The Hardness of Sampling Connected Subgraphs

Andrew Read-McFarland(✉) and Daniel Štefankovič

University of Rochester, Rochester, NY, USA
areadmcf@ur.rochester.edu, stefanko@cs.rochester.edu

Abstract. We consider the problem of sampling connected induced subgraphs of a given input graph G. Our first result is that an efficient algorithm to approximately sample connected induced subgraphs of a given size (the size is specified in the input) does not exist unless $\mathbf{RP} = \mathbf{NP}$. We then focus on the problem of approximately sampling connected induced subgraphs with a bias, more precisely we consider a distribution where the probability of a connected subgraph induced by $S \subseteq V(G)$ is proportional to $\lambda^{|S|}$. When the input graph G has maximum degree d we identify a threshold $\lambda_d = \frac{(d-1)^{(d-1)}}{d^d}$. For $0 < \lambda < \lambda_d$ there exists a trivial efficient sampler for the problem, and for $\lambda_d < \lambda < 1$ an efficient approximate sampler does not exist unless $\mathbf{RP} = \mathbf{NP}$. Finally, we show local Markov chains are unlikely to be effective at approximately sampling connected subgraphs.

1 Introduction

Sampling a subgraph allows us to examine small sections of a graph without having to look at the potentially massive graph as a whole [4,11,13]. When we can approximately sample connected subgraphs, we gain information about the occurrence of configurations in the graph [2,11]. There are several variants of what sampling a connected subgraph means, the most common of which is to examine spanning subgraphs as done in [5], where we sample edges such that the graph is connected and every vertex is reachable. Another variant is counting the number of induced subgraphs of a graph G that are isomorphic to another graph H [17]. Exactly counting these connected induced copies is essential for polynomial time execution for Barvinok's algorithm, as done in [16].

In this paper we are concerned with fully polynomial approximate samplers (FPAS), rather than the more common fully polynomial randomized approximation scheme (FPRAS) as we wish to sample connected subgraphs induced by vertices rather than count them. Within this paradigm we consider two models of sampling: fixed size and with bias $\lambda > 0$ (where each graph of size k is sampled with probability proportional to λ^k). In Sect. 2 we look at the fixed size case, where the connected subgraph we sample always has k vertices (for a given k). This has been studied in an applied setting by [13] with various algorithms given. The related problem of exactly counting connected induced subgraphs

on k vertices is $\#W[1]$-hard [7] and is also considered in combinatorial settings [9,23]. We show in Theorem 1 that if there is an FPAS for the uniform distribution of fixed sized connected subgraphs, then $\mathbf{RP} = \mathbf{NP}$. We then prove the even stronger result that an FPAS on graphs of maximum degree three implies $\mathbf{RP} = \mathbf{NP}$ in Theorem 2.

Next we consider sampling with bias λ. Specifically, Theorem 3 shows sampling with bias λ is efficient on a graph with maximum degree d for any $\lambda < \lambda_d = \frac{(d-1)^{d-1}}{d^d}$ and Theorem 4 proves an FPAS for connected induced subgraphs with bias $\lambda \in (\lambda_d, 1)$ implies $\mathbf{RP} = \mathbf{NP}$.

Finally, in Sect. 5 we give a tree such that no local Markov chain can efficiently sample connected subgraphs of fixed size, and similarly with bias $1 > \lambda > 0$ with Theorems 5 and 6 respectively. This hints that local Markov chains likely are not effective, as they do not perform well even on trees, where we know the problem to be easy (using dynamic programming).

The following examples are mentioned to motivate sampling with bias λ and the study of computational thresholds in this setting. The variant with sampling biased by size is considered, for example, in the hardcore model in statistical physics [6,12]. Weitz shows that on a graph of maximum degree d for all $\lambda < \lambda_c = \frac{(d-1)^{d-1}}{(d-2)^d}$ we can efficiently approximately count independent sets [24]. Sly then showed for all $\lambda > \lambda_c$ we cannot efficiently approximately count independent sets unless $\mathbf{RP} = \mathbf{NP}$ [21]. Closer to our setting, Savoie et al. sample simply connected subgraphs (that is, connected subgraphs with no "holes") on a grid, with bias λ on the perimeter [19].

We generally take a subgraph of G to be induced by a subset of the vertices of G. However, in the proofs of Theorems 1, 2, and 4 we also induce subgraphs of G induced by edges of G. We formally define both below.

Definition 1. *For a graph G and $S \subseteq V(G)$, let $G[S]$ denote the subgraph of G induced by S. Formally, $V(G[S]) = S$ and $E(G[S]) = \{\{u, v\} \mid \{u, v\} \in E(G)$ and $u, v \in S\}$.*

Similarly, let $G[R]$ for $R \subseteq E(G)$ be defined as $V(G[R]) = \bigcup_{\{u,v\} \in R} \{u, v\}$ and $E(G[R]) = R$.

We will use the following formal definition of FPAS (see, e.g. [3]).

Definition 2. *An algorithm $\mathcal{A}$ is a Fully Polynomial Approximate Sampler (FPAS) for a problem $\mathcal{B}$ if for any $\delta > 0$ and input to $\mathcal{B}$ the distribution of the output of $\mathcal{A}$ is within δ of the distribution of $\mathcal{B}$ (on the given input) and $\mathcal{A}$ runs in time polynomial with respect to its input and $\log \delta^{-1}$. By distance we mean the total variation distance, $d_{TV}(\mu, \nu) = \frac{1}{2}||\mu - \nu||_1$.*

2 Sampling Fixed Size Connected Subgraphs

In this section we show an FPAS for connected subgraphs of a given size is possible only if $\mathbf{RP} = \mathbf{NP}$. Now we give the formal definition of our sampling problem, which asks for a uniformly random sample from the set of all connected subgraphs of a given size.

Definition 3. *Let* ***Connected Induced Subgraphs of Given Size*** *or* **CISGS** *be the problem that on input* (G, K) *(for a graph* G *and non-negative integer* K*) outputs uniformly random* $L \subseteq V(G)$ *such that* $|L| = K$ *and* $G[L]$ *is connected.*

We show a FPAS for CISGS solves the Steiner Tree problem (see below).

Definition 4. *(see [10]) Let* ***Steiner Tree*** *or* **ST** *be the decision problem of whether there is a connected subgraph of* G *such that all vertices of a set* S *(the vertices in* S *are called terminals) are included and the total weight of all the edges used is no more than* ℓ*. Formally,* $\mathrm{ST} = \{(G, S, \phi, \ell) \mid \exists R \subseteq E(G)$ *such that* $G[R]$ *is connected,* $S \subseteq V(G[R])$*, and* $\sum_{r \in R} \phi(r) \leq \ell\}$.

We will use the **NP**-Hardness of ST many times throughout the paper, which Karp shows [10]. However we use the stronger result that when $\phi(e) = 1$ for each edge (often called the Cardinality Steiner Problem) ST is hard [10,25].

Theorem 1. *If an* FPAS *exists for* CISGS*, then* $\mathbf{RP} = \mathbf{NP}$.

Proof. Let $(G,\ S,\ \phi,\ \ell)$ be an instance of ST, where G has n vertices and m edges, and $\phi(e) = 1$ for all $e \in E(G)$. Now let us give a brief outline of the proof. Given an instance of ST we construct a graph G' such that an FPAS for CISGS on G' allows us to obtain a solution to ST with high probability. We have 4 main sections of this proof to accomplish this.

1. First we construct G' and pick K based upon G, S, and ℓ.
2. Then we create a function f that maps connected subgraphs of G' to connected subgraphs of G.
3. Next we give a combinatorial argument to show at least 2/3 of the connected subgraphs of G' of size K map to solutions of ST on G.
4. Finally we make the complexity argument to show the Theorem's claim.

Now let us construct G' such that if we can sample connected subgraphs of G' in polynomial time we will solve ST in polynomial time on G (using a randomized algorithm). Let $k = n^3$, $c = k^2$, and $K = |S| \cdot k/2 + \ell \cdot c$.

Intuitively, G' replaces vertices in S with complete graphs of size k whose connected subgraphs provide high entropy (so that typical subgraphs will include these vertices). On the other hand, it also replaces edges with paths of length c so that including long paths consumes many vertices and therefore lowers the entropy (hence typical solutions will avoid long paths). Let A be the set of nodes in the "S-gadgets" (each a complete graph on k vertices), B be the vertices forming the elongated edges, and let C be the unchanged vertices of $V(G)$. Formally, let G' be such that $V(G') = A \cup B \cup C$ where

1. $A = \bigcup_{s \in S, i \in [1,k]} \{v_{s,i}\}$ (where $[1, k] = \{1, ..., k\}$),
2. $B = \bigcup_{e \in E(G), i \in [1,c]} \{v_{e,i}\}$, and
3. $C = (V(G) - S)$.

Note we use $v_{s,i}$ and $v_{e,i}$ to give names to vertices we are creating. Now let A' be the edges between the nodes in A, B' be the edges between the nodes in B, and C' be the edges between A, B, and C. Thus $E(G') = A' \cup B' \cup C'$, where

1. $A' = \bigcup_{s \in S} \{\{v_{s,i}, v_{s,j}\} \mid i \neq j,\ i, j \in [1, k]\}$ and
2. $B' = \bigcup_{e \in E(G), i \in [1, c-1]} \{\{v_{e,i}, v_{e,i+1}\}\}$ (making paths for each $e \in E(G)$).
3. Finally, we need to connect the paths to the original vertices and the complete graphs. Fix an arbitrary ordering of $V(G')$ and let

$$C' = \bigcup_{\{u,v\}=e \in E(G)} \{\{v_{e,1}, \min(u', v')\}, \{v_{e,c}, \max(u', v')\}\}$$

where $u' = v_{u,1}$ if $u \in S$ and otherwise is simply u, and the same holds for v'. Note the min and max are with respect to the ordering we picked.

Let us now define a function f that maps a connected subset L of $V(G')$ (that is, $G[L]$ is connected) to $R \subseteq E(G)$ such that $G[R]$ is connected (note $G[R]$ is a subgraph induced by edges rather than vertices). Informally, if a path of vertices corresponding to an edge in G is fully included in L, then we will include that edge, and otherwise we will not. Formally, let L be a subset of $V(G')$ such that $G[L]$ is connected. Then $f(L) = R$ where $R \subseteq E(G)$ such that $e \in R$ if and only if $v_{e,1}, v_{e,2}, ..., v_{e,c} \in L$.

A subset of vertices L falls into one of two cases:

(1) $\sum_{e \in G[f(L)]} \phi(e) \leq \ell$ and $G[f(L)]$ includes all points in S
(2) $G[f(L)]$ excludes some point in S.

Note that for $\sum_{e \in G[f(L)]} \phi(e) > \ell$ we must use $c(\ell + 1)$ vertices on edges in G'. This is impossible, as we sample $K = |S| \cdot k/2 + \ell \cdot c < \ell \cdot c + c$ vertices.

Note a subset L from case (1) yields a solution to ST, whereas a subset L from case (2) does not. For convenience, let C_1 be the set of all L such that $G[f(L)]$ falls into case (1), and C_2 be likewise for case (2). Note that if $(G, S, \phi, \ell) \notin \text{ST}$, then $C_1 = \emptyset$. With that in mind, let $(G, S, \phi, \ell) \in \text{ST}$ and let us bound the size of C_1. Since $(G, S, \phi, \ell) \in \text{ST}$, there is a subset of edges with weight less than or equal to ℓ such that all nodes in S are included and the graph is connected. Thus in G' we can include the paths that correspond to those edges, which requires $\ell \cdot c$ vertices, and then use the remaining $|S| \cdot k/2$ nodes in the clusters for each $s \in S$. We can use $k/2$ in each of the complete graphs created for vertices in S, and so $|C_1| \geq \binom{k}{k/2}^{|S|}$.

Now let us show $|C_2| \leq 2^{k(|S|-1)}|S| \cdot c^{2m}$. This follows since there are $|S|$ ways to pick an $s \in S$ to omit, and then $2^{k(|S|-1)}$ ways to include or exclude the $k(|S| - 1)$ points in the remaining $|S| - 1$ gadgets. Note that the number of ways to allocate any amount of vertices to edges is at most c^{2m} as for each of the m edges we can choose the length of the partial paths on either side, which can both be at most length c. Now let us show $|C_1| \geq 2|C_2|$.

Since for $n \geq 75$ we have $n^3 \geq 1 + \log_2(n) + n \log_2(n^3 + 1) + 4n^2 \log_2(n^3)$. Thus, we can substitute in $k = n^3$, $|S| \leq n$, and $m \leq n^2$ to get

$k \geq 1 + \log_2(|S|)|S| \log_2(k+1) + 4m \log_2(k)$.
Then, by exponentiating both sides and multiplying by $\frac{2^{k(|S|-1)}}{(k+1)^{|S|}}$ we obtain $\left(\frac{2^k}{k+1}\right)^{|S|} \geq 2 \cdot 2^{k(|S|-1)}|S| \cdot c^{2m}$. Since $\binom{k}{k/2} \geq \frac{2^k}{k+1}$, this gives $|C_1| \geq 2|C_2|$.

Thus a random sample of a connected subgraph of size K from G' falls into case (1) with probability $\geq 2/3$ (recall we assume $(G, S, \phi, \ell) \in \mathrm{ST}$). However, we are using an FPAS, and so our distribution is within δ of the uniform distribution. Thus, we obtain a sample from case (1) with probability $\geq 2/3 - \delta$, and so any $\delta < 1/6$ is sufficient. Our reduction at this point is quite simple; we sample L from G', and then accept if and only if $G[f(L)]$ has weight $\leq \ell$ and includes all terminals. Thus, if a solution with weight $\leq \ell$ does not exist, we will never accept, and if one does we accept with probability $> 1/2$. Finally, to show that the size of G' is polynomial with respect to (G, S, ϕ, ℓ), note G' has $O(mn^6)$ vertices. Thus, if an FPAS exists for CISGS, we have an **RP** algorithm that solves ST, and so $\mathbf{RP} = \mathbf{NP}$. □

Since ST is hard for planar graphs [18] (with maximum degree 4), so the above proof shows hardness for planar graphs as well by simply modifying the complete graphs on k vertices to be a single vertex with k adjacent vertices. Now let us extend this further to show if an FPAS exists for CISGS on graphs of maximum degree three, then $\mathbf{RP} = \mathbf{NP}$.

First, note Steiner Tree is hard for graphs of maximum degree three by a simple reduction of splitting vertices and connecting them with a 0 weight edge. In the proof of Theorem 2 we require all terminals to have maximum degree two (as we will attach a tree to each), so note any vertex with degree three can be split into two, one of which has degree two (and we consider that one to be a terminal and the other not to be).

The basic idea of the proof is the same as Theorem 1 but instead of complete graph gadgets, we will have binary trees of size k. Let us give some definitions for use in analyzing the number of connected subgraphs of a tree.

Definition 5. *For a graph G and $v \in V(G)$, the connected **rooted subgraphs** of G at v are the subgraphs of G that include v, together with the empty subgraph.*

Definition 6. *Let GT_d be an infinite d-ary tree with root vertex v_{GT_d}.*

Definition 7. *Let G be a tree with arbitrary root v. Then for all $w \in V(G)$ let the **height** of w be the length of the path between w and v.*

Definition 8. *Let $T_{h,k}$ denote the number of connected subtrees of GT_2 rooted at v_{GT_2} of size k with maximum height h.*

Suppose that $h = \lfloor \log_2(n) \rfloor$, then we can compute $T_{h,k}$ for any k in polynomial time, as we can recursively compute $T_{h,k} = \sum_{i=0}^{k} T_{h-1,i} T_{h-1,k-i-1}$ (note that the numbers have polynomially many (in n) bits). Thus for a given h, we can compute the k such that $T_{h,k}$ is maximal.

Definition 9. *Let k_h denote the index such that $T_{h,k_h} \geq T_{h,k}$ for all $0 \leq k \leq 2^{h+1} - 1$.*

Definition 10. *Let T_h denote the number of connected subtrees of GT_2 rooted at v_{GT_2} with maximum height h. That is, $T_h = \sum_{k=0}^{2^{h+1}-1} T_{h,k}$.*

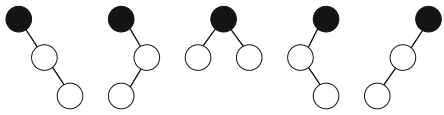

Fig. 1. The 5 configurations of $T_{2,3}$, with v_{GT_2} being the black vertex.

Figure 1 shows the 5 configurations of $T_{2,3}$. Also note we can iteratively calculate $T_h = 1 + T_{h-1}^2$ as we can either include no vertices (by omitting the root), or include the root and have any height $h-1$ subtree on either side. Now let us move on to the proof.

Theorem 2. *If an* FPAS *to* CISGS *exists on graphs of maximum degree three, then* **RP** = **NP**.

Proof. The proof follows similarly to that of Theorem 1 except let $h = \lfloor \log_2(n^3) \rfloor$, $k = 2^h$, and $K = T_{h,k_h} \cdot |S| + \ell \cdot c + n^2 + |S|$.

We assume that $\phi(e)$ is 0 or 1 for each edge e and G has maximum degree 3. As mentioned above, we assume every terminal has degree ≤ 2. Now, we construct G' with the same idea as that of Theorem 1, but using trees instead of complete graph gadgets. If $\phi(e) = 1$, then as before we replace it with a path of length c, but if $\phi(e) = 0$, e is a single node in G' rather than a path of length c (note there are at most n^2 edges with weight 0). The sets B, C, B', and C' are essentially the same as in the proof of Theorem 1 (except for the additional zero weight edges). However, we use trees for gadgets rather than complete graphs so let

$$A = S \cup \bigcup_{s \in S, i \in [1,k-1]} \{v_{s,i}\} \text{ and}$$
$$A' = \bigcup_{s \in S} \left(\{\{v_{s,1}, s\}\} \cup \{\{v_{s,i}, v_{s,\lfloor i/2 \rfloor}\} \mid i \in [2, k-1]\} \right).$$

Let f be as in Theorem 1 and note that a sample L falls into one of 2 cases as before (where again, $G[f(L)]$ is a subgraph induced by edges):

1. $\phi(G[f(L)]) \leq \ell$ and $G[f(L)]$ includes all points in S
2. $G[f(L)]$ excludes some point in S.

Let C_1 be the set of such L that fall into case (1) and C_2 be likewise for case (2). Let us first bound $|C_1| \geq \left(\frac{T_h}{2^{h+1}}\right)^{|S|}$. This is since $\sum_{k=0}^{2^{h+1}-1} T_{h,k} = T_h$ and since T_{h,k_h} has maximum value, it must be at least the average value. Thus we

can allocate $c \cdot \ell$ vertices to the edges of G' and $k_h + 1$ to each of the "S-trees" (k_h for the tree, 1 for s) and so there are at least $\left(\frac{T_h}{2^{h+1}}\right)^{|S|}$ distinct connected subgraphs created in this manner. Note that we might need to use weight 0 edges in this construction which equates to using a single vertex for each weight 0 edge. However, we have n^2 "extra" vertices in K to be used specifically for this (as there are no more than n^2 weight 0 edges).

Now let us show $|C_2| \leq T_h^{|S|-1}|S| \cdot c^{2m}$. The logic for this is the same as in Theorem 1 except we use T_h instead of 2^k as there are T_h ways to allocate the vertices to a tree. Note that the zero weight edges (that are represented by a single vertex) are accounted as they are edges in G and so contribute to the value of m. Finally let us conclude by showing $|C_1| \geq 2|C_2|$. Since for $n \geq 73$

$$n^3 \geq 1 + n + n\lfloor \log_2(n^3) \rfloor + \log_2(n) + 4n^2 \lfloor \log_2(n^3) \rfloor,$$

by substituting in terms and exponentiating both sides we get

$$2^{2^h} \geq 2 \cdot 2^{|S|(h+1)} \cdot |S| \cdot c^{2m}.$$

Now note $T_h \geq 2^{2^h}$ as $T_0 = 2$ and $T_h = 1 + T_{h-1}^2 \geq T_{h-1}^2$. Thus by making another substitution and multiplying by $\frac{T_h^{|S|-1}}{2^{|S|(h+1)}}$ we have

$$\left(\frac{T_h}{2^{h+1}}\right)^{|S|} \geq 2 \cdot T_h^{|S|-1}|S| \cdot c^{2m}$$

Therefore $|C_1| \geq 2|C_2|$ and so if $(G, S, \phi, \ell) \in \mathrm{ST}$ then the probability that the sampler gives a solution with weight $\leq \ell$ is at least $2/3 - \delta$ as in Theorem 1. Additionally if no such solution exists, this algorithm will never give one. Since the whole process runs in polynomial time, we have an **RP** algorithm that solves ST, and so **RP** = **NP** if an FPAS exists for CISGS on graphs with maximum degree three. □

3 Trees and Efficient Sampling with Bias

We showed earlier that it is hard to sample connected subgraphs of general graphs, planar graphs, and even for bounded degree graphs. In this section we will show that for bounded degree graphs as long as the bias parameter λ is small enough, we can sample connected subgraphs with bias λ allowing arbitrarily small error ε in time polynomial in n and $1/\varepsilon$ (for a fixed λ).

We analyzed earlier $T_{h,k}$, but let us now extend this definition to letting h be unbounded for a fixed k.

Definition 11. *Let $\tilde{T}_{k,d}$ be the number of connected subtrees of GT_d of size k rooted at v_{GT_d}.*

We will need the following result of Stanley [22] reformulated in our setting.

Lemma 1. $\tilde{T}_{k,d} = \binom{dk}{k} \frac{1}{(d-1)k+1}$.

Proof. Letting a full d-ary tree mean every node either has d children or is a leaf, in his Proposition 6.2.2 [22, p. 172] Stanley shows the number of full d-ary trees with n vertices and m leaves is equal to $\frac{1}{n}\binom{n}{j}$ if $n = dj+1$ and $m = (d-1)j+1$ for some j and 0 otherwise. Note that if a full d-ary tree has a nodes each with d children, then there are $(d-1)a+1$ leaves. Thus, we wish $n = dk+1$ and $m = (d-1)k+1$, so it is clear to see that if we remove all leaves from such trees we can obtain every d-ary tree on k vertices, and likewise every d-ary tree can have every node without d children add leaves until it has d children to obtain all such full trees. Thus the number of d-ary trees on k vertices is $\binom{dk+1}{k}/(dk+1) = \binom{dk}{k}\frac{1}{(d-1)k+1}$. □

Now we move on to show that a d-ary tree has more rooted connected subgraphs of a fixed size than any maximum degree d graph.

Definition 12. *Let $C_{k,d,G,v}$ be the number of connected subgraphs on k vertices with maximum degree d of a graph G rooted at vertex $v \in V(G)$. Let $C_{k,d,G}$ be as above, but for unrooted subgraphs.*

Lemma 2. *For all k, d, G, v, $C_{k,d,G,v} \leq \tilde{T}_{k,d}$.*

Proof. Let G be a graph with n nodes, and fix k, d, v. Let T be the SAW (self-avoiding walk) tree of G rooted at v (see [8,24]). That is, each node in T corresponds to a path in G starting at v. Thus T has maximum height n, has maximum degree $\leq d$, and since it has no cycles it must be a subtree of GT_d.

Now let $S \subseteq V(G)$ such that $G[S]$ is connected and $v \in S$, and let T' be a spanning tree of $G[S]$. Note T' is a subtree of T as every node in T' is a node in T, and T' is unique as any other S' cannot generate T' because it must necessarily omit some vertex in S. Thus $C_{k,d,G,v} \leq C_{k,d,T,v} \leq \tilde{T}_{k,d}$. □

Now we will shall show that for small enough λ, the total weight, $\sum_{k=0}^{\infty} \tilde{T}_{k,d}\lambda^k$ converges.

Lemma 3. *Fix d, and let $\lambda = c\frac{(d-1)^{d-1}}{d^d}$ where $c < 1$. Then for any $s \geq 0$ $\sum_{k=s}^{\infty} \tilde{T}_{k,d}\lambda^k \leq \frac{c^s}{1-c}$.*

Proof. By Lemma 1 we have that $\sum_{k=0}^{\infty} \tilde{T}_{k,d}\lambda^k = \sum_{k=0}^{\infty} \binom{dk}{k}\frac{\lambda^k}{(d-1)k+1}$. Then $\binom{dk}{k} \cdot \lambda^k = c^k\binom{dk}{k}(d-1)^{(d-1)k}/((d-1)+1)^{dk}$ which by the binomial theorem is

$$c^k \frac{\binom{dk}{k}(d-1)^{(d-1)k}}{\sum_{i=0}^{dk}\binom{dk}{i}(d-1)^i}.$$

Since the numerator occurs in the sum in the denominator, the fraction is less than 1. Thus, $\sum_{k=0}^{\infty} \tilde{T}_{k,d}\lambda^k \leq \sum_{k=0}^{\infty} \frac{c^k}{(d-1)k+1} \leq \sum_{k=0}^{\infty} c^k = \frac{1}{1-c}$.

Additionally, for $s \geq 0$, $\sum_{k=s}^{\infty} \tilde{T}_{k,d}\lambda^k \leq \sum_{k=s}^{\infty} \frac{c^k}{(d-1)k+1} \leq \frac{c^s}{1-c}$, by the same logic as above, giving us our result. □

We now show there is a sampler for any $\lambda < \frac{(d-1)^{d-1}}{d^d}$ and $\epsilon > 0$ that runs in time polynomial with respect to n and $1/\epsilon$. Note that this is not a FPAS.

Theorem 3. *For any $c < 1$ and any constant d the following is true. There exists an algorithm that for any $\lambda < c\frac{(d-1)^{d-1}}{d^d}$ and graph G with maximum degree d samples connected subgraphs with size bias λ in polynomial (in n and $1/\varepsilon$) time with error at most $\varepsilon > 0$.*

Proof. We will pick some s such that the probability that we would obtain a graph with size $\geq s$ is less than ε, and so we can only consider graphs of size $< s$. Thus, we want

$$\frac{\sum_{k=s}^{\infty} C_{k,d,G,v}\lambda^k}{\sum_{k=0}^{\infty} C_{k,d,G,v}\lambda^k} < \varepsilon.$$

This term is less than $\sum_{k=s}^{\infty} C_{k,d,G,v}\lambda^k$ as the denominator is at least 1 (because of the empty set). By Lemma 2 we have that this term is again less than $\sum_{k=s}^{\infty} \tilde{T}_{k,d}\lambda^k$. By Lemma 3 we have that this sum is no more than $\frac{c^s}{1-c}$. Now we simply need to pick s such that $\frac{c^s}{1-c}$ is less than ε. Therefore, as long as $s > \log_c(\varepsilon(1-c))$ we have that the chance of randomly sampling a subgraph of size greater than s is less than ε. So, we can have a sampling algorithm that only samples up to size s and since $C_{k,d,G,v} \leq \tilde{T}_{k,d} \leq \binom{dk}{k} \leq (\frac{\mathrm{e}\cdot d\cdot k}{k})^k = (\mathrm{e}\cdot d)^k$, there are $O((\mathrm{e}\cdot d)^{\log_c(\varepsilon(1-c))}) = O((\varepsilon(1-c))^{\log_c(\mathrm{e}d)})$ graphs we need to sample allowing error ε. Since this is a polynomial number of graphs, we can inductively enumerate them (up to size s) to calculate their weights and approximate Ω.

Additionally, if we wish to remove the rooted aspect of the subgraphs, note that $C_{k,d,G} \leq \sum_{v\in V(G)} C_{k,d,G,v} \leq n \max_{v\in V(G)} C_{k,d,G,v}$, and since $\tilde{T}_{k,d} \geq C_{k,d,G,v}$ for any v, we have $C_{k,d,G} \leq n\tilde{T}_{k,d} \leq n(\mathrm{e}\cdot d)^k$. Then we only need to consider sampling from $O(n(\varepsilon(1-c))^{\log_c(\mathrm{e}d)})$ subgraphs. Thus, sampling unrooted connected subgraphs still only requires examining a polynomial (in n and $1/\varepsilon$) number of subgraphs. □

4 Hardness of Sampling with Bias

We will now show that even when we sample connected subgraphs with bias λ rather than having fixed size, the problem is hard for $1 > \lambda > \frac{(d-1)^{d-1}}{d^d}$. However we need to give the analogous definition for CISGS.

Definition 13. *Let **Connected Induced Subgraphs With Bias** or* **CISWB** *be the problem that on input (G, λ) (for a graph G and $\lambda \in \mathbb{R}_{\geq 0}$) outputs L such that $L \subseteq V(G)$, $G[L]$ is connected, and L occurs with probability $\lambda^{|L|}/Z$ where $Z = \sum_{L'\subseteq V(G), G[L']\text{is connected}} \lambda^{|L'|}$.*

Now let us show that an efficient algorithm for CISWB would give an effective solution to ST.

Theorem 4. *If there is an* FPAS *to* CISWB *for (G, λ) where G has maximum degree d and $1 > \lambda > (d-1)^{d-1}/d^d$, then* **RP** = **NP**.

The proof is extremely similar to those of Theorems 1 and 2, so for brevity we have removed the proof of Theorem 4.

5 Markov Chains and Sampling

Now we will show that Markov chains are not likely to be useful in sampling connected subgraphs of either fixed size or with bias λ. To do this we will show that local Markov chains cannot be rapidly mixing while also sampling connected subgraphs from the desired distribution. Here we use local to mean neighboring states share $k-1$ vertices, that is for two neighboring states X and Y, $X-\{x\} = Y-\{y\}$ for some $x \in X$ and $y \in Y$ (see, e.g., [15]). The notion of local can be extended to mean neighboring states must share at least one vertex and our proofs would follow accordingly, but we use sharing $k-1$ vertices for simplicity. Our proof uses conductance (see, e.g., [20]) to show slow mixing, the standard definition is given below. We use the standard notions of $P(i,j)$ to mean the probability we move from state i to state j and $\pi(i)$ to be the probability of being in state i according to the stationary distribution.

Definition 14. *The* ***conductance*** *of a Markov chain M on state space Ω is $\Phi_M = \min_{U \subseteq \Omega, C_U \leq 1/2} \Phi(U)$ where $\Phi(U) = F_U/C_U$ and $F_U = \sum_{i \in U, j \in \overline{U}} P(i,j)\pi(i)$, $C_U = \sum_{i \in U} \pi(i)$.*

We use conductance to bound τ, the mixing time of M. Formally, $\tau = \min\{t : \sum_{j \in \Omega} |P^{t'}(i,j) - \pi(j)| \leq 1/\mathrm{e}$ for all $t' \geq t$ and $i \in \Omega\}$. It is a well known result that $1/\tau \leq 8\Phi(M)$ for an ergodic chain M (see, e.g., [1], we use this result so that M can be non-reversible, see [14] for a similar argument). Therefore, in the following proofs we give a tree G such that for any local ergodic chain M, $\Phi(M)$ is tiny.

Theorem 5. *There is a tree G with maximum degree 3 such that the following is true. Let M be a local ergodic Markov chain whose states are $S \subseteq V(G)$ such that $|S| = k$ and $G[S]$ is connected and the stationary distribution is uniform. Then the mixing time of M is exponential in k.*

Proof. Let G be a graph on $4n$ vertices consisting of 2 binary trees on n nodes with a path of length $2n$ in between them. Nodes 1 through n are in one tree, nodes $n+1$ to $3n$ are a path from the first tree to the second tree, and nodes $3n+1$ through $4n$ are the second tree. We will use a conductance argument to show slow mixing, and so we will give some U such that $\Phi(U) \leq \frac{2}{2^{k/2}}$.

Let $U = \{U' \mid U' \in \Omega, \forall v \in U', v \leq 2n, |U'| = k\}$. Clearly $|U|$ is no more than $1/2$ of the total number of connected subsets of size k as we can see $|\bar{U}| \geq |U|$, thus $C_U \leq 1/2$. Note that the only set in U that can move out of U in 1 move is $\{2n, 2n-1, ..., 2n-(k-1)\}$ as we require the vertex $2n$ to be included to add the vertex $2n+1$. Thus, $F_U \leq \frac{1}{|\Omega|}$.

Now let us give a lower bound on $|\Omega|$ by counting the number of configurations in the trees alone. Consider taking a connected subset of size $k/2$ rooted at the root of the tree such that no leaves are in the subset. Thus there are

k vertices left to choose from (as each vertex has degree 3 and $k/2-1$ edges are used internally), and so there are at least $\binom{k}{k/2}$ connected subsets with this specific configuration. Thus $|\Omega| \geq \binom{k}{k/2} \geq 2^{k/2}$ and so $\Phi(U) \leq \frac{2}{2^{k/2}}$. Therefore the mixing time is exponential in k. □

Note that Theorem 5 implies that the chain is not rapidly mixing for $k = \omega(\log n)$. Now let us give an analogous proof for sampling with bias λ.

Theorem 6. *Fix d and $1 > \lambda > \frac{(d-1)^{(d-1)}}{d^d}$, then there is a graph G with maximum degree $d+1$ such that the following is true. Let M be a local ergodic Markov chain whose states are $S \subseteq V(G)$ and $G[S]$ is connected and the stationary distribution is such that S occurs with probability $\lambda^{|S|}/Z$. Then the mixing time of M is exponential in n.*

The proof is very similar to that of Theorem 5 and so in consideration of space we have removed it.

Further Questions

- We showed hardness for CISWB on a general graph for $1 > \lambda > \frac{(d-1)^{d-1}}{d^d}$. Is there a polynomial solution for CISWB on an infinite grid (rooted at an arbitrary vertex) for some $1 > \lambda > \frac{(d-1)^{d-1}}{d^d}$?
- We have hardness results for CISWB with $\lambda < 1$. Is there a similar threshold for $\lambda > 1$?
- Similarly, we can sample connected subgraphs of a bounded degree graph with bias $\lambda < \frac{(d-1)^{d-1}}{d^d}$ for any error ε. Is there some threshold for $\lambda > 1$ where this is also true?
- In Sect. 5 we showed Markov chains are likely not useful in randomly sampling trees. What sets of graphs can they randomly sample and rapidly mix?

References

1. Aldous, D., Fill, J.A.: Reversible Markov chains and random walks on graphs (2002). https://www.stat.berkeley.edu/users/aldous/RWG/book.pdf. Unfinished monograph, recompiled 2014
2. Baskerville, K., Grassberger, P., Paczuski, M.: Graph animals, subgraph sampling, and motif search in large networks. Phys. Rev. E **76**(3), 036107, 13 (2007)
3. Frieze, A.: Notes on Counting and rapidly mixing Markov chains. http://www.math.cmu.edu/~af1p/Mixing.html
4. Grochow, J.A., Kellis, M.: Network motif discovery using subgraph enumeration and symmetry-breaking. In: Speed, T., Huang, H. (eds.) RECOMB 2007. LNCS, vol. 4453, pp. 92–106. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71681-5_7
5. Guo, H., Jerrum, M.: A polynomial-time approximation algorithm for all-terminal network reliability. In: Proceedings of the 45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, Prague, Czech Republic, 9–13 July 2018, pp. 68:1–68:12 (2018)

6. Ising, E.: Contribution to the theory of ferromagnetism. Z. Phys. **31**, 253–258 (1925)
7. Jerrum, M., Meeks, K.: The parameterised complexity of counting connected subgraphs and graph motifs. J. Comput. Syst. Sci. **81**(4), 702–716 (2015)
8. Jung, K., Shah, D.: Inference in binary pair-wise Markov random fields through self-avoiding walks. arXiv e-prints p. cs/0610111 (2006)
9. Kangas, K., Kaski, P., Koivisto, M., Korhonen, J.H.: On the number of connected sets in bounded degree graphs. In: Kratsch, D., Todinca, I. (eds.) WG 2014. LNCS, vol. 8747, pp. 336–347. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12340-0_28
10. Karp, R.M.: Reducibility among combinatorial problems. In: Miller, R.E., Thatcher, J.W. (eds.) Complexity of Computer Computations, pp. 85–103. Plenum Press, Boston (1972)
11. Kashtan, N., Milo, R., Itzkovitz, S., Alon, U.: Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics **20**(11), 1746–1758 (2004)
12. Lenz, W.: Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern. Z. Phys. **21**, 613–615 (1920)
13. Lu, X., Bressan, S.: Sampling connected induced subgraphs uniformly at random. In: Ailamaki, A., Bowers, S. (eds.) SSDBM 2012. LNCS, vol. 7338, pp. 195–212. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31235-9_13
14. Łuczak, T., Vigoda, E.: Torpid mixing of the Wang-Swendsen-Kotecký algorithm for sampling colorings. J. Discret. Algorithms **3**(1), 92–100 (2005)
15. Mossel, E., Weitz, D., Wormald, N.: On the hardness of sampling independent sets beyond the tree threshold. Probab. Theory Relat. Fields **143**(3), 401–439 (2009)
16. Patel, V., Regts, G.: Deterministic polynomial-time approximation algorithms for partition functions and graph polynomials. SIAM J. Comput. **46**(6), 1893–1919 (2017)
17. Patel, V., Regts, G.: Computing the number of induced copies of a fixed graph in a bounded degree graph. Algorithmica **81**(5), 1844–1858 (2018)
18. Garey, M.R., Johnson, D.: The rectilinear steiner tree problem is NP-complete. SIAM J. Appl. Math. **32**, 826–834 (1977)
19. Savoie, W., et al.: Phototactic supersmarticles. Artif. Life Robot. **23**(4), 459–468 (2018). https://doi.org/10.1007/s10015-018-0473-7
20. Sinclair, A.: Algorithms for Random Generation and Counting: A Markov Chain Approach. Birkhauser Verlag, Basel (1993)
21. Sly, A.: Computational transition at the uniqueness threshold. In: Proceedings of the 51st IEEE Annual Symposium on Foundations of Computer Science, FOCS, pp. 287–296 (2010)
22. Stanley, R.P.: Enumerative Combinatorics: vol. 2, 1st edn. Cambridge University Press, New York (1999)
23. Vince, A.: Counting connected sets and connected partitions of a graph. Australas. J. Comb. **67**(2), 281–293 (2017)
24. Weitz, D.: Counting independent sets up to the tree threshold. In: Proceedings of the 38th Annual ACM Symposium on Theory of Computing, STOC, pp. 140–149. ACM, New York (2006)
25. White, K., Farber, M., Pulleyblank, W.: Steiner trees, connected domination and strongly chordal graphs. Networks **15**(1), 109–124 (1985)