DIFFERENTIALLY PRIVATE ALGORITHMS FOR GRAPH SEQUENCES

YIWEI FU, MENTOR: PROF. MARTIN STRAUSS

1. Introduction

The extent and usage of huge datasets have seen significant growth in the past few decades, along which comes the challenges of preserving data privacy. Differential privacy is a concept that helps define the way information is extracted from the dataset such that small changes in inputs produce similar outputs.

Graph datasets is an area where differentially private algorithms have started to gather interest. While it is difficult to apply statistical methods to complicated graph algorithms where the output belongs to a much more complex probability space, there has been progress in relatively simple statistics release such as counting of edges or certain subgraphs. We investigate these kinds of algorithms and try to improve upon existing results.

2. Background

Definition 2.1. A randomized algorithm $\mathcal{A}$ is ε-differetially private if given two adjacent datasets B, B', for any subset $S \subseteq \operatorname{Im} \mathcal{A}$,

$$\mathbb{P}[\mathcal{A}(B) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{A}(B') \in S].$$

A randomized algorithm $\mathcal{A}$ is (ε, δ)-differetially private if given two adjacent datasets B, B', for any subset $S \subseteq \operatorname{Im} \mathcal{A}$,

$$\mathbb{P}[\mathcal{A}(B) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{A}(B') \in S] + \delta.$$

Remark. Note that B and B' can be interchanged in the inequality above, so we actually have

$$e^{-\varepsilon} \cdot \mathbb{P}[\mathcal{A}(B') \in S] \leq \mathbb{P}[\mathcal{A}(B) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{A}(B') \in S]$$

in the case of ε-differential privacy.

A graph sequence $\mathcal{G}$ is a sequence of graphs $(G_1, G_2, \ldots, G_T)$ where each graph $G_t = (V_t, E_t)$ is obtained by applying changes in nodes and edges on the previous G_{t-1} in the sequence. We denote $\partial V_t^+, \partial V_t^-, \partial E_t^+, \partial E_t^-$

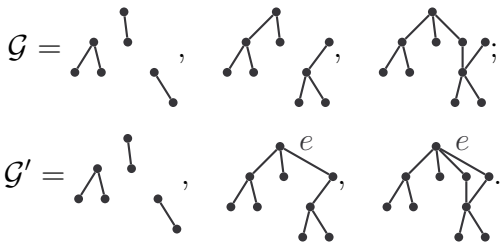
as the set of added vertices, deleted vertices, added edges, and deleted edges at timestep t respectively. With this notation, we can now define adjacent graph sequences. At $t = 1$, we have $\partial V_t^+ = V_1$ and $\partial E_t^+ = E_1$ and $\partial V_t^- = \partial E_t^- = \emptyset$.

Definition 2.2. Suppose $\mathcal{G}, \mathcal{G}'$ are two graph sequences characterized by $\{\partial V_t^+, \partial V_t^-, \partial E_t^+, \partial E_t^-\}$ and $\{\partial V_t^{'+}, \partial V_t^{'-}, \partial E_t^{'+}, \partial E_t^{'-}\}$ respectively. Let $\partial V_t^+ = \partial V_t^{'+}$ and $\partial V_t^- = \partial V_t^{'-}$ for all t. Then $\mathcal{G}, \mathcal{G}'$ are *edge-adjacent* on e^* if $|\mathcal{G}| = |\mathcal{G}'|$ and one the following conditions hold:

(1) $\partial E_t^+ = \partial E_t^{'+}$ for all t, $\partial E_t^- = \partial E_t^{'-}$ for all $t \neq t^*$, and $\partial E_{t^*}^- \setminus \partial E_{t^*}^{'-} = \{e^*\}$ for $t = t^*$,

(2) $\partial E_t^- = \partial E_t^{'-}$ for all t, $\partial E_t^+ = \partial E_t^{'+}$ for all $t \neq t^*$, and $\partial E_{t^*}^+ \setminus \partial E_{t^*}^{'+} = \{e^*\}$ for $t = t^*$.

Example 2.1.



We see that $\mathcal{G}$ and $\mathcal{G}'$ are edge-adjacent on e. Notice that in edge-adjacent graph sequences $\mathcal{G}$ and $\mathcal{G}'$, G_t and G_t' are adjacent graphs for any t.

The definition of node-adjacency is a little more complicated as the difference in node insertion/deletion also affects edge insertion/deletion.

Definition 2.3. Suppose $\mathcal{G}, \mathcal{G}'$ are two graph sequences characterized by $\{\partial V_t^+, \partial V_t^-, \partial E_t^+, \partial E_t^-\}$ and $\{\partial V_t^{'+}, \partial V_t^{'-}, \partial E_t^{'+}, \partial E_t^{'-}\}$ respectively. $\mathcal{G}, \mathcal{G}'$ are *node-adjacent* on v^* if $|\mathcal{G}| = |\mathcal{G}'|$ and one the following conditions hold:

(1) $\partial V_t^+ = \partial V_t^{'+}$ for all t, $\partial V_t^- = \partial V_t^{'-}$ for all $t \neq t^*$, and $\partial V_{t^*}^- \setminus \partial V_{t^*}^{'-} = \{v^*\}$ for $t = t^*$,

(2) $\partial V_t^- = \partial V_t^{'-}$ for all t, $\partial V_t^+ = \partial V_t^{'+}$ for all $t \neq t^*$, and $\partial V_{t^*}^+ \setminus \partial V_{t^*}^{'+} = \{v^*\}$ for $t = t^*$.

In addition, $\partial E_t^+, \partial E_t^{'+}$ and $\partial V_t^-, \partial V_t^{'-}$ does not differ other than edges that are incident from v^*.

Example 2.2.

$\mathcal{G} =$  ;

$\mathcal{G}' =$  .

Here $\mathcal{G}$ and $\mathcal{G}'$ are node-adjacent datasets but not edge-adjacent.

3. Related Work

Chan, Shi, and Song [2011] studied the private streaming of statistics. Fichtenberger, Henzinger, and Ost [2021] has then worked on differentially private algorithms on dynamic graph sequences, where difference sequence base techniques is used for monotone sequences and SVT based techniques by Lyu, Su, and Li [2016] in other scenarios.

4. Problem

4.1. p-sum Algorithm.

The following noisy p-sum by Fichtenberger et al. [2021] shows how

Algorithm 4.1 p-sum

Input: Privacy loss ε, global sensitivity Γ, graph sequence $\mathcal{G}$, graph function f

Output: noisy p-sums $a \in \mathbb{R}^k$, released over T time steps

function Count$(\varepsilon, \Gamma, \mathcal{G}, f)$
  $f(0) \leftarrow 0, \Delta f(0) \leftarrow 0$
  for each $t \in \{1, \ldots, T\}$ **do**
    Compute $f(t), \Delta f(t) \leftarrow f(t) - f(t-1)$
    Compute real partial sum $p_i, \ldots, p_j$
    for $\ell = i, \ldots, j$ **do**
      $\gamma_\ell \leftarrow \mathrm{Lap}(\Gamma \varepsilon^{-1}), \hat{p}_\ell \leftarrow p_\ell + \gamma_\ell$
    end for
  end for
  Release $\hat{p}_i, \ldots, \hat{p}_j$
end function

Corollary 4.1 (Corollary 2.9 from Chan et al. [2011])**.** *Suppose γ_i are independent random variables, where each γ_i has Laplace distribution*

$\mathrm{Lap}(b_i)$. *Let* $Y := \sum_i \gamma_i$ *and* $b_M := \max b_i$. *Let* $\nu \geq \sqrt{\sum_i b_i^2}$. *Suppose* $0 < \delta < 1$ *and* $\nu > \max\left\{\sqrt{\sum_i b_i^2}, b_M\sqrt{\ln\frac{2}{\delta}}\right\}$. *Then* $\Pr\left[|Y| > \nu\sqrt{8\ln\frac{2}{\delta}} \leq \delta\right]$.

Theorem 4.1 (Fichtenberger et al. [2021])**.** *Let f be a graph function whose difference sequence has continuous global sensitivity Γ. Let $0 < \delta < 1$ and $\varepsilon > 0$. Let $\mathcal{A}$ be a mechanism to estimate f as in Algorithm 4.1 that releases k noisy p-sums and satisfies the following conditions:*

(1) at any time step the value of a graph function f can be estimated as the sum of at most y noisy p-sums,

(2) $\mathcal{A}$ adds independent noise from $\mathrm{Lap}(\Gamma/\varepsilon)$ to every p-sum,

(3) the set P of p-sums computed by the algorithm can be partitioned into at most x subsets, such that in each partition all p-sums cover disjoint time intervals.

Then $\mathcal{A}$ is $(x\cdot\varepsilon)$-differentially private, and the error is $O(\Gamma\varepsilon^{-1}\sqrt{y}\log\frac{1}{\delta})$ with probability $1-\delta$.

To realistically implement the algorithm, we need to precompute the global sensitivity of the graph sequence using graph statistics. Some global sensitivities are known for graph functions on undirected graphs with bounded degree.

Lemma 4.1 (Fichtenberger et al. [2021])**.** *Suppose $\mathcal{G}$ is a partially dynamic (undirected) graph sequence with bounded degree D, then for edge-adjacent graph sequences, we have*

(1) Global sensitivity of triangle count is D;

(2) Global sensitivity of k-star count is $2\cdot\left(\binom{D}{k} - \binom{D-1}{k}\right)$.

4.2. **Global Sensitivities on Directed Graph Sequences.** For directed graphs, the count certain subgraphs would be more nuanced since we have edge orientations. We first define these subgraphs of interest.

Definition 4.1. For a directed graph G,

(1) the sequential triangle count function counts the number of subgraphs in form of Figure 4.1;

FIGURE 4.1. A sequential triangle

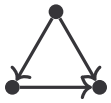


FIGURE 4.2. An alternate triangle

(2) the alternate triangle count function counts the number of subgraphs in form of Figure 4.2;

(3) the k-instar function counts the number of subgraphs where all the edges of a k-star ends at the center vertex, as shown in Figure 4.3;

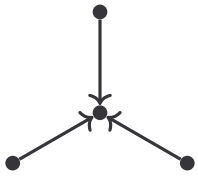


FIGURE 4.3. A 3-instar

(4) the k-outstar function counts the number of subgraphs where all the edges of a k-star begin at the center vertex, as shown in Figure 4.4.

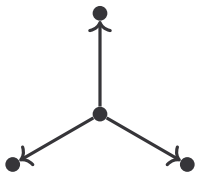


FIGURE 4.4. A 3-outstar

Lemma 4.2. *Suppose $\mathcal{G}$ is a partially dynamic* directed *graph sequence with bounded in-degree D_{in} and out-degree D_{out}, then for edge-adjacent graph sequences, we have*

(1) Global sensitivity of sequential triangle count is $\min\{D_{in}, D_{out}\}$*;*

(2) *Global sensitivity of alternate triangle count* $\leq \min\{D_{in}, D_{out}\} + D_{in} + D_{out}$;

(3) *Global sensitivity of k-instar count is* $\binom{D_{in}}{k} - \binom{D_{in}-1}{k}$;

(4) *Global sensitivity of k-outstar count is* $\binom{D_{out}}{k} - \binom{D_{out}-1}{k}$;

Proof. Suppose two graph sequence $\mathcal{G}$ and $\mathcal{G}'$ with a difference on edge (u, v).

(1) The difference edge (u, v) would contribute to the difference through the sequential triangles containing u and v. Suppose a sequential triangle containing u, v, w and (u, v), it should contain edge (v, w) and (w, u). So the number of such node w is strictly bounded above by the minimum of max in-degree and out-degree.

(2) The difference edge (u, v) would contribute to the difference through the alternative triangles containing u and v. Suppose a sequential triangle containing u, v, w and (u, v), then there are three scenarios for remaining edges involving vertex w.

   (a) $(w, u), (w, v)$.

   (b) $(u, w), (v, w)$.

   (c) $(w, v), (u, w)$.

   The total number of such w's in case (a) is bounded by the indegree of u, which is bounded by D_{in}; the total number of such w's in case (b) is bounded by the outdegree of u, which is bounded by D_{out}; the total number of such w's in case (c) is bounded $\min\{D_{in}, D_{out}\}$.

(3) The difference edge (u, v) would contribute to the difference through in-star of vertex v. v may be the center of up to $\binom{D_{in}}{k}$-many k-instars with edge added and $\binom{D_{in}-1}{k}$-many k-instars without such edge. So the global sensitivity is $\binom{D_{in}}{k} - \binom{D_{in}-1}{k}$.

(4) The difference edge (u, v) would contribute to the difference through in-star of vertex u. u may be the center of up to $\binom{D_{out}}{k}$-many k-outstars with edge added and $\binom{D_{out}-1}{k}$-many k-outstars without such edge. So the global sensitivity is $\binom{D_{out}}{k} - \binom{D_{tou}-1}{k}$.

■

4.3. **Batch Update Algorithm.**

Definition 4.2. For adjacent graph sequences $\mathcal{G}, \mathcal{G}'$ of length T, the j-step continuous global sensitivity $\mathrm{GS}_j(f)$ is the maximum value of $\sum_{t=1}^{j} |\Delta f_{\mathcal{G}}(t) - \Delta f_{\mathcal{G}'}(t)|$. That is the global sensitivity as if the graph sequence only contains the first t elements.

Remark. When $j = T$ the j-step continuous global sensitivity is the same as continuous global sensitivity for the whole sequence.

Algorithm 4.2 Batch p-sum for partially dynamic graph sequences

Input: Privacy loss ε, graph sequence $\mathcal{G}$, graph function f
Output: noisy p-sums $a \in \mathbb{R}^k$, released over T time steps

function COUNT($\varepsilon, \mathcal{G}, f$)
  $f(0) \leftarrow 0, \Delta f(0) \leftarrow 0$
  for each $t \in \{1, \ldots, T\}$ **do**
    Compute $f(t), \Delta f(t) \leftarrow f(t) - f(t-1)$
    Compute real partial sum $p_i, \ldots, p_j$
    Calculate the t-step global sensitivity Γ_t
    for $\ell = i, \ldots, j$ **do**
      $\gamma_\ell \leftarrow \mathrm{Lap}(\Gamma_t \cdot \varepsilon^{-1}), \hat{p}_\ell \leftarrow p_\ell + \gamma_\ell$
    end for
  end for
  Release $\hat{p}_i, \ldots, \hat{p}_j$
end function

Theorem 4.2. *Let f be a graph function whose difference sequence has t-step continuous global sensitivity Γ_t for each t. Let $0 < \delta < 1$ and $\varepsilon > 0$. Let $\mathcal{A}$ be a mechanism to estimate f as in Algorithm 4.2 that releases k noisy p-sums, computes sensitivity dynamically, and satisfies the following conditions:*

(1) at any time step the value of a graph function f can be estimated as the sum of at most y noisy p-sums,

(2) $\mathcal{A}$ adds independent noise from $\mathrm{Lap}(\Gamma_t/\varepsilon)$ to every p-sum at each timestep t,

(3) the set P of p-sums computed by the algorithm can be partitioned into at most x subsets, such that in each partition all p-sums cover disjoint time intervals. Equivalently, for all $P_i \in \{P_1, \cdots, P_x\}$ and all $j, k \in P_i$, $j \neq k$, it holds that

$start(j) \neq start(k)$ *and* $start(j) < start(k) \implies end(j) < start(k)$.

Then $\mathcal{A}$ *is* $(x \cdot \varepsilon)$*-differentially private, and the error is at each time step* t *is* $O(\Gamma_t \varepsilon^{-1} \sqrt{y} \log \frac{1}{\delta})$ *with probability* $1 - \delta$.

Proof. The error follows from Corollary 4.1 and condition (1). We now show that the algorithm is differentially private.

Let $\mathcal{G} = (G_1, \ldots, G_T), \mathcal{G}' = (G'_1, \ldots, G'_T)$ be two adjacent graph sequences and denote $f(t) = f(G_t)$, $f'(t) = f(G'_t)$. By definition of adjacent graph sequences we have $f(0) = f'(0) = f((V_0, E_0))$. At every time step the algorithm computes the difference sequence of $\mathcal{G}$ and $\mathcal{G}'$ by $\Delta f_{\mathcal{G}}(t) = f_{\mathcal{G}}(t) - f_{\mathcal{G}}(t-1)$.

The noisy p-sums computed by $\mathcal{A}$ are independent continuous random variables with joint distribution

$$p(z) = \prod_{i=1}^{k} p_i(z_i) \quad \text{and} \quad p'(z) = \prod_{i=1}^{k} p'_i(z_i)$$

for sequence $\mathcal{G}$ and $\mathcal{G}'$ respectively.

Let $c = (c_1, \ldots, c_k)^T$ and $c' = (c'_1, \ldots, c'_k)^T$ be the noiseless p-sums calculated by $\mathcal{A}$ on inputs $\mathcal{G}$ and $\mathcal{G}'$, respectively. For each time step $t \in \{1, \ldots, T\}$, we define $\delta(t) = \Delta f_{G'}(t) - \Delta f_G(t)$. We use $start(i)$ and $end(i)$ to denote the beginning and end of the time interval corresponding to the p-sums with index i. For each $i \in \{1, \ldots, k\}$, we define $\delta_i = c'_i - c_i = \sum_{t=start(i)}^{end(i)} \delta(t)$. Let $I = \{i : \delta i \neq 0\}$ be the indices of p-sums where the values for the two graph sequences are different. Now suppose a certain result $r = (r_1, \ldots, r_k)^T \in \mathrm{Im}(\mathcal{A})$. For any time

step t, the probability of obtaining the same output

$$\begin{aligned}
\frac{p(s)}{p'(s)} &= \prod_{i=1}^{k} \frac{p_i(s_i)}{p'_i(s_i)} \\
&= \prod_{i \in I} \frac{p_i(s_i)}{p'_i(s_i)} \\
&= \prod_{i \in I} \frac{\exp\left(\frac{\varepsilon(c_i - s_i)}{\Gamma_t}\right)}{\exp\left(\frac{\varepsilon(c_i + \delta_i - s_i)}{\Gamma_t}\right)} \\
&= \prod_{i \in I} \exp\left(\frac{\varepsilon}{\Gamma_t}\left(|c_i + \delta_i - s_i| - |c_i - s_i|\right)\right) \\
&\leq \prod_{i \in I} \exp\left(\frac{\varepsilon}{\Gamma_t} \cdot |\delta_i|\right)
\end{aligned}$$

We use condition 3 and partition I into sets of indices $I_1, \ldots, I_x$ such that for all $j \in \{1, \ldots, x\}$ the p-sums corresponding to indices in I_j cover disjoint time intervals. For each set I_j we then have

$$\sum_{i \in I_j} |\delta_i| \leq \sum_{i \in I_j} \sum_{t=start(i)}^{end(i)} |\delta(t)| \leq \sum_{t=1}^{T} \leq \Gamma_\tau, \forall \tau \in \{1, \ldots, T\}.$$

The above inequalities combined we have

$$\begin{aligned}
\frac{p(s)}{p'(s)} &\leq \prod_{i \in I} \exp\left(\frac{\varepsilon}{\Gamma_t} \cdot |\delta_i|\right) \\
&\leq \prod_{j=1}^{x} \prod_{i \in I_j} \exp\left(\frac{\varepsilon}{\Gamma_t} \cdot |\delta_i|\right) \\
&\leq \prod_{j=1}^{x} \exp\left(\frac{\varepsilon}{\Gamma_t} \sum_{i \in I_j} |\delta_i|\right) \\
&\leq \prod_{j=1}^{x} \exp\left(\frac{\varepsilon}{\Gamma_t} \Gamma_t\right) = \exp(x \cdot \varepsilon).
\end{aligned}$$

Hence $\mathcal{A}$ is $x \cdot \varepsilon$-differentially private. ■

Remark. This error bound is tighter as the t-step global sensitivity $\Gamma_t \leq \Gamma$ for all $t \in \{1, \ldots, T\}$.

5. Experiments

5.1. **Method.** We use Algorithm 4.1, Algorithm 4.2, and algorithm that outputs real statistics with no privacy constraints.

5.2. **Datasets.** We used real world datasets from SNAP [Leskovec and Krevl, 2014], especially temporal datasets. In addition to the real world directed graph datasets, the direction is also ignored to make an undirected version.

Math Overflow Database. [Paranjape, Benson, and Leskovec, 2017] This is a real temporal network of interactions on the stack exchange website Math Overflow. There are three different types of interactions represented by a directed edge (u, v, t):

- user u answered user v's question at time t
- user u commented on user v's question at time t
- user u commented on user v's answer at time t

Each kind of the edge makes up an individual dataset, and a union of all three datasets make a full dataset.

EU Mail Database. [Leskovec, Kleinberg, and Faloutsos, 2007] The network was generated using email data from a large European research institution. For a period from October 2003 to May 2005 (18 months) we have anonymized information about all incoming and outgoing email of the research institution.

MOOC(Massive open online course) User Action Dataset. [Kumar, Zhang, and Leskovec, 2019] This is a real world bipartite graph dataset. The MOOC user action dataset represents the actions taken by users on a popular MOOC platform.

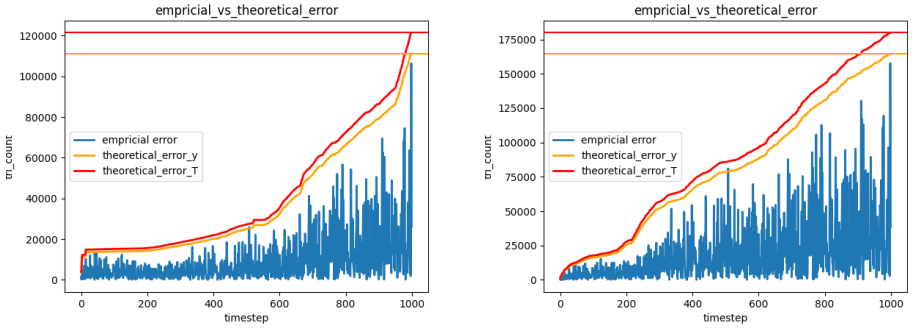
5.3. **Results.**

Lower Empirical Error Bound. Fichtenberger et al. [2021] has shown that there exists ε-differntially algorithms with $O(\log^{3/2} T)$ error and all such algorithms have $\Omega(\log T)$ error bound. We observe that the algorithms exhibits $O(\log T)$ error bound on all datasets that we have tested.

Theorem 5.1 (Fichtenberger et al. [2021])**.** *Let f be a graph function whose difference sequence has continuous global sensitivity Γ. Let $0 < \delta < 1$ and $\varepsilon > 0$. For each $T \in \mathbb{N}$ there exists an ε-differentially*

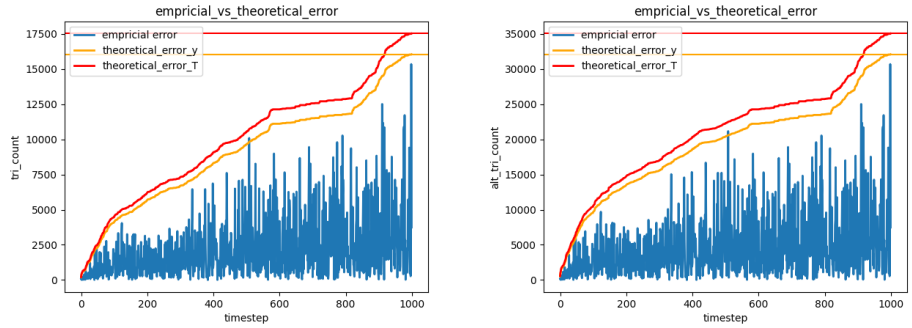
private algorithm to estimate f on a graph sequence which has error $O(\Gamma\varepsilon^{-1} \cdot \log^{3/2} T \cdot \log \delta^{-1})$ with probability at least $1 - \delta$.

Corollary 5.1. *Let f be a graph function whose difference sequence has $t-$step continuous global sensitivity Γ_t. Let $0 < \delta < 1$ and $\varepsilon > 0$. For each $T \in \mathbb{N}$ there exists an ε-differentially private algorithm to estimate f on a graph sequence which has error $O(\Gamma_t\varepsilon^{-1} \cdot \log^{3/2} T \cdot \log \delta^{-1})$ with probability at least $1 - \delta$.*



(A) EU Mail Database (B) Math Overflow Database

FIGURE 5.1. Triangle count of an undirected version of graph sequences of length $1000, \varepsilon = 1$.



(A) Sequential triangle count (B) Alternate triangle count

FIGURE 5.2. Triangle count of the directed version of graph sequences from Math Overflow Database of length $1000, \varepsilon = 1$.

In Figure 5.1 and Figure 5.2, the orange curve indicates the error bound proposed by Theorem 4.2, while the red curve indicates the error bound

proposed by Corollary 5.1, with $\log^{3/2} T$ replaced with $\log T$. The blue line graph indicates the absolute value of the error at each timestep. As we can see, the $O(\log T)$ error bounds the estimation algorithms pretty well in real world scenarios.

References

T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur.*, 14(3), nov 2011. ISSN 1094-9224. doi: 10.1145/2043621.2043626. URL https://doi.org/10.1145/2043621.2043626.

Hendrik Fichtenberger, Monika Henzinger, and Wolfgang Ost. Differentially private algorithms for graphs under continual observation. *CoRR*, abs/2106.14756, 2021. URL https://arxiv.org/abs/2106.14756.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1269–1278. ACM, 2019.

Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2–es, mar 2007. ISSN 1556-4681. doi: 10.1145/1217299.1217301. URL https://doi.org/10.1145/1217299.1217301.

Min Lyu, Dong Su, and Ninghui Li. Understanding the sparse vector technique for differential privacy, 2016. URL https://arxiv.org/abs/1603.01699.

Ashwin Paranjape, Austin R. Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, feb 2017. doi: 10.1145/3018661.3018731. URL https://doi.org/10.1145%2F3018661.3018731.