

# Private Hierarchical Clustering in Federated Networks

Aashish Kolluri  
aashish7@comp.nus.edu.sg  
School Of Computing  
National University Of Singapore  
Singapore

Teodora Baluta  
teobaluta@comp.nus.edu.sg  
School Of Computing  
National University Of Singapore  
Singapore

Prateek Saxena  
prateeks@comp.nus.edu.sg  
School Of Computing  
National University Of Singapore  
Singapore

## ABSTRACT

Analyzing structural properties of social networks, such as identifying their clusters or finding their central nodes, has many applications. However, these applications are not supported by federated social networks that allow users to store their social contacts locally on their end devices. In the federated regime, users want access to personalized services while also keeping their social contacts private. In this paper, we take a step towards enabling analytics on federated networks with differential privacy guarantees about protecting the user's social contacts. Specifically, we present the first work to compute hierarchical cluster trees using local differential privacy. Our algorithms for computing them are novel and come with theoretical bounds on the quality of the trees learned. Empirically, our differentially private algorithms learn trees that are of comparable quality (with at most about 10% utility loss) to the trees obtained from the non-private algorithms, while having reasonable privacy ( $0.5 \leq \epsilon \leq 2$ ). Private hierarchical cluster trees enable new application setups where a service provider can query the community structure around a target user without having their social contacts. We show the utility of such queries by redesigning two state-of-the-art social recommendation algorithms for the federated social network setup. Our recommendation algorithms significantly outperform the baselines that do not use social contacts.

## CCS CONCEPTS

- Information systems → *Hierarchical data models*; • Security and privacy → *Privacy-preserving protocols; Social network security and privacy*; • Networks → *Network privacy and anonymity*.

## KEYWORDS

Federated social networks; Private hierarchical clustering; Local differential privacy; Social recommendation

### ACM Reference Format:

Aashish Kolluri, Teodora Baluta, and Prateek Saxena. 2021. Private Hierarchical Clustering in Federated Networks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21), November 15–19, 2021, Virtual Event, Republic of Korea*. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3460120.3484822>

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*CCS '21, November 15–19, 2021, Virtual Event, Republic of Korea*

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8454-4/21/11.

<https://doi.org/10.1145/3460120.3484822>

## 1 INTRODUCTION

Millions of users are moving towards more decentralized or federated services due to trust and privacy concerns of centralized data storage [3, 35]. Federated social networks are a popular alternative to the centralized networks as the user's social connections are kept private and stored on user-controlled devices. For instance, Signal, a federated social network, has recently seen tens of millions of users joining the platform after WhatsApp announced a controversial privacy policy update [82]. This movement towards decentralization has incentivized companies to develop techniques to port conventional end applications to the federated setup [2, 11, 72].

Conventional end applications for social and communication networks, such as personalized recommendations and online advertising, require finding similar users on the network that have an influence over a target user [39, 92]. For instance, a user's network neighbors and other users in their close community are known to influence the user's behavior [31]. Therefore, the ability to probe the close community of a target user is important to analyze. In the centralized setup, such queries are trivial because the users' data is stored in the servers of a centralized service provider and thus the whole network is available. However, in the federated setup this network structure is not available to the service provider. Moreover, the users of a service often derive value and benefits from conventional end applications, while expecting a reasonable privacy guarantee on their sensitive data [11]. This new paradigm raises the problem of redesigning conventional applications for the federated setup, which in turn requires supporting queries such as "Who are the users in the close community of a target user?"

It is useful to think of a federated social network as a graph. Users form vertices of the graph and the outgoing links of a user are only kept locally with it. A starting point for answering queries over a federated network that preserve *privacy of the individual links* is the local differential privacy (LDP) framework [73]. The LDP regime eliminates trust in a centralized authority and thus naturally fits the federated setup [29, 54, 66]. In this setup, each user locally adds noise to the query outputs computed on its raw data to preserve privacy before releasing it. Users communicate with an untrusted authority which combines the noisy outputs of the users, possibly with more than one round of communication, to compute the final result. In the LDP model, the noise added by every user in multiple rounds of communication adversely affects the utility of a query. Consequently, queries in the LDP regime have largely been restricted to simple statistical queries such as counts and histograms over categorical, set-valued data, or key-value pairs [21, 25, 32, 37, 95], with only a few exceptions [71]. So, more complex queries that probe the sensitive link or community structure around a target user remain a challenge in the LDP regime.

As our first contribution, we propose learning a well-known data structure called a *hierarchical cluster tree* (HCT) over a federated network of users in the LDP regime. An HCT is a tree representation of the network which clusters similar vertices at various levels. At lower levels, small tightly-connected clusters emerge; and, at higher levels, larger clusters and structural hubs are captured. An HCT naturally captures the communities around a user at various granularities. In the centralized setup, HCTs are commonly used in recommender systems [8, 9, 78, 91], intrusion detection [44], link prediction [49] and even in applications beyond computer science [16, 17, 46, 60]. Learning HCTs with LDP guarantees unlocks such applications in the federated setup. The only existing approach to privately learn HCTs is for the centralized setup, where the central authority is entrusted with the whole network [89]. However, this approach incurs large amounts of noise if it were directly adapted to fit the LDP regime as it requires multiple queries on the network structure. Therefore, we ask: *Can we compute HCTs over federated networks (graphs) with acceptable privacy and utility?*

We present the first algorithm called PRIVACT<sup>1</sup> to learn HCTs over federated graphs in the LDP framework. In the process, we also design a novel randomized algorithm called GENTREE, to compute HCTs in the federated setup without differential privacy. To achieve this, we make the key observation that a recently proposed quality function to measure the quality of HCTs couples well with a known differentially private construct called degree vectors. Our design choices are principled and guided throughout by theoretical utility analysis. Specifically, we show that GENTREE creates HCTs within  $O(\frac{\log n}{n^2})$  approximation error of the ideal HCT in expectation, where  $n$  is the number of vertices in the given federated graph. Its differential private version called PRIVACT has an additive approximation error term bounded by a quantity that depends only on  $n$ , not on the edge structure of the graph. Therefore, once we fix a graph with  $n$  vertices, the expected error (or loss in utility) can be analytically bounded for various choices of the privacy budget  $\epsilon$  [23]. Further, PRIVACT requires querying each user only once.

Finally, we show a concrete application that directly benefits from our private HCTs: social recommendation systems. To alleviate issues such as lack of data when a new user joins the centralized recommendation service, or to increase the recommendation quality, social recommender systems use additional cross-site data such as users' social network [20, 64, 67, 76]. The key insight is to use the private HCTs to identify users in the close communities of a target user. We show that PRIVACT can be readily integrated into both traditional [83] and state-of-the-art [85] algorithms to provide recommendations in a setup where users' privacy is important and they do not trust the central entity with their social contacts. On 3 commonly used social recommendation datasets, we demonstrate that PRIVACT-based private social recommendation algorithms perform 0.9-1000× better than the algorithms that do not use social information across all evaluated settings. Therefore, PRIVACT can be used to improve recommendation quality by utilizing the social information from federated networks at a low privacy budget ( $\epsilon = 1$ ) [12, 55].

To summarize, we claim the following contributions:

- **Conceptual:** We propose the first work that learns hierarchical cluster trees from a federated network in the local differential privacy model to the best of our knowledge.
- **Technical:** We propose two novel algorithms (Section 4) with utility guarantees: 1) GENTREE learns hierarchical cluster trees which are close to the ideal in expectation and 2) PRIVACT learns private hierarchical cluster trees whose utility loss can be bounded for various choices of privacy budget. We evaluate the quality of our learned cluster trees by comparing them with the state-of-the-art solution developed for differential privacy in a centralized setup, showing empirical improvements in the obtained utility of the hierarchical cluster trees (Section 6). Further, the observed difference in empirical utility of our differentially private and non-private versions of our algorithms is small (< 7% at  $\epsilon = 1$ ).
- **End application:** Using PRIVACT-based HCTs, we redesign the state-of-the-art social recommendation algorithms such that users in the federated setup can be served high-quality recommendations without having to share their contacts with the service provider. We evaluate our algorithms on 3 popular social recommendation datasets with a privacy budget  $\epsilon = 1$ . Our algorithms significantly outperform the baselines that do not use social contacts in almost all evaluated settings and in some settings perform on-par with non-private baselines that use raw social contacts (Section 7).

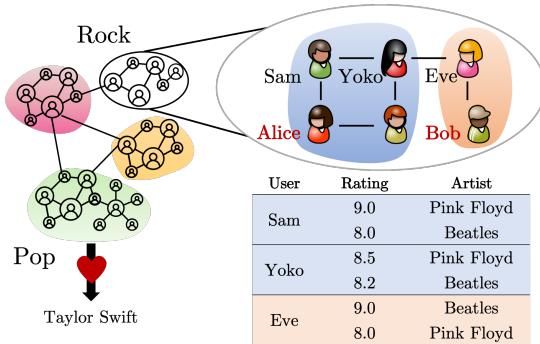
## 2 MOTIVATION & PROBLEM

We illustrate our problem setup in the context of networks where edges are private information such as communication and social networks. The structure of social networks captures influence relationships between users which are instrumental in conventional applications such as serving recommendations [31] and link prediction [50]. At the heart of these applications lies the idea that users' preferences are influenced by other users on the network. A user's influence is projected beyond his immediate neighbors into the broader community and decreases as the distance on the network increases. Ideally, having the network structure allows us to understand a user's preferences based on their close community consisting of their immediate neighbors, one-hop neighbors, and so on. Therefore, the queries that we are specifically interested in answering are: *"Which users are in the close community of a target user?"* where closeness dictates the granularity of exploration [20, 64]. We show how these queries are useful in social recommendations.

### 2.1 Motivating Example

Recommender systems are widely used to help users find interesting products, services, or other users from a large set of choices. These systems are crucial for the user experience and user retention [4]. Despite their success, recommender systems suffer from the problem of data sparsity. For example, when a new user joins the recommendation system there is no history to base the recommendations on (known as the cold start problem [77]). To mitigate this problem, recommender systems benefit from cross-site data such as online social networks [48, 67, 83]. For example, YouTube recommender accuracy was reported to increase when information from Google+ and Twitter networks is used [20] and Amazon's

<sup>1</sup>Our tool is publicly available at <https://github.com/ashgeek/privaCT-public>



**Figure 1: To recommend Alice and Bob an artist, we use the preferences of users in their close communities. Therefore, Alice is recommended Pink Floyd, then Beatles, and vice-versa for Bob. The amount of “closeness” can be set to just first-degree neighbors or to a higher degree depending on the scenario. Both of them are not recommended Taylor Swift, a pop artist popular in another community.**

recommendations were reportedly improved with network information from Cheetah Mobile [64, 68]. These techniques constitute the realm of *social recommendation* [30] and are based on the idea that a user’s preferences are similar to or influenced by their friends and other users in the close community. The reason behind this observation can be explained by social correlation theories [81].

Consider a music streaming service that recommends artists to users for illustration. On the streaming service, each user rates artists from 1 to 10 either explicitly or implicitly by the number of times they listened to that particular artist. Hence, the service stores user’s item preferences on its centralized servers. The recommendation algorithm additionally uses social contacts to make recommendations. The social contacts can be obtained by asking users to share their contacts directly while signing up (e.g., Google or Facebook contacts). In Figure 1, we illustrate such a service where we zoom into the community of users interested in the rock music genre. When new users Alice and Bob sign up for the streaming service, they connect with their friends who belong to this broad community. A good social recommendation algorithm recommends Alice the artists Pink Floyd and the Beatles, in that order, and the same artists for Bob, but in reverse order. The intuition for this is captured by the community structure around these users. Observe that Alice’s close community consisting of her first- and second-degree neighbors has rated Pink Floyd higher than the Beatles. For example, Sam, her direct neighbor, has rated Pink Floyd 9.0 and the Beatles, 8.0. In the real world, this means that Alice’s friends and friends of her friends have rated Pink Floyd higher than the Beatles. Alice’s next closest community consisting of third-degree neighbors (Eve) rates the Beatles higher than Pink Floyd but they still listen to and rate Pink Floyd highly  $> 8.0$ . However, Taylor Swift, a popular artist in another part of the network, is rated lower in this community. Thus, Taylor Swift does not appear in the top recommendations for either Alice or Bob.

Currently, centralized streaming services query for close communities around a target user. In this work, we take a step towards

supporting such queries in the federated setup. This would enable conventional end applications in the federated setup without the need for users to share their contacts.

## 2.2 Problem Setup

A federated social network is a graph  $G : (V, E)$  with  $n = |V|$  vertices. Each vertex in  $V$  corresponds (say) to a separate user who stores the list of its neighbors locally. We assume that the graph is static thus there is no change in the set of vertices or edges. There is an untrusted central *authority* who knows the registered users on the network. However, the central authority does not know the edges of any user. Users want to keep their edges private. This setup is common in the real-world federated social networks [29, 54]. We assume that the users are *honest* and they wish to have good service from the central authority. Therefore, the central authority can query the users on their private edge information and they respond as long as they are guaranteed an acceptable level of privacy.

In this setup, our goal is to enable the aforementioned queries without the central authority knowing the edges of a user. At the same time, the user should not incur a prohibitive computation or communication cost. We turn to a local differential privacy framework as it is widely regarded as a strong privacy guarantee that can be given to a user while allowing the central authority to extract useful information from the user.

*Differential Privacy.* The differential privacy (DP) framework helps to bound the privacy loss incurred by a user by participating in a computation.

*Definition 1 (Differential Privacy).* For any two datasets,  $D$  and  $D'$  such that  $(|D - D'| \leq 1)$ , a randomized query  $M : D \rightarrow S$  satisfies  $\epsilon$ -DP if

$$Pr(M(D) \in s) \leq e^\epsilon Pr(M(D') \in s)$$

Here,  $s \subseteq S$  is any possible subset of all outputs of  $M$ .

The  $\epsilon$  parameter is called the privacy loss and it bounds the ratio of probabilities of observing any chosen output with any two input datasets  $D$  and  $D'$  that differ in one element. The parameter  $\epsilon$  captures the maximum loss in privacy resulting from one run of the algorithm  $M$ . We can see that the lower the value of  $\epsilon$ , the closer the two output probability distributions are, therefore, the higher the privacy. Every time  $M$  is run and its outputs are released the value of  $\epsilon$  increases implying more loss of privacy. However, once the outputs of  $M$  are public then they can be reused many times without losing additional privacy as given by the *post-processing property* [23]. In our setup, the central authority is untrusted which corresponds to the local differential privacy (LDP) regime where each user stores the edge information locally and uses a DP mechanism on it. Therefore, the input dataset for the  $i$ -th vertex in our problem is a binary array  $D_i : \{0, 1\}^n$  with 1 in the  $j$ -th index if an edge exists between vertices  $i$  and  $j$ , and 0 otherwise. Since users do not release their edge information but the identities are known to the central authority, we work with *edge local differential privacy* [71]. In this framework, the privacy of edges (each bit in the  $D_i$ ) of a user ( $v_i$ ) is preserved.

*Enabling Queries in LDP.* Our LDP setup poses a significant challenge to design queries that explore the community around a user.

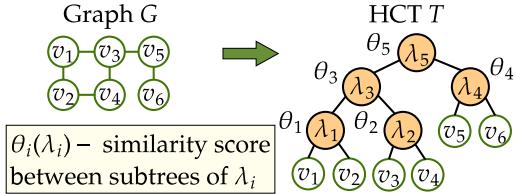


Figure 2: The network from Figure 1 and its candidate HCT.

For instance, even a simple query such as “reporting the immediate contacts” is privacy-sensitive and requires a user to add significant amounts of noise. To illustrate this, consider the non-private answer for a user  $v_i$  which would be the original bit vector  $D_i$ . To preserve LDP, we can use a popular strategy called randomized response for this query [23]. In randomized response, for each bit in  $D_i$  the user reports the original value of  $D_i[j]$  with a high probability, say 70%, and reports the flipped value otherwise. This strategy confers plausible deniability: even if the user  $v_i$  reported that  $v_j$  is their immediate contact, there is a 30% chance that there is in fact no edge between  $v_i$  and  $v_j$ . Observe that by using this strategy a user will claim to have 300,000 contacts for a network of size 1,000,000 even if they have just 70 contacts. Consequently, most of the existing works provide LDP queries for less sensitive numerical queries such as “number of contacts” or aggregate statistics. Our goal is to support more complex queries that ask for the community around a user at various levels of closeness (i.e., second-degree, third-degree, and so on). If the private algorithm requires users to release their data for every degree of closeness, additional noise is required. Therefore, exploring the community around users at different granularity on a federated graph without adding significant noise is challenging.

### 3 HIERARCHICAL CLUSTER TREES

To support complex community queries at different granularities without additional noise, we propose to use a well-known data structure called hierarchical cluster tree (HCT). HCTs are used in many applications [8, 9, 16, 17, 44, 46, 49, 60, 78, 91]. An HCT captures the idea that the network is a large cluster, which recursively splits into smaller clusters of vertices, until only individual vertices remain in each cluster. Thus, one key advantage is that we can learn the *private* HCT of the network once and then publicly release it. Any subsequent query on it for various levels of granularity does not require additional noise, due to the post-processing property.

To illustrate this, in Figure 2 we draw the HCT for the sub-community from our motivating example (Figure 1). Each internal node of the HCT splits the network into two communities, one comprising of the leaves in the left subtree and the other from the right subtree. For instance, at the root the whole network is split into two communities with Alice ( $v_2$ ) in the left one and Bob ( $v_6$ ) on the right. Observe that with the increasing depth of the HCT we get users in closer communities. To avoid confusion, we will henceforth refer to constituents of the HCT as nodes/links and that of the original network as vertices/edges.

Let  $G : (V, E)$  be the original network. An HCT for  $G$  has to group the vertices in  $V$  by a measure of similarity/dissimilarity. Different applications can define different similarity notions. For example, one natural definition is based on neighbor information—two vertices in  $V$  are highly similar if they have many common neighbors in  $G$ . Degree similarity states that two vertices with a larger difference in degrees are more dissimilar. Modularity measures the similarity of two clusters by counting how many more edges between them exist than that predicted in a random graph [62]. It is easy to see that members of an isolated clique in  $G$  will be highly similar to each other by all these definitions. In this work, we design all our techniques using *dissimilarity* scores; however, our techniques can be extended to the analogous measures of similarity easily.

Generically, an HCT ( $T$ ) has two components,  $\Lambda$  and  $\theta$ .  $\Lambda$  is the set of internal nodes and  $\theta$  is a dissimilarity measure defined as a function that maps internal nodes to a real value. Each internal node  $\lambda_i \in \Lambda$  is associated with two subtrees  $L_i$  and  $R_i$ . Since the leaves of each subtree represent the vertices of the original graph, each internal node represents two sets of clusters, corresponding to the left and right subtrees. Therefore  $\theta(\lambda_i)$  can also be thought of as the dissimilarity score between two clusters represented by  $\lambda_i$ . For instance, in Figure 2 the internal node  $\lambda_3$  has two clusters  $\{(v_1, v_2), (v_3, v_4)\}$ . The  $\theta(\lambda_3)$  represents the dissimilarity between the two clusters at  $\lambda_3$ . Intuitively, the clusters at lower levels of the hierarchical structure should be more similar to each other than at the higher levels i.e., closer to the tree root. This can be seen in the graph  $G$  (sub-community) as well, as  $v_1, v_2, v_3, v_4$  are better clustered together than the other two vertices  $v_5, v_6$ .

Therefore, an HCT is meaningful only if the dissimilarity score at an internal node  $\lambda$  is lower than the dissimilarity score at all of its ancestors. We call this property as *ideal clustering property*. Next, we describe traditional algorithms to learn HCTs in the centralized setup with and without differential privacy, and the challenges of extending them to the federated setup with LDP.

#### 3.1 Learning an HCT in the Centralized Setup

In the centralized setup, the authority is trusted and knows the entire graph, i.e., it can query the raw edges in  $G$  directly. The centralized setup offers a reasonable private baseline to compare our eventual solution to learn an HCT in the federated setup.

Even with no DP guarantee, how do we compute  $T$ ? Observe that there are combinatorially many (in  $n$ ) cluster trees possible for  $G$ , since each cluster tree corresponds to a unique way of partitioning the vertices in the graph. The goal is to find a tree that preserves the ideal clustering property, and among those which do, find the one that provides the best clustering at each level. Many traditional methods like average linkage [61], which are natural to use in the centralized setup, are ad-hoc and do not provide any quantitative way of measuring the quality of trees produced. A more systematic way is the algorithm proposed by Clauset et al. [13, 14] which we refer to as the CMN algorithm. This algorithm is one of the most popular for computing hierarchical structures [45, 69, 94], and its DP version is known [89].

CMN defines a *quality function*, which quantifies how good is a cluster tree at clustering similar vertices at each level [14]. For a given graph  $G$  and a probability assignment function  $\pi$ , the quality

of a computed cluster tree  $T$  is as follows:

$$Q_M(T) = \prod_{i=1}^{n-1} (\pi_i)^{E_i} (1 - \pi_i)^{L_i R_i - E_i}, \text{ where } \pi_i = \pi(\lambda_i)$$

$L_i$  and  $R_i$  are the number of leaves in the left and right sub-trees at the  $i$ -th internal node  $\lambda_i$  in  $T$ .  $E_i$  is the number of edges between the two clusters represented by  $\lambda_i$ . The probability function  $\pi$  assigns a probability score at each internal node  $\lambda_i$  which signifies the probability of an edge existing between a leaf (vertex) in the left subtree at  $\lambda_i$  and another vertex in the right. We will refer to edges that have one vertex in the left sub-tree and one in the right sub-tree as edges "crossing the clusters" rooted at  $\lambda_i$ . Subsequently,  $\pi_i$  becomes the probability of an edge crossing the clusters rooted at  $\lambda_i$  for any graph (not necessarily  $G$ ) that can be sampled using  $T$  and  $\pi$ . Therefore,  $Q_M$  also measures the *likelihood* of sampling a given graph using  $T$  and  $\pi$ . A tree that optimizes this quality function given the underlying graph  $G$ , will sample  $G$  that has the *maximum likelihood* among all graphs with  $n$  vertices [13]. Here,  $\pi_i$  is computed using  $\pi_i = \frac{E_i}{L_i R_i}$  as this value maximizes the  $Q_M$ .

Therefore, the CMN algorithms employ the principle of maximum likelihood estimation to find a  $T$  and  $\pi$  conditioned on the given graph  $G$  as evidence. Maximizing  $Q_M$  by enumeratively evaluating it on the space of all possible cluster trees is intractable. Therefore, the CMN algorithm optimizes for  $Q_M$  using a Markov chain Monte Carlo (MCMC) sampling procedure [58]. This sampling procedure is shown to converge in expectation to the desired tree for which  $G$  maximizes the likelihood.

**CMN-DP.** The DP version of this algorithm is the state-of-the-art solution in the computation of differentially private HCTs [89]. This is our *centralized DP* baseline method and we call it CMN-DP. Specifically, the CMN-DP simulates the exponential mechanism of differential privacy [23] by following a similar MCMC procedure to maximize  $Q_M$  and adds noise to the edge counts  $E_i$  after convergence for computing the probabilities  $\pi$ . Details of this algorithm are elided here; we refer interested readers to prior work [89]. The key point is that both the computation of CMN and CMN-DP assume access to the raw edges in  $G$ . We will show how to compute  $T$  in our LDP setup, without direct access to the raw edges.

### 3.2 Challenges: from CMN-DP to LDP

The quality function  $Q_M$  depends on the probability assignment function  $\pi$  to measure the "closeness" between two clusters and  $\pi$  uses fine-grained private information such as computing the exact number of interconnecting edges between the two clusters. To compute this fine-grained information in the federated setup, the users from one of the clusters can be asked to report the counts of their neighbors in the other cluster after adequately noising them for satisfying LDP. For instance, in Figure 2, the number of edges ( $E_3$ ) crossing the internal node  $\lambda_3$  can be calculated using the edge counts reported by users  $v_1, v_2, v_3, v_4$  that cross  $\lambda_3$ . However, searching for an optimal  $T$  that optimizes for  $Q_M$  requires computing the interconnecting edges for all possible sets of clusters in the worst case. Concretely, CMN-DP implements an iterative search procedure that queries a user for edge information *thousands* of times before converging to the optimal  $T$  (see Section 6). While this search works in the centralized setup where the graph is available,

it is not feasible in the federated setup for two reasons. First, answering every differentially private query leads to a privacy loss, and over many such queries the aggregated privacy loss will be high [23]. To exemplify, a thousand queries with an  $\epsilon = 0.1$  for each query will lead to an aggregated epsilon  $\epsilon > 20$  with 99% probability, as given by the advanced composition theorem [23]. We consider  $\epsilon \leq 2$  as reasonable [38]<sup>2</sup>, although prior works have considered up to  $\epsilon = 8$  [1]. Second, the search procedure for finding the optimal  $T$  is iterative so the users have to be available for many iterations of the search to compute the closeness.

## 4 OUR SOLUTION

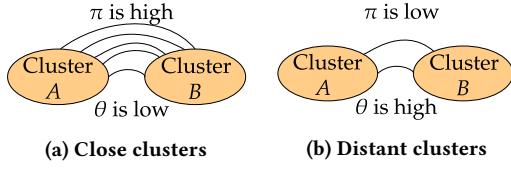
Our key ideas to tackle the aforementioned challenges are two-fold. First, instead of using a fine-grained probability assignment function  $\pi$ , we propose a coarse-grained method to compute the closeness between two clusters such that the users need not be queried repeatedly for their neighbors. Second, we propose to replace the quality function  $Q_M$  with another quality function that can work with any coarse-grained method that captures the closeness between two clusters. These two observations enable us to design a novel hierarchical clustering algorithm which only queries the users once and the rest of the iterative search for an optimal tree happens on the server side. We detail our insights next.

### 4.1 Key Insights

Our first insight is to use an easy to compute coarse-grained approximation for closeness that can be obtained for a low privacy budget. We start from a construct called degree vectors, previously proposed in the LDP setup [71]. The degree vectors are a generalized version of degree counts, wherein the vertices are randomly partitioned into  $K$  bins of almost same size, and each user is asked to report how many neighbors it has in each bin. For  $K = 1$  this degree vector has one element and it yields just the degrees of the vertices in  $G$ . For  $K = n$ , all nodes would have a unique bin, therefore the degree vectors will encode the original neighbor list for all vertices. For any  $1 < K < n$  the idea is to preserve more edge information than just degrees and less than the exact edges. Then, the degree vector is noised by the user; a random Laplacian noise  $Lap(0, \frac{1}{\epsilon})$  is added to each bin count before sending it to the untrusted authority. The  $K$  is small compared to the network size, typically  $\leq \log n$  or a small constant [71]. This is intuitively good, since the noise added to the degree vector is proportional to  $K$ . Now we ask: What can we compute with degree vectors?

It is not straightforward to compute  $Q_M$  using degree vectors. Nevertheless, observe that if we take two close clusters with respect to  $\pi$ , then on average each user in the left cluster has a lot of neighbors in the right cluster and vice versa. This notion is readily captured by degree vectors. It is very likely for two neighboring vertices (say friends) to have a similar set of common neighbors (common friends) hence similar degree vectors. Consequently, if we measure the dissimilarity as a  $L_1$ -norm between their degree vectors then the dissimilarity for such vertices should be low. Therefore, the average dissimilarity across vertices of two *close* clusters will be lower than for two far clusters. We illustrate our intuition in

<sup>2</sup>Usually, an  $\epsilon = 2$  allows an attacker to infer a random bit of the training sample (in this case an edge for each user) with 86% probability.



**Figure 3:** (a) Two close clusters, high  $\pi$ , have more interconnecting edges than two distant ones. (b) Average dissimilarity  $\theta$  using  $L_1$ -norm of degree vectors also captures closeness.

Figure 3. Figure 3a shows two clusters with high interconnecting edges, hence the  $\pi$  associated with the two is high. Consequently, the average  $L_1$ -norm distance will be less. A similar argument can be made for Figure 3b. Therefore, we propose using the average  $L_1$ -norm distance using degree vectors as a coarse-grained replacement for  $\pi$  to measure the closeness between clusters. Observe that the degree vectors can be constructed just once for each user which can then be used to compute the average  $L_1$ -norm distance for any two clusters. This key insight allows us to propose a different quality function that optimizes for the average  $L_1$ -norm distance.

We point to a recently introduced quality function<sup>3</sup> function by Dasgupta [18]. Dasgupta’s quality function takes a dissimilarity matrix and a tree as inputs and measures the quality of the tree for that dissimilarity matrix. Our first observation is that this quality function does not need the fine-grained edge counts between clusters (like  $Q_M$  does). The second and even more important observation is that a tree that optimizes Dasgupta’s quality function has a specific dissimilarity measure  $\theta$  (defined in Section 3). The measure represents the average dissimilarity between the vertices of two clusters computed using the dissimilarity matrix. In fact, we formally show that such a tree satisfies the ideal clustering property as well (Section 4.2). Thus, if we use the  $L_1$ -norm of degree vectors to compute the dissimilarity matrix, then by optimizing for the Dasgupta’s quality function we get our desired coarse-grained average  $L_1$ -norm distance as the  $\theta$ . So, we directly arrive at a tree which minimizes the distance between the close clusters and maximizes the distance between the far ones.

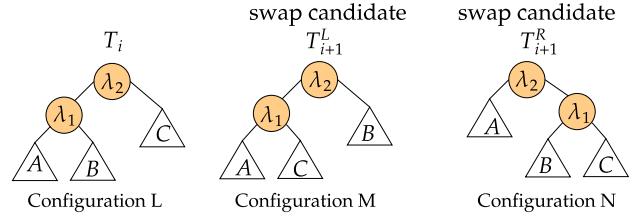
Using these insights, we design a novel randomized algorithm to sample a differentially private HCT that optimizes Dasgupta’s quality function after querying a degree vector from each user. To summarize, we have shown a way to avoid multiple privacy-violating queries in learning an HCT by designing another learning strategy that allows us to replace the fine-grained queries with a coarse-grained one that preserves the ideal clustering property.

## 4.2 Dasgupta’s Quality Function

We now formulate the Dasgupta’s quality function considering only full binary trees, and when a dissimilarity matrix is given as input.

*Definition 2 (Dasgupta’s Quality Function ( $Q_D$ )).* The quality of a tree  $T : (\Lambda, \theta)$  with respect to a graph  $G$  with a non-negative

<sup>3</sup>Originally, it was introduced as a “cost” function and it used a similarity matrix. That cost function has to be minimized for producing better trees. Here we use dissimilarities, therefore, the function becomes a quality function that has to be maximized.



**Figure 4:** The possible swaps of a tree.

dissimilarity matrix  $S$  is given by

$$Q_D(T) = \sum_{x \in V, y \in V} S(x, y) \cdot |\text{leaves}(T[x \vee y])|$$

where  $T[x \vee y]$  is least common ancestor of  $x, y$ .

The expression is a weighted sum of dissimilarities of each pair of vertices, weighted by the number of leaves in the subtree of the least common ancestor of the pair of vertices. The idea of maximizing  $Q_D$  is intuitive. Any pair of vertices  $(x, y)$  which are highly dissimilar to each other will have a high dissimilarity score  $S(x, y)$  and therefore, their least common ancestor node (which we denote by  $T[x \vee y]$ ) should have more leaves in order to maximize the product  $S(x, y) \cdot |\text{leaves}(T[x \vee y])|$ . It has been shown that an optimal tree with respect to  $Q_D$  has several nice properties such as: 1) Two disconnected components in the graph will be separated completely into two different clusters; 2) The quality of every tree for a clique is the same; and, 3) it represents the clusters well in the planted partition model [18]. Property 2) is the most relevant to us, as we will use it to bound the utility loss.

Originally, the quality function was designed for a non-negative similarity matrix with no restriction on the tree space. In the dissimilarity case, if we allow all possible trees then there is always a trivial tree that maximizes this function, i.e., the star graph with only one internal root node. Therefore we restrict the space of trees to optimize over to only full-binary trees. All the above properties continue to hold in the dissimilarity case when the tree space is restricted to the full-binary trees [15].

Note that the quality function does not require any edge information of the graph but only a dissimilarity matrix, a key requirement for the design of our algorithm in the decentralized setting. As stated previously, we show that the tree  $T$  that maximizes  $Q_D$  also satisfies the ideal clustering property. Further, the  $\theta$  of such a tree will just be the average dissimilarity between the nodes of two subtrees at each internal node  $\lambda$ . Formally, Theorem 3 captures both these observations.

**Theorem 3.** Given a graph  $G : (V, E)$  and a dissimilarity matrix  $S$ , the tree  $T^{OPT} : (\Lambda, \theta)$  that maximizes  $Q_D$  preserves the ideal clustering property with

$$\theta(\lambda) = \frac{\sum_{x \in L, y \in R} S(x, y)}{|L| \cdot |R|}$$

where  $L, R$  are the left and right subtrees of  $\lambda$ .

The full proof of Theorem 3 is provided in the Appendix A.1. It is a proof by contradiction. Assuming that there exists a  $Q_D$ -optimal

---

**Algorithm 1:** Outline for GENTREE

---

**Input:** Dissimilarities matrix  $S$ , max iterations  $I$   
**Output:** Hierarchical cluster tree  $T^*$

```

1 Randomly sample a tree  $T_0$ ;
2  $i = 0$ ;
3 while not converged and  $i < I$  do
4   Sample an internal node  $\lambda_i$  from  $T_i$ ;
5   Construct two local swap trees  $T_{i+1}^L, T_{i+1}^R$ ;
6   Pick one of them at random, call it  $T_{i+1}$ ;
7   Transition to  $T_{i+1}$  with probability
       $\min(1, e^{Q_D(T_{i+1}) - Q_D(T_i)})$ ;
8   if not Transitioned then
9     |  $T_{i+1} = T_i$ ;
10  end
11   $T^* = T_{i+1}$ ;
12   $i = i + 1$ ;
13 end
14 return  $T^*$ 

```

---

tree  $T$  with two internal nodes  $(\lambda_1, \lambda_2)$  such that  $\lambda_2 = \text{ancestor}(\lambda_1)$  and  $\theta(\lambda_1) > \theta(\lambda_2)$ , we can always construct another tree  $T'$  with a higher quality than  $T$ . In fact, if  $T$  were the configuration  $L$  in Figure 4, then  $T'$  would be one of the other configurations.

Theorem 3 constitutes our first key analytical result. It explains why we choose Dasgupta’s quality function and the  $L_1$ -norm of the degree vectors together. Next, we describe our algorithm GENTREE to learn a tree that maximizes this quality function. GENTREE is an independent contribution of this work which can be used with any dissimilarity metric that captures closeness between clusters.

### 4.3 The GENTREE Algorithm

Finding a hierarchical cluster tree that maximizes  $Q_D$  is known to be NP-hard [18]. Let  $\mathbb{T}$  be the set of all possible full binary trees that have  $V$  as leaves. We want to find the optimal tree  $T^{OPT} \in \mathbb{T}$  that maximizes  $Q_D$ . The number of possible trees including all the permutations from internal nodes to leaves could be intractably large to enumerate.

We propose a randomized algorithm GENTREE which will sample from the distribution of all possible trees such that the sample probability is exponential in the quality of the tree as measured by  $Q_D$ . Therefore, if the quality of the tree is high then the sample probability is exponentially high. This distribution is known as Boltzmann (Gibbs) distribution and we choose it since it enables bounding the utility loss as we show later. Hence, GENTREE samples a tree  $T' \in \mathbb{T}$  with probability

$$Pr(T' \in \mathbb{T}) = \frac{e^{Q_D(T')}}{\sum_{T'' \in \mathbb{T}} e^{Q_D(T'')}}$$

GENTREE creates samples from this distribution by using a Metropolis-Hastings (MH) algorithm based on a Markov chain with states as trees and state-transition probabilities as the ratio of qualities between the states. The outline for GENTREE is similar to the CMN algorithm in the centralized setup and is given in Algorithm 1. First, it starts with a randomly sampled tree  $T_0$ . In each iteration,

---

**Algorithm 2:** Outline for PRIVACT

---

**Input:** Vertices  $V$ , dissimilarities matrix  $S$   
**Output:** Private Hierarchical cluster tree  $T_{dp}^*$

```

1 Aggregator: Randomly partition  $V$  to  $K = \lfloor \log n \rfloor$  bins;
2 Aggregator: Show the  $K$  partitions to the user;
3 User: Send DP degree vectors with  $\text{Lap}(0, \frac{1}{\epsilon})$  noise;
4 Aggregator: Compute dissimilarities  $S$  using  $L_1$ -norm;
5 Aggregator: Compute  $T_{dp}^* = \text{GENTREE}(S)$  and release  $T_{dp}^*$ 

```

---

it samples a random internal node  $\lambda_i$  and does a local swap into one of the configurations as shown in Figure 4. Observe that in a local swap, a subtree at the chosen internal node is detached and swapped with the subtree at the parent internal node leading to two possible local swaps. This state transition (swap) between the trees  $T_i, T_{i+1}$  is done with a probability  $\min(1, e^{Q_D(T_{i+1}) - Q_D(T_i)})$  which is the ratio of the probabilities of sampling the respective trees in the stationary distribution. This follows the standard MH procedure.

Our chosen stationary distribution ensures that every state (tree) has a positive probability of being sampled. Further, every full binary tree can be obtained from another by a sequence of swaps therefore the entire chain is connected. Consequently, a standard analysis [58] leads to the conclusion that the Markov chain induced by GENTREE is ergodic and reversible. Therefore, GENTREE will converge to its stationary distribution, which in our case is the Boltzmann distribution.

The convergence criteria used by the prior works, for their MCMC algorithms [14, 89], heuristically compares the average quality of sampled HCTs over an interval of 65536 iterations to the average in the previous interval. The algorithms stop when the average quality ceases to increase across intervals. Using this criteria, CMN-DP is reported to converge in  $1000 \times n$  iterations [89]. Therefore, we also run our algorithms for  $1000 \times n$  iterations. We find that it is sufficient for convergence of our algorithms on all the evaluated datasets (see Figure 5).

### 4.4 The PRIVACT Algorithm

GENTREE only requires a dissimilarity matrix to operate on. Therefore, a differentially private hierarchical cluster tree can be learned using GENTREE if the input dissimilarity matrix is also differentially private. Recall that any computation on a differentially private output is also differentially private using the post-processing property. Using the same property, we can construct a private dissimilarity matrix from private degree vectors. For that purpose, the central aggregator first sends a random partition of the users, with each bin having a set of users, to every user. Using these bins, every user constructs a private degree vector by counting their neighbors in each bin and adding Laplacian noise  $\text{Lap}(0, \frac{1}{\epsilon})$  to these counts. Finally, the users send their vectors to the aggregator so that the dissimilarity of every pair of users can be computed as measured by the  $L_1$ -norm of their respective degree vectors. The outline of PRIVACT is summarized in Algorithm 2. Notice that the aggregator only requires the users to compute degree vectors privately so that GENTREE can be run on the server side to compute a private HCT.

*Communication cost.* Our protocol has only 2 rounds. To each user, the server sends a mapping between  $n$  user ids and  $K$  partition (bin) ids. Each user maps to one bin. The size of one user id will be  $\log n$  bits and the size of each partition id will be  $\log K$  bits. Therefore, the number of bits sent by the server to each user is  $O(n(\log n + \log K))$ . Each user then replies to the server with a degree vector which constitutes  $K$  counts, one for each bin. Each count may have a maximum size of  $\log n$  bits since the count can be as high as  $n$ . Therefore, a user sends  $O(K \log n)$  bits to the server.

## 5 THEORETICAL BOUND ON UTILITY LOSS

GENTREE and PRIVACT are randomized algorithms. So, we need to show that they learn trees that are close to the ideal tree that optimizes  $Q_{\mathcal{D}}$  given the dissimilarity matrix computed using non differentially private degree vectors. The utility loss is given by the quality difference between the ideal tree and the trees sampled by our algorithms. Therefore, we bound the following:

- (1) GENTREE: The utility loss for the tree output by the algorithm when dissimilarity matrix is not noised.
- (2) PRIVACT: The utility loss for the tree output by the algorithm when the dissimilarity matrix is noised.

For ease of analysis, we enforce the dissimilarity between two vertices of the graph to be at least  $1^4$ , i.e., the  $L_1$ -norm between two degree vectors is at least 1 instead of 0. Therefore, a clique will have a dissimilarity of 1 for all pairs of nodes. Following this, we state a fact that is proved in the original Dasgupta's work [18].

**Theorem 4.** (Dasgupta [18]) The quality of a clique with all dissimilarities equal to 1 is same for all possible trees and is given by

$$Q_{\mathcal{D}}(T_{clique}) = \frac{n^3 - n}{3}$$

This is the least maximum quality tree that can be produced with any dissimilarity matrix under our assumption, i.e.,  $S(v_i, v_j) \geq 1$ . For real-world networks, the optimal tree quality can be many times higher than  $Q_{\mathcal{D}}(T_{clique})$  depending on the dissimilarity matrix. Theorem 4 is useful in our analysis presented later and we will use  $\rho$  to represent  $Q_{\mathcal{D}}(T_{clique})$  for brevity from here on.

### 5.1 Utility Loss: GENTREE

There is always an ideal full binary tree that maximizes the Dasgupta quality function given the dissimilarity matrix, say  $T^{OPT}$  with  $OPT = Q_{\mathcal{D}}(T^{OPT})$ . GENTREE is a randomized algorithm that samples a different tree  $T$  at convergence in every run. Therefore, the expected utility loss is the difference between  $OPT$  and the expected value of the quality of the obtained tree  $\mathbb{E}_T[Q_{\mathcal{D}}(T)]$ . We now show that the expected  $Q_{\mathcal{D}}$  quality of the sampled tree at convergence is close to  $OPT$ . The expected quality of a sampled tree is less than  $OPT$  only by at most a small factor  $\frac{c \cdot \log n}{n^2 - 1}$  of the  $OPT$ , where  $c \ll n$ . The complete proofs for the theorems are provided in the Appendix A.1. Here, we explain the key ideas used in them.

<sup>4</sup>This condition is a technicality for obtaining a multiplicative factor in the utility bound. We enforce this restriction in our implementation for correctness. We observe that in our evaluated datasets, only 2 in a million possible pairs have dissimilarity lower than 1. Thus, our technical restriction is not a significant practical concern.

**Theorem 5.** GENTREE outputs a tree  $T$  whose expected quality  $\mathbb{E}_T[Q_{\mathcal{D}}(T)]$  is a  $(1 - \frac{\log n}{n^2 - 1})$  multiplicative factor of  $OPT$ .

$$\mathbb{E}_T[Q_{\mathcal{D}}(T)] \geq \left(1 - c \cdot \frac{\log n}{n^2 - 1}\right) \cdot OPT$$

To prove this, we first show that the probability of sampling a sub-optimal tree is exponentially decreasing.

**Lemma 5.1.** The probability of sampling a tree  $T$  with quality  $OPT - c' \cdot n \cdot \log n$  decreases exponentially in  $n$  (where constant  $c' \ll n$ ).

$$\Pr(Q_{\mathcal{D}}(T) \leq OPT - c' \cdot n \cdot \log n) \leq e^{-n \cdot \log n}$$

Lemma 5.1 implies Theorem 5 and this is a consequence of choosing Boltzman distribution as the stationary distribution for GENTREE. The probability of sampling any sub-optimal tree exponentially reduces with its quality distance from  $OPT$ .

### 5.2 Utility Loss: PRIVACT

PRIVACT samples trees based on the probabilities that are exponential in the quality of tree computed on the differentially private dissimilarities. Hence, the probability of sampling a tree  $T$  will now depend on the noisy quality, say  $g(T) = \bar{Q}_{\mathcal{D}}(T)$  instead of the actual quality  $f(T) = Q_{\mathcal{D}}(T)$ . Let the variables  $S$  and  $\bar{S}$  be the dissimilarity and noisy dissimilarity matrices that store dissimilarities between vertices of the network. Recall that the dissimilarity between two vertices is computed as  $L_1$ -norm of their degree vector counts. The quality computed with the original dissimilarity matrix is  $f(T) = \sum_{ij} S(v_i, v_j) \cdot \text{leaves}(T[v_i \vee v_j])$  whilst the quality computed by the noisy dissimilarity matrix is  $g(T) = \sum_{ij} \bar{S}(v_i, v_j) \cdot \text{leaves}(T[v_i \vee v_j])$ . If  $T_{dp}$  is the sampled tree at convergence then the expected utility loss is computed by taking the difference between  $OPT$  and  $\mathbb{E}_{T_{dp}}[f(T_{dp})]$ . In essence, we are saying that the tree sampled with differentially private dissimilarity matrix should not be too far away from the optimal tree, with respect to the quality computed using the non-noised dissimilarities. The next theorem states this formally.

**Theorem 6.** Let  $T_{dp}$  be the output of PRIVACT and  $\rho$  be the least maximum quality of any hierarchical cluster tree. The expected utility loss of PRIVACT with a high probability ( $> 8/9$ ) is given by

$$OPT - \mathbb{E}_{T_{dp}}[f(T_{dp})] \leq \frac{2K}{\epsilon} \cdot \left(\frac{3}{2} + \frac{6}{\sqrt{K}}\right) \cdot \rho$$

In order to prove the bound, we start by first bounding the expected value of  $g(T)$  over the randomness in degree vectors, in terms of  $f(T)$ . Let  $R_i$  denote the random variable that represents the Laplacian noise added by  $i^{th}$  vertex to its degree vectors and  $R = (R_1, R_2, \dots, R_n)$ . The next lemma bounds the expectation  $|\mathbb{E}_R[g(T)] - f(T)|$ .

**Lemma 6.1.** The expected value of the noisy quality  $g(T)$  over randomness  $R$  is bounded in terms of the actual quality  $f(T)$  by

$$|\mathbb{E}_R[g(T)] - f(T)| \leq \frac{3K}{2\epsilon} \cdot \rho$$

Then we bound the variance of  $g(T)$ .

**Table 1: Dataset graph statistics**

Network	Domain	#Nodes	#Edges	Density
<i>lastfm</i>	social graph	1843	12668	0.0075
<i>delicious</i>	social graph	1503	6350	0.0056
<i>douban</i>	social graph	2848	25185	0.0062

**Lemma 6.2.** *The variance of the noisy quality  $g(T)$  is bounded by*

$$\text{Var}_R[g(T)] \leq \frac{4K}{\epsilon^2} \cdot \rho^2$$

Finally, we use the Chebyshev's inequality to bound the value of  $g(T)$  in terms of  $f(T)$ .

**Lemma 6.3.** *The noisy quality  $g(T)$  is in the interval  $[f(T) - P, f(T) + P]$  with a high probability ( $> 8/9$ ) where*

$$P = \frac{K}{\epsilon} \cdot \left( \frac{3}{2} + \frac{6}{\sqrt{K}} \right) \cdot \rho$$

We then use this lemma to prove Theorem 6. Note that the utility loss does not depend on the dissimilarity matrix but only depends on the parameters  $K$  and  $\epsilon$ . It can be seen that if  $\epsilon$  is very high then the utility loss converges to that of the GENTREE and even if  $\epsilon$  tends to zero, our artificial bounding of the dissimilarities to be greater than 1 will ensure that the final tree quality never goes below  $\rho$ . Recall that  $OPT$  is several times more than  $\rho$  for the real-world graphs as it depends on the dissimilarity matrix, i.e., the structure of the graph. We empirically observe that the utility loss is smaller than our bound for all of our evaluated graphs (see Section 6).

## 6 EVALUATION: QUALITY OF PRIVATE HCT

In this section, our goal is to evaluate the utility of PRIVACT as measured by the quality of the HCTs it produces.

First, we want to empirically measure the utility loss due to adding noise in PRIVACT. For that purpose, we compare the quality of the private trees produced by PRIVACT with the non-private trees generated by GENTREE. The quality is measured using the Dasgupta quality function  $Q_{\mathcal{D}}$ .

Second, we want to evaluate the utility-privacy trade-off of the LDP regime vs. the centralized DP setup. Thus, we compare the private trees that PRIVACT generates with the trees generated by the centralized DP algorithm (CMN-DP) for the same privacy budget. In order to compare these trees, we use the quality function  $Q_M$  which measures the closeness between two clusters using the fine-grained edge counts as opposed to our newly proposed coarse-grained degree vectors (see Section 3.1)—we do so to show how our algorithms perform on the quality criterion proposed by previous work instead of ours. In summary, we evaluate the following:

**(EQ1)** What is the empirical utility loss of the private HCTs produced by PRIVACT vs. the non-DP HCTs produced by GENTREE?

**(EQ2)** What is the quality of the private HCTs produced by PRIVACT vs. the centralized DP algorithm CMN-DP?

*Experimental Setup.* We use 3 real-world networks which are commonly used for evaluating recommender systems [53, 71, 78, 85]<sup>5</sup>. We detail the networks in Table 1. We use privacy budget

<sup>5</sup><https://grouplens.org/datasets/hetrec-2011/>

**Table 2: The quality of the HCTs is evaluated using the Dasgupta quality function. The empirical utility loss of the private trees that PRIVACT generates is computed relative to the non-private trees generated by GENTREE and it is at most 10.87% for  $\epsilon = 0.5$ .**

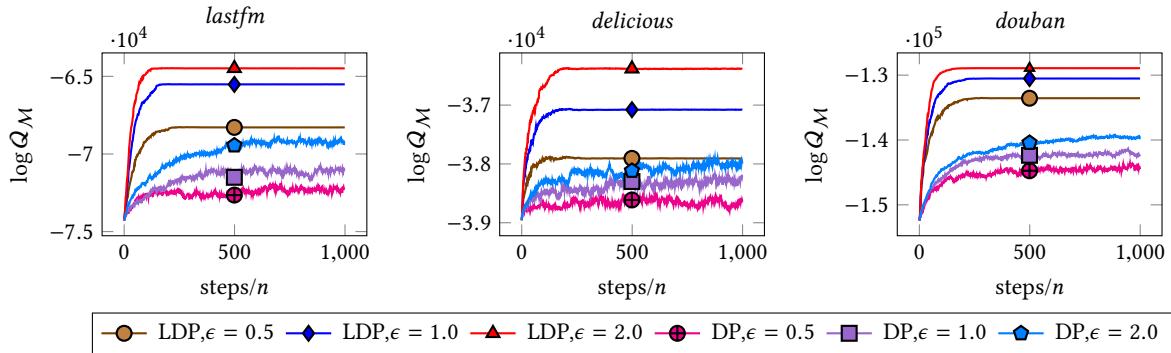
Network	$\epsilon$	Relative Utility	Empirical Utility Loss (%)
<i>lastfm</i>	0.5		9.57
	1.0	22.82	4.05
	2.0		1.45
<i>delicious</i>	0.5		10.87
	1.0	12.20	6.61
	2.0		3.29
<i>douban</i>	0.5		7.09
	1.0	28.83	3.27
	2.0		1.14

values of  $\epsilon \in \{0.5, 1.0, 2.0\}$  as they have been used in the prior work for both centralized as well as the local differential privacy setup [1, 36, 89]. We have two tunable parameters: the number of bins  $K$  and the convergence criteria for MCMC in GENTREE. The  $K$  is chosen such that it minimizes the noise in the final degree vector as well as it minimizes the number of collisions in the degree vector calculation. A higher  $K$  minimizes the collisions but incurs more noise on aggregate and vice versa. Previous work has shown that the value of  $K$  is usually small and it depends on the graph structure. We therefore heuristically choose  $K = \lfloor \log n \rfloor$  as it is not too low to not capture any edge information (see Section 4.1) and not too high to end up adding a lot of noise (see Section 2.2). If we substitute this  $K$  value in our utility bound for PRIVACT in Theorem 6, we get a loss scaling with  $\frac{\log n}{n \cdot \epsilon}$ . Further, our values of  $K$  agree with the ones used in prior work [71].

**(EQ1) Quality of LDP vs. non-DP HCTs.** We show that the observed quality of the differentially private tree generated by PRIVACT, for all values of  $\epsilon$  is very close to the quality of the tree generated by the non-DP version of the algorithm GENTREE.

We define the *empirical utility loss* as  $\frac{|Q_{\mathcal{D}}(T_{dp}) - Q_{\mathcal{D}}(T)|}{Q_{\mathcal{D}}(T)}$ , where  $Q_{\mathcal{D}}(T_{dp})$  is the quality of the private HCT and  $Q_{\mathcal{D}}(T)$  is the quality of the non-private tree. In Table 2, we observe that the empirical utility loss is less than 10.87% for all networks and values of  $\epsilon$  we evaluated. The empirical utility loss is approximately 1.14 – 3.29% on average for  $\epsilon = 2.0$ . Since the empirical utility loss is small, we empirically confirm that the quality of our private tree ( $T_{dp}$ ) is close to that of the non-private tree ( $T$ ).

We also compute the *relative utility* of the non-private tree  $T$  as  $\frac{Q_{\mathcal{D}}(T)}{\rho}$ . When relative utility is high, our non-private trees generated by GENTREE have quality higher than the  $\rho$  (the least maximum quality tree) which is used in Theorem 6 to theoretically bound the utility loss. The higher the relative utility, the better it is since the upper bound on the expected utility loss will be lower. We experimentally observe that the higher the relative utility, the lower the empirical utility loss is (see Table 2). Finally, we observe that the



**Figure 5: The PRIVACT (marked as LDP) and the CMN-DP (marked as DP) evaluated with  $\log Q_M$  at each step of the MCMC process for 3 values of  $\epsilon \in \{0.5, 1.0, 2.0\}$ . Higher values of  $\log Q_M$  signify smaller loss, so better trees. PRIVACT produces trees with a better  $Q_M$  quality than the baseline CMN-DP for all values. We observe that PRIVACT converges within 400 MCMC steps.**

empirical utility loss is within the theoretical bounds implied by Theorem 6. Tightening the bounds remains promising future work.

*Result 1: The empirical utility loss between the private trees produced by PRIVACT and non-private trees generated by GENTREE is less than 10.87% for all the graphs and  $\epsilon$  values evaluated.*

**(EQ2) Quality of LDP vs. centralized DP HCTs.** Our main result is that our LDP algorithm produces trees that have a better utility than the ones produced by the centralized DP algorithm as measured by the baseline  $Q_M$  quality function. In Figure 5, we show that the  $\log Q_M$  of the trees generated by PRIVACT have better utility than trees generated by CMN-DP for all evaluated configurations.

*Result 2: The quality of the trees (measured using  $Q_M$ ) generated by PRIVACT is better than the trees generated by the baseline CMN-DP for all evaluated graphs and  $\epsilon$  values.*

## 7 APPLICATION: RECOMMENDER SYSTEMS

We revisit the motivating example of social recommendation from Section 2 and ask “Can PRIVACT be utilized for providing recommendations in the federated setup?” The goal is to accurately predict the top- $k$  recommendations for the target users. The target users share their preferences/tastes with the recommender system but do not want to share their private contacts. Therefore, this naturally falls in the edge-LDP setup.

We consider two scenarios: cold start and existing users. Recall that the cold-start scenario occurs when a new user with no history of interactions with the service provider has to be served recommendations. In contrast, existing users have a history of interactions that convey their interests. Specifically, we evaluate the utility of the PRIVACT algorithm for social recommendations by asking the following questions in both scenarios:

- Are PRIVACT-based top- $k$  recommendations better than the non-social baselines?
- Are PRIVACT-based top- $k$  recommendations comparable to the non-private social baselines?

**Datasets.** We use the same datasets mentioned in Section 6 as they are commonly used for social recommendations. For *lastfm* and *delicious* which are implicit feedback ratings we normalize the weight of each user-artist and user-bookmark link by dividing the maximum number of plays/bookmarks of that user across all artists/bookmarks. For *douban* with explicit feedback ratings we normalize the ratings by dividing with the maximum rating (5.0).

**Experimental Setup.** We split the datasets into training and test sets. We perform k-fold cross-validation for  $k = 5$  splits of the datasets. For the HCT computation, we use a privacy budget of  $\epsilon = 1$  and we average over HCTs computed with 3 random seeds.

**Evaluation Metrics.** To evaluate our recommendations, we use mean average precision (MAP) and normalized discounted cumulative gain (NDCG). These two metrics are widely used to evaluate the quality of the top- $k$  recommendations [48, 85]. Higher values of these metrics imply better recommendations. Due to space constraints, we explain how these are computed in Appendix A.2.

Next, we describe our baseline and PRIVACT-based algorithm, and then present our results on our social recommendation datasets.

### 7.1 Collaborative Filtering: Setup & Baseline

We implement a widely used technique for recommender systems called collaborative filtering (CF) [74, 75]. In particular, we consider memory-based CF techniques that operate on user-item rating matrices to make top- $k$  recommendations. We choose this technique because it yields good performance and forms the basis of real-world recommender systems in Netflix [74] and Facebook [24].

**CF for cold start.** In this scenario, the target users have no history of rated items. We choose two popular strategies: “recommend highest rated items” (*itemAvg*) and “recommend highest rated items among your friends” (*friendsCF*). Notice that these two methods correspond to two different granularities of the “close” community. The *itemAvg* strategy considers all users in the network whereas the *friendsCF* strategy takes only the neighbors of the target user.

Let  $S : U \times U \rightarrow \{0, 1\}$  be the social relation function where  $S(u', u) = 1$  if the users  $(u', u)$  are friends and 0 otherwise. In

the *friendsCF* strategy, the predicted rating for a target user  $u$  is computed as follows:

$$r_{u,i} = \bar{r}_u + \frac{\sum_{u' \in \mathcal{U}} S(u', u)(r_{u',i} - \bar{r}_{u'})}{\sum_{u' \in \mathcal{U}} S(u', u)}$$

where  $\bar{r}_u = \frac{\sum_{i \in \mathcal{I}} r_{u,i}}{|\mathcal{I}|}$  represents the average item rating for user  $u$ ,  $\mathcal{I}$  is the set of items with a rating from user  $u'$  and  $\mathcal{U} \subseteq U$  is the set of users that have a rating for item  $i$ . For the *itemAvg* strategy, we just replace  $S(u', u)$  with 1 for all  $u' \in \mathcal{U}$ . Note that the *itemAvg* strategy is a private baseline as no one shares their social relations whereas the *friendsCF* strategy is a non-private baseline.

*CF for existing users.* When the user is present on the platform, we can leverage both their ratings and social connections to make better recommendations. In this case, we use one of the state-of-the-art social recommendation algorithm called SERec [85]. SERec first models the user exposures to the items and then uses these exposures to guide the recommendation. The exposures are computed based on the social information rather than the rating history. Due to space constraints, we refer the reader to the paper for more details. We use a popular Python implementation<sup>6</sup> which follows the original implementation. This represents the non-private baseline for social recommendations for existing users. The performance of our non-private baselines in the cold start setup and in the existing users setup is shown in Tables 4 and 3, respectively.

## 7.2 Collaborative Filtering with PRIVACT

We explain how to recommend items to target users based on their communities using collaborative filtering. Observe that the community to which a target user belongs can be queried from an HCT. Recall that the clusters at the lower levels of the HCT form the close community around the target user, due to the ideal clustering property. We can thus query the HCT for “Who are the  $m$  nearest neighbors for user  $u$ ?”<sup>7</sup>. The vertices corresponding to the immediate parent of the leaf node for  $u$  in the HCT constitute the closest community of  $u$ . The parent of that parent node defines the second-closest community, and so on. Therefore, the query for “ $m$  closest vertices” returns a set  $N(u)$  which has users from the closest community to  $u$ , the second-closest community, and so on.

To do the collaborative filtering, we simply replace the non-private algorithms used in our baselines of Section 7.1 with our differentially private algorithms, as follows:

- (1) For the case of cold start, we select  $m$  closest users  $N(u)$  for each target user  $u$  by querying our HCT and then compute the rating  $r_{u,i}$ , as done in the *friendsCF* strategy, with  $S(u', u) = 1$  when  $u' \in N(u)$ , and  $S(u', u) = 0$  otherwise.
- (2) For the case of existing users, we do the same as above, i.e., we replace the immediate social contacts with the top- $m$  closest users in HCT and feed it to the SERec algorithm.

We call the first strategy as PRIVACT-CF and the second one as PRIVACT-SERec. We choose  $m$  as the degree of the target user  $u$  in order to consider similar number of neighbors to the baseline strategies to be fair. Recall that the degree of a user can be estimated

<sup>6</sup><https://github.com/Coder-Yu/RecQ>

<sup>7</sup>Popularly known as  $k$ -nearest neighbors. We use  $m$  because  $k$  is used elsewhere.

**Table 3:** PRIVACT-SERec is  $2.1 - 17.9\times$  better than the non-social baseline (Basic SERec), as per NDCG, and has similar performance to the non-private baseline SERec.

	Top-100	Basic SERec	SERec	PRIVACT-SERec
<i>lastfm</i>	NDCG	0.1223	0.3249	0.3270
	MAP	0.0240	0.1405	0.1419
<i>delicious</i>	NDCG	0.0058	0.0135	0.0150
	MAP	0.0021	0.0048	0.0045
<i>douban</i>	NDCG	0.0349	0.2466	0.2469
	MAP	0.0052	0.0929	0.0930

**Table 4:** Results for the two cold start setup. PRIVACT-CF is  $1.1 - 286\times$  better than the *itemAvg* and worse than *friendsCF* by at most  $9.2\times$  as per NDCG. This suggests that PRIVACT-CF can be used for private social recommendation at  $\epsilon = 1$ .

	Top-100	<i>itemAvg</i>	<i>friendsCF</i>	PRIVACT-CF
<i>lastfm</i>	NDCG	5.33E-04	1.92E-01	5.54E-02
	MAP	2.56E-05	5.18E-02	7.68E-03
<i>delicious</i>	NDCG	8.82E-04	8.91E-03	9.67E-04
	MAP	7.18E-05	1.03E-03	6.91E-05
<i>douban</i>	NDCG	2.00E-04	7.30E-02	5.73E-02
	MAP	1.46E-05	1.87E-02	1.47E-02

from its degree vector we computed for PRIVACT. In the case of the *friendsCF* and PRIVACT-CF strategies, there might be some items which are not rated by any of the suggested top- $m$  users. For such items, we replace the score with its average rating across users outside of top- $m$  (referred to as the *itemAvg*) as it is a common practice to do so<sup>8</sup>. Please see Appendix A.2 for how we compute our NDCG and MAP scores for this scenario. We use  $\epsilon = 1$  for our private collaborative filtering algorithms.

## 7.3 Results

*Comparison to Non-social Baselines.* In both evaluated scenarios, cold start and existing users, our PRIVACT-based algorithms significantly outperform *non-social* baselines that do not use the social link information. For all evaluated networks, PRIVACT-based algorithms perform about  $1.1-286\times$  better (as per NDCG) than the *non-social* recommendation baselines. Similarly, as per MAP scores PRIVACT-based algorithms perform about  $0.9-1000\times$  better than the non-social baselines. Tables 3 and 4 detail the results.

*Result 3:* For cold-start users, PRIVACT-CF’s recommendations are better than the non-social baselines for most evaluated configurations.

*Comparison to Non-private Social Baselines.* PRIVACT-SERec performs on par with the non-private baseline (SERec) that uses the social links for the case of existing users (Table 3). One may expect that a non-private algorithm with raw social links would perform better than a differentially private one significantly; however, we observe that our private algorithms do comparably well. These

<sup>8</sup><https://github.com/Coder-Yu/RecQ>

results show that the top- $m$  similar users as captured by the communities of our private HCTs are a good supplement to the existing rating profiles of the users.

*Result 4: For existing users, PRIVACT-SERec has similar performance to non-private baseline SERec that uses raw social contacts.*

In the cold start setup, the non-private *friendsCF* strategy performs better than our private algorithms as one may expect. The baseline *friendsCF* strategy has  $3.5\times$  better NDCG score and  $6.7\times$  better MAP score compared to the PRIVACT-CF strategy for the *lastfm* dataset. Similarly, it has  $9.2\times$  better NDCG score than the PRIVACT-CF and  $14.9\times$  better MAP score for the *delicious* dataset. For *douban*, the *friendsCF* strategy has  $1.3\times$  better NDCG score than the PRIVACT-CF strategy and  $1.3\times$  better MAP score.

#### 7.4 When Does PRIVACT-CF Work Well?

PRIVACT-CF significantly outperforms non-social baselines empirically. We investigate why this is the case. Recall that the common hypothesis for social recommendation to work is that users have ratings similar to their close social contacts [31, 34]. We investigate this phenomenon with respect to our datasets and with respect to the scores given by PRIVACT-CF in its final recommendation.

We perform 2 experiments. First, we measure the correlation between the similarity of users' rating profile and the distance between them on the network in our used datasets. This tests whether the above common hypothesis holds in our datasets. Second, we measure whether the recommendation performance score given by PRIVACT-CF for a target user correlates with distances of the top- $m$  users who are considered for the score computation. This second experiment shows that PRIVACT-CF is selecting from users at a closer distance from the target (with noise) when it does well.

*User Similarity vs. Distance Correlation.* There are two ways to measure the similarity between users. First, we consider the partial profile of users, i.e., we consider only the items that have been rated by the users. Second, we complete the rating profiles by giving a rating of zero to the non-rated items. Full profiles capture the fact that similar users not only have similar ratings for specific items but also rate the same items. Finally, we compute the correlation between the shortest path distances and the similarities computed between the users. All correlations and user similarities are measured using the Pearson correlation. The results are summarized in the Appendix A.4, Table 6. For partial profiles, we observe that the correlation between shortest path distances and user similarity is negative for about 77%, 96% and 95% of the users in the three datasets *lastfm*, *delicious*, and *douban*. For full rating profiles, the correlation is negative for all users. This shows that *as the distance increases between users, their similarity of ratings (interests) decreases* in all our datasets. So, users in tighter communities have closer interests in our datasets.

*NDCG Scores vs. Close Communities.* Our goal is to measure whether relevant items that are ranked higher for a target user are recommended by other users that are at short distances from the target user on average by PRIVACT-CF. Recall that in the cold start scenario, we choose  $m$  users as the close community, where  $m$  is the degree of the target user. So, we measure how the average shortest

path distances of the top- $m$  users, as suggested by PRIVACT-CF, who rated the top-100 items, as predicted by PRIVACT-CF, for each target user vary with the computed NDCG scores. We first compute the shortest paths between the target user and other users in their  $m$ -sized community who have rated the same items and average all of them. Thus for each target user we obtain an average shortest path distance. Finally, we compute the Pearson's correlation between the NDCG scores<sup>9</sup> of all users and their corresponding mean shortest path distances of the top- $m$  users who rated the top-100 recommended items as suggested by PRIVACT-CF. For *lastfm* and *delicious* we find that the correlations for almost all seeds are negative. Specifically, in *lastfm* we find that the correlations range from  $[-0.31, -0.03]$ . In *delicious*, they range from  $[-0.25, -0.006]$  with four small positive outliers of up to 0.08. This means that for the top item predictions, the HCT-generated close community mostly consists of users that are at a shorter distance on the original network. This suggests why PRIVACT-CF is performing better than the *itemAvg* strategy wherein users at all distances (short and long) are considered for predictions. For *douban*, however, the correlations are positive for all seeds. In this network, we observe that the users at distances of 3 are as similar to the target user as the users at distances of 1 (unlike in the other two networks). Therefore, this allows many users at a distance of 3 to contribute more to the NDCG score, leading to an unclear correlation.

A similar analysis of PRIVACT-CF performance for collaborative filtering with existing users, with the baseline of SERec, is not straight-forward to perform using correlations because SERec uses both user ratings and social network distances. We leave the design and evaluation of an involved analysis for this case to future work.

#### 7.5 Additional Remarks

*Remark 1.* In practical applications, selecting  $m$  is likely to be application-dependent and require domain expertise or online tuning. Note that our insights can be used in any algorithm that depends on social contacts for recommendations. Furthermore, recommendation algorithms can be designed using the private HCT with the availability of additional information such as node attributes. We consider such extensions promising future work.

*Remark 2.* We are aware of only one prior work that aims to design social recommendation algorithms in the LDP setup. LDGen, which appeared at CCS 2017, uses differentially private synthetic graphs for recommendations [71]. It uses the direct graph edge information in the synthetically generated graphs for recommendations, unlike our work. The data structure they compute (synthetic graph) is fundamentally different from ours (hierarchical cluster tree) and therefore, their recommendation algorithm as well. However, a direct comparison to LDGen is not possible. LDGen does not have a publicly available implementation. After several attempts, we were unable to reproduce findings reported therein. Furthermore, the presented theoretical analysis in the LDGen paper has potential inaccuracies, which we have disclosed in a private communication to the authors of LDGen and discussed in more detail with one of its authors. We outline our the observations about LDGen in Appendix A.3 with the goal of advancing research in this domain.

<sup>9</sup>We expect MAP scores to have similar performance to NDCG scores.

## 8 RELATED WORK

Modeling networks as hierarchical clusters has been widely studied for more than 70 years [28, 80]. The algorithms to mine hierarchical clusters can be classified into two groups: 1) agglomerative/divisive algorithms based on vertex similarities, and 2) algorithms optimizing quality functions based on probabilistic modeling of the graph. The first group consists of traditional algorithms such as single, average and complete linkage, given in the book Pattern Classification [22]. Ward’s algorithm optimizes for the minimum variance of cluster similarities [88]. While these algorithms are simple, they are considered to be ad-hoc and there is no analytical way to compare them. We refer the reader to this survey [61].

Unlike the ad-hoc algorithms, more systematic ones have been proposed that optimize certain quality functions. The examples include Hierarchical Random Graph model [13, 14] and, more recently, the Hierarchical Stochastic Block Model (HSBM) [52]. All the aforementioned algorithms are not designed with privacy as a concern and require access to the complete graph for functioning.

Consequently, plenty of works try to address the privacy problem in graphs. These works [19, 43, 63, 96] release degree distributions, minimum spanning trees and subgraph counts under edge/node differential privacy. More works generated synthetic graphs privately by using graph statistics such as degree counts or edge information as inputs [42, 51, 57, 87, 89]. All of them, however, work in the trusted central-aggregator model.

In the LDP setup, many works have been proposed for frequency estimation and heavy-hitters for categorical or set-valued data [6, 21, 25, 27, 32, 37, 70, 90]. Rappor [25] has been used by Google in the past to collect app statistics. Fanti et al. extend this model to a situation when the number of categories is not known beforehand [27]. Qin et al. give a multi-round approach like ours to mine heavy hitters from set-valued data [70]. The mechanisms are extended to numerical values and multi-dimensional data by Wang et al. [86]. Further, Ye et al. present the first frequency and mean estimation algorithms for key-valued data [95]. Bassily has a more general setup of estimating linear queries [5]. Our work shares the idea of using minimal tools at our disposal with them.

LDP mechanisms for graphs are less explored. Qin et al. [71] proposes a multi-round decentralized synthetic graph generation technique. They use degree vectors to flat cluster users and then generate edges between users based on their cluster assignments. Our work takes their degree vector insight. However, there are important differences in both problem setup and hence, the techniques used. Our problem setup is different wherein we mine hierarchical cluster trees as opposed to flat clustering of vertices. HCT requires clustering at multiple levels, i.e., each vertex is assigned to many hierarchical clusters, multiple rounds of refining flat clusters cannot be directly applied in our setup. The challenges in learning HCT are unique and require a principled design that builds on some existing insights and many new ones that we talked about in Section 4.1.

GENTREE differs from other algorithms that have been proposed after the introduction of Dasgupta’s quality function. All of them are greedy approaches and some of them require the edge information of the graph. For similarity-based clustering, Dasgupta proposed a greedy sparsest-cut based algorithm with  $O(cn \log n)$  approximation. Charikar and Chatziafratis have improved it to a  $O(\sqrt{\log n})$

approximation by using SDP. Addad et al. summarize the algorithms and shows that average linkage achieves a 0.5 approximation in the dissimilarity setting [15], coinciding with the findings of Moseley et al. [59]. In comparison, although slower than greedy, GEN TREE achieves better approximation  $1 - O(\frac{\log n}{n^2-1})$  in the dissimilarity setting due to its MCMC design. Further, none of the other works have been designed with a privacy objective and it is not straightforward to argue about their utility under differential privacy (e.g., using degree vectors). In contrast, we propose a privacy-aware and easy to implement algorithm that performs better than the existing greedy approaches. Our design also allows for analyzing its utility guarantees under differential privacy. In fact, our final algorithm turns out to have comparable simplicity to the popular algorithm implemented by open-source libraries, e.g., R packages [47].

Social recommender systems are popular and have been studied for over two decades [7, 83]. The correlation between social contacts and their influence on user’s interests has been observed and theoretically modeled in real-world networks [31, 34, 65, 81]. When available, the most popular social recommendation algorithms combine both item ratings data as well as social data for collaborative filtering [10, 39]. The collaborative filtering techniques range from matrix decomposition-based [33, 40, 41, 85, 93] to deep learning-based [26]. All of these algorithms, however, require the exact social graph. Differential private matrix factorization for recommender systems is popular in the centralized setup [56, 97]. Recently, this has been studied in the LDP setup as well [79, 84]. All these techniques have a different privacy setup wherein they preserve the privacy of the user’s ratings. Further, they do not have any social component associated with them. Instead, in our work we do social recommendation while preserving the privacy of the inherent social network rather than user ratings.

## 9 CONCLUSION & FUTURE WORK

With millions of users moving towards federated services and social networks, studying how to enable existing analytics applications on these platforms is a new challenge. In this work, we provide the first algorithm to learn a differentially private hierarchical clustering of in a federated network without a trusted authority. Our approach is principled and follows theoretical analysis which explains why our design choices expect to yield good utility. We apply our new algorithms in social recommendation for the federated setup, replacing privacy-invasive ones, and show promising results. We hope our work encourages future work on supporting richer queries and the full gamut of conventional analytics for federated networks with no trusted coordinators. Enabling queries on graphs with private node attributes and on interest graphs are promising next steps.

## ACKNOWLEDGMENTS

We thank Shruti Tople, Shiqi Shen, Kareem Shehata, Jason Zhi-jingcheng Yu, Ahmad Soltani, Sasi Kumar Murakonda, Vipul Arora and Yash Pote for their valuable feedback on our paper. This work is supported by Crystal Center at NUS. Further, Teodora Baluta is supported by the National Research Foundation Singapore under its NRF Fellowship Programme [NRFNRAI1-2019-0004].

## REFERENCES

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Conference on Computer and Communications Security (CCS)*.
- [2] Google Research & Ads. 2020. Evaluation of Cohort Algorithms for the FLoC API. (2020). <https://github.com/google/ads-privacy/blob/master/proposals/FLoC-FLOC-Whitepaper-Google.pdf>
- [3] Anmol Alphonso. 2019. Twitter has become toxic. Can Mastodon provide a saner, safer alternative? *Scroll.in* (2019). <https://scroll.in/article/945163/twitter-has-become-toxic-can-mastodon-provide-a-safer-alternative>
- [4] Xavier Amatriain and Justin Basilico. 2012. Netflix Recommendations: Beyond the 5 stars (Part 1). (2012). <https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-5583468f429>
- [5] Raef Bassily. 2019. Linear queries estimation with local differential privacy. In *Conference on Artificial Intelligence and Statistics (AISTATS)*.
- [6] Raef Bassily and Adam Smith. 2015. Local, Private, Efficient Protocols for Succinct Histograms. In *Symposium on Theory of Computing (STOC)*.
- [7] Chunki Basu, Haym Hirsh, William Cohen, et al. 1998. Recommendation as classification: Using social and content-based information in recommendation. In *AAAI Conference on Artificial Intelligence (AAAI)*.
- [8] Hossein Bateni and Kevin Aydin. 2018. Balanced Partitioning and Hierarchical Clustering at Scale. (2018). <https://ai.googleblog.com/2018/03/balanced-partitioning-and-hierarchical.html>
- [9] Mohammadhossein Bateni, Soheil Behnezhad, Mahsa Derakhshan, Mohammad-Taghi Hajiaghayi, Raimondas Kiveris, Silvio Lattanzi, and Vahab Mirrokni. 2017. Affinity Clustering: Hierarchical Clustering at Scale. In *Neural Information Processing Systems (NeurIPS)*.
- [10] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems (JAIS)* 6, 3 (2005), 4.
- [11] Brave. 2020. An Introduction to BraveâŽs In-Browser Ads. (2020). <https://brave.com/intro-to-brave-ads/>
- [12] Ruichuan Chen, Alexey Reznichenko, Paul Francis, and Johannes Gehrke. 2012. Towards statistical queries over distributed private user data. In *9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*. 169–182.
- [13] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. 2007. Structural Inference of Hierarchies in Networks. In *ICML Workshop on Statistical Network Analysis*.
- [14] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* (2008).
- [15] Vincent Cohen-Addad, Varun Kanade, Frederik Mallmann-Trenn, and Claire Mathieu. 2018. Hierarchical Clustering: Objective Functions and Algorithms. In *Journal of the ACM (JACM)*.
- [16] Shaun Cole and Cedric lacey. 1996. The structure of dark matter haloes in hierarchical clustering models. *Monthly Notices of the Royal Astronomical Society* (1996).
- [17] Florence Corpet. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research* (1988).
- [18] Sanjoy Dasgupta. 2016. A Cost Function for Similarity-Based Hierarchical Clustering. In *Symposium on Theory of Computing (STOC)*.
- [19] Wei-Yen Day, Ninghui Li, and Min Lyu. 2016. Publishing Graph Degree Distribution with Node Differential Privacy. In *Conference on Management of Data (SIGMOD)*.
- [20] Zhengyu Deng, Jitao Sang, and Changsheng Xu. 2013. Personalized video recommendation based on cross-platform user modeling. In *International Conference on Multimedia and Expo (ICME)*.
- [21] John C Duchi, Michael I Jordan, and Martin J Wainwright. 2013. Local Privacy and Statistical Minimax Rates. In *Symposium on Foundations of Computer Science (FOCS)*.
- [22] Richard O Duda, Peter E Hart, et al. 1973. *Pattern classification and scene analysis*. Vol. 3. Wiley New York.
- [23] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (2014), 211–407.
- [24] Facebook Engineering. 2015. Recommending items to more than a billion people. (2015). <https://engineering.fb.com/2015/06/02/core-data/recommending-items-to-more-than-a-billion-people/>
- [25] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Conference on Computer and Communications Security (CCS)*.
- [26] Wensi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *International World Wide Web Conference (WWW)*.
- [27] Giulia Fanti, Vasyl Pihur, and Úlfar Erlingsson. 2016. Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries. In *Privacy Enhancing Technologies Symposium (PETS)*.
- [28] Kazimierz Florek, Jan Łukaszewicz, Julian Perkal, Hugo Steinhaus, and Stefan Zubrzycki. 1951. Sur la liaison et la division des points d'un ensemble fini. In *Colloquium mathematicum*, Vol. 2. 282–285.
- [29] Diaspora Foundation. 2021. Diaspora. (2021). <https://diasporafoundation.org>.
- [30] Jennifer Golbeck. 2006. Generating predictive movie recommendations from trust in social networks. In *International Conference on Trust Management*. Springer.
- [31] Jennifer Ann Golbeck. 2005. *Computing and applying trust in web-based social networks*. Ph.D. Dissertation.
- [32] Xiaolan Gu, Ming Li, Yang Cao, and Li Xiong. 2019. Supporting Both Range Queries and Frequency Estimation with Local Differential Privacy. In *Conference on Communications and Network Security (CNS)*.
- [33] Guibing Guo, Jie Zhang, and Neil Yorke-Smith. 2015. TrustSVD: Collaborative Filtering with both the explicit and implicit influence of user trust and of item ratings. In *AAAI Conference on Artificial Intelligence (AAAI)*.
- [34] Chung-Wei Hang, Yonghong Wang, and Munindar P Singh. 2008. *Operators for propagating trust and their evaluation in social networks*. Technical Report, North Carolina State University, Dept. of Computer Science.
- [35] Alex Hern. 2021. WhatsApp Loses Millions of Users After Terms Update. *The Guardian* (2021). <https://www.theguardian.com/technology/2021/jan/24/whatsapp-loses-millions-of-users-after-terms-update>
- [36] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C Pierce, and Aaron Roth. 2014. Differential Privacy: An Economic Method for Choosing Epsilon. In *Computer Security Foundations Symposium (CSF)*.
- [37] Justin Hsu, Sanjeev Khanna, and Aaron Roth. 2012. Distributed Private Heavy Hitters. In *International Colloquium on Automata, Languages and Programming (ICALP)*.
- [38] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing Differentially Private Machine Learning: How Private is Private SGD? *arXiv preprint arXiv:2006.07709* (2020).
- [39] Mohsen Jamali and Martin Ester. 2009. TrustWalker: A Random Walk Model for Combining Trust-based and Item-based Recommendation. In *Conference on Knowledge Discovery and Data Mining (KDD)*.
- [40] Mohsen Jamali and Martin Ester. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In *Recommender Systems Conference (RecSys)*.
- [41] Meng Jiang, Peng Cui, Rui Liu, Qiang Yang, Fei Wang, Wenwu Zhu, and Shiqiang Yang. 2012. Social contextual recommendation. In *Conference on Information and Knowledge Management (CIKM)*.
- [42] Vishesh Karwa and Aleksandra Slavkovic. 2012. Differentially Private Graphical Degree Sequences and Synthetic Graphs. *Privacy in Statistical Databases*.
- [43] Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2013. Analyzing Graphs with Node Differential Privacy. In *Theory of Cryptography Conference (TCC)*.
- [44] Latifur Khan, Mamoun Awad, and Bhavani Thuraisingham. 2007. A new intrusion detection system using support vector machines and hierarchical clustering. *The VLDB Journal* (2007).
- [45] Ho-Yon Kim and Jin H Kim. 2001. Hierarchical random graph representation of handwritten characters and its application to Hangul recognition. *Pattern Recognition* 34, 2 (2001), 187–201.
- [46] David Krackhardt. 1999. The Ties That Torture: Simmelian Tie Analysis in Organizations. *Research in the Sociology of Organizations* (1999).
- [47] R language. 2021. Hierarchical Random Graph package. (2021). <http://search.r-project.org/library/igraph/html/hrg-methods.html>.
- [48] Chenghao Liu, Xin Wang, Tao Lu, Wenwu Zhu, Jianling Sun, and Steven Hoi. 2019. Discrete social recommendation. In *AAAI Conference on Artificial Intelligence (AAAI)*.
- [49] Linyuan Lü and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* (2011).
- [50] Linyuan Lü and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* 390, 6 (2011), 1150–1170.
- [51] Wentian Lu and Jerome Miklau. 2014. Exponential Random Graph Estimation under Differential Privacy. In *Conference on Knowledge Discovery and Data Mining (KDD)*.
- [52] Vince Lyzinski, Minh Tang, Avanti Athreya, Youngser Park, and Carey Priebe. 2015. Community Detection and Classification in Hierarchical Stochastic Block-models. *Transactions on Network Science and Engineering* (2015).
- [53] Hao Ma, Dengyong Zhou, Chao Liu, Michael R Lyu, and Irwin King. 2011. Recommender systems with social regularization. In *Conference on Web Search and Data Mining (WSDM)*.
- [54] Mastodon. 2021. Mastodon network. (2021). <https://joinmastodon.org>.
- [55] Frank McSherry and Ratul Mahajan. 2010. Differentially-private network trace analysis. *ACM SIGCOMM Computer Communication Review* 40, 4 (2010), 123–134.
- [56] Frank McSherry and Ilya Mironov. 2009. Differentially Private Recommender Systems: Building Privacy into the Netflix Prize Contenders. In *Conference on Knowledge Discovery and Data Mining (KDD)*.
- [57] Darakhshan Mir and Rebecca N. Wright. 2012. A Differentially Private Estimator for the Stochastic Kronecker Graph Model. In *Joint EDBT/ICDT Workshops*.

- [58] Michael Mitzenmacher and Eli Upfal. 2005. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press.
- [59] Benjamin Moseley and Joshua Wang. 2017. Approximation Bounds for Hierarchical Clustering: Average Linkage, Bisecting K-means, and Local Search. In *Neural Information Processing Systems (NeurIPS)*.
- [60] Sara Mostafavi, Debjayoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. 2008. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology* (2008).
- [61] Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Data Mining and Knowledge Discovery* (2012).
- [62] Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* (2006).
- [63] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth Sensitivity and Sampling in Private Data Analysis. In *Symposium on Theory of Computing (STOC)*.
- [64] Miles Osborne, Saša Petrović, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2012. Biebee no more: First story detection using Twitter and Wikipedia. In *SIGIR 2013 Workshop on Time Aware Information Access (TAIA)*.
- [65] Rina Panigrahy, Marc Najork, and Yinglian Xie. 2012. How user behavior is related to social affinity. In *Conference on Web Search and Data Mining (WSDM)*.
- [66] Peepeth. 2021. A blockchain-powered social network. (2021). <https://peepeth.com/welcome>.
- [67] Dilruk Perera and Roger Zimmermann. 2017. Exploring the use of time-dependent cross-network information for personalized recommendations. In *International Conference on Multimedia (MM)*.
- [68] Dilruk Perera and Roger Zimmermann. 2020. Towards Comprehensive Recommender Systems: Time-Aware Unified Recommendations Based on Listwise Ranking of Implicit Cross-Network Data. In *AAAI Conference on Artificial Intelligence (AAAI)*.
- [69] Marcin Piekarczyk. 2010. Hierarchical Random Graph Model for Off-line Handwritten Signatures Recognition. In *Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*.
- [70] Zhan Qin, Yin Yang, Ting Yu, Issa Khalil, Xiaokui Xiao, and Kui Ren. 2016. Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy. In *Conference on Computer and Communications Security (CCS)*.
- [71] Zhan Qin, Ting Yu, Yin Yang, Issa Khalil, Xiaokui Xiao, and Kui Ren. 2017. Generating Synthetic Decentralized Social Graphs with Local Differential Privacy.
- [72] Daniel Ramage and Stefano Mazzocchi (Google Research). 2020. Federated Analytics: Collaborative Data Science without Data Collection. (2020). <https://ai.googleblog.com/2020/05/federated-analytics-collaborative-data.html>
- [73] Sofya Raskhodnikova, Adam Smith, Homin K Lee, Kobbi Nissim, and Shiva Prasad Kasiviswanathan. 2008. What Can We Learn Privately?. In *Symposium on Foundations of Computer Science (FOCS)*.
- [74] Netflix Research. 2021. Recommenders: Figuring out how to bring unique joy to each member. (2021). <https://research.netflix.com/research-area/recommendations>
- [75] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Conference on Computer Supported Cooperative Work*.
- [76] Suman Deb Roy, Tao Mei, Wenjun Zeng, and Shipeng Li. 2012. SocialTransfer: Cross-Domain Transfer Learning from Social Streams for Media Applications. In *International Conference on Multimedia (MM)*. 649–658.
- [77] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web*. Springer, 291–324.
- [78] Andriy Shepitzen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. 2008. Personalized recommendation in social tagging systems using hierarchical clustering. In *Recommender Systems Conference (RecSys)*.
- [79] Hyejin Shin, Sungwook Kim, Junbum Shin, and Xiaokui Xiao. 2018. Privacy Enhanced Matrix Factorization for Recommendation with Local Differential Privacy. *Transactions on Knowledge and Data Engineering (TKDE)* 30, 9 (2018), 1770–1782.
- [80] Robin Sibson. 1973. SLINK: an optimally efficient algorithm for the single-link cluster method. *The computer journal* 16, 1 (1973), 30–34.
- [81] Parag Singla and Matthew Richardson. 2008. Yes, there is a correlation - from social networks to personal behavior on the web. In *International World Wide Web Conference (WWW)*.
- [82] Nick Statt. 2021. WhatsApp clarifies itâ€Žs not giving all your data to Facebook after surge in Signal and Telegram users. *The Verge* (2021). <https://www.theverge.com/2021/1/12/22226792/whatsapp-privacy-policy-response-signal-telegram-controversy-clarification>
- [83] Jiliang Tang, Xia Hu, and Huan Liu. 2013. Social recommendation: a review. *Social Network Analysis and Mining* 3, 4 (2013), 1113–1133.
- [84] Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. 2020. LDP-Fed: Federated learning with local differential privacy. In *ACM International Workshop on Edge Systems, Analytics and Networking (EdgeSys)*.
- [85] Menghan Wang, Xiaolin Zheng, Yang Yang, and Kun Zhang. 2018. Collaborative Filtering with Social Exposure: A Modular Approach to Social Recommendation. In *AAAI Conference on Artificial Intelligence (AAAI)*.
- [86] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. 2019. Collecting and Analyzing Multidimensional Data with Local Differential Privacy. In *International Conference on Data Engineering (ICDE)*.
- [87] Yue Wang, Xintao Wu, and Leting Wu. 2013. Differential Privacy Preserving Spectral Graph Analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.
- [88] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.
- [89] Qian Xiao, Rui Chen, and Kian-Lee Tan. 2014. Differentially Private Network Data Release via Structural Inference. In *Conference on Knowledge Discovery and Data Mining (KDD)*.
- [90] Sijie Xiong, Anand D Sarwate, and Narayan B Mandayam. 2016. Randomized requantization with local differential privacy. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- [91] Dana Yakoobinsky and Dafang He. 2020. Using machine learning to auto-organize boards. (2020). <https://medium.com/pinterest-engineering/using-machine-learning-to-auto-organize-boards-13a12b22bf5>
- [92] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. 2009. How much can behavioral targeting help online advertising?. In *International World Wide Web Conference (WWW)*.
- [93] Bo Yang, Yu Lei, Jiming Liu, and Wenjie Li. 2016. Social collaborative filtering by trust. *Transactions on Pattern Analysis and Machine Intelligence* 39, 8 (2016), 1633–1647.
- [94] Yanli Yang, Hao Guo, Tian Tian, and Haifang Li. 2015. Link prediction in brain networks based on a hierarchical random graph model. *Tsinghua Science and Technology* 20, 3 (2015), 306–315.
- [95] Qingqing Ye, Haibo Hu, Xiaofeng Meng, and Huadi Zheng. 2019. PrivKV: Key-Value Data Collection with Local Differential Privacy. In *Symposium on Security and Privacy (S&P)*.
- [96] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2015. Private Release of Graph Statistics Using Ladder Functions. In *Conference on Management of Data (SIGMOD)*.
- [97] Tianqing Zhu, Gang Li, Yongli Ren, Wanlei Zhou, and Ping Xiong. 2013. Differential privacy for neighborhood-based collaborative filtering. In *Advances in Social Networks Analysis and Mining (ASONAM)*.

## A APPENDIX

### A.1 Proofs

In this section, we give the detailed proofs for the theorems and lemma presented throughout the paper.

**Theorem 3.** *Given a graph  $G : (V, E)$  and a dissimilarity matrix  $S$ , the tree  $T^{OPT} : (\Lambda, \theta)$  that maximizes  $Q_{\mathcal{D}}$  preserves the ideal clustering property with*

$$\theta(\lambda) = \frac{\sum_{x \in L, y \in R} S(x, y)}{|L| \cdot |R|}$$

where  $L, R$  are the left and right subtrees of  $\lambda$ .

**PROOF.** We will prove the claim by contradiction.

Let  $T^{OPT}$  maximize  $Q_{\mathcal{D}}$ ; assume  $\lambda_1$  and  $\lambda_2$  be two internal nodes where  $\lambda_2 \in ancestors(\lambda_1)$  such that the  $\theta(\lambda_1) > \theta(\lambda_2)$ .

Then, the three configurations of a tree with  $\lambda_1, \lambda_2$  are shown in the Figure 4. If configuration  $L$  is  $T^{OPT}$ , then one of the other two configurations,  $M$  or  $N$ , will have a higher quality than  $T^{OPT}$ . Let  $|A| = a, |B| = b, |C| = c$  and  $x \in A, y \in B, z \in C$ .

Given,  $\theta(\lambda_2) < \theta(\lambda_1)$  (1)

$$\theta(\lambda_2) < \theta(\lambda_1) \quad (2)$$

$$\text{i.e., } \frac{\sum S(x, z) + \sum S(y, z)}{(a+b) \cdot c} < \frac{\sum S(x, y)}{a \cdot b} \quad (3)$$

$$\implies (a \cdot c \sum S(x, y) - a \cdot b \sum S(x, z)) \quad (4)$$

$$+ (b \cdot c \sum S(x, y) - a \cdot b \sum S(y, z)) > 0 \quad (5)$$

(6)

Next,

$$\begin{aligned} Q_{\mathcal{D}}(M) - Q_{\mathcal{D}}(L) &= (a+c) \cdot \sum S(x, z) \\ &\quad + (a+b+c) \cdot \left( \sum S(x, y) + \sum S(y, z) \right) \\ &\quad - (a+b) \cdot \sum S(x, y) \\ &\quad - (a+b+c) \cdot \left( \sum S(x, z) + \sum S(y, z) \right) \\ &= c \cdot \sum S(x, y) - b \cdot \sum S(x, z) \end{aligned} \quad (7)$$

Similarly,

$$Q_{\mathcal{D}}(N) - Q_{\mathcal{D}}(L) = c \cdot \sum S(x, y) - a \cdot \sum S(y, z) \quad (8)$$

Observe that (5) implies at least one of (7) and (8) has to be greater than 0. Hence the contradiction.  $\square$

**Lemma 5.1.** *The probability of sampling a tree  $T$  with quality  $OPT - c' \cdot n \cdot \log n$  decreases exponentially in  $n$  (where constant  $c' \ll n$ ).*

$$\Pr(Q_{\mathcal{D}}(T) \leq OPT - c' \cdot n \cdot \log n) \leq e^{-n \cdot \log n}$$

**PROOF.** Let  $n = |V|$ ,  $Q_{\mathcal{D}}(T)$  represent the quality of the sampled tree  $T$ , from the stationary distribution and let  $N$  be the total number of possible trees/states. The quality of the optimal tree  $T^{OPT}$  is  $OPT$ .

$$\begin{aligned} \Pr(Q_{\mathcal{D}}(T) \leq m) &\leq \frac{\Pr(Q_{\mathcal{D}}(T) \leq m)}{\Pr(Q_{\mathcal{D}}(T) = OPT)} \\ &\leq \frac{N \cdot \frac{e^m}{\sum e^{Q_{\mathcal{D}}(T)}}}{1 \cdot \frac{e^{OPT}}{\sum e^{Q_{\mathcal{D}}(T)}}} \\ &\leq N \cdot e^{m - OPT} \end{aligned}$$

Since at maximum only  $N$  states can have less quality than  $m$  and assume only one state has  $OPT$ . The maximum value of  $N$  is  $2^{cn \log n}$ , therefore, if we substitute  $m = OPT - \log N - t$ , then

$$\begin{aligned} \Pr(Q_{\mathcal{D}}(T) \leq OPT - \log N - t) &\leq N \cdot e^{-\log N - t} \\ &\leq e^{-t} \end{aligned}$$

$$\Pr(Q_{\mathcal{D}}(T) \leq OPT - c \cdot n \cdot \log n - t) \leq e^{-t}$$

$$\Pr(Q_{\mathcal{D}}(T) \leq OPT - c' \cdot n \cdot \log n) \leq e^{-n \cdot \log n} \quad (9)$$

Observe that while  $Q_{\mathcal{D}}$  is a polynomial in  $n$ , the probability of being far from  $OPT$  falls exponentially in  $n$ . Therefore, the expected

loss in utility is smaller than  $c' \cdot n \cdot \log n = \frac{c' \log n}{n^2 - 1} \cdot \rho$  which implies it is also smaller than  $\frac{c' \log n}{n^2 - 1} \cdot OPT$ .  $\square$

**Lemma 6.1.** *The expected value of the noisy quality  $g(T)$  over randomness  $R$  is bounded in terms of the actual quality  $f(T)$  by*

$$|\mathbb{E}_R[g(T)] - f(T)| \leq \frac{3K}{2\epsilon} \cdot \rho$$

**PROOF.** Let the degree vector of vertex  $v_i$  be  $dv_i = (c_i^1, c_i^2, \dots, c_i^K)$  and the noisy degree vector  $\bar{d}v_i = (\bar{c}_i^1, \bar{c}_i^2, \dots, \bar{c}_i^K)$  such that  $\bar{c}_i^l = c_i^l + R_i^l$ . Similarly, for vertex  $v_j$   $dv_j = (c_j^1, c_j^2, \dots, c_j^K)$  and  $\bar{d}v_j = (\bar{c}_j^1, \bar{c}_j^2, \dots, \bar{c}_j^K)$ . Recall that the dissimilarity between two vertices is computed as  $L_1$ -norm of their degree vector counts. Let's bound the expected difference between original and noisy dissimilarities for two nodes  $v_i, v_j$  i.e.,  $\mathbb{E}_R[|\bar{S}(v_i, v_j) - S(v_i, v_j)|]$ .

$$\begin{aligned} \mathbb{E}_R[|\bar{S}(v_i, v_j) - S(v_i, v_j)|] &= \\ &\mathbb{E}_R\left[\left|\sum_{l=1}^{l=K} |\bar{c}_i^l - \bar{c}_j^l| - \sum_{l=1}^{l=K} |c_i^l - c_j^l|\right|\right] \\ &\leq \mathbb{E}_R\left[\sum_{l=1}^{l=K} |R_i^l - R_j^l|\right] \end{aligned}$$

Note that if  $z = R_i^l - R_j^l$ , then PDF of  $z$ ,  $f_{R_i^l - R_j^l}(z)$  is

$$\begin{aligned} f_{R_i^l - R_j^l}(z) &= \frac{1}{4 \cdot \sigma} [e^{\frac{-|z|}{\sigma}} + \frac{|z|}{\sigma} \cdot e^{\frac{-|z|}{\sigma}}] \\ \therefore \mathbb{E}_{|z|}[f_{R_i^l - R_j^l}(z)] &= \frac{3\sigma}{2} \end{aligned}$$

Finally, we bound the difference between  $g(T)$  and  $f(T)$

$$\begin{aligned} |\mathbb{E}_R[g(T)] - f(T)| &= \\ &\left| \sum_{i,j} (\bar{S}(v_i, v_j) - S(v_i, v_j)) \cdot p_{i,j} \right| \\ \text{where } p_{i,j} &= |\text{leaves}(T[i \vee j])| \\ &\leq \sum_{i,j} \left| \mathbb{E}_R[|\bar{S}(v_i, v_j) - S(v_i, v_j)|] \cdot p_{i,j} \right| \\ &\leq \frac{3K}{2\epsilon} \cdot \sum_{i,j} (1 \cdot |\text{leaves}(T[i \vee j])|) \\ &\leq \frac{3K}{2\epsilon} \cdot \rho \end{aligned}$$

$\square$

**Lemma 6.2.** *The variance of the noisy quality  $g(T)$  is bounded by*

$$\text{Var}_R[g(T)] \leq \frac{4K}{\epsilon^2} \cdot \rho^2$$

**PROOF.** Let  $L_{i,j}$  denote  $|\text{leaves}(T[i \vee j])|$ . The variance of  $g(T)$  is given by,

$$\begin{aligned}\text{Var}[g(T)] &= \text{Var}\left[\sum_{i,j} \bar{S}(i,j) \cdot L_{i,j}\right] \\ &= L_{i,j}^2 \cdot \sum_{i,j} \text{Var}[\bar{S}(i,j)] \\ &\quad + 2 \sum_{i,j} \sum_{i,k: i \neq j \neq k} \text{Cov}[\bar{S}(i,j), \bar{S}(i,k)]\end{aligned}\quad (10)$$

First let's look at  $\text{Var}[\bar{S}(i,j)]$ ,

$$\text{Var}[\bar{S}(i,j)] = \text{Var}\left[\sum_{l=1}^{l=K} |\bar{c}_i^l - \bar{c}_j^l|\right]$$

If  $X_l = |\bar{c}_i^l - \bar{c}_j^l|$  then observe that  $X_l \forall l \in \{1, 2, \dots, K\}$  are independent. Therefore,

$$\begin{aligned}\text{Var}[\bar{S}(i,j)] &= K \cdot \text{Var}[|\bar{c}_i^l - \bar{c}_j^l|] \\ &\quad \text{for some } l \in \{1, 2, \dots, K\} \\ &= K \cdot \text{Var}[|c_i^l - c_j^l + R_i^l - R_j^l|]\end{aligned}\quad (11)$$

Fact,  $\text{Var}[|X|] = \text{Var}[X] + \text{E}[X]^2 - \text{E}[|X|]^2$  for any random variable  $X$ . In our case  $X = |c_i^l - c_j^l + R_i^l - R_j^l|$  therefore, evaluating each of the terms in R.H.S

$$\begin{aligned}\text{Var}[X] &= \text{Var}[c_i^l - c_j^l + R_i^l - R_j^l] \\ &= \text{Var}[R_i^l - R_j^l] \\ &= \text{Var}[R_i^l] + \text{Var}[R_j^l] \\ &= 4 \cdot \sigma^2 (\sigma = \frac{1}{\epsilon}) \\ \text{E}[X]^2 &= (c_i^l - c_j^l)^2 (\text{E}[R_i^l] = 0 \forall i \in [n]) \\ \text{E}[|X|]^2 &\geq \text{E}[X]^2 \\ &\geq (c_i^l - c_j^l)^2 : \\ \text{Var}[|X|] &\leq 4 \cdot \sigma^2\end{aligned}\quad (12)$$

From Equations (11),(12),

$$\text{Var}[\bar{S}(i,j)] \leq 4K\sigma^2\quad (13)$$

Now let's bound the covariance term from Equation 10 i.e.,  $\text{Cov}[\bar{S}(i,j), \bar{S}(i,k)]$ . Similar to  $X_l$  assume  $Y_l = |\bar{c}_i^l - \bar{c}_k^l|$

$$\bar{S}(i,j) = \sum_{l=1}^{l=K} X_l, \quad \bar{S}(i,k) = \sum_{l=1}^{l=K} Y_l$$

Observe that  $X_i, Y_j$  when  $i \neq j$  are independent. Therefore,

$$\text{Cov}[\bar{S}(i,j), \bar{S}(i,k)] = \sum_{l=1}^{l=K} \text{Cov}[X_l, Y_l]\quad (14)$$

We will investigate  $\text{Cov}[X_l, Y_l]$  for some  $l$ ,

$$\begin{aligned}\text{Cov}[X_l, Y_l] &= \text{Cov}[|\bar{c}_i^l - \bar{c}_j^l|, |\bar{c}_i^l - \bar{c}_k^l|] \\ &= \text{Cov}\left[|c_i^l - c_j^l + R_i^l - R_j^l|, |c_i^l - c_k^l + R_i^l - R_k^l|\right]\end{aligned}$$

Observe that  $R_i^l - R_j^l, R_i^l - R_k^l$  are identical random variables but not independent, therefore, we replace both of them with a random variable  $X_{id}$ .

$$\begin{aligned}\text{Cov}[X_l, Y_l] &= \left[|c_i^l - c_j^l + X_{id}|, |c_i^l - c_k^l + X_{id}|\right] \\ &= [|a + X_{id}|, |b + X_{id}|] \\ &\quad \text{replacing } a + X_{id} \text{ with } X'_{id} \\ &= [|X'_{id}|, |(b-a) + X'_{id}|]\end{aligned}\quad (15)$$

Fact,  $\text{Cov}[|X|, |a+X|] \leq \text{Cov}[|X|, |X|]$ .

$$\begin{aligned}\therefore \text{Cov}[X_l, Y_l] &\leq [|X'_{id}|, |X'_{id}|] \\ &= \text{Var}[|X'_{id}|] \\ &= \text{Var}[|a + X_{id}|] \text{ From 15} \\ &= \text{Var}[|a + R_i^l - R_j^l|] \\ &\leq 4 \cdot \sigma^2 \text{ From Equation 12}\end{aligned}\quad (16)$$

Similar analysis works for all  $l$ , therefore,  $\text{Cov}[X_l, Y_l] \leq 4 \cdot \sigma^2 \forall l \in \{1, 2, \dots, K\}$ . Therefore, from Equation 14,16

$$\text{Cov}[\bar{S}(i,j), \bar{S}(i,k)] \leq 4K\sigma^2\quad (17)$$

From Equations 10,13,17 this follows

$$\begin{aligned}\text{Var}[g(T)] &\leq 4K\sigma^2 \cdot \left(\sum_{i,j} L_{i,j}^2 + 2 \cdot \sum_{i,j} \sum_{i,k} L_{i,j} L_{i,k}\right) \\ &\leq 4K\sigma^2 \cdot \left(\sum_{i,j} 1 \cdot L_{i,j}\right)^2 \\ &\leq 4K\sigma^2 \cdot \rho^2\end{aligned}$$

Hence Proved.  $\square$

**Lemma 6.3.** *The noisy quality  $g(T)$  is in the interval  $[f(T)-P, f(T)+P]$  with a high probability ( $> 8/9$ ) where*

$$P = \frac{K}{\epsilon} \cdot \left(\frac{3}{2} + \frac{6}{\sqrt{K}}\right) \cdot \rho$$

**PROOF.** We know from Lemma 6.1,

$$\left|\mathbb{E}_R[g(T) - f(T)]\right| \leq \frac{3K}{2\epsilon} \cdot \rho$$

Also, we know a bound on variance of  $g(T)$  from theorem 6.3.

$$\text{Var}[g(T)] \leq 4K\sigma^2 \cdot \rho^2$$

Applying Chebyshev's inequality,

$$\Pr \left( |g(T) - \mathbb{E}_R[g(T)]| \geq \delta \right) \leq \frac{\text{Var}[g(T)]}{\delta^2}$$

Substituting,  $\delta = \frac{6\sqrt{K}}{\epsilon} \cdot \rho$

$$\Pr \left( |g(T) - \mathbb{E}_R[g(T)]| \geq \frac{6\sqrt{K}}{\epsilon} \cdot \rho \right) \leq \frac{1}{9}$$

$$\therefore |g(T) - f(T)| \leq \frac{3K}{2\epsilon} \cdot \rho + \frac{6\sqrt{K}}{\epsilon} \cdot \rho \text{ w.h.p} \quad (18)$$

Hence proved.  $\square$

**Theorem 6.** Let  $T_{dp}$  be the output of PRIVACT and  $\rho$  be the least maximum quality of any hierarchical cluster tree. The expected utility loss of PRIVACT with a high probability ( $> 8/9$ ) is given by

$$OPT - \mathbb{E}_{T_{dp}}[f(T_{dp})] \leq \frac{2K}{\epsilon} \cdot \left( \frac{3}{2} + \frac{6}{\sqrt{K}} \right) \cdot \rho$$

**PROOF.** In PRIVACT, the sampling is done based on the noisy probabilities  $\Pr(T_{dp}) = \frac{e^{g(T_{dp})}}{\sum_{T'} e^{g(T')}}.$  In order to measure utility loss, the quality of sampled tree  $T_{dp}$  is computed using  $f(T_{dp}).$  Let  $T$  be a tree with  $f(T) = m$  and the set of trees with quality less than  $f(T)$  are  $T_{set}$  and  $T''$  has the highest probability to be sampled among them. If

$$P = \frac{2K}{\epsilon} \cdot \left( \frac{3}{2} + \frac{6}{\sqrt{K}} \right) \cdot \rho$$

then,

$$\begin{aligned} \Pr(f(T_{dp}) \leq m) &\leq \frac{\Pr(f(T_{dp}) \leq m)}{\Pr(f(T_{dp}) = OPT)} \\ &\leq |T_{set}| \cdot e^{g(T'') - g(T^{OPT})} \\ &\leq N \cdot e^{(f(T'') + P) - (f(T^{OPT}) - P)} \\ &\text{w.h.p. from Lemma 6.3} \\ &\leq N \cdot e^{2P} \cdot e^{f(T) - OPT} \text{ replace } T'' \text{ with } T \end{aligned}$$

Now, given Lemma 6.3 w.h.p., substitute  $m = OPT - 2P - 2 \cdot \log N$

$$\Pr(f(T_{dp}) \leq OPT - 2P - 2 \cdot \log N) \leq e^{-n \log n}$$

since,  $\log N = c \cdot n \log n \quad (19)$

$\therefore$  Expected Utility Loss  $\leq 2P$

$$= \frac{2K}{\epsilon} \cdot \left( \frac{3}{2} + \frac{6}{\sqrt{K}} \right) \cdot \rho$$

the  $\log N$  is  $\ll P$

Hence proved.  $\square$

## A.2 Evaluation Metrics for Social Recommendations

We use mean average precision (MAP) and normalized discounted cumulative gain (NDCG) for evaluating the performance of our recommendation algorithms. These two metrics give a sense of how relevant the top- $k$  recommendations are based on their presence

**Table 5: Results for the cold start setup with arbitrary preference order for items not rated by any of the suggested top- $m$  users in PRIVACT-CF and friendsCF.**

	Top-100	itemAvg	friendsCF	PRIVACT-CF
<i>lastfm</i>	NDCG	5.33E-04	1.58E-01	6.21E-02
	MAP	2.56E-05	4.03E-02	8.63E-03
<i>delicious</i>	NDCG	8.80E-04	6.70E-03	9.00E-04
	MAP	7.18E-05	8.00E-04	6.59E-05
<i>douban</i>	NDCG	2.00E-04	7.49E-02	7.800E-02
	MAP	1.46E-05	1.87E-02	2.000E-02

as well as the order in which they are presented to the user. MAP is the mean of average precisions that are computed on predicted recommendations for each user. Average precision ( $AP_k$ ) is the weighted average of precisions measured at every cut-off point in the predictions. The precision  $p_i$  at the cut-off point of  $i$  is defined as the number of true positives in the sequence of  $i$  recommendations divided by  $i.$  The relevance  $r_i$  of an item  $i$  is an indicator function which equals 1 if the item at rank  $i$  is relevant and 0 otherwise. The  $AP_k$  and  $MAP_k$  for  $N$  target users are thus given by:

$$AP_k = \frac{1}{k} \sum_{i=1}^{i=k} p_i \cdot r_i; MAP_k = \frac{1}{N} \sum_{N \text{ users}} AP_k^{\text{user}}$$

The normalized discounted gain is the discounted cumulative gain (DCG) normalized by the ideal DCG (IDCG).  $DCG_k$  is a metric that evaluates the gain of an item based on its position in the top- $k$  recommendation. In  $DCG_k$ , the highly relevant items that appear lower in the recommendation list are penalized by reducing their relevance  $r_i.$   $IDCG_k$  represents the  $DCG_k$  for ground truth items. We compute the mean of  $NDCG_k$  across all user's predictions.

$$DCG_k = \sum_{i=1}^{i=k} \frac{r_i}{\log_2(i+1)}; NDCG_k = \frac{DCG_k}{IDCG_k}$$

**Score computation for edge cases.** In friendsCF and PRIVACT-CF, there are items that have not been rated by any of the suggested top- $m$  users. In such a scenario, their predicted rating will be computed by itemAvg instead. Therefore, NDCG and MAP have to be computed for both items that are rated by top- $m$  users as well as the items that have not been rated. We give preference to the items that have been rated by top- $m$  users and keep them before the other category of items irrespective of their predicted rating. For instance, consider the items and their predicted ratings  $(a, 4.5), (b, 2.5)$  as rated by top- $m$  suggested users (in friendsCF or PRIVACT-CF). However, say  $(c, 3.0)$  is not rated by any of them and 3.0 is the average rating across all users outside of top- $m$  suggested users (itemAvg). Then  $\{(a, 4.5), (b, 2.5), (c, 3.0)\}$  will be the final order of recommendations rather than  $\{(a, 4.5), (c, 3.0), (b, 2.5)\}.$

**Note.** Instead of using itemAvg for such unrated items one could also give preference to them in any arbitrary order which may not be sensible. This might lead to unintended results depending on the datasets. For instance, in a previous version of this paper the results, as given in Table 5, were computed by fixing their preference order according to the order in which they appeared in the dataset. Coincidentally, these items affected the NDCG and MAP scores

**Table 6: The correlations between the user-user similarities and the shortest path distances between them is given here.**

	Rating profile	% negative corr.	corr. range
<i>lastfm</i>	Partial	76.8	[-0.25, 0.17]
	Full	100.0	[-0.23, 0.02]
<i>delicious</i>	Partial	96.6	[-0.26, 0.04]
	Full	100.0	[-0.23, 0.02]
<i>douban</i>	Partial	95.1	[-0.14, 0.05]
	Full	100.0	[-0.28, -0.01]

even though they were not intended to. By using *itemAvg* for such items we alleviate this problem by rating and presenting them to the user in an order that is backed by a meaningful metric.

### A.3 Comparision with LDGen

We detail our efforts to reproduce both experiments and theoretical analysis of prior work that is closely related to ours [71]. The paper provides good insights, including but not limited to using degree vectors to capture edge information. However, we find some issues with the paper and knowledge of these issues might be helpful to the reader. LDGen implementation was not available at the time of this writing, so direct comparison or refutation of these issues has not been feasible for us.

*The formula for number of clusters is not valid for their datasets.* The paper has derived a formula for number of bins ( $K$ ) to choose

for obtaining degree vectors in Section 4.2 (page 431). However, their assumption that the degree of nodes is between 0-50 is not valid for their evaluated networks. So, it is not clear if the derived formula can be used for these networks.

*The formula for the probability of an edge between two nodes  $u, v$  does not seem to be correct.* The formula given in Section 4.4 is not symmetric, i.e.,  $p_{u,v}$  is not equal to  $p_{v,u}$ . Therefore, for undirected graphs it is not clear how to choose the probability. Therefore, we tried to replace the formula with our own expertise, however, we could not reproduce their results using our formulas.

*Recommender system performance seems high.* To the best of our ability, we have looked for state-of-the-art social recommendation systems that are not private and they seem to record an NDCG score of 0.3 on *lastfm* [85]. However, LDGen based differentially private technique records an NDCG score close to 0.8. LDGen reported results are better than the previous non-private state-of-the-art.

*The epsilon values reported are half of what they should be.* In the LDP setup, the same edge is queried twice, once by each user. Therefore, the privacy loss is twice the amount used for one query.

We have informed the authors of LDGen about our findings and have discussed the above with one of the authors in more detail.

### A.4 Observations that help PRIVACT-CF

The correlations between user-user similarities and their shortest path distances on the network are given in Table 6.