Anonymizing List Data

Hossein Esfandiari
esfandiari@google.com

Alessandro Epasto
aepasto@google.com

Vahab Mirrokni
mirrokni@google.com

Andres Muñoz Medina
ammedina@google.com

Umar Syed
usyed@google.com

Sergei Vassilvitskii
sergeiv@google.com

Abstract

We provide novel approximation algorithms for anonymizing list data. List data is ubiquitous in ML applications, for instance in graph adjacency lists or in sparse matrix representations. We introduce a novel notion of anonymity in list datasets, and present polynomial time approximation algorithm for anonymizing such datasets. We show how this notion anonymity allows us to improve over current approximation results for k-anonymization.

1 Introduction

List data is ubiquitous in modern ML applications. Graphs can be uniquely defined by their adjacency lists. Sparse binary matrix representations consist of lists of non-zero entries. Processing lists of documents, interests, links, etc. is central in recommendation systems.

In all such applications, modifying the lists to make them k-anonymous [11] is a popular preprocessing technique that can be used to make data not identifiable (i.e. ensuring that each list is not unique in the dataset). While it is true that k-anonymity can be vulnerable to certain attacks [9], it still provides meaningful guarantees when adversaries have limited access to side information [3]. Moreover, in cases where a data analyst cannot withstand noise, it still represents one of the stronger ways to achieve privacy.

Achieving k-anonymity in a arbitrary list dataset by suppressing the minimum number of entries possible is an NP-hard problem[1]. Current approximation algorithms offer the guarantee of removing at most $O(\log(k))$ times more items than it is need in an optimal k-anonymous solution. Unfortunately, this implies that unless the optimum removes less than a $1 - O(\frac{1}{\log(k)})$ fraction of the dataset, the guarantee is trivial (as the algorithm might remove all entries).

In this paper, we improve over the state of the art of this area. We introduce a novel notion of k-anonymization (which we call k-majority-anonymity) that can be seen as a relaxation of the the notion of anonymization by suppression. Then we provide a polynomial time approximation algorithm for k-majority-anonymity in list datasets. We show how this new notion allows to improve over current approximation results for k-anonymization.

Setup. For ease of notation, we will define the problem using the language of bipartite graphs. We have a set of n users $U = \{u_1, \ldots, u_n\}$ and a set of m items $L = \{l_1, \ldots, l_m\}$. Each item is associated with a list of users. With abuse of notation we will use l_i to represent as well the list of users associated with the item, such that for all i, $l_i \subset U$. We can represent the input using the following bipartite graph. On one side we have n nodes each corresponding to users in U, and on the other side we have m nodes each corresponding to items in L. There is an edge between a user u_i and a item l_j if and only if user u_i is associated to item l_j, i.e. $u_i \in l_j$. We refer to this graph as the input graph and denote it by $G = (U \cup L, E)$.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Our goal is to increase the privacy of the users in U by making their adjacency lists in $G = (U \cup L, E)$ anonymous.

We first define the notion of k-anonymization by suppression in this context.

Definition 1.1 (k-anonymization by suppression)**.** *A graph $G' = (U \cup L, E')$ is k-anonymized by suppression from $G = (U \cup L, E)$ if: 1) (anonymity) for each user $u \in U$, at least k users (including u) have the same neighborhood in G' and 2) (suppression) if $E' \subseteq E$.*

Next, we define a relaxed notion of anonymity which we call k-majority-anonymity.

Definition 1.2 (k-majority-anonymity)**.** *A dataset $G' = (U \cup L, E')$ is a k-majority-anonymity modification of dataset $G = (U \cup L, E)$, if*

- *(anonymity) for each user $u \in U$, at least k users (including u) have the same neighborhood in G';*
- *(majority) If the edge $(u, l) \in E'$ then there is a set of least $k/2$ vertices $v_1, \ldots, v_{k/2}$ such that all vertices v_i have the neighbourhood of u in G', and $\forall v_i$ $(v_i, l) \in E$*

Notice that k-anonymization by suppression implies k-majority-anonymity. The main difference between the two definition is that k-majority-anonymity allows us to add edges u, l in the graph, as long as the *majority* of the k nodes that have the same neighborhood of u have an edge to l.

Problem definition. Given a graph $G = (V, E)$, we want to output another graph $G' = (V, E')$ such that G' is k-majority-anonymous version of G and E' is similar as possible to E. More precisely, given that our dataset is a set of edges E, we use the common Jaccard similarity to measure the similarity of E' and E. Given two sets A, B we define $J(A, B) := \frac{|A \cap B|}{|A \cup B|}$. So formally our problem consists in maximizing the Jaccard similarity $J(E', E)$ subject to G' being k-majority-anonymous version of G.

We will denote by $A \oplus B$ the symmetric difference of A and B.

2 Related work

Releasing an anonymized graph is closely related to the problem of releasing an anonymous data base. However, instead of removing or adding edges, there, anonymity is achieved by generalizing *quasi-identifiers*. In this line of research, k-anonymity [11] was the first proposed technique for protecting the identity of uses in a data base. To improve upon k-anonymity [8] introduced l-diversity and later [6] proposed the notion of t-closeness. While they do provide stronger privacy guarantees, there is no real equivalent of these notions for our setup since by definition we want the user information left in our graph to be completely homogeneous for each set of k users. Finally, in recent years the notion of differential privacy [4] has become the defacto tool for answering queries from a database privately. The notion of differential privacy however is often not amenable to releasing a full data set (or in our case a graph) as the amount of noise needed to guarantee differential privacy would seriously compromise the quality of the database.

3 Technical

In this section we provide an approximation algorithm to find a k-majority-anonymization of G. We say an algorithm alg is α-approximation if $J(E, E_{alg})/J(E, E_{\text{Opt}}) \geq \alpha$, where E_{alg} is the output of alg, and $J(\cdot, \cdot)$ is the Jaccard similarity function.

All of the previous works [1, 5] provide non-trivial guarantees when $J(E, E_{\text{Opt}}) \geq 1 - O(\frac{1}{\log k})$. However, they provide no guarantees when $J(E, E_{\text{Opt}}) < 1 - \omega(\frac{1}{\log k})$. In this work we aim to push this boundary further and design a constant approximation algorithm for when $J(E, E_{\text{Opt}}) \geq 1 - \phi$ for some constant ϕ. Next theorem states the main result of this paper.

Theorem 3.1. *Assume $J(E, E_{Opt}) \geq 0.9$. There exists an algorithm that finds a constant approximation k-majority-anonymization of G in polynomial time.*

Next we describe our algorithm. At a high level, we decompose users into clusters, each of size at least k. Then in each cluster c, for each item l, if the majority of the vertices in c have an edge to l, we add all edges between nodes in c and l, otherwise we remove all edges from nodes in c to l. This by the pigeonhole principle satisfies the k-majority-anonymity.

3.1 Preliminary

Let us start with some preliminary notions and lemmas. Note that we can represent each user u with a point in a m dimensional space, where the i-th dimension is 1 if $u \in l_i$ and the i-dimension is 0 otherwise. In this space we define the distance of two points, u and v to be the number of positions that u and v are different (a.k.a., Hamming distance). It is easy to see that this space is metric.

Let Δ_u^{Alg} be the number of positions that the algorithm Alg changes in the binary vector corresponding to the user u. Intuitively, $\sum_u \Delta_u^{\text{Opt}}$ should be related to $J(E, E_{\text{Opt}})$. The following lemma formalizes this intuition.

Lemma 3.2. *Assume $J(E, E_{Opt}) \geq 1 - \phi$. We have $\sum_u \Delta_u^{Opt} \leq \frac{2\phi}{1-\phi}|E|$.*

Proof available in the Supplemental material in appendix.

To compliment Lemma 3.2, the following lemma lower bounds $J(E, E_{\text{Alg}})$ given an upper bound on $\sum_u \Delta_u^{\text{Alg}}$.

Lemma 3.3. *Assume $\sum_u \Delta_u^{Alg} \leq \phi'|E|$. We have $J(E, E_{Alg}) \geq 1 - \frac{\phi'}{2}$.*

3.2 Initial algorithm

We now provide an approximation algorithm using a reduction to *lower-bounded r-median*[1]. Later, in the next subsection we improve this algorithm using a slightly more complicated algorithm. In lower-bounded r-median we select at most r centers from n points and assign each point to one center such that 1) the number of points assigned to each center is at least k, 2) the total distance of the points from their assigned centers is minimized. We refer to each set of the points that are assigned to the same center as a cluster. In this paper we let $r = n$, which means that the algorithm may use as many centers as it needs, however, it must assign at least k points to each center[2]. Here we use a 82.6 approximation algorithm for lower-bounded r-median [2], which is the best known result to the best of our knowledge. Below is our first algorithm. We refer to this algorithm as Alg_1.

1. Embed each user in $\mathbb{R}^m$ as described at the beginning of this section.
2. Approximately lower-bounded r-median on the points[3].
3. For each cluster c, for each item l, if most vertices in c have an edge to l, add all edges between nodes in c and l, otherwise remove all edges from nodes in c to l.

Note that by definition of lower-bounded r-median, each cluster contains at least k points. Moreover, the data that we output for users that belong to the same cluster are the same. Hence, the output satisfies anonymity part of the k-majority-anonymity condition. Moreover, by the pigeonhole principle it is easy to see that the output satisfies the majority part of the k-majority-anonymity assumptions as well. Next we bound $J(E, E_{\text{Alg}_1})$ assuming $J(E, E_{\text{Opt}}) \geq 1 - \phi$.

For analysis sake, we introduce *relaxed lower-bounded r-median* in which we are allowed to select any possible discrete point in the space as a center (as opposed to being restricted to select centers only from the points that appear in the input). Note that, if we take a solution to relaxed lower-bounded r-median and move each center to its closest point (that appear in the input), by triangle inequality the cost of the solution increases by at most a factor 2. Therefore the cost of lower-bounded r-median is at most twice that of relaxed lower-bounded r-median.

By lemma 3.2 we have $\sum_u \Delta_u^{\text{Opt}} \leq \frac{2\phi}{1-\phi}|E|$. We now prove there exists a solution to relaxed lower-bounded r-median with cost at most $\frac{2\phi}{1-\phi}|E|$. Take an optimal anonymous solution and consider the equivalence classes of nodes with same neighborhood. This induces a clustering of the nodes with clusters of size at least k. Now, it is possible to see that the total number of entries changed is equal

[1] We use r-median instead of k-median to avoid the confusion with the parameter k in k-anonymity.

[2] This implicitly means there are at most n/k centers.

[3] For $r = n$

to the sum of distances from the output neighborhood (of each class) and the original nodes. So this shows that there exists a clustering with sizes at least k with total cost $\sum_u \Delta_u^{\text{Opt}}$.

Hence, there exists a solution to lower-bounded r-median with cost at most $\frac{4\phi}{1-\phi}|E|$. Note that we are using a 82.6-approximation algorithm to find lower-bounded r-median. Hence, the total cost of our solution is at most $\frac{330.4\phi}{1-\phi}$. It is easy to see that the last line of the algorithm does not increase the total cost (since it selects the best center for each cluster). Hence we have $\sum_u \Delta_u^{\text{Alg}_1} \leq \frac{330.4\phi}{1-\phi}|E|$. By applying this to Lemma 3.3 we have $J(E, E_{\text{Alg}_1}) \geq 1 - \frac{165.2\phi}{1-\phi}$. This is a positive constant for any $\phi \leq 0.006$. This implies the following theorem.

Theorem 3.4. *Assume $J(E, E_{Opt}) \geq 0.994$. There exists an algorithm that finds a constant approximation k-majority-anonymization of G in polynomial time.*

In fact, $J(E, E_{\text{Opt}}) \geq 0.994$ is a very strong assumption. In the next subsection we provide an algorithm with a weaker assumption.

3.3 Improved algorithm

In this subsection we provide an algorithm that requires a weaker assumption on the input. In this section as a subroutine we use a 1.488 approximation algorithm for the metric facility location problem [7]. In the metric facility location problem we have a set of points and a set of facilities in a metric, and for each facility we have an opening cost. The objective is to select a set of facilities and assign each point to a facility such that the total cost of selected facilities plus the the total distance of points from their assigned facilities is minimized. Again here, we refer to the set of points assigned to each facility as a cluster. Below is our first algorithm. We refer to this algorithm as Alg_2.

1. Embed each user in $\mathbb{R}^m$ as before.
2. For each point u_i define a facility with the same coordinates and opening cost $2\sum_{u' \in U_i^k} \text{Dist}(u', u_i)$, where U_i^k is the set of k closest point to i.
3. Approximately solve facility location.
4. Arbitrary pair the clusters with size less than k and merge them.
5. For each cluster c, for each item l, if most vertices in c have an edge to l, add all edges between nodes in c and l, otherwise remove all edges from nodes in c to l.

Svitkina [10] showed that Lines 2 and 3 gives solution in which 1) size of each cluster is at least k, and 2) the cost of the solution is at most $3 \times 1.488 = 4.464$ times that of lower-bounded r-median. This means that after Line 4 size of each cluster is at least k. Similar to our previous algorithm, since size size of each cluster is at least k, the last line of the algorithm guarantees k-majority-anonymity. Next we bound $J(E, E_{\text{Alg}_2})$ assuming $J(E, E_{\text{Opt}}) \leq 1 - \phi$.

Lets refer to the first three line of the algorithm Alg_2 as Alg_2'. Similar to the previous subsection we know that there exists a solution to lower-bounded r-median with cost at most $\frac{4\phi}{1-\phi}|E|$. Therefore the cost of the solution to the facility location problem is at most $\frac{17.86\phi}{1-\phi}|E|$. Again similar to the previous subsection and by applying Lemma 3.3 we have $J(E, E_{\text{Alg}_2'}) \geq 1 - \frac{8.93\phi}{1-\phi}$. It is easy to see that Line 4 decreases the Jaccard similarity by at most a factor 3. Hence we have $J(E, E_{\text{Alg}_2}) \geq \frac{1}{3}\left(1 - \frac{8.93\phi}{1-\phi}\right)$, which is a positive constant for $\phi \leq 0.1$. This implies the following theorem.

Theorem 3.5. *Assume $J(E, E_{Opt}) \geq 0.9$. There exists an algorithm that finds a constant approximation k-majority-anonymization of E in polynomial time.*

4 Conclusion

We presented a novel notion of anonymity for graphs. This notion is a relaxation of k-anonymity and allows us to obtain better approximation guarantees to the optimum. Moreover, our algorithms derive from a connection we established between the novel notion of anonymity and well-known clustering problems. Therefore, any improvements in clustering algorithms can translate to improvement in our approximation guarantees.

References

[1] Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology (JOPT)*, 2005.

[2] Sara Ahmadian and Chaitanya Swamy. Improved approximation guarantees for lower-bounded facility location. In *International Workshop on Approximation and Online Algorithms*, pages 257–271. Springer, 2012.

[3] Raef Bassily, Adam Groce, Jonathan Katz, and Adam D. Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 439–448, 2013.

[4] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[5] Batya Kenig and Tamir Tassa. A practical approximation algorithm for optimal k-anonymity. *Data Mining and Knowledge Discovery*, 25(1):134–168, 2012.

[6] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 106–115, 2007.

[7] Shi Li. A 1.488 approximation algorithm for the uncapacitated facility location problem. *Information and Computation*, 222:45–58, 2013.

[8] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. *L*-diversity: Privacy beyond *k*-anonymity. *TKDD*, 1(1):3, 2007.

[9] Ganta Srivatsava Ranjit, Kasiviswanathan Shiva Prasad, and Smith Adam. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 265–273. ACM, 2008.

[10] Zoya Svitkina. Lower-bounded facility location. *ACM Transactions on Algorithms (TALG)*, 6(4):69, 2010.

[11] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

A Supplemental material

Proof of Lemma 3.2. Since $J(E, E_{\text{Opt}}) \geq 1 - \phi$, we have $\frac{|E \cap E_{\text{Opt}}|}{|E \cup E_{\text{Opt}}|} \geq 1 - \phi$. Thus $\frac{|E \oplus E_{\text{Opt}}|}{|E \cup E_{\text{Opt}}|} \leq \phi$, which gives us $|E \oplus E_{\text{Opt}}| \leq \phi |E \cup E_{\text{Opt}}|$. On the other hand $\frac{|E \cap E_{\text{Opt}}|}{|E \cup E_{\text{Opt}}|} \geq 1 - \phi$ implies that $|E \cup E_{\text{Opt}}| \leq \frac{1}{1-\phi}|E \cap E_{\text{Opt}}| \leq \frac{1}{1-\phi}|E|$. Therefore we have

$$\sum_u \Delta_u^{\text{Opt}} = 2|E \oplus E_{\text{Opt}}| \leq 2\phi |E \cup E_{\text{Opt}}| \leq \frac{2\phi}{1-\phi}|E|.$$

□

Proof of Lemma 3.3. Note that $\sum_u \Delta_u^{\text{Alg}} \leq \phi' |E|$ means $2|E \oplus E_{\text{Opt}}| \leq \phi' |E|$. Hence we have

$$2\big(|E \cup E_{\text{Alg}}| - |E \cap E_{\text{Alg}}|\big) \leq \phi' |E| \leq \phi' |E \cup E_{\text{Alg}}|.$$

By some rearrangement we have $\frac{|E \cap E_{\text{Alg}}|}{|E \cup E_{\text{Alg}}|} \geq 1 - \frac{\phi'}{2}$, which means $J(E, E_{\text{Alg}}) \geq 1 - \frac{\phi'}{2}$ as desired. □