Optimal Lower Bound for Differentially Private Multi-party Aggregation

T-H. Hubert Chan[1], Elaine Shi[2], and Dawn Song[3]

[1] The University of Hong Kong
hubert@cs.hku.hk
[2] University of Maryland, College Park
elaine@cs.umd.edu
[3] UC Berkeley
dawnsong@cs.berkeley.edu

Abstract. We consider distributed private data analysis, where n parties each holding some sensitive data wish to compute some aggregate statistics over all parties' data. We prove a tight lower bound for the private distributed summation problem. Our lower bound is strictly stronger than the prior lower-bound result by Beimel, Nissim, and Omri published in CRYPTO 2008. In particular, we show that any n-party protocol computing the sum with sparse communication graph must incur an additive error of $\Omega(\sqrt{n})$ with constant probability, in order to defend against potential coalitions of compromised users. Furthermore, we show that in the client-server communication model, where all users communicate solely with an untrusted server, the additive error must be $\Omega(\sqrt{n})$, regardless of the number of messages or rounds. Both of our lower-bounds, for the general setting and the client-to-server communication model, are strictly stronger than those of Beimel, Nissim and Omri, since we remove the assumption on the number of rounds (and also the number of messages in the client-to-server communication model). Our lower bounds generalize to the (ϵ, δ) differential privacy notion, for reasonably small values of δ.

1 Introduction

Dwork *et al.* [DMNS06] proposed (information theoretical) differential privacy, which has become a de-facto standard privacy notion in private data analysis. In this paper, we investigate the setting of *distributed private data analysis* [BNO08], in which n parties each holds some private input, and they wish to jointly compute some statistic over all parties' inputs in a way that respects each party's privacy.

In a seminal work by Beimel, Nissim, and Omri [BNO08], they demonstrate a lower bound result for distributed private data analysis. Specifically, they consider the *distributed summation* problem, namely, computing the sum of all parties' inputs. They prove that any differentially-private multi-party protocol with *a small number of rounds* and *small number of messages* must have large error.

This paper proves a strictly stronger lower bound than the result by Beimel, Nissim, and Omri [BNO08]. We show that for the distributed summation

L. Epstein and P. Ferragina (Eds.): ESA 2012, LNCS 7501, pp. 277–288, 2012.

problem, any differentially private multi-party protocol *with a sparse communication graph* must have large error, where two nodes are allowed to communicate only if they are adjacent in the communication graph. In comparison with the previous lower bound by Beimel *et al.* [BNO08], our lower bound relaxes the constraint on the small number of messages or rounds. In this sense, our lower bound is strictly stronger than that of Beimel *et al.* [BNO08].

We also consider a special setting in which only client-server communication is allowed (i.e., the communication graph is the star graph with the server at the center). Beimel *et al.* [BNO08] referred to this communication model as *local model*. In the client-server communication setting, we prove a lower bound showing that any differentially-private protocol computing the sum must have large error. This lower bound has no restriction on the number of messages or the number of rounds, and is also strictly stronger than [BNO08], who showed that in the client-server setting, any differentially-private protocol with *a small number of rounds* must have large error.

Furthermore, our lower-bound results hold for (ϵ, δ)-differential privacy where δ is reasonably small. Since ϵ-differential privacy is a special case of this with $\delta = 0$, our lower bounds are also more general than those of Beimel *et al.* who considered ϵ differential privacy.

The lower bounds proven in this paper hold for information theoretic differential privacy. By contrast, previous works have demonstrated the possibility of constructing multi-party protocols with $O(1)$ error and small message complexity in the computational differential privacy setting [DKM$^+$06, RN10, SCR$^+$11]. Therefore, our lower-bound results also imply a gap between computational and information theoretic differential privacy in the multi-party setting.

1.1 Informal Summary of Main Results

Lower Bound for the General Setting (Corollary 2). Informally, we show that any n-party protocol computing the sum, which consumes at most $\frac{1}{4}n(t+1)$ messages must incur $\Omega(\sqrt{n})$ additive error (with constant probability), in order to preserve differentially privacy against coalitions of up to t compromised users.

Lower Bound for Client-Server Model (Corollary 1). Informally, we show that in the client-server model, an aggregator would make an additive error $\Omega(\sqrt{n})$ on the sum from any n-user protocol that preserves differential privacy. This lower-bound holds regardless of the number of messages or number of rounds.

Tightness of the Lower Bounds. Both of the above lower bounds are tight in the following sense. First, for the client-server model, there exists a naive protocol, in which each user perturbs their inputs using Laplace or geometric noise with standard deviation $O(\frac{1}{\epsilon})$, and reveals their perturbed inputs to the aggregator. Such a naive protocol has additive error $O(\sqrt{n})$; so in some sense, the naive protocol is the best one can do in the client-server model.

To see why the lower bound is tight for the general multi-party setting, we combine standard techniques of secure function evaluation [CK93] and

distributed randomness [SCR+11] and state in Section 5 that there exists a protocol which requires only $O(nt)$ messages, but achieves $o(\sqrt{n})$ error.

Techniques. To prove the above-mentioned lower-bounds, we combine techniques from communication complexity and measure anti-concentration techniques used in the metric embedding literature. Our communication complexity techniques are inspired by the techniques adopted by McGregor *et al.* [MMP+10] who proved a gap between information-theoretic and computational differential privacy in the 2-party setting. The key observation is that independent inputs remain independent even after conditioning on the transcript of the protocol. This eliminates the dependence on the number of rounds of communication in the lower bound.

As argued by in [BNO08], if a party communicates with only a small number of other parties, then there must still be sufficient randomness in that party's input. Then, using anti-concentration techniques, we show that the sum of these independent random variables is either much smaller or much larger than the mean, both with constant probability, thereby giving a lower bound on the additive error. The anti-concentration techniques are inspired by the analysis of the square of the sum of independent sub-Gaussian random variables [IN07], which generalizes several Johnson-Lindenstrauss embedding constructions [DG03, Ach03]. Moreover, we generalize the techniques to prove the lower bound for (ϵ, δ)-differentially private protocols (as opposed to just ϵ-differential privacy). The challenge is that for $\delta > 0$, it is possible for some transcript to break a party's privacy and there might not be enough randomness left in its input. However, we show that for small enough δ, the probability that such a transcript is encountered is small, and hence the argument is still valid.

2 Related Work

Differential privacy [DMNS06, Dwo06, Dwo10] was traditionally studied in a setting where a trusted curator, with access to the entire database in the clear, wishes to release statistics in a way that preserves each individual's privacy. The trusted curator is responsible for introducing appropriate perturbations prior to releasing any statistic. This setting is particularly useful when a company or a government agency, in the possession of a dataset, would like to share it with the public.

In many real-world applications, however, the data is distributed among users, and users may not wish to entrust their sensitive data to a centralized party such as a cloud service provider. In these cases, we can employ distributed private data analysis – a problem proposed and studied in several recent works [MMP+10, BNO08, DKM+06, RN10, SCR+11] – where participating parties are mutually distrustful, but wish to learn some statistics over their joint datasets. In particular, the client-server communication model [BNO08, RN10, SCR+11] where all users communicate solely with an untrusted server, is especially desirable in real-world settings.

This work subsumes the distributed private data analysis setting previously studied by Beimel, Nissim, and Omri [BNO08], and improves their lower-bounds for information-theoretic differentially private multi-party protocols.

While this work focuses on lower bounds for information theoretic differential privacy, computational differential privacy is an alternative notion first formalized by Mironov *et al.* [MPRV09], aiming to protect individual user's sensitive data against polynomially-bounded adversaries. Previous works have shown the possibility of constructing protocols with $O(1)$ error and small message complexity in the computational differential privacy setting [DKM$^+$06, RN10, SCR$^+$11]. This demonstrates a gap between information theoretic and computational differential privacy in the multi-party setting. In particular, the constructions by Rastogi *et al.* [RN10] and Shi *et al.* [SCR$^+$11] require only client-server communication, and no peer-to-peer interactions.

3 Problem Definition and Assumptions

Consider a group of n parties (or nodes), indexed by the set $[n] := \{1, 2, \ldots n\}$. Each party $i \in [n]$ has private data $x_i \in \mathcal{U}$, where $\mathcal{U} := \{0, 1, 2, \ldots, \Delta\}$ for some positive integer Δ. We use the notation $\mathbf{x} := (x_1, x_2, \ldots, x_n) \in \mathcal{U}^n$ to denote the vector of all parties' data, also referred to as an *input configuration*. The n parties participate in a protocol such that at the end at least one party learns or obtains an estimate of the sum, denoted $\mathsf{sum}(\mathbf{x}) := \sum_{i \in [n]} x_i$. For a subset $S \subseteq [n]$, we denote $\mathsf{sum}(x_S) := \sum_{i \in S} x_i$.

Given a protocol Π and an input $\mathbf{x} \in \mathcal{U}^n$, we use $\Pi(\mathbf{x})$ to denote the execution of the protocol on the input. A *coalition* is a subset T of nodes that share their information with one another in the hope of learning the other parties' input. The view $\Pi(\mathbf{x})_{|T}$ of the coalition T consists of the messages, any input and private randomness viewable by the nodes in T. In contrast, we denote by $\pi(\mathbf{x})$ the transcript of the messages and use $\pi(\mathbf{x})_{|T}$ to mean the messages sent or received by nodes in T.

Trust and Attack Model. As in Beimel *et al.* [BNO08], we assume that all parties are semi-honest. A subset T of parties can form a *coalition* and share their input data, private randomness and view of the transcript with one another in order to learn the input data of other parties. Since we adopt the semi-honest model, all parties, whether within or outside the coalition, honestly use their true inputs and follow the protocol. The *data pollution* attack, where parties inflate or deflate their input values, is out of the scope of this paper. Defense against the data pollution attack can be considered as orthogonal and complementary to our work, and has been addressed by several works in the literature [PSP03].

Communication Model. Randomized *oblivious protocols* are considered in [CK93, BNO08], where the communication pattern (i.e., which node sends message to which node in which round) is independent of the input and the randomness. We relax this notion by assuming that for a protocol Π, there is a *communication graph* G_Π (independent of input and randomness) on the nodes such that only adjacent nodes can communicate with each other. For a node i, we denote by $N_\Pi(i)$ its set of neighbors in G_Π. The subscript Π is dropped when there is no risk of ambiguity. Observe that the number of messages sent in each round is

only limited by the number of edges in the communication graph, and to simply our proofs, we only assume that there is some finite upper bound on the number of rounds for all possible inputs and randomness used by the protocol.

3.1 Preliminaries

Intuitively, differential privacy against a coalition guarantees that if an individual outside the coalition changes its data, the view of the coalition in the protocol will not be affected too much. In other words, if two input configurations $\mathbf{x}$ and $\mathbf{y}$ differ only in 1 position outside the coalition, then the distribution of $\Pi(\mathbf{x})_{|T}$ is very close to that of $\Pi(\mathbf{y})_{|T}$. This intuition is formally stated in the following definition.

Definition 1 (Differential Privacy Against Coalition). *Let $\epsilon > 0$ and $0 \leq \delta < 1$. A (randomized) protocol Π preserves (ϵ, δ)-differential privacy against coalition T if for all vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{U}^n$ that differ by only 1 position corresponding to a party outside T, for all subsets S of possible views by T, $\Pr[\Pi(\mathbf{x})_{|T} \in S] \leq \exp(\epsilon) \cdot \Pr[\Pi(\mathbf{y})_{|T} \in S] + \delta$.*

A protocol Π preserves ϵ-differential privacy against a coalition if it preserves $(\epsilon, 0)$-differential privacy against the same coalition.

Two noise distributions are commonly used to perturb the data and ensure differential privacy, the Laplace distribution [DMNS06], and the Geometric distribution [GRS09]. The advantage of using the geometric distribution over the Laplace distribution is that we can keep working in the domain of integers.

Definition 2 (Geometric Distribution). *Let $\alpha > 1$. We denote by $\mathsf{Geom}(\alpha)$ the* symmetric geometric distribution *that takes integer values such that the probability mass function at k is $\frac{\alpha-1}{\alpha+1} \cdot \alpha^{-|k|}$.*

Proposition 1. *Let $\epsilon > 0$. Suppose u and v are two integers such that $|u - v| \leq \Delta$. Let r be a random variable having distribution $\mathsf{Geom}(\exp(\frac{\epsilon}{\Delta}))$. Then, for any integer k, $Pr[u + r = k] \leq \exp(\epsilon) \cdot Pr[v + r = k]$.*

The above property of Geom distribution is useful for designing differentially private mechanisms that output integer values. In our setting, changing one party's data can only affect the sum by at most Δ. Hence, it suffices to consider $\mathsf{Geom}(\alpha)$ with $\alpha = e^{\frac{\epsilon}{\Delta}}$. Observe that $\mathsf{Geom}(\alpha)$ has variance $\frac{2\alpha}{(\alpha-1)^2}$. Since $\frac{\sqrt{\alpha}}{\alpha-1} \leq \frac{1}{\ln \alpha} = \frac{\Delta}{\epsilon}$, the magnitude of the error added is $O(\frac{\Delta}{\epsilon})$.

Naive Scheme. As a warm-up exercise, we describe a Naive Scheme, where each party generates an independent $\mathsf{Geom}(e^{\frac{\epsilon}{\Delta}})$ noise, adds the noise to its data, and sends the perturbed data to one special party called an aggregator, who then computes the sum of all the noisy data. As each party adds one copy of independent noise to its data, n copies of noises would accumulate in the sum. It can be shown that the accumulated noise is $O(\frac{\Delta\sqrt{n}}{\epsilon})$ with high probability. In comparison with our lower-bound, this shows that under certain mild assumptions, if one wishes to guarantee small message complexity, the Naive Scheme is more or less the best one can do in the information theoretic differential privacy setting.

4 Lower Bound for Information-Theoretic Differential Privacy

This section proves lower-bounds for differentially private distributed summation protocols. We consider two settings, the general settings, where all nodes are allowed to interact with each other; and the client-server communication model, where all users communicate only with an untrusted server, but not among themselves.

We will prove the following main result, and then show how to extend the main theorem to the afore-mentioned two communication models.

Theorem 1 (Lower Bound for Size-t Coalitions). *Let $0 < \epsilon \leq \ln 99$ and $0 \leq \delta \leq \frac{1}{4n}$. There exists some $\eta > 0$ (depending on ϵ) such that the following holds. Suppose n parties, where party i ($i \in [n]$) has a secret bit $x_i \in \{0, 1\}$, participate in a protocol Π to estimate $\sum_{i \in [n]} x_i$. Suppose further that the protocol is (ϵ, δ)-differentially private against any coalition of size t, and there exists a subset of m parties, each of whom has at most t neighbors in the protocol's communication graph. Then, there exists some configuration of the parties' bits x_i's such that with probability at least η (over the randomness of the protocol), the additive error is at least $\Omega(\frac{\sqrt{\gamma}}{1+\gamma} \cdot \sqrt{m})$, where $\gamma = 2e^{\epsilon}$.*

Note that the assumption that $0 \leq \delta \leq \frac{1}{4n}$ is not a limitation. Typically, when we adopt (ϵ, δ) differential privacy, we wish to have $\delta = o(\frac{1}{n})$, to ensure that no individual user's sensitive data is leaked with significant probability.

The following corollaries are special cases of Theorem 1, corresponding to the client-server communication model, and the general model respectively. In both settings, our results improve upon the lower bounds by Beimel *et al.* [BNO08]. We will first show how to derive these corollaries from Theorem 1. We then present a formal proof for Theorem 1.

Corollary 1 (Lower Bound for Client-Server Communication Model). *Let $0 < \epsilon \leq \ln 99$ and $0 \leq \delta \leq \frac{1}{4n}$. Suppose n parties, each having a secret bit, participate in a protocol Π with a designated party known as the aggregator, with no peer-to-peer communication among the n parties. Suppose further that the protocol is (ϵ, δ)-differentially private against any single party (which forms a coalition on its own). Then, with constant probability (depending on ϵ), the aggregator estimates the sum of the parties' bits with additive error at least at least $\Omega(\frac{\sqrt{\gamma}}{1+\gamma} \cdot \sqrt{n})$, where $\gamma = 2e^{\epsilon}$.*

Proof. The communication graph is a star with the aggregator at the center. The protocol is also differentially private against any coalition of size 1, and there are n parties, each of which has only 1 neighbor (the aggregator). Therefore, the result follows from Theorem 1. □

Corollary 2 (Lower Bound for General Setting). *Let $0 < \epsilon \leq \ln 99$ and $0 \leq \delta \leq \frac{1}{4n}$. Suppose n parties participate in a protocol that is (ϵ, δ)-differentially private against any coalition of size t. If there are at most $\frac{1}{4}n(t+1)$ edges in the*

communication graph of the protocol, then with constant probability (depending on ϵ), the protocol estimates the sum of the parties' bits with additive error at least $\Omega(\frac{\sqrt{\gamma}}{1+\gamma} \cdot \sqrt{n})$, where $\gamma = 2e^{\epsilon}$.

Proof. Since there are at most $\frac{1}{4}n(t+1)$ edges in the communication graph, there are at least $\frac{n}{2}$ nodes with at most t neighbors (otherwise the sum of degrees over all nodes is larger than $\frac{1}{2}n(t+1)$). Hence, the result follows from Theorem 1. □

Proof Overview for Theorem 1. We fix some $\epsilon > 0$ and $0 \leq \delta \leq \frac{1}{4n}$, and consider some protocol Π that preserves (ϵ, δ)-differential privacy against any coalition of size t.

Suppose that the bits X_i's from all parties are all uniform in $\{0,1\}$ and independent. Suppose M is the subset of m parties, each of whom has at most t neighbors in the communication graph. For each $i \in M$, we consider a set $\mathcal{P}^{(i)}$ of *bad* transcripts for i, which intuitively is the set of transcripts π under which the view of party i's neighbors can compromise party i's privacy.

We consider the set $\mathcal{P} := \cup_{i \in M} \mathcal{P}^{(i)}$ of bad transcripts (which we define formally later), and show that the probability that a bad transcript is produced is at most $\frac{3}{4}$. Conditioning on a transcript $\pi \notin \mathcal{P}$, for $i \in M$, each X_i still has enough randomness, as transcript π does not break the privacy of party i. Therefore, the conditional sum $\sum_{i \in M} X_i$ still has enough variance like the sum of $m = |M|$ independent uniform $\{0,1\}$-random variables. Using anti-concentration techniques, we can show that the sum deviates above or below the mean by $\Omega(\sqrt{m})$, each with constant probability. Since the transcript determines the estimation of the final answer, we conclude that the error is $\Omega(\sqrt{m})$ with constant probability.

Notation. Suppose that each party i's bit X_i is uniform in $\{0,1\}$ and independent. We use $\mathbf{X} := (X_i : i \in [n])$ to denote the collection of the random variables. We use a probabilistic argument to show that the protocol must, for some configuration of parties' bits, make an additive error of at least $\Omega(\sqrt{m})$ on the sum with constant probability.

For convenience, given a transcript π (or a view of the transcript by certain parties) we use $\Pr[\pi]$ to mean $\Pr[\pi(\mathbf{X}) = \pi]$ and $\Pr[\cdot|\pi]$ to mean $\Pr[\cdot|\pi(\mathbf{X}) = \pi]$; given a collection $\mathcal{P}$ of transcripts (or collection of views), we use $\Pr[\mathcal{P}]$ to mean $\Pr[\pi(\mathbf{X}) \in \mathcal{P}]$.

We can assume that the estimate made by the protocol is a deterministic function on the whole transcript of messages, because without loss of generality we can assume that the last message sent in the protocol is the estimate of the sum.

We will define some event $\mathcal{E}$ where the protocol makes a large additive error.

Bad Transcripts. Denote $\gamma := 2e^{\epsilon}$. For $i \in M$, define $\mathcal{P}_0^{(i)} := \{\pi : \Pr[\pi_{|N(i)}|X_i = 0] > \gamma \cdot \Pr[\pi_{|N(i)}|X_i = 1]\}$ and $\mathcal{P}_1^{(i)} := \{\pi : \Pr[\pi_{|N(i)}|X_i = 1] > \gamma \cdot \Pr[\pi_{|N(i)}|X_i = 0]\}$. We denote by $\mathcal{P}^{(i)} := \mathcal{P}_0^{(i)} \cup \mathcal{P}_1^{(i)}$ the set of *bad* transcripts with respect to party i. Let $\mathcal{P} := \cup_{i \in M} \mathcal{P}^{(i)}$.

Proposition 2 (Projection of Events). *Suppose U is a subset of the views of the transcript by the neighbors of i, and define the subset of transcripts by $\mathcal{P}_U := \{\pi : \pi_{|N(i)} \in U\}$. Then, it follows that $\Pr_{\mathbf{X},\Pi}[\pi(\mathbf{X}) \in \mathcal{P}_U] = \Pr_{\mathbf{X},\Pi}[\pi(\mathbf{X})_{|N(i)} \in U]$.*

Lemma 1 (Most Transcripts Behave Well). *Let $\epsilon > 0$ and $0 \leq \delta \leq \frac{1}{4n}$. Suppose the protocol is (ϵ, δ)-differentially private against any coalition of size t, and $\mathcal{P}$ is the union of the bad transcripts with respect to parties with at most t neighbors in the communication graph. Then,* $\Pr_{\mathbf{X},\Pi}[\mathcal{P}] \leq \frac{3}{4}$.

Proof. From definition of $\mathcal{P}_0^{(i)}$ and using Proposition 2, we have $\Pr[\mathcal{P}_0^{(i)}|X_i = 0] > \gamma \cdot \Pr[\mathcal{P}_0^{(i)}|X_i = 1]$. Since the protocol is (ϵ, δ)-differentially private against any coalition of size t, we have for each $i \in M$, $\Pr[\mathcal{P}_0^{(i)}|X_i = 0] \leq e^{\epsilon} \Pr[\mathcal{P}_0^{(i)}|X_i = 1] + \delta$. Hence, we have $(\gamma - e^{\epsilon}) \Pr[\mathcal{P}_0^{(i)}|X_i = 1] \leq \delta$, which implies that $\Pr[\mathcal{P}_0^{(i)}|X_i = 1] \leq e^{-\epsilon}\delta$, since $\gamma = 2e^{\epsilon}$.

Hence, we also have $\Pr[\mathcal{P}_0^{(i)}|X_i = 0] \leq e^{\epsilon} \Pr[\mathcal{P}_0^{(i)}|X_i = 1] + \delta \leq 2\delta$. Therefore, we have $\Pr[\mathcal{P}_0^{(i)}] = \frac{1}{2}(\Pr[\mathcal{P}_0^{(i)}|X_i = 0] + \Pr[\mathcal{P}_0^{(i)}|X_i = 1]) \leq \frac{3\delta}{2}$.

Similarly, we have $\Pr[\mathcal{P}_1^{(i)}] \leq \frac{3\delta}{2}$. Hence, by the union bound over $i \in M$, we have $\Pr[\mathcal{P}] \leq 3n\delta \leq \frac{3}{4}$, since we assume $0 \leq \delta \leq \frac{1}{4n}$. □

We perform the analysis by first conditioning on some transcript $\pi \notin \mathcal{P}$. The goal is to show that $\Pr_{\mathbf{X}}[\mathcal{E}|\pi] \geq \eta$, for some $\eta > 0$. Then, since $\Pr[\mathcal{P}] \leq \frac{3}{4}$, we can conclude $\Pr_{\mathbf{X}}[\mathcal{E}] \geq \frac{\eta}{4}$, and hence for some configuration $\mathbf{x}$, we have $\Pr[\mathcal{E}|\mathbf{x}] \geq \frac{\eta}{4}$, as required.

Conditioning on Transcript π. The first step (Lemma 2) is analogous to the techniques of [MMP+10, Lemma 1]. We show that conditioning on the transcript $\pi \notin \mathcal{P}$, the random variables X_i's are still independent and still have enough randomness remaining.

Definition 3 (γ-random). *Let $\gamma \geq 1$. A random variable X in $\{0, 1\}$ is γ-random if $\frac{1}{\gamma} \leq \frac{\Pr[X=1]}{\Pr[X=0]} \leq \gamma$.*

Lemma 2 (Conditional Independence and Randomness). *Suppose each party's bit X_i is uniform and independent, and consider a protocol to estimate the sum that is (ϵ, δ)-differentially private against any coalition of size t, where $0 \leq \delta \leq \frac{1}{4n}$. Then, conditioning on the transcript $\pi \notin \mathcal{P}$, the random variables X_i's are independent; moreover, for each party $i \in M$ that has at most t neighbors in the communication graph, the conditional random variable X_i is γ-random, where $\gamma = 2e^{\epsilon}$.*

Proof. The proof is similar to that of [MMP+10, Lemma 1]. Since our lower bound does not depend on the number of rounds, we can without loss of generality sequentialize the protocol and assume only one node sends a message in each round. The conditional independence of the X_i's can be proved by induction on the number of rounds of messages. To see this, consider the first message m_1 sent by the party who has input X_1, and suppose X' is the joint input of all other parties. Observe that (X_1, m_1) is independent of X'. Hence, we have $\Pr[X_1 = a, X' = b|m_1 = c] = \frac{\Pr[X_1=a, X'=b, m_1=c]}{\Pr[m_1=c]} = \frac{\Pr[X_1=a, m_1=c] \Pr[X'=b]}{\Pr[m_1=c]}$ $= \Pr[X_1 = a|m_1 = c] \cdot \Pr[X' = b|m_1 = c]$, which means conditioning on m_1, the random variables X_1 and X' are independent. After conditioning on m_1, one can

view the remaining protocol as one that has one less round of messages. Therefore, by induction, one can argue that conditioning on the whole transcript, the inputs of the parties are independent.

For each party i having at most t neighbors, the γ-randomness of each conditional X_i can be proved by using the uniformity of X_i and that $\pi \notin \mathcal{P}^{(i)}$ is not bad for i.

We first observe that the random variable X_i has the same conditional distribution whether we condition on π or $\pi_{|N(i)}$, because as long as we condition on the messages involving node i, everything else is independent of X_i.

We next observe that if party $i \in M$ has at most t neighbors in the communication graph and $\pi \notin \mathcal{P}^{(i)}$, then by definition we have $\frac{\Pr[\pi_{|N(i)}|X_i=1]}{\Pr[\pi_{|N(i)}|X_i=0]} \in [\gamma^{-1}, \gamma]$.

Hence, $\frac{\Pr[X_i=1|\pi]}{\Pr[X_i=0|\pi]} = \frac{\Pr[X_i=1|\,\pi_{|N(i)}]}{\Pr[X_i=0|\,\pi_{|N(i)}]} = \frac{\Pr[\pi_{|N(i)}|X_i=1]\cdot\Pr[X_i=1]}{\Pr[\pi_{|N(i)}|X_i=0]\cdot\Pr[X_i=0]} = \frac{\Pr[\pi_{|N(i)}|X_i=1]}{\Pr[\pi_{|N(i)}|X_i=0]} \in [\gamma^{-1}, \gamma]$. □

We use the superscripted notation X' to denote the version of the random variable X conditioning on some transcript π. Hence, Lemma 2 states that the random variables X_i''s are independent, and each X_i' is γ-random for $i \in M$. It follows that the sum $\sum_{i\in M} X_i'$ has variance at least $\frac{m\gamma}{(1+\gamma)^2}$.

The idea is that conditioning on the transcript π, the sum of the parties' bits (in M) has high variance, and so the protocol is going to make a large error with constant probability. We describe the precise properties we need in the following technical lemma, whose proof appears in Section 4.1, from which Theorem 1 follows.

Lemma 3 (Large Variance Dichotomy). *Let $\gamma \geq 1$. There exists $\eta > 0$ (depending on γ) such that the following holds. Suppose Z_i's are m independent random variables in $\{0,1\}$ and are all γ-random, where $i \in [n]$. Define $Z := \sum_{i\in[m]} Z_i$ and $\sigma^2 := \frac{m\gamma}{2(1+\gamma)^2}$. Then, there exists an interval $[a,b]$ of length $\frac{\sigma}{2}$ such that the probabilities $\Pr[Z \geq b]$ and $\Pr[Z \leq a]$ are both at least η.*

Proof of Theorem 1: Using Lemma 3, we set $\gamma := \exp(\epsilon)$ and $Z_i := X_i'$, for each $i \in M$. Suppose $\eta > 0$ (depending on γ and hence on ϵ), $\sigma^2 := \frac{m\gamma}{2(1+\gamma)^2}$ and the interval $[a,b]$ are as guaranteed from the lemma. Suppose s is the sum of the bits of parties outside M. Let $c := \frac{a+b}{2} + s$.

Suppose the protocol makes an estimate that is at most c. Then, conditioning on π, the system still has enough randomness among parties in M, and with probability at least η, the real sum is at least $b + s$, which means the additive error is at least $\frac{\sigma}{4}$. The case when the protocol makes an estimate greater than c is symmetric. Therefore, conditioning on $\pi \notin \mathcal{P}$, the protocol makes an additive error of at least $\frac{\sigma}{4}$ with probability at least η in any case. Note that this is true even if the protocol is randomized.

Let $\mathcal{E}$ be the event that the protocol makes an additive error of at least $\frac{\sigma}{4}$. We have just proved that for $\pi \notin \mathcal{P}$, $\Pr_{\mathbf{X},\Pi}[\mathcal{E}|\pi] \geq \eta$, where the probability is over the $\mathbf{X} = (X_i : i \in [n])$ and the randomness of the protocol Π.

Observe that $\Pr_{\mathbf{X},\Pi}[\mathcal{E}|\pi] \geq \eta$ for all transcripts $\pi \notin \mathcal{P}$, and from Lemma 1, $\Pr[\mathcal{P}] \leq \frac{3}{4}$. Hence, we conclude that $\Pr_{\mathbf{X},\Pi}[\mathcal{E}] \geq \frac{\eta}{4}$. It follows that there must

exist some configuration $\mathbf{x}$ of the parties' bits such that $\Pr_{\Pi}[\mathcal{E}|\mathbf{x}] \geq \frac{\eta}{4}$. This completes the proof of Theorem 1. □

4.1 Large Variance Dichotomy

We prove Lemma 3. For $i \in M$, let $p_i := \Pr[Z_i = 1]$. From the γ-randomness of Z_i, it follows that $\frac{1}{1+\gamma} \leq p_i \leq \frac{\gamma}{1+\gamma}$. Without loss of generality, we assume that there are at least $\frac{m}{2}$ indices for which $p_i \geq \frac{1}{2}$; otherwise, we consider $1 - Z_i$. Let $J \subseteq M$ be a subset of size $\frac{m}{2}$ such that for each $i \in J$, $p_i \geq \frac{1}{2}$.

Define for $i \in J$, $Y_i := Z_i - p_i$. Let $Y := \sum_{i \in J} Y_i$, and $Z' := \sum_{i \in J} Z_i$. It follows that $E[Y_i] = 0$ and $E[Y_i^2] = p_i(1 - p_i) \geq \frac{\gamma}{(1+\gamma)^2}$. Denote $\sigma^2 := \frac{m\gamma}{2(1+\gamma)^2}$, $\mu := E[Z'] = \sum_{i \in J} p_i$ and $\nu^2 := E[Y^2] = \sum_{i \in J} p_i(1 - p_i)$. We have $\nu^2 \geq \sigma^2$.

The required result can be achieved from the following lemma.

Lemma 4 (Large Deviation). *There exists $\eta_0 > 0$ (depending only on γ) such that $\Pr[|Y| \geq \frac{9\sigma}{10}] \geq \eta_0$.*

We show how Lemma 4 implies the conclusion of Lemma 3. Since $\Pr[|Y| \geq \frac{9\sigma}{10}] = \Pr[Z' \geq E[Z'] + \frac{9\sigma}{10}] + \Pr[Z' \leq E[Z'] - \frac{9\sigma}{10}]$, at least one of the latter two terms is at least $\frac{\eta_0}{2}$. We consider the case $\Pr[Z' \geq E[Z'] + \frac{9\sigma}{10}] \geq \frac{\eta_0}{2}$; the other case is symmetric.

By Hoeffding's Inequality, for all $u > 0$, $\Pr[Z' \geq E[Z'] + u] \leq \exp(-\frac{2u^2}{n})$. Setting $u := \frac{2\sigma}{5}$, we have $\Pr[Z' < E[Z'] + \frac{2\sigma}{5}] \geq 1 - \exp(-\frac{8\gamma}{25(1+\gamma)^2}) =: \eta_1$.

We set $\eta := \frac{1}{2}\min\{\frac{\eta_0}{2}, \eta_1\}$. Let $\widehat{Z} := \sum_{i \in M \setminus J} Z_i$. Observe that $\widehat{Z}$ and Z are independent. Hence we can take the required interval to be $[\mathsf{median}(\widehat{Z}) + E[Z] + \frac{2\sigma}{5}, \mathsf{median}(\widehat{Z}) + E[Z] + \frac{9\sigma}{10}]$, which has width $\frac{\sigma}{2}$.

Hence, it remains to prove Lemma 4.

Proof of Lemma 4: We use the method of sub-Gaussian moment generating function in the way described in [IN07, Remark 3.1].

First, for each $i \in M$, for any real h,

$$\begin{aligned} E[e^{hY_i}] &= p_i \cdot e^{h(1-p_i)} + (1 - p_i) \cdot e^{h(0-p_i)} \\ &= \exp(-p_i h) \cdot (1 + p_i(e^h - 1)) \leq \exp(p_i h^2), \end{aligned}$$

where the last inequality follows from $1 + p(e^h - 1) \leq \exp(ph^2 + ph)$, for all real h and $\frac{1}{2} \leq p \leq 1$.

Let g be a standard Gaussian random variable, i.e., it has density function $x \mapsto \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}x^2}$. It is assumed that g is independent of all other randomness in the proof. Recall that $E[e^{hg}] = e^{\frac{1}{2}h^2}$ and for $h < \frac{1}{2}$, $E[e^{hg^2}] = \frac{1}{\sqrt{1-2h}}$.

For $0 \leq h \leq \frac{1}{8\mu}$, we have

$$\begin{aligned} E[e^{hY^2}] &= E_Y[E_g[e^{\sqrt{2h}gY}]] = E_g[E_Y[e^{\sqrt{2h}g\sum_i Y_i}]] \\ &= E_g[\prod_i E_{Y_i}[e^{\sqrt{2h}gY_i}]] \leq E_g[\prod_i e^{2hp_i g^2}] \\ &= E_g[\exp(2\mu h g^2)] = \frac{1}{\sqrt{1 - 4\mu h}} \leq \sqrt{2}. \end{aligned}$$

For $-\frac{1}{8\mu} \leq h \leq \frac{1}{8\mu}$, we have

$$\begin{aligned} E[e^{hY^2}] &\leq 1 + hE[Y^2] + \sum_{m\geq 2} \frac{1}{m!}(8\mu|h|)^m (\frac{1}{8\mu})^m E[Y^{2m}] \\ &\leq 1 + h\nu^2 + (8\mu h)^2 \sum_{m\geq 2} \frac{1}{m!}(\frac{1}{8\mu})^m E[Y^{2m}] \\ &\leq 1 + h\nu^2 + (8\mu h)^2 E[\exp(\frac{Y^2}{8\mu})] \\ &\leq 1 + h\nu^2 + 100\mu^2 h^2 \leq \exp(h\nu^2 + 100\mu^2 h^2) \end{aligned}$$

Let $0 < \beta < 1$. For $-\frac{1}{8\mu} \leq h < 0$, we have

$$\begin{aligned} \Pr[Y^2 \leq (1-\beta)\nu^2] &= \Pr[hY^2 \geq h(1-\beta)\nu^2] \\ &\leq \exp(-h(1-\beta)\nu^2) \cdot E[\exp(hY^2)] \\ &\leq \exp(h\beta\nu^2 + 100\mu^2 h^2). \end{aligned}$$

Observe that $\frac{1}{1+\gamma} \leq \frac{\nu^2}{\mu} = \frac{\sum_i p_i(1-p_i)}{\sum_i p_i} \leq \frac{\gamma}{1+\gamma}$.

We can set $h := -\frac{\beta\nu^2}{200\mu^2} \geq -\frac{1}{8\mu}$, and we have $\Pr[Y^2 \leq (1-\beta)\nu^2] \leq \exp(-\frac{\beta^2\nu^4}{400\mu^2}) \leq \exp(-\frac{\beta^2}{400(1+\gamma)^2})$.

Setting $\beta := \frac{19}{100}$ and observing that $\nu^2 \geq \sigma^2$, we have $\Pr[|Y| \geq \frac{9}{10}\sigma] \geq 1 - \exp(-(\frac{19}{2000(1+\gamma)})^2)$. □

5 Differentially Private Protocols against Coalitions

We show that the lower bound proved in Section 4 is essentially tight. As noted by Beimel *et al.* [BNO08], one can generally obtain differentially private multi-party protocols with small error, by combining general (information theoretic) Secure Function Evaluation (SFE) techniques with differential privacy. Although our upper-bound constructions use standard techniques from SFE and differential privacy, we include the main result here for completeness. The details are given in the full version.

Theorem 2 (Differentially Private Protocols Against Coalitions). *Given $\epsilon > 0$, $0 < \delta < 1$ and a positive integer t, there exists an oblivious protocol among n parties each having a secret input $x_i \in \mathcal{U} := \{0, 1, 2, \ldots, \Delta\}$, such that the protocol uses only $O(nt)$ messages to estimate the sum $\sum_{i\in[n]} x_i$; the differential privacy guarantees and error bounds of the protocols are given as follows.*

(a) For ϵ-differential privacy against any coalition of size t, with probability at least $1-\eta$, the additive error is at most $O(\frac{\Delta}{\epsilon} \cdot \exp(\frac{\epsilon}{2\Delta})\sqrt{t+1}\log\frac{1}{\eta})$.

(b) For (ϵ, δ)-differential privacy against any coalition of size t, with probability at least $1-\eta$, the additive error is at most $O(\frac{\Delta}{\epsilon} \cdot \exp(\frac{\epsilon}{2\Delta})\sqrt{\frac{n}{n-t}\log\frac{1}{\delta}\log\frac{1}{\eta}})$.

References

[Ach03] Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins. J. Comput. Syst. Sci. 66(4), 671–687 (2003)

[BNO08] Beimel, A., Nissim, K., Omri, E.: Distributed Private Data Analysis: Simultaneously Solving How and What. In: Wagner, D. (ed.) CRYPTO 2008. LNCS, vol. 5157, pp. 451–468. Springer, Heidelberg (2008)

[CK93] Chor, B., Kushilevitz, E.: A communication-privacy tradeoff for modular addition. Information Processing Letters 45, 205–210 (1993)

[DG03] Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algorithms 22(1), 60–65 (2003)

[DKM+06] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our Data, Ourselves: Privacy Via Distributed Noise Generation. In: Vaudenay, S. (ed.) EUROCRYPT 2006. LNCS, vol. 4004, pp. 486–503. Springer, Heidelberg (2006)

[DMNS06] Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)

[Dwo06] Dwork, C.: Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)

[Dwo10] Dwork, C.: A firm foundation for private data analysis. In: Communications of the ACM (2010)

[GRS09] Ghosh, A., Roughgarden, T., Sundararajan, M.: Universally utility-maximizing privacy mechanisms. In: STOC 2009 (2009)

[IN07] Indyk, P., Naor, A.: Nearest-neighbor-preserving embeddings. ACM Transactions on Algorithms 3(3) (2007)

[MMP+10] McGregor, A., Mironov, I., Pitassi, T., Reingold, O., Talwar, K., Vadhan, S.: The limits of two-party differential privacy. In: FOCS (2010)

[MPRV09] Mironov, I., Pandey, O., Reingold, O., Vadhan, S.: Computational Differential Privacy. In: Halevi, S. (ed.) CRYPTO 2009. LNCS, vol. 5677, pp. 126–142. Springer, Heidelberg (2009)

[PSP03] Przydatek, B., Song, D., Perrig, A.: Sia: secure information aggregation in sensor networks. In: ACM Sensys (2003)

[RN10] Rastogi, V., Nath, S.: Differentially private aggregation of distributed time-series with transformation and encryption. In: SIGMOD 2010, pp. 735–746 (2010)

[SCR+11] Shi, E., Chan, H., Rieffel, E., Chow, R., Song, D.: Privacy-preserving aggregation of time-series data. In: NDSS (2011)