Over-the-Air Histogram Estimation

Henrik Hellström[†*], Jiwon Jeong[*], Wei-Ning Chen[*], Ayfer Özgür[*], Viktoria Fodor[†], Carlo Fischione[†]
[*]Electrical Engineering Department, Stanford University, California, USA
Emails: hhells@stanford.edu, jeongjw@stanford.edu, wnchen@stanford.edu, aozgur@stanford.edu
[†]School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
Emails: hhells@kth.se, vjfodor@kth.se, carlofi@kth.se

ICC 2024 - IEEE International Conference on Communications | 978-1-7281-9054-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICC51166.2024.10622573

***Abstract*—We consider the problem of secure histogram estimation, where n users hold private items x_i from a size-d domain and a server aims to estimate the histogram of the user items. Previous results utilizing orthogonal communication schemes have shown that this problem can be solved securely with a total communication cost of $O(n^2 \log(d))$ bits by hiding each item x_i with a mask. In this paper, we offer a different approach to achieving secure aggregation. Instead of masking the data, our scheme protects individuals by aggregating their messages via a multiple-access channel. A naïve communication scheme over the multiple-access channel requires d channel uses, which is generally worse than the $O(n^2 \log(d))$ bits communication cost of the prior art in the most relevant regime $d \gg n$. Instead, we propose a new scheme that we call Over-the-Air Group Testing (AirGT) which uses group testing codes to solve the histogram estimation problem in $O(n \log(d))$ channel uses. AirGT reconstructs the histogram exactly with a vanishing probability of error $P_{\text{error}} = O(d^{-T})$ that drops exponentially in the number of channel uses T.**

***Index Terms*—Over-the-Air Computation, Histogram Estimation, Group Testing, Non-Coherent, Goal-Oriented Communications.**

I. Introduction

In this paper, we introduce the idea of over-the-air computation (AirComp) for federated analytics (FA). Introduced in [1], FA is the practice of applying data science methods to the analysis of raw data that is stored locally on users' devices. This often involves the estimation of population-level statistics (histogram, ranges, heavy hitters, quantiles, mean/median, etc.) from distributed data for the analysis of user behaviour and system performance. It also provides critical input for more sophisticated downstream tasks such as training of machine learning models.

In this paper, we focus on one of the most canonical tasks in FA for discrete data: histogram estimation. Assume we have a set of n user devices, each holding an item $x_i \in \{1, \dots, d\}$ for $i = 1, \dots, n$. The server wants to estimate the histogram of the items held by the n users. For example, x_i can represent a word typed by user i, in which case the server wishes to estimate the empirical frequency distribution of the words, and the domain size d is the set of words in the English dictionary. As another example, x_i can represent the location trace of user i, in which case the server wishes to estimate the distribution of users in a given geographical area. Note that in both cases x_i (words typed by user i or their location) represents sensitive private information, hence it is essential to ensure that the underlying data remains private and secure while the server learns the population distribution.

A popular method to ensure privacy is the secure aggregation (SecAgg) scheme [2]. SecAgg is a cryptographic multi-party computation (MPC) scheme that ensures only population-level information (such as the summation of inputs) is revealed to the server while individual information remains secret. This is typically achieved by having all user devices agree a priori on pairwise masks that can be used to obscure their items. By communicating the masked items to the server, any individual's x_i is hidden from the server yet it can still recover the histogram of the items. This method has recently been analyzed in [3], however as shown in this paper it results in a significant increase in communication cost. Without concern for privacy, each user can communicate its item to the server by using only $\log d$ bits. With secure aggregation, [3] shows that the communication cost increases to $O(n \log d)$ bits per user.

In this paper, we study the problem of histogram estimation over multiple access channels (MACs). Our core idea is to leverage the natural superposition of simultaneously transmitted electromagnetic waves to achieve secure aggregation, without requiring masks such as in SecAgg. This general concept is known as AirComp [4] and already has a rich literature [5]. However, applying AirComp to federated analytics, and histogram estimation in particular, requires a rethinking of this paradigm. This is because a naïve application of AirComp requires d channel uses to communicate the histogram (by representing each x_i as a one-hot vector of length d). This is generally worse than the total communication cost for SecAgg, $O\left(n^2 \log(d)\right)$ bits, especially in the most relevant regime $d \gg n$ for practical applications. For example, in the natural language processing application mentioned above, $d \approx 500,000$ while n is limited by the number of users that connect to the same access point (typically in the tens or hundreds). Our proposal reconstructs the histogram support in $O(n \log(d))$ channel uses by posing histogram estimation as a distributed group testing (GT) problem. Once the support is estimated, the full histogram can be securely computed with SecAgg.

GT was originally introduced during World War 2 in the

Henrik Hellström and Jiwon Jeong contributed equally to this work. We acknowledge the financial support of the Ericsson Research Foundation, the Swedish Foundation for Strategic Research: AICOM, KTH Digital Futures, NSF Award #CCF-2213223, and Stanford University fellowship.

context of testing patients for syphilis [6]. The problem formulation involves a large population of people with a small unknown subset of infected individuals. The goal is to identify the infected individuals with a minimal number of tests. The basic idea is to mix blood samples from different individuals and test the mixed sample to reveal whether at least one person is infected in the group. In this way, the infected subset can be identified with exponentially fewer tests, provided that the number of infected people is much smaller than the population size. We show that these codes can be used to construct efficient communication schemes in our setting in the regime $d \gg n$. Moreover, their inherent coding power can be leveraged to mitigate the channel noise in AirComp and reconstruct the histogram exactly with high probability. In Section II, we detail how group testing can be leveraged for histogram estimation.

A. Contributions

The main contributions of our paper can be summarized as follows:

- We propose a new protocol for histogram estimation over MACs referred to as over-the-air group testing (AirGT);
- AirGT only requires $O(n\log(d))$ channel uses to securely reconstruct histograms, which is a factor n improvement over the $O\left(n^2\log(d)\right)$ bits required for SecAgg. Moreover, it leverages the additive nature of the channel to reveal only the histogram to the server eliminating the need for cryptographic protocols;
- AirGT can be viewed as a distributed channel code for computation. In particular, AirGT offers exact recovery of the histogram, with a vanishing probability of error that decreases exponentially in the number of channel uses.

B. Related Work

Since its introduction in 1943 [6], GT has been an active area of research. Different types of order-optimal or nearly optimal constructions are known for GT including adaptive [7], nonadaptive [8], combinatorial [9], and probabilistic [10] GT codes.

GT has been applied to wireless communications before, such as in neighbor discovery [11] and massive random access [12]. However, as far as we are aware, our work is the first one that applies GT to either AirComp or histogram estimation.

Within AirComp, we leverage non-coherent methods with hypothesis testing at the receiver. We have taken inspiration from similar methods in majority-vote distributed gradient descent [13] and aggregation in low power wide area networks [14]. However, unlike our work, none of these related works offer a reconstruction of the desired function with a vanishing probability of error.

II. Problem Setup

In this section, we introduce the notation of AirGT and discuss group testing in the context of AirComp.

Assume we have a set of n user devices and each device holds an arbitrary item x_i from the set $\{1,\ldots,d\}$. We will represent x_i as a one-hot vector $\mathbf{x}_i \in \{0,1\}^d$, i.e., $\mathbf{x}_i$ is a one-sparse vector consisting of all zero entries but a single 1 where the location of 1 indicates the corresponding item. The GT algorithm is aimed at recovering the histogram support $\mathbf{h}_s = \bigvee_{i=1}^n \mathbf{x}_i$, where $\bigvee$ denotes bitwise OR (disjunction). Note that the cardinality of the histogram support is $\|\mathbf{h}_s\|_1 \leq n$, with equality if and only if (iff) all devices carry unique items, i.e., $\mathbf{x}_i \neq \mathbf{x}_j$ for all $i \neq j$.

At timeslot t, the server administers a group test by polling all n devices using a (potentially dense) bitmask $\mathbf{m}_t \in \{0,1\}^d$. A non-zero entry $m_t[j] = 1$ indicates that item $j \in \{1,\ldots,d\}$ is tested in timeslot t. The devices reply with a single bit of information $b_i[t] = \langle \mathbf{x}_i, \mathbf{m}_t\rangle$ which equals one if their item is part of test t. Just as in binary GT, the server only needs to know if at least one device responded positively, i.e., it only needs $\bigvee_{i=1}^n b_i[t]$. This is where the AirComp protocol is used. We view $f(\mathbf{b}) = \bigvee_{i=1}^n b_i[t]$ as the desired function and estimate it with a non-coherent AirComp protocol. In particular, we design a communication scheme such that the server receives the noisy superposition

$$v[t] := \left(\bigvee_{i=1}^{n} b_i[t]\right) \oplus z[t], \tag{1}$$

where $z[t] \sim \mathcal{B}(q)$ represents a bit flip with probability q, and $\oplus$ is the XOR operation.

In this paper, we restrict ourselves to nonadaptive GT, since the adaptive version requires a downlink communication step for each group test, whereas nonadaptive GT can be completed in a single communication round. In particular, consider that the server administers a total of T tests. All group tests $\mathbf{m}_t$ can be expressed as a GT matrix $M \in \{0,1\}^{T\times d}$ which can be shared offline as a software update. With this setup, an instance of AirGT can be realized as

1) The server broadcasts a pilot to initiate group testing;
2) Using the predetermined GT matrix M, the user devices compute a vector of bits as $\mathbf{b}_i = M\mathbf{x}_i$;
3) All user devices communicate their vector $\mathbf{b}_i$ over (1) with T channel uses and the server receives $\mathbf{v} = \left(\bigvee_{i=1}^n \mathbf{b}_i[t]\right) \oplus \mathbf{z}[t] \in \{0,1\}^T$.

The server is then tasked with a linear inverse problem to recover the high-dimensional but sparse histogram support $\mathbf{h}_s$ given the noisy low-dimensional test results $\mathbf{v}$. After $\mathbf{h}_s$ is recovered, it is cheap to find $\mathbf{h}$ since the problem has been reduced from dimension d to (at most) dimension n. $\mathbf{h}$ can then be found with SecAgg by querying each histogram bin corresponding to $\mathbf{h}_s$. If desired, AirGT can be generalized to quantitative GT [15]. In that case, the full histogram $\mathbf{h}$ could be recovered directly over-the-air. However, in this paper, we focus on recovering the support $\mathbf{h}_s$. The overall problem setup is illustrated in Fig. 1.

The rest of the paper is organized as follows. In Section III, we present the design of the GT algorithm and give explicit upper bounds on the number of channel uses T required to reconstruct the histogram. In Section IV, we introduce the non-coherent communication scheme that realizes the disjunction

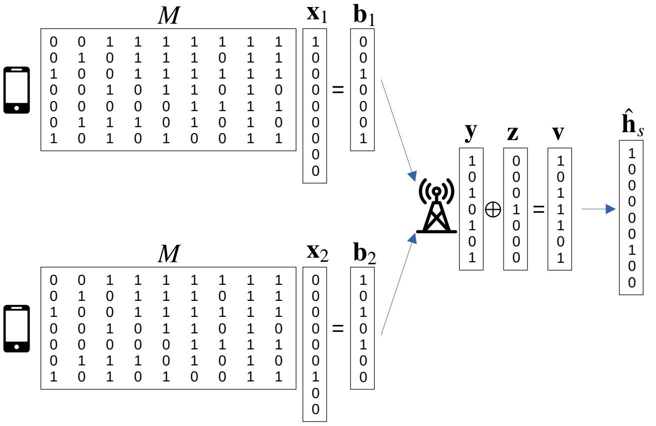

Fig. 1. Illustration of the AirGT setup over the disjunction MAC. All notation follows Section II with the addition of $\mathbf{y} := \mathbf{b}_1 + \mathbf{b}_2$ for brevity. The devices compute $\mathbf{b}_i = M\mathbf{x}_i$ and communicate it over the MAC. The server receives a noisy and distorted sum of the transmitted messages $\mathbf{v} = (\mathbf{b}_1 \bigvee \mathbf{b}_2) \oplus \mathbf{z}$. Finally, the server solves a linear inverse problem to estimate the histogram. In this example, $T = 7$, $d = 9$, and $n = 2$.

MAC from (1), and give a closed-form expression of q. In Section V, we present numerical results to evaluate the tightness of our bounds. Finally, in Section VI, we conclude the paper.

III. Group Testing Analysis

In this section, we describe the GT algorithm that we use in AirGT, how the GT matrix is constructed, and prove an upper bound on the number of tests required to reliably reconstruct the histogram.

In this paper, we leverage the Noisy Combinatorial Orthogonal Matching Pursuit (NCOMP) algorithm [10] to solve the noisy GT problem. For a detailed exposition of NCOMP, such as a process of setting parameters, we refer to [10] but we also give a brief overview here for completeness. In NCOMP, the group-testing matrix is a random binary Bernoulli matrix, where every element $M_{i,j}$ is independently selected to be one with some probability $p \le 1/2$. Upon receiving the test results $\mathbf{v} \in \{0,1\}^T$ via the disjunction-MAC (1), the server initiates a decoding algorithm to recover $\mathbf{h}_s$ from $\mathbf{v}$. The decoding algorithm is based around matching columns $\mathbf{m}_j \in \{0,1\}^{T\times 1}$ with the received vector $\mathbf{v} \in \{0,1\}^{T\times 1}$, where each column $j \in \{1,\ldots,d\}$ corresponds to one bin in the histogram. In particular, a column $\mathbf{m}_j$ is considered to be matched with $\mathbf{v}$ if

$$\left|\mathcal{S}_j\right| \ge \|\mathbf{m}_j\|_1(1 - q(1+\Delta)), \tag{2}$$

where $\mathcal{S}_j$ is the set of indices i where both $m_j[i] = 1$ and $v[i] = 1$, q is the bit flip probability in (1), and Δ is a design parameter for the algorithm which will be specified later. The estimated histogram is then simply constructed as

$$\hat{h}_s[j] = \begin{cases} 0 & \text{if } \left|\mathcal{S}_j\right| < \|\mathbf{m}_j\|_1(1 - q(1+\Delta)) \\ 1 & \text{if } \left|\mathcal{S}_j\right| \ge \|\mathbf{m}_j\|_1(1 - q(1+\Delta)) \end{cases}. \tag{3}$$

In addition to simplicity, this algorithm achieves order-wise optimality. We are now ready to state the first proposition of the paper.

Proposition 1. *Consider that the result of test t is communicated over the disjunction-MAC* (1) *with bit flip probability q. The GT matrix M is selected as a random Bernoulli matrix with probability $p = 0.5n^{-1}$. The number of user devices n is assumed to be much smaller than the histogram dimension d, i.e., $n = o(d)$. Upon receiving the noisy test results $\mathbf{v}$, the server applies the decoding algorithm from* (3) *with parameter*[1]

$$\Delta = \frac{\sqrt{\delta}e^{-0.5}(1-2q)}{q\left(\sqrt{\delta} + \sqrt{1+\delta}\right)}, \tag{4}$$

where $\delta \in \mathbb{R}^+$ is a design parameter. Let $\beta := 4.36\left(\sqrt{\delta} + \sqrt{1+\delta}\right)^2(1-2q)^{-2}$. The histogram support $\mathbf{h}_s$ can then be exactly reconstructed in

$$T \le \beta n \log(d), \tag{5}$$

tests with error probability $\epsilon_{GT} := \Pr\left(\hat{\mathbf{h}}_s \neq \mathbf{h}_s\right) \le d^{-\delta}$.

Proof. The proof for this proposition closely follows the proof in [10, Theorem 6], with one additional step. This step is needed because the proof in [10, Theorem 6] assumes that the true histogram $\mathbf{h}_s$ is exactly n-sparse ($|\mathbf{h}_s|_1 = n$), but in our case it can be less than n-sparse. Therefore, we need to show that the error probability ϵ_{GT} is increasing in $|\mathbf{h}_s|_1$ for a fixed number of tests T.

The error probability is upper bounded by the sum of the probability of false negatives and the probability of false positives $\epsilon_{\text{GT}} \le \epsilon^-_{\text{GT}} + \epsilon^+_{\text{GT}}$. From [10, Theorem 6], we know that the probability of false negatives is bounded by

$$\epsilon^-_{\text{GT}} \le |\mathbf{h}_s|_1 e^{-0.5\beta(1-e^{-2})(q\Delta)^2 \log(d)} =: b^-, \tag{6}$$

where $\beta > 0$. The probability of false positives is bounded by

$$\epsilon^+_{\text{GT}} \le (d - |\mathbf{h}_s|_1)e^{-0.5\beta(1-e^{-2})(e^{-0.5}(1-2q)-\Delta q)^2 \log(d)} =: b^+, \tag{7}$$

which increases with decreasing $|\mathbf{h}_s|_1$. To ensure that a reduction in $|\mathbf{h}_s|_1$ does not break the bound, it is sufficient to verify that the sum of the bounds on ϵ^-_{GT} and ϵ^+_{GT} satisfies

$$\frac{\partial b^-}{\partial |\mathbf{h}_s|_1} + \frac{\partial b^+}{\partial |\mathbf{h}_s|_1} \ge 0. \tag{8}$$

Inspection of (6) and (7) shows that (8) is equivalent to

$$e^{e^{-1}(1-2q)^2 - 2q\Delta e^{-1/2}(1-2q)} \ge 1. \tag{9}$$

At this point, we substitute for Δ from (4) and do some algebra to get

$$\sqrt{\delta} + \sqrt{1+\delta} \ge 2\left(q\sqrt{1+\delta} + \sqrt{\delta}(1-q)\right), \tag{10}$$

which always holds since $q \le 1/2$. With this additional step, [10, Theorem 6] gives our Proposition. □

Given Proposition 1, we know that reliable histogram estimation is achievable in $T = \beta n \log(d)$ channel uses over the disjunction MAC. The constant β is lower-bounded by 4.36

[1] Our Δ from (4) does not match with the one in [10]. This is because the Δ from [10] has a parameter controlled by n and d. If we use the fact that $n = o(d)$ to approximate that parameter, the Δ from [10] results in the same value as ours.

and will not be larger than 30 for reasonable values of d, δ and q. For instance, consider a histogram dimension of $d = 10^6$, a bit flip probability of $q = 5\%$, and $\delta = 1/2$. Proposition 1 will then guarantee a GT error probability $\epsilon_{\text{GT}} \leq 0.1\%$ with $\beta \approx 20.1$.

In the next section, we present the communication scheme that leads to the disjunction MAC.

IV. Communication Scheme

In this section we describe how the disjunction-MAC from (1) can be realized using a Rayleigh fading wireless MAC with additive white Gaussian noise (AWGN). In particular, we consider that the user devices transmit the analog value $s_i[t] \in \mathbb{R}^+$ and that the server receives

$$y[t] = \sum_{i=1}^{n} h_i[t] s_i[t] + z[t], \tag{11}$$

where $z[t] \sim \mathcal{CN}(0, 2\sigma_z^2)$ and $h_i[t] \sim \mathcal{CN}(0, 2\sigma_h^2)$. Since the devices $i \in \{1, ..., n\}$ do not have access to a shared clock, it is notoriously difficult to achieve phase alignment at the receiver [4], [5], [16]. To reflect this, we consider that neither transmitting devices nor the server have access to channel state information (CSI). As such, the fading distribution will be zero-mean regardless of the modulation scheme and the server only has access to the zero-mean random variable $y[t]$ (11). Given this information, the server is tasked with distinguishing between the null hypothesis $H_0 := \bigvee_{i=1}^n b_i[t] = 0$ and $H_1 := \bigvee_{i=1}^n b_i[t] = 1$. The outcome of this hypothesis test yields the disjunction-MAC defined in (1).

Given this model, the most energy-efficient modulation scheme is on-off keying [17, Appendix B]. Therefore, we consider that the devices transmit

$$s_i[t] = \sqrt{P_{\max}} b_i[t] = \sqrt{P_{\max}} \langle \mathbf{m}_t, \mathbf{x}_i \rangle, \tag{12}$$

where $P_{\max}$ is a maximum power constraint placed on the device and $b_i[t]$ is the bit defined in Section I. The server will then receive a superposition of the $n_p[t] \leq n$ participating devices for which $b_i[t] = 1$

$$y[t] = \sqrt{P_{\max}} \sum_{i=1}^{n_p[t]} h_i[t] + z[t], \tag{13}$$

where we have committed some abuse of notation since the index i is no longer the same as in (12). Since the tests are independent, we will occasionally drop t for brevity. As mentioned previously, $\mathbb{E}[y] = 0$, and therefore the relevant information for the hypothesis testing problem lies in the energy of y. Note that

$$|y|^2 = \left(\sqrt{P_{\max}} \sum_{i=1}^{n_p} h_i^I + z^I \right)^2 + \left(\sqrt{P_{\max}} \sum_{i=1}^{n_p} h_i^Q + z^Q \right)^2, \tag{14}$$

where the superscripts I and Q indicate in-phase and quadrature components, respectively. It is clear that the IQ-components of (14) are squared independent Gaussian variables with mean $\mu_y = 0$ and variance $\sigma_y[t]^2 = n_p[t] P_{\max} \sigma_h^2 + \sigma_z^2$. By [17, Theorem 2], the optimal decision rule by the server can be reduced to a threshold test on $|y[t]|^2$. Therefore, we consider that the server makes its decision as

$$v[t] = \begin{cases} 0 & \text{if } |y[t]|^2 < \gamma \\ 1 & \text{if } |y[t]|^2 \geq \gamma. \end{cases}, \tag{15}$$

where γ is a predefined threshold. With this setup, we achieve the disjunction MAC from (1) with explicit bit flip probability q according to the following lemma.

Lemma 1. *Consider that* $\mathbf{x}_i$ *is a one-hot vector, where the non-zero position is independently and uniformly generated for all* $i \in \{1, \ldots, n\}$*. Additionally, consider that the GT matrix* $M \in \{0,1\}^{T \times d}$ *is generated via a Bernoulli Process with probability* $p = 0.5n^{-1}$ *for each element of* M *to be one. Then, by communicating* $s_i[t] \in \mathbb{R}$ *according to* (12) *and estimating* $v[t] \in \{0,1\}$ *according to* (15)*, the Rayleigh Fading MAC from* (11) *can be turned into the disjunction MAC from* (1) *with*

$$q \leq \max \left(e^{-0.5\gamma/\sigma_z^2}, \frac{\sum_{l=1}^n \left(1 - e^{-0.5\gamma/(l P_{\max} \sigma_h^2 + \sigma_z^2)}\right) \Pr\left(n_p = l\right)}{\sum_{l=1}^n \Pr\left(n_p = l\right)} \right), \tag{16}$$

where

$$\Pr\left(n_p = l\right) = \sum_{k=0}^{d} f_B\left(l; n, \frac{k}{d}\right) f_B\left(k; d, \frac{1}{2n}\right), \tag{17}$$

$f_B(k; n, p)$ *is the Binomial PMF of getting exactly* k *successes in* n *trials with probability* p*, and* $\gamma \in \mathbb{R}^+$ *is a constant.*

Proof. Since the IQ-components of $|y|^2$ are squared independent Gaussian variables with mean $\mu_y = 0$ and variance $\sigma_y^2 = n_p P_{\max} \sigma_h^2 + \sigma_z^2$, the server can apply the linear transformation

$$A(n_p) = \frac{|y|^2}{n_p P_{\max} \sigma_h^2 + \sigma_z^2} \tag{18}$$

to attain a chi-squared distribution with 2 degrees of freedom[2]. With $A(n_p)$, it is clear that the estimator from (15) has a decision error probability

$$\begin{aligned} \epsilon_{\text{com}} &= \Pr\left(|y|^2 \geq \gamma \,\middle|\, H_0\right) \Pr\left(H_0\right) + \Pr\left(|y|^2 < \gamma \,\middle|\, H_1\right) \Pr\left(H_1\right) \\ &= \Pr\left(A(0) \geq \frac{\gamma}{\sigma_z^2}\right) \Pr\left(n_p = 0\right) \\ &\quad + \sum_{l=1}^{n} \Pr\left(A(l) < \frac{\gamma}{l P_{\max} \sigma_h^2 + \sigma_z^2}\right) \Pr\left(n_p = l\right), \end{aligned} \tag{19}$$

where

$$\begin{aligned} &\Pr\left(A(0) \geq \frac{\gamma}{\sigma_z^2}\right) = e^{-0.5\gamma/\sigma_z^2}, \text{ and} \\ &\Pr\left(A(l) < \frac{\gamma}{l P_{\max} \sigma_h^2 + \sigma_z^2}\right) = \left(1 - e^{-0.5\gamma/(l P_{\max} \sigma_h^2 + \sigma_z^2)}\right). \end{aligned} \tag{20}$$

For any given device, the probability of participating follows a Bernoulli trial of probability $p = 0.5n^{-1}$, since the device is

[2] In practice, the server can not perform this linear transformation since it does not know n_p. However, it does not have to, it can operate directly on $|y|^2$ to distinguish between H_0 and H_1.

equally likely to select any one-hot vector and the sparsity of any given group test is $0.5n^{-1}$. However, note that the Bernoulli trials of any two devices are not independent for a fixed GT matrix M, therefore n_p is not a Binomial random variable. Instead, the PMF of n_p is given by

$$\begin{aligned}\Pr(n_p = l) &= \sum_{k=0}^{d} \Pr\left(n_p = l \,\middle|\, j = k\right) \Pr\left(j = k\right) \\ &= \sum_{k=0}^{d} f_B\left(l; n, \frac{k}{d}\right) f_B\left(k; d, \frac{1}{2n}\right),\end{aligned} \tag{21}$$

where $f_B(k; n, p)$ is the Binomial PMF of getting exactly k successes in n trials with probability p and j is the number of non-zero entries in one group test. To clarify, $\Pr(j = k)$ is the probability that there are k non-zero values in any row of M, and $\Pr(n_p = l | j = k)$ is the probability that l out of n devices has $b_i[t] = 1$ given that the corresponding GT row has k non-zero values. With Eqs. (19)-(21), we have a closed form expression for ϵ_{com}. However, this does not directly translate to a bit flip probability q for the disjunction MAC in (1), since the disjunction MAC considers symmetric bit flip probability of false negatives q^- and false positives q^+. Therefore, we require that neither q^- nor q^+ exceeds q. Equivalently,

$$\begin{aligned} q &\le \max\left(\Pr\left(|y[t]|^2 \ge \gamma \,\middle|\, H_0\right), \Pr\left(|y[t]|^2 < \gamma \,\middle|\, H_1\right)\right) \\ &= \max\left(e^{-0.5\gamma/\sigma_z^2}, \frac{\sum_{l=1}^{n}\left(1 - e^{-0.5\gamma/(lP_{\max}\sigma_h^2 + \sigma_z^2)}\right)\Pr\left(n_p = l\right)}{1 - \Pr\left(n_p = 0\right)}\right). \end{aligned} \tag{22}$$

Next, γ can be selected according to

$$\underset{\gamma}{\operatorname{argmax}} \quad 1 - q(\gamma) \qquad \text{s.t. } \gamma \ge 0, \tag{23}$$

where $q(\gamma)$ is the right-hand side of (22). Since q^- is an increasing function of γ and q^+ is a decreasing function, (23) is equivalent to

$$e^{-0.5\gamma/\sigma_z^2} = \frac{\sum_{l=1}^{n}\left(1 - e^{-0.5\gamma/(lP_{\max}\sigma_h^2 + \sigma_z^2)}\right)\Pr\left(n_p = l\right)}{1 - \Pr\left(n_p = 0\right)}, \tag{24}$$

which yields a symmetric bit flipping probability q. By combining Eqs. (19)-(22), with γ according to (24), the lemma follows. □

We are now ready to state the main result of this paper.

Theorem 1. *Consider that n user devices share a Rayleigh fading wireless MAC with a single-antenna server. In channel use t, each user device sends $s_i[t] \in \mathbb{R}^+$ and the server receives*

$$y[t] = \sum_{i=1}^{n} h_i[t] s_i[t] + z[t], \tag{25}$$

where $z[t] \sim \mathcal{CN}(0, 2\sigma_z^2)$ and $h_i[t] \sim \mathcal{CN}(0, 2\sigma_h^2)$. Each device carries a one-hot vector $\mathbf{x}_i \in \{0,1\}^d$, where the non-zero location is independently generated over a uniform distribution $\mathcal{U}(1, d)$. Then, without any CSI, the server can exactly reconstruct the support $\mathbf{h}_s = \bigvee_i \mathbf{x}_i$ with error probability $\epsilon_{GT} := \Pr\left(\hat{\mathbf{h}}_s \neq \mathbf{h}_s\right) \le d^{-\delta}$. The communication cost is upper-bounded (in terms of channel uses) by

$$T \le \beta n \log(d), \tag{26}$$

where

$$\beta = 4.36\left(\sqrt{\delta} + \sqrt{1+\delta}\right)^2 (1 - 2q)^{-2}, \tag{27}$$

$$q = \max\left(e^{-0.5\gamma/\sigma_z^2}, \frac{\sum_{l=1}^{n}\left(1 - e^{-0.5\gamma/(lP_{max}\sigma_h^2 + \sigma_z^2)}\right)\Pr\left(n_p = l\right)}{1 - \Pr\left(n_p = 0\right)}\right), \tag{28}$$

$$\Pr\left(n_p = l\right) = \sum_{k=0}^{d} f_B\left(l; n, \frac{k}{d}\right) f_B\left(k; d, \frac{1}{2n}\right), \tag{29}$$

$f_B(k; n, p)$ is the Binomial PMF of getting exactly k successes in n trials with probability p, and $\gamma \in \mathbb{R}^+$, $\delta \in \mathbb{R}^+$ are design parameters.

Proof. Proposition 1 and Lemma 1 yield the theorem. □

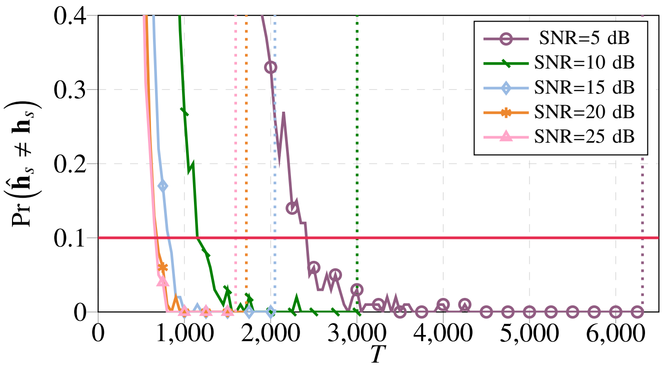

Fig. 2. The error probability for AirGT as a function of the number of channel uses T under $n = 10$ and $d = 10^4$. It is empirically calculated by running 100 Monte-Carlo simulations, and each plot corresponds to a different SNR value. The value of δ was set to 1/4 so that the upper bound of the error probability is $d^{-\delta} = 0.1$. Each vertical dotted line represents the upper bound of T calculated with (26).

With Theorem 1, we have proven that AirGT can reconstruct the histogram support $\mathbf{h}_s$ in $O(n\log(d))$ channel uses. Also, Theorem 1 tells us that the probability of group testing error is exponentially decreasing in the number of channel uses as $\epsilon_{\text{GT}} = O(d^{-T})$ since $T = O(\delta)$.

Before introducing the numerical results, we also wish to mention an alternative viewpoint of AirGT as a distributed channel code for aggregation. Since $\mathbf{x}_i$ is a one-hot vector, the vector-matrix multiplication $\mathbf{b}_i = M\mathbf{x}_i$ is equivalent to selecting one column in M, i.e., if $x_i[j] = 1$ then $\mathbf{b}_i = \mathbf{m}_j$, where $\mathbf{m}_j$ is column j of M. Every user device selects one column of M, and the server can reconstruct $\mathbf{h}_s$ using $\bigvee_i \mathbf{b}_i$. If we consider $\bigvee_i \mathbf{b}_i$ as a separate codeword, we can observe that these codewords are likely to be well-separated from each other, i.e., the disjunction of any n columns is well separated from the disjunction of any other n columns.

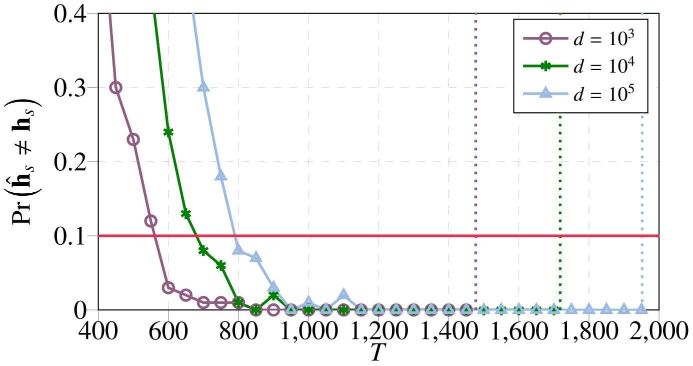

Fig. 3. The error probability for AirGT as a function of the number of channel uses T under $n = 10$ and SNR=20 dB. It is empirically calculated by running 100 Monte-Carlo simulations, and each plot corresponds to a different d value. The value of δ was set according to d so that the upper bound of the error probability is $d^{-\delta} = 0.1$. Each vertical dotted line represents the upper bound of T calculated with (26).

V. Numerical Results

In this section, we evaluate the end-to-end performance of AirGT. Most of the results in this paper are given in exact closed form, so the mathematical expressions give the full picture. However, the number of channel uses T and the error probability $\epsilon_{\text{GT}} := \Pr\left(\hat{\mathbf{h}}_s \neq \mathbf{h}_s\right)$ are given as upper bounds, which we evaluate numerically.

All simulations are created by randomly generating the items $\mathbf{x}_i$, the GT matrix M, the AWGN $\mathbf{z}$, and the fading coefficients $\mathbf{h}$ for 100 Monte Carlo trials. In each trial, $\mathbf{y}$ is computed according to (11), $\mathbf{v}$ is computed based on $\mathbf{y}$ according to (15), $\mathcal{S}_j$ is computed with $\mathbf{v}$ and M, and finally the estimate $\hat{\mathbf{h}}_s$ is formed with $\mathcal{S}_j$ according to (3). The empirical error is the fraction of Monte Carlo trials that resulted in $\hat{\mathbf{h}}_s \neq \mathbf{h}_s$. The simulation code is available at https://github.com/henrikhellstrom93/AirGT.

In Figure 2, we fix $d = 10^4$ and $\delta = 1/4$ such that the GT error probability is bounded to be $\epsilon_{\text{GT}} \leq 10\%$ according to Theorem 1. The goal of this simulation is to illustrate how the empirical error probability is affected by the SNR and the number of channel uses T. From the results, it is clear that T can be selected as approximately half of the upper bound, while still maintaining the desired error probability. In fact, if T is equal to the bound, the empirical error probability is less than 1%, i.e., we get zero errors in 100 Monte Carlo trials.

In Figure 3, we are interested in evaluating the level of compression achieved by the GT scheme. We define the compression ratio $r := d/T$, i.e., if a histogram of dimension 10^5 is communicated using 10^3 channel uses, we consider that to be a compression ratio of $r = 100$. For $d = 10^3$, we see that almost no compression is possible ($r \approx 1$). When $d = 10^4$, a ratio of $r \approx 10$ is possible, and for $d = 10^5$, around $r \approx 80$. Larger dimensions are prohibitively expensive to simulate via Monte Carlo simulation, but our bound in Theorem 1 can be used to get a minimum guarantee on r. For example, if $d = 10^7$, $n = 100$, $\delta = 1/7$ and SNR=20 dB, the upper bound of T becomes 24021, resulting in $r \geq 400$. Finally, it is worth noting that the numerical results are in the same order of magnitude as the bounds, suggesting that they can be viewed as an approximation for system performance.

VI. Conclusion

In this paper, we have demonstrated that secure histogram estimation can be performed over MACs in $T/n = O(\log(d))$ channel uses per user with a vanishing probability of error. In each time slot, the user devices elect to participate based on a common group testing matrix M and their data item $\mathbf{x}_i$. Based on the superimposed received waveform, the server solves a linear inverse problem to recover the histogram support $\mathbf{h}_s \in \{0, 1\}^d$. Our upper bound on T is exact, where $T \leq \beta n \log(d)$ with a relatively modest constant $\beta \geq 4.36$.

References

[1] D. Ramage and S. Mazzocchi, "Federated Analytics: Collaborative Data Science Without Data Collection," *URL https://ai. googleblog. com/2020/05/federated-analytics-collaborative-data.*, 2020.

[2] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical Secure Aggregation for Federated Learning on User-Held Data," *arXiv preprint arXiv:1611.04482*, 2016.

[3] W.-N. Chen, A. Ozgur, G. Cormode, and A. Bharadwaj, "The Communication Cost of Security and Privacy in Federated Frequency Estimation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 4247–4274.

[4] M. Goldenbaum and S. Stanczak, "Robust Analog Function Computation via Wireless Multiple-Access Channels," *IEEE Transactions on Communications (TCOM)*, vol. 61, no. 9, pp. 3863–3877, 2013.

[5] A. Şahin and R. Yang, "A Survey on Over-the-Air Computation," *IEEE Communications Surveys & Tutorials (COMST)*, 2023.

[6] R. Dorfman, "The Detection of Defective Members of Large Populations," *The Annals of Mathematical Statistics*, 1943.

[7] I.-H. Wang, S.-L. Huang, and K.-Y. Lee, "Extracting Sparse Data via Histogram Queries," in *54th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2016, pp. 39–45.

[8] O. Johnson, M. Aldridge, and J. Scarlett, "Performance of Group Testing Algorithms With Near-Constant Tests-per-Item," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 707–723, 2018.

[9] D. Du, F. K. Hwang, and F. Hwang, *Combinatorial Group Testing and its Applications*. World Scientific, 2000, vol. 12.

[10] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama, "Non-Adaptive Probabilistic Group Testing with Noisy Measurements: Near-Optimal Bounds with Efficient Algorithms," in *49th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2011.

[11] J. Luo and D. Guo, "Neighbor Discovery in Wireless ad hoc Networks Based on Group Testing," in *46th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2008, pp. 791–797.

[12] H. A. Inan, P. Kairouz, and A. Ozgur, "Sparse Group Testing Codes for Low-Energy Massive Random Access," in *55th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2017, pp. 658–665.

[13] A. Sahin and X. Wang, "Majority Vote Computation With Complementary Sequences for Distributed UAV Guidance," *arXiv preprint arXiv:2308.06372*, 2023.

[14] A. Gadre, F. Yi, A. Rowe, B. Iannucci, and S. Kumar, "Quick (and dirty) Aggregate Queries on Low-Power WANs," in *19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 2020, pp. 277–288.

[15] A. E. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, and M. I. Jordan, "Decoding from Pooled Data: Sharp Information-Theoretic Bounds," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 1, pp. 161–188, 2019.

[16] H. Hellström, S. Razavikia, V. Fodor, and C. Fischione, "Optimal Receive Filter Design for Misaligned Over-the-Air Computation," in *IEEE Global Communications Conference (GLOBECOM)*, 2023.

[17] F. Li, J. S. Evans, and S. Dey, "Decision Fusion over Noncoherent Fading Multiaccess Channels," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4367–4380, 2011.