MEng Thesis Proposal

by

Yaateh Richardson

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MEng in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Electrical Engineering and Computer Science
Dec 16, 2020

Certified by .....
Ramesh Raskar
Associate Professor
Thesis Supervisor

Accepted by .....
Asu Ozdaglar
Chairman, Department Committee on Graduate Theses

MEng Thesis Proposal

by

Yaateh Richardson

Submitted to the Department of Electrical Engineering and Computer Science
on Dec 16, 2020, in partial fulfillment of the
requirements for the degree of
MEng in Computer Science and Engineering

Abstract

The Abstract presents the overall topic of the thesis, briefly presenting the motivation for the study, challenges, and the proposed work that will be done

This work surveys the current state of the art in Local Differential Privacy counting queries, specifically for epidemiological use cases. In particular we focus on Hadamard Response techniques [?], and how *Relaxations* of *HR*s can be parameterized by via static and data driven priors on the underlying distribution. Thus far, experiments have shown improvements in power law distributions in ideal settings, and fairly re-liable performance in mock epidemiological settings.

One perhaps un-intuitive results is that *Relaxations* don't even need to reduce privacy for performance gains. Uniform and multi-modal distributions pose a bigger challenge but are arguably more rewarding. Finally, extensive literature review has inspired variants of the *HR* which we hope to formalize, implement, and test - time permitting.

Thesis Supervisor: Ramesh Raskar
Title: Associate Professor

Contents

List of Figures

Chapter 1

Proposal

1.1 Intro

The introduction section details motivation, the problem and challenges, and the proposed work the author will do. It also serves to explain and introduce the concepts and the basics necessary to understand the proposal.

Local Differential Privacy (*LDP*) is the ideal tool for Private Frequency Estimation and Counting Problems, since it does not assume a centralized trusted administrator/server, and provides perpetual privacy to user records. However, this privacy comes at the cost of *utility* (accuracy) and *sample complexity* - the number of participants required to provide accuracy within a certain degree of confidence. The more private an LDP scheme is the less utility it can provide.

The COVID-19 pandemic has highlighted the potential for LDP in major public health monitoring and aggregation services - specifically symptom reporting and heatmap generation [14, 15]. However, implementing LDP algorithms effectively requires strong assumptions about, or prior knowledge of underlying data. The privacy of an LDP algorithm is a function of two components. The *sensitivity* of the aggregate query - ie how much a single user can affect the output of particular query - and the security hyper-parameter ϵ. The value of ϵ is inversely proportional to the security of the algorithm - commonly denoted as ϵ-DP Differential Privacy - and ϵ-LDP in LDP.

Optimal choice of ϵ is crucial to maximizing utility. Low ϵ may guarantee user privacy, but aggregate statistics may also render query results un-usable. In Frequency

Estimation this has an adverse effect on less frequent elements of the domain. For example a variant of a largely successful and widely deployed LDP implementation - Google's RAPPOR Zero - discards elements below a certain count due to low confidence. This is acceptable in the case of web traffic estimation. But in public health the diversity and complexity of the problem space make it harder to decide which trends are statistically insignificant. Thus retaining as much signal as possible from these tail elements is desirable.

This work seeks to address the pain points of LDP that may arise in standard epidemiological use-cases by exploiting prior knowledge about the underlying distribution. We pay special attention to online learning algorithms for scenarios with limited prior knowledge. Our fundamental question is: how can we use limited prior knowledge to increase utility, while losing minimal privacy? To do this, we focus on the *counting query sub-problem*, a generalized version of the symptom vector problem for COVID-19. Our preliminary results suggest that certain distributions don't even require reduction in overall ϵ to see significant utility gains. We hope to take those results a step further via our proposed iterative bootstrapping algorithms and input-dependent LDP techniques [3].

High-level overview of the goals of the proposed work

1.2 Background and Related Work

This section includes the background needed to understand the problem, including the terms and notations that will be used. It also includes previous work on the subject, limitations of that work, and how the author plans to build on it.

1.2.1 COVID-19 Background and Data

State of the art Exposure Notification Systems like *GEAN*, Google and Apple's Exposure Notification platform emphasize privacy, but *GEAN* and other purely bluetooth based protocols are not complete solutions . They don't contain information about the context of encounters, and require active peer to peer connections to function [15]. They also don't provide public health officials with useful statistics to combat the spread of disease. Securely providing histograms of these statistics (namely symptom vectors) is precisely the point of *private counting queries.*

To simulate a realistic Covid-19 symptom vector setting, we use Google's Symptom Search Query dataset [4] to generate underlying symptom distributions that vary by region. From these distributions we draw samples to simulate LDP user responses. For methods that show promise on synthetic data-sets, we will perform further testing on this realistic synthetic data to build confidence in our methods before moving on to live-data scenarios (if possible).

Brief description of how the author's proposed work differs from the related work

1.2.2 Differential Privacy and Counting Queries

In this section we introduce notation and formal definitions of Differential Privacy from [16]. All Differential Privacy methods require an arregator (server) to perform algorithm A on a collection of user inputs D (aka. a database). The result t is an obfuscated aggregate statistic. Formally:

Definition 1.2.1 (Differential Privacy)**.** An algorithm $\mathbf{A}$ is ϵ-differentially private (ϵ-DP) if and only if for all $\epsilon \geq 0$ and datasets D, D' that differ in at most one element (row)

$$\forall t \in \text{Range}(\mathbf{A}) : \Pr[\mathbf{A}(D) = t] \leq e^{\mathcal{E}} \Pr\left[\mathbf{A}\left(D'\right) = t\right] \tag{1.1}$$

In an ϵ-LDP scheme, users perform two steps before sending their data to the aggregator: First the *Encode* step, which maps an input value to a data-domain. Second the *Perturb* step which adds a calibrated amount of noise to the data encoding such that it satisfies ϵ-DP at a *per-record level* instead of at the database level.

Key concept for understanding the work described

One of the most fundamental aggregation objectives is a counting query, which essentially asks the aggregator: "How many elements in the database satisfy a given property?". Formally a counting query consists of n users. Each user reports once, and can only contribute once per domain element $\{1, 2, \ldots, d\}$ (which we will abbreviate as $[d]$). In this setup there are d possible queries, and each query has a sensitivity of 1. This can be generalized to a sensitivity of k by allowing users k contributions

per domain element.

We focus our efforts on the Hadamard Response, which boasts the best run-time, computational complexity and sample complexity of all LDP frequency estimation variants [2]. Other counting query techniques documented in [16] include Warner's classic (k-)Randomized Response, variants of the Google Rappor [6], and the foundations for Apple's *Private Count Min Sketch* [1].

1.2.3 Relaxations on Differential Privacy

ϵ-LDP mechanisms require more noise in total than ϵ-DP methods to achieve the same amount of privacy, and yield poorer results. However generic ϵ-LDP algorithms make a strong assumption: **That all pairs of elements in the domain are assumed to be equally sensitive**. This is not always the case and has subtle implications. Consider privacy in geo-spatial queries, where densely populated areas require less noise to obfuscate a user's location to the same extent as would be necessary for a sparsely populated area. Locally Lipschitz privacy was introduced in [10] as a possible solution to this problem. The authors provide a data derived "privacy map" of ϵ parameters for each domain element or subgroup. Their method does not always yield a solution, but the notion of semantic security Lipschitz Privacy proposes is essential to using ϵ-LDP in practice. This thesis does not address that problem.

When one considers that not all pairs of elements in the domain are equally sensitive, a natural conclusion is that one could improve overall utility by reducing privacy for less sensitive domain elements. Acharya et al. seek to exploit that in [3] by introducing two input dependent relaxations of Hadamard Response: 1) High Low LDP (*HLLDP*) which emulates Mangat's Randomized Response[13] in which only some elements are obfuscated 2) Block Structured LDP (*BSLDP*) which could be used to implement the Lipshcitz privacy map from [16]. We hope to extend the contributions from [3] by showing that these relaxations can be realistically implemented (ie with bootstrapping methods). We also see strong potential for a new ϵ-LDP method

inspired by the *HLLDP* scheme.

1.2.4 Theoretical work in relaxing privacy

Formal privacy guarantees for all algorithms and techniques we propose can be derived with the composition theorems of Differential Privacy. Specifically the sequential composition theorem [9], the parallel composition theorem, and the k-fold adaptive composition theorem [5]. We may also draw upon prior theoretical work on the gradual relaxation of differential privacy to explain empirical behavior [11].

1.2.5 Frequency Estimation

This work was heavily inspired by research on Learned Data Structures. Hsu et al. used Learned Data Structures for a homologous objective - *Count Min Sketch* frequency estimation [7]. They exploit information about the underlying distribution through features of the input data via learned classifiers. These are typically Neural nets with the task of predicting "Heavy Hitters" (most frequent) elements. However [7] does not condider privacy. Learning a classifier adds an attack vector for membership inference since the model would be publicly available to users in an LDP scheme. Other work on Learned Index Structures can be found in this seminal paper [12].
Prior and gained knowledge about the distribution can also be exploited with *post processing*, which preserves privacy and can be added to any LDP algorithm [8]. *Post processing* methods for frequency estimation are often Bayesian, however the Calibrate method from [8] uses the predictions themselves as a data dependent prior for smoothing, yielding significant gains in utility.

1.3 Proposed Work

This section details the proposed work and identifies the steps and methods that will be followed. It includes both preliminary results and next steps.
Also note that this section is almost 70% of the whole proposal.

Highlights again in a very specific way the problem that the author will study and the reason to do so

We focus on the Hadamard Response (*HR*) [2], and *HR* based LDP techniques due to their superior performance (see section on Differential Privacy and Counting Queries). We use the term **Relaxations** to refer to modifications in the *Encode* and *Perturb*

steps of ϵ-LDP algorithms based on our data dependent prior. Our Proposed contributions are currently planned around the following relaxations (items with a † are potential contributions):

- Split domain - Splitting domain elements into 2 separate *HR*s.
- High Low LDP (*HLLDP*)[1] [3].
- Block Structured LDP (*BSLDP*) [3].
- Oracle Relaxation[2].

† Meta Hadamard Response - Inspired by [14].

† Tiered K Randomized Randomized Response - Possible variant on *HLLDP*.

All relaxations require at least 2 rounds of estimation. The first round can be considered a ramp up round for the prior, from which relaxation parameters are derived. We propose the following contributions in this thesis (from highest to lowest priority).

List of specific contributions proposed in this thesis

1. Empirical proof of the benefits of various LDP relaxations.
2. Empirical proof of the effectiveness of iterative bootstrapping with LDP to parameterize these relaxations.
3. Formal privacy guarantees for 1 & 2.
4. Analysis and theoretical justification of 1 & 2.
5. An open source implementation of the High Low LDP relaxation from [3].
6. Developing a new variant of context aware LDP based on the Hadamard Response.

[1] Does not necessarily preserve the privacy of all elements, awaiting confirmation from authors
[2] Doesn't not preserve the privacy of certain elements

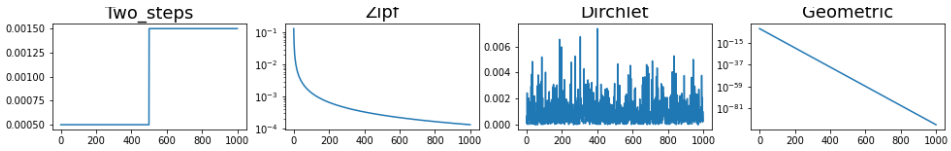

Figure 1-1: Examples of ground truth distributions over 1000 domain items [d]. Note the log-scale on the power law distributions.

To measure utility we use the ℓ_2 loss (error) between the estimated and ground truth distributions. We evaluate the following frequency distributions: Uniform, Dirchlet, Two-Step, Zipf, and Geometric (see fig. 1-1). Privacy is denoted as ϵ. The following subsections detail the goals, results, and next steps for salient contributions of this work. A proposed timeline of all contribution milestones can be found at the end of this section.

1.3.1 Empirical Study: Existing LDP Methods/Relaxations

Goal: To evaluate the proposed ϵ-LDP relaxations and any the conditions under which they consistently outperform standard ϵ-LDP. Successful relaxations will show decreases in loss while maintaining privacy or allowing for a pre-designated relaxation factor.

Subsection headings make it clear what work has been done and what work has yet to be completed

Preliminary Results: In preliminary testing, the *Split Domain* relaxation has shown significant improvements in the first two rounds of experimental trials without reducing ϵ at all. However it only shows significant improvement in power law distributions (Zipf and Geometric 1-2). Distributions without distinct sets of Heavy Hitters may require an alternat approach, such as *HLLDP* or other input specific methods. We observed similar gains in non-private Learned Count Min Sketch experiment based on [7].

Next Steps: 1) A logical next step is to begin analyzing how results correspond with existing theory. Specifically, using sample complexity bound on the Hadamard Response to explain the effectiveness of the *split domain* relaxation on power law distributions. 2) Implementing the other existing relaxations, and determining their effectiveness could be crucial to improving performance on other distributions. 3)

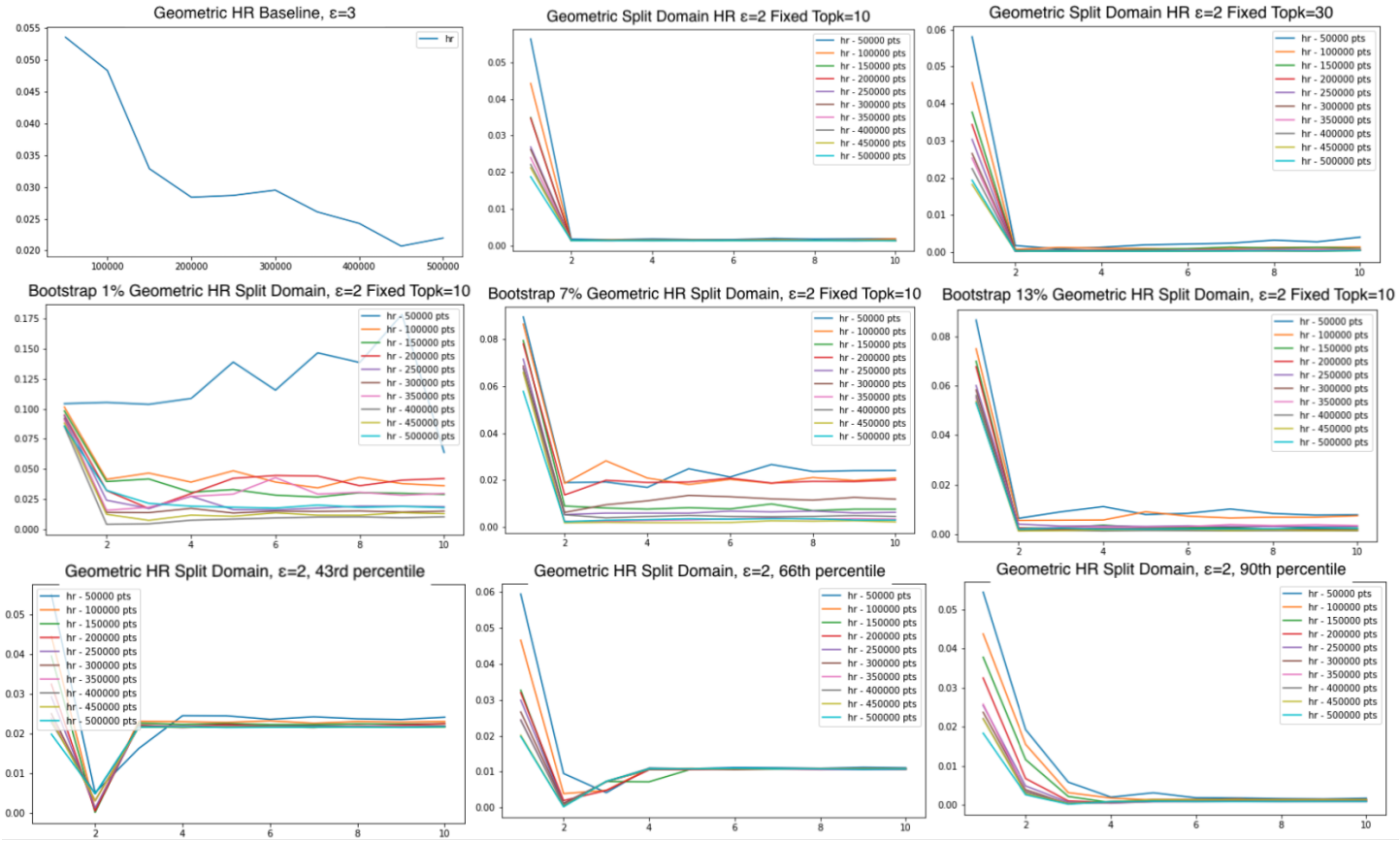

Figure 1-2: Preliminary results for the Split Domain relaxation. The top left chart shows the ℓ_2 loss of a Hadamard Sketch along the y axis as the number of inputs changes along the x axis. For the remaining charts the Y axis remains ℓ_2 loss, but the x axis becomes iteration. Each iteration releases the predicted distribution to the public. The prediction then becomes the data driven prior for the next round of estimation. Note that the first datapoint is equivalent to the baseline singe *HR* case. The first and last row depict preliminary results from Relaxations in Empirical Study: Existing LDP Methods/Relaxations. The middle row shows the preliminary results for Iterative Relaxation 1.3.2. See each section experimental setup details.

Adapting our current implementation from the high privacy variant of the Hadamard Response to a general purpose algorithm [2] will be key to discovering new performance gains. Relaxing ϵ did not consistently show improvements, which could be due to the high privacy variant only being effective for low values of ϵ. 4) Depending on early results on synthetic Covid-19 Data, we hope to submit our findings to the *ACM* conference in early January.

1.3.2 Iterative Relaxation Algorithm

Goal: To determine whether the relaxations can practically and reliably be deployed in real world scenarios while preserving privacy. Since ϵ-LDP techniques cannot be used repeatedly on the same data without reductions in privacy, bootstrapping methods must be used to parameterize the relaxation methods of 1.3.1. This contribution is slightly more open ended than the first, but success could include either an algorithm that converges to optimal utility for one or more ϵ-LDP relaxations or analyzing the conditions under which the best ϵ-LDP relaxation can be effectively bootstrapped. We are currently experimenting with the following algorithm for *split domain*.

Algorithm 1: Simulate a *split domain* ϵ-LDP relaxation

Result: Θ^* for an ϵ-LDP relaxation

Sample n_0 elements X^0 from the distribution with initial params Θ;

Predict distribution $P_0 = A(X^0, \Theta)$, where A satisfies ϵ-LDP specified by Θ;

Using P_0 define $\Theta^1 = (\Theta^1_{HH},\ \Theta^1_{reg})$.;

while $i < max\_iterations$ **do**

  Sample n_i elements X^i where $n_i \geq n_{i-1}$;

  Split elements into sub-domains $X^i_{HH}, X^i_{reg} = \text{split}(X^i, \Theta^i)$;

  Predict new distribution $P_i = A(X^i_{HH}, \Theta^i_{HH}) \circ A(X^i_{reg}, \Theta^i_{reg})$;

  Define $\Theta^{i+1} = \Theta^{i+1}_{HH},\ \Theta^{i+1}_{reg}$;

end

Note that in **Algorithm 1** we include the domain elements d and ϵ in the parameters Θ for each round. Let Θ^t_{HH} include the set of predicted Heavy hitter elements

predicted in round $t-1$ and likewise for Θ^t_{reg}

Preliminary Results: The middle row of fig. 1-2 shows results of **Algorithm 1** with an initial sample size n_0 as a percentage of total number of samples N. The following T points grew exponentially as follows: $n_i = n_0 + (N - n_0) * 2^{-(T-i)}$. With an n_0 of 1%, most choices of N converged roughly to pre-relaxation loss by $T = 10$ but did not show uniform convergence. Results were notably worse than in Section 1.3.1.

Next Steps: 1) Further experimentation in this setting requires varying ϵ, hence adapting the general purpose Hadamard Response will also be crucial for this contribution. 2) Incorporating information from previous rounds to get a somewhat noisy, but reasonable estimate for the entire population in the first round will give us a better idea of utility in a scenario with limited resources. Empirical proof of this is crucial to our *ACM* submission. 3) If sampling with replacement is necessary, we can bound losses in privacy by using the k-fold adaptive composition theorem from [5]. 4) For both of these contributions, empirically determining their effective ϵ is computationally intractable, but there may be research to give empirical upper bounds and validate our theoretical contributions WHP.

Author includes a "stretch goal" which will be pursued if the work up until this point goes well

1.3.3 Stretch Goal: New LDP Mechanism

Goal: I have highlighted 2 possibilities for new LDP mechanisms. These may prove essential in case existing relaxations are not as promising as we hoped on non-power law distributions. The first is inspired by this recent meta estimation paper [14]. The second seeks to improve the High Low LDP mechanism from [3] by adding tiers of security. It is still not clear to me whether the High low LDP actually provides no privacy for nonsensitive elements given their scheme. That will require some analysis. Finally, if either of these algorithms is theoretically sound, I would implement it and use my current test-bed for evaluation.

1.3.4 Proposed Timeline

Clear timeline outlines deliverables and the expected completion date month by month

1. IAP 2021
   (a) December 2020
      i. Neurips Literature Review
      ii. Complete analysis for Split Domain Relaxation and Iterative Relaxation
      iii. Incorporate sample complexity into above experiments
      iv. Build Synthetic Covid dataset Generator
   (b) January 2021
      i. ACM abstract due Jan 7th
      ii. IJCAI abstract due Jan 12th
      iii. Complete General Purpose HR analysis
      iv. Theoretical analysis and privacy guarantees for the above.
2. Spring 2021
   (a) February 2021
      i. Deep dive on HLLDP
      ii. Theoretical proof of concepts for *Meta HR* and *Tiered KRR* (new LDP algorithms)
      iii. Implement and sanity check one or more of the above Relaxations.
   (b) March 2021
      i. Analyze empirical results on syntehtic Covid Data or Real data-sets
      ii. Provide formal privacy guarantees and draw parallels to theory.
      iii. Derive meta theorems about analyzed techniques
   (c) April 2021
      i. Prepare Final Thesis Report

Bibliography

[1] Learning with privacy at scale differential. 2017.

[2] *Hadamard Response: Estimating Distributions Privately, Efficiently, and with Little Communication*, 2018.

[3] *Context-Aware Local Differential Privacy*, 2020.

[4] Shailesh Bavadekar, Andrew Dai, John Davis, Damien Desfontaines, Ilya Eckstein, Katie Everett, Alex Fabrikant, Gerardo Flores, Evgeniy Gabrilovich, Krishna Gadepalli, Shane Glass, Rayman Huang, Chaitanya Kamath, Dennis Kraft, Akim Kumok, Hinali Marfatia, Yael Mayer, Benjamin Miller, Adam Pearce, and Masrour Zoghi. Google covid-19 search trends symptoms dataset: Anonymization process description (version 1.0), 09 2020.

[5] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010.

[6] Úlfar Erlingsson, A. Korolova, and V. Pihur. Rappor: Randomized aggregatable privacy-preserving ordinal response. *ArXiv*, abs/1407.6981, 2014.

[7] C. Hsu, P. Indyk, D. Katabi, and A. Vakilian. Learning-based frequency estimation algorithms. In *ICLR*, 2019.

[8] Jinyuan Jia and Neil Zhenqiang Gong. Calibrate: Frequency estimation and heavy hitter identification with local differential privacy via incorporating prior knowledge. 2018.

[9] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy, 2015.

[10] F. Koufogiannis and G. J. Pappas. Location-dependent privacy. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 7586–7591, 2016.

[11] Fragkiskos Koufogiannis, Shuo Han, and George J. Pappas. Gradual release of sensitive data under differential privacy, 2015.

[12] Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. The case for learned index structures. *CoRR*, abs/1712.01208, 2017.

[13] N. S. Mangat. An improved randomized response strategy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):93–95, 1994.

[14] Ramesh Raskar Praneeth Vepakomma, Subha Nawar Pushpita. Dams: Meta-estimation of private sketch data structures for differentially private covid-19 contact tracing. October 2020.

[15] Ramesh Raskar, Deepti Pahwa, and Robson Beaudry. Contact tracing: Holistic solution beyond bluetooth. *IEEE Data Eng. Bull.*, 43:67–70, 2020.

[16] Tianhao Wang, Milan Lopuhaä-Zwakenberg, Zitao Li, Boris Skoric, and Ninghui Li. Locally differentially private frequency estimation with consistency. 2020.