+1) = (2p+1)f(p) < (2p+2)f(p)$ and $g(p+1) = (2p+2)g(p)$. Notice that $f(2) = g(2) = 3$. Therefore, if $f(q) \leq g(q)$ for all $q \in \{2, 3, \ldots, p\}$, then $f(p+1) \leq g(p+1)$ and the result follows by induction. □

References

Naman Agarwal, Ananda Theertha Suresh, Felix Yu, Sanjiv Kumar, and Brendan McMahan. cpSGD: Communication-efficient and differentially-private distributed SGD. In *Neural Information Processing Systems*, 2018. URL https://arxiv.org/pdf/1805.10559.pdf.

Amos Beimel, Kobbi Nissim, and Eran Omri. Distributed private data analysis: Simultaneously solving how and what. In *Annual International Cryptology Conference*, pages 451–468. Springer, 2008.

Andrew C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941. ISSN 00029947. URL http://www.jstor.org/stable/1990053.

A.Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29, 1997. doi: 10.1109/SEQUEN.1997.666900.

Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15676–15688. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b53b3a3d6ab90ce0268229151c9bde11-Paper.pdf.

Nishanth Chandran, Nishka Dasgupta, Divya Gupta, Sai Lakshmi Bhavana Obbattu, Sruthi Sekar, and Akash Shah. Efficient Linear Multiparty PSI and Extensions to Circuit/Quorum PSI. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 1182–1204, 2021.

Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, page 380388, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134959. doi: 10.1145/509907.509965. URL https://doi.org/10.1145/509907.509965.

Melissa Chase and Peihan Miao. Private set intersection in the internet setting from lightweight oblivious PRF. In *Annual International Cryptology Conference*, pages 34–63. Springer, 2020.

Ronald Cramer, Ivan Bjerre Damgård, et al. *Secure multiparty computation*. Cambridge University Press, 2015.

Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. Fast and private computation of cardinality of set intersection and union. In *International Conference on Cryptology and Network Security*, pages 218–231. Springer, 2012.

Ivan Damgård, Yuval Ishai, and Mikkel Krøigaard. Perfectly secure multiparty computation and the computational overhead of cryptography. In *Advances in Cryptology–EUROCRYPT 2010: 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, French Riviera, May 30–June 3, 2010. Proceedings 29*, pages 445–465. Springer, 2010.

Sumit Kumar Debnath, Pantelimon Stnic, Nibedita Kundu, and Tanmay Choudhury. Secure and efficient multiparty private set intersection cardinality. *Advances in Mathematics of Communications*, 15(2):365, 2021.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

Samuel Dittmer, Yuval Ishai, Steve Lu, Rafail Ostrovsky, Mohamed Elsabagh, Nikolaos Kiourtis, Brian Schulte, and Angelos Stavrou. Function secret sharing for psica: With applications to private contact tracing. *arXiv preprint arXiv:2012.13053*, 2020.

Changyu Dong and Grigorios Loukides. Approximating private set union/intersection cardinality with logarithmic complexity. *IACR Cryptol. ePrint Arch.*, 2018:495, 2018.

Thai Duong, Duong Hieu Phan, and Ni Trieu. Catalic: delegated PSI cardinality with applications to contact tracing. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 870–899. Springer, 2020.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

Rolf Egert, Marc Fischlin, David Gens, Sven Jacob, Matthias Senker, and Jörn Tillmanns. Privately computing set-union and set-intersection cardinality via bloom filters. In *Australasian Conference on Information Security and Privacy*, pages 413–430. Springer, 2015.

David Evans, Vladimir Kolesnikov, Mike Rosulek, et al. A pragmatic introduction to secure multi-party computation. *Foundations and Trends® in Privacy and Security*, 2(2-3): 70–246, 2018.

Michael J Freedman, Kobbi Nissim, and Benny Pinkas. Efficient private matching and set intersection. In *International conference on the theory and applications of cryptographic techniques*, pages 1–19. Springer, 2004.

Michael J Freedman, Carmit Hazay, Kobbi Nissim, and Benny Pinkas. Efficient set intersection with simulation-based security. *Journal of Cryptology*, 29(1):115–155, 2016.

Gayathri Garimella, Benny Pinkas, Mike Rosulek, Ni Trieu, and Avishay Yanai. Oblivious key-value stores and amplification for private set intersection. In *Annual International Cryptology Conference*, pages 395–425. Springer, 2021.

Oded Goldreich. Secure multi-party computation. *Manuscript. Preliminary version*, 78(110), 1998.

Adam Groce, Peter Rindal, and Mike Rosulek. Cheaper private set intersection via differentially private leakage. *Proceedings on Privacy Enhancing Technologies*, 2019 (3), 2019.

Changhui Hu, Jin Li, Zheli Liu, Xiaojie Guo, Yu Wei, Xuan Guang, Grigorios Loukides, and Changyu Dong. How to make private distributed cardinality estimation practical, and get differential privacy for free. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 965–982, 2021.

Mihaela Ion, Ben Kreuter, Ahmet Erhan Nergiz, Sarvar Patel, Shobhit Saxena, Karn Seth, Mariana Raykova, David Shanahan, and Moti Yung. On deploying secure computing: Private intersection-sum-with-cardinality. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 370–389. IEEE, 2020.

Bailey Kacsmar, Basit Khurram, Nils Lukas, Alexander Norton, Masoumeh Shafieinejad, Zhiwei Shang, Yaser Baseri, Maryam Sepehri, Simon Oya, and Florian Kerschbaum. Differentially private two-party set operations. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 390–404. IEEE, 2020.

Vladimir Kolesnikov, Ranjit Kumaresan, Mike Rosulek, and Ni Trieu. Efficient batched oblivious PRF with applications to private set intersection. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 818–829, 2016.

Yehuda Lindell. Secure multiparty computation (mpc). *Cryptology ePrint Archive*, 2020.

Frank McSherry. How many secrets do you have? `https://github.com/frankmcsherry/blog/blob/master/posts/2017-02-08.md`, February 2017. Blog Post.

Benny Pinkas, Mike Rosulek, Ni Trieu, and Avishay Yanai. Spot-light: Lightweight private set intersection from sparse ot extension. In *Annual International Cryptology Conference*, pages 401–431. Springer, 2019a.

Benny Pinkas, Thomas Schneider, Oleksandr Tkachenko, and Avishay Yanai. Efficient circuit-based psi with linear communication. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 122–153. Springer, 2019b.

Benny Pinkas, Mike Rosulek, Ni Trieu, and Avishay Yanai. PSI from PaXoS: fast, malicious private set intersection. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 739–767. Springer, 2020.

Peter Rindal and Phillipp Schoppmann. VOLE-PSI: fast OPRF and circuit-psi from vector-ole. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 901–930. Springer, 2021.

Mike Rosulek and Ni Trieu. Compact and malicious private set intersection for small sets. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 1166–1181, 2021.

Irina Shevtsova. On the absolute constants in the berry-esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*, 2011.

Hagen Sparka, Florian Tschorsch, and Björn Scheuermann. P2kmv: A privacy-preserving counting sketch for efficient and accurate set intersection cardinality estimations. *Cryptology ePrint Archive*, 2018.

Ni Trieu, Kareem Shehata, Prateek Saxena, Reza Shokri, and Dawn Song. Epione: Lightweight contact tracing with strong privacy. *arXiv preprint arXiv:2004.13293*, 2020.

Ni Trieu, Avishay Yanai, and Jiahui Gao. Multiparty private set intersection cardinality and its applications. *Cryptology ePrint Archive*, 2022.