Information Processing Letters 152 (2019) 105851

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

# Mixing time bounds for graphlet random walks

Matteo Agostini, Marco Bressan *, Shahrzad Haddadan

*Sapienza University of Rome, Italy*

ARTICLE INFO

*Article history:*
Received 13 February 2019
Received in revised form 15 August 2019
Accepted 18 August 2019
Available online 27 August 2019
Communicated by Benjamin Doerr

*Keywords:*
Graph algorithms
Motif mining
Random walks
MCMC

ABSTRACT

A popular technique to sample fixed-size connected induced subgraphs of a graph, also known as graphlets, is based on running a certain random walk designed over the space of all graphlets in the graph. This technique requires knowledge of the mixing time of the walk but, unfortunately, no satisfying bounds are known. In this paper we provide upper and lower bounds on such a mixing time, showing how it is intimately tied to the mixing time of the original graph, and to its maximum and minimum degree.



## 1. Introduction

Sampling small subgraphs of a graph, or *graphlets*, is a central problem in graph mining and network analysis. The two most popular techniques are color coding [1–3], which gives strong guarantees but is computationally intensive, and random walks [4–7], which are lightweight but whose guarantees are poorly understood. Given a graph $G$, the random walk method works by running a simple random walk over the virtual graph $\mathcal{G}_k = \mathcal{G}_k(G)$ defined as follows. The vertices are all the induced connected $k$-node subgraphs of $G$, called its $k$-graphlets; and there is an edge between two $k$-graphlets $g$ and $g'$ if and only if $g$ can be obtained from $g'$ by replacing one vertex. With some adaptations, the simple random walk on $\mathcal{G}_k$ converges to a distribution where each $k$-graphlet of $G$ has probability proportional to its degree in $\mathcal{G}_k$, and since that degree is easy to compute, we can finally sample $k$-graphlets uniformly via e.g. rejection sampling. The key obstacle of the technique is that one needs knowledge of how long it takes for the walk to approach stationarity, that is, of the *mixing time* $t(\mathcal{G}_k)$ of $\mathcal{G}_k$. Indeed, any graphlet observed before the mixing time follows a distribution potentially far from the stationary, making the samples far from uniform.

Despite the popularity of the random walk technique, no good bounds on $t(\mathcal{G}_k)$ are known. The sole results available [2] say that there are graphs where $t(\mathcal{G}_k) = \Omega(n^{k-1})$, and that if $G$ has $n$ nodes and maximum degree $\Delta$ then $t(\mathcal{G}_k) = O(n^2\Delta^{2k})$. These bounds fail to capture the nature of $t(\mathcal{G}_k)$, and except for very small graphs they are useless in practice, too. In this work we develop novel bounds on $t(\mathcal{G}_k)$ that improve over those of [2] and that establish an intimate connection between $t(\mathcal{G}_k)$ and $t(G)$. On the one hand, we show graph families with $t(\mathcal{G}_k) = t(G) \cdot \Omega(\Delta^{k-1}/\delta^k)$, where $\delta$ is the minimum degree of $G$. On the other hand, we show that $t(\mathcal{G}_k) = t(G) \cdot \tilde{O}(\Delta^{2(k-1)})$ for every $G$[1]; this is useful in social graphs, where $t(G)$ is typically small [8], and is almost always tighter than the bound of [2] since $t(G) = \tilde{O}(n^2\Delta^2)$ for any connected $G$.

* Corresponding author.
*E-mail addresses:* agostini.mat@gmail.com (M. Agostini), bressan@di.uniroma1.it (M. Bressan), shahrzad.haddadan@gmail.com (S. Haddadan).

[1] The $\tilde{O}(\cdot)$, $\tilde{\Theta}(\cdot)$, $\tilde{\Omega}(\cdot)$ notations hide $\mathrm{poly}\log(\cdot)$ factors.

https://doi.org/10.1016/j.ipl.2019.105851

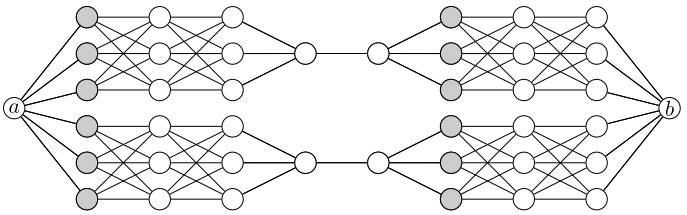


**Fig. 1.** The graph $G$ of Theorem 1, for $n = 42$, $k = 4$, $\Delta = 6$, $\delta = 3$, and $\psi = 1$. The highlighted nodes are layers $L_1$ (on the left) and $R_3$ (on the right).

### 1.1. Preliminaries and notation

Consider a graph $G = \langle V, E\rangle$. The simple lazy random walk on $G$ starts at an arbitrary vertex $X_0 \in V$, and for every $t \geq 0$ moves from $X_t$ to $X_{t+1}$ as follows: if $X_t = v$, then $X_{t+1} = v$ with probability $1/2$, and $X_{t+1} = u$ with probability $1/2d_v$, where $d_v$ is the degree of $v$ and $u$ is any neighbor of $v$. Let $\pi_t$ be the distribution of $X_t$. By a folklore theorem, if $G$ is connected then $\pi_t$ converges to a unique stationary distribution $\pi$ where $\pi(v) = d_v/2|E|$. The number of steps needed for $\pi_t$ to approach $\pi$ is called the mixing time. Formally:

**Definition 1.** The $\epsilon$-mixing time of the simple walk on $G$ is $t_\epsilon(G) = \min\{t : \forall t' \geq t : \|\pi_{t'} - \pi\|_1 \leq 2\epsilon\}$. When we drop the subscript we assume $\epsilon = 1/4$.

To bound $t(G)$ we use the strictly related *conductance* $\Phi(G)$ of $G$, defined as follows. For any subset $U \subseteq V$, define its volume $\mathrm{vol}(U) = \sum_{u\in U} d_u$. The cut induced by $U$ is $C(U) = \{uv \in E : u \in U, v \notin U\}$, and we let $c(U) = |C(U)|$. The conductance of $U$ in $G$ is $\Phi(U) = c(U)/\mathrm{vol}(U)$. Then:

**Definition 2.** The *conductance* of $G$ is

$$\Phi(G) = \min_{U\subset V : \mathrm{vol}(U) \leq \frac{1}{2}\mathrm{vol}(V)} \Phi(U).$$

Crucially, $\Theta(\Phi(G)^{-1}) \leq t_\epsilon(G) \leq \tilde{\Theta}(\Phi(G)^{-2})$. Therefore we can bound $t_\epsilon(G)$ by bounding $\Phi(G)$. All definitions and claims reported above can be found in [9].

Finally, a $k$-graphlet in $G$ is a connected induced subgraph $g = (V_g, E_g)$ of $G$ with $|V_g| = k$. To lighten the notation we write $g$ for $V_g$. Two $k$-graphlets $g, g'$ are *adjacent* if $|g \cap g'| = k - 1$. We let $\mathcal{V}_k$ be the set of all $k$-graphlets of $G$, and $\mathcal{E}_k$ be the set of (unordered) pairs of adjacent $k$-graphlets. The $k$-graphlet graph of $G$ is defined as $\mathcal{G}_k = (\mathcal{V}_k, \mathcal{E}_k)$. One can immediately adapt the definitions of lazy simple random walk to $\mathcal{G}_k$, and similarly define $t_\epsilon(\mathcal{G}_k)$ and $\Phi(\mathcal{G}_k)$. Our goal is to bound $\Phi(\mathcal{G}_k)$ in terms of $\Phi(G)$, and thus $t(\mathcal{G}_k)$ in terms of $t(G)$. To avoid trivialities we assume that $G$ is connected and $n \geq k \geq 2$. Note that in this paper $k$ is a constant; in particular, $k$ is independent of $G$.

## 2. A lower bound on the mixing time of $\mathcal{G}_k$

**Theorem 1.** *Fix any functions $\Delta(n), \delta(n), \phi(n)$ with $\Delta(n) \in \Omega(1) \cap O(n)$, $\delta(n) \in \Omega(1) \cap O(\Delta(n))$, and $t(n) \in \Omega(\delta(n)) \cap O(\delta(n)^2)$. There is a family of arbitrarily large graphs whose generic element $G$ on $n$ nodes satisfies $\Delta = \Theta(\Delta(n))$, $\delta = \Theta(\delta(n))$, $t(G) = \Omega(t(n))$, and $t(\mathcal{G}_k) = t(G) \cdot \Omega(\Delta^{k-1}/\delta^k)$.*

**Proof.** The graph $G = (V, E)$ is as follows (see Fig. 1). Fix $\Delta \in \Theta(\Delta(n))$, $\delta \in \Theta(\delta(n))$ so that $d = \Delta/\delta \in \mathbb{N}$. Fix $\psi \in \Theta(\delta(n)/\sqrt{t(n)})$ with $\psi \in [\delta]$ (note that this is always possible). The graph contains two nodes $a, b$ that are connected by $d$ parallel identical "fat" paths. For each $j \in [d]$, the $j$-th path consists of $2k$ layers $L_{1,j}, \ldots, L_{k,j}, R_{k,j}, \ldots, R_{1,j}$. Each layer has exactly $\delta$ nodes, except for $L_{k,j}, R_{k,j}$ that have $\psi$ nodes each. Every node is connected to all nodes in the immediately preceding/succeeding layers. Finally, let $L_i = \cup_{j=1}^d L_{i,j}$ and $R_i = \cup_{j=1}^d R_{i,j}$; then $a$ is connected to all nodes in $L_1$ and $b$ to all nodes in $R_1$. One can check that if $G$ has $n$ nodes then $\Delta = \Theta(\Delta(n))$ and $\delta = \Theta(\delta(n))$.

First, we show $t(G) = O(t(n))$. Suppose first the walk starts at $X_0 \in \{a, b\}$. Note that, conditioned on $X_t \in L_i$ ($R_i$), the distribution of $X_t$ is uniform over $L_i$ ($R_i$). Thus, as far as $t(G)$ is concerned, we can equivalently analyse a walk on the line graph $a, l_1, \ldots, l_{k-1}, l_k, r_k, r_{k-1}, \ldots, r_1, b$ with edge weights $\delta, \delta^2, \ldots, \delta\psi, \psi^2, \delta\psi, \ldots, \delta^2, \delta$ (i.e. the cuts between the layers, all divided by $d$). Standard calculations show that $t(G) = O(k^2 \max(\delta, \delta^2/\psi^2))$. For the case $X_0 \notin \{a, b\}$, we add a bound on the worst-case hitting time $h_0$ from $X_0 \in V \setminus \{a, b\}$ to $\{a, b\}$. Obviously, $X_0 \in V \setminus \{a, b\}$ means $X_0 \in L_{i,j}$ ($R_{i,j}$) for some $i, j$. But then, before hitting $\{a, b\}$ the walk is again equivalent to the weighted line graph above, for which standard calculations show $h_0 = O(k^2\delta)$. Hence in any case $t(G) = O(k^2 \max(\delta, \delta^2/\psi^2))$, which one can check is in $O(\delta^2/\psi^2) = O(t(n))$.

We now bound $t(\mathcal{G}_k)$ using $t(\mathcal{G}_k) = \Omega(1/\Phi(\mathcal{G}_k))$. Let $U = \{a\} \cup (\cup_{i=1}^k L_i)$ and $\overline{U} = \{b\} \cup (\cup_{i=1}^k R_i)$. Let $\mathcal{U} = \{g \in \mathcal{V}_k : |g \cap U| \geq \frac{k}{2}\}$, and $\overline{\mathcal{U}} = \mathcal{V}_k \setminus \mathcal{U}$. Clearly $\mathcal{U}$ ($\overline{\mathcal{U}}$) contains the $\binom{\Delta}{k-1} = \Omega(\Delta^{k-1})$ stars centered in $a$ ($b$), each of which has degree $\Omega(\Delta)$, so $\mathrm{vol}(\mathcal{U}), \mathrm{vol}(\overline{\mathcal{U}}) \in \Omega(\Delta^k)$. Also, $\min(\mathrm{vol}(\mathcal{U}), \mathrm{vol}(\overline{\mathcal{U}})) \leq \mathrm{vol}(\mathcal{V}_k)/2$. Therefore, in any case $\Phi(\mathcal{G}_k) = O(c(\mathcal{U})/\Delta^k)$ and so $t(\mathcal{G}_k) = \Omega(\Delta^k/c(\mathcal{U}))$. We thus bound $c(\mathcal{U})$ from below. First, $c(\mathcal{U}) = |\{gg' \in \mathcal{E}_k : g \in \mathcal{U}, g' \in \overline{\mathcal{U}}\}|$. Now, for each such $gg'$ observe that $g \cup g'$ in $G$ must be spanned by a $(k+1)$-tree containing only nodes of degree $\leq \delta$ and including some edge $uv$ in some cut $c(L_{k,j}, R_{k,j})$. But there are $O(d\psi^2\delta^{k-1})$ such trees, since $uv$ can be chosen in $O(d\psi^2)$ ways and, with $uv$ fixed, the remaining $k-1$ nodes can be chosen in $O(\delta^{k-1})$ ways. So $c(\mathcal{U}) = O(d\psi^2\delta^{k-1}) = O(\Delta\psi^2\delta^{k-2})$, and consequently $t(\mathcal{G}_k) = \Omega(\Delta^{k-1}/\psi^2\delta^{k-2})$. Comparing this bound to $t(G)$ proves the thesis. □

## 3. An upper bound on the mixing time of $\mathcal{G}_k$

This section is devoted to prove:

**Theorem 2.** $\Phi(\mathcal{G}_k) \geq \Phi(G)\frac{1}{4k^3(4\Delta)^{k-1}}$.

By the inequalities between $t(G)$ and $\Phi(G)$ (Section 1.1), Theorem 2 implies $t(\mathcal{G}_k) = t(G) \cdot \tilde{O}((4\Delta)^{2(k-1)})$. To prove the theorem, we start with two ancillary lemmata. For every $v \in V$ we let $\mathcal{V}_k(v) = \{g \in \mathcal{V}_k : v \in g\}$. More generally, for every $A \subseteq V$ we let $\mathcal{V}_k(A) = \{g \in \mathcal{V}_k : A \subseteq g\}$ and $\mathcal{G}_k(A) = \mathcal{G}_k[\mathcal{V}_k(A)]$.

**Lemma 1.** *Consider any nonempty set $A \subseteq V$ such that $G[A]$ is connected. Then in $\mathcal{G}_k(A)$ any two graphlets $g, \bar{g}$ are connected by a path $g = g_0, \ldots, g_\ell = \bar{g}$ where $g_i \cap g_{i+1}$ is connected for all $i = 0, \ldots, \ell - 1$.*

**Proof.** We assume $|\mathcal{V}_k(A)| > 1$, otherwise the claim is trivial. Fix a node $r \in A \subseteq g \cap \bar{g}$. Let $C(r)$ be the connected component of $r$ in $g \cap \bar{g}$; notice that $A \subseteq C(r)$ since $A$ is connected. Consider any spanning tree $T$ of $g$ that is a supertree of a spanning tree of $C(r)$. Clearly $T$ must have a leaf $z \notin C(r)$. Moreover, since $\bar{g}$ is connected, there exists $x \in \bar{g} \setminus g$ adjacent to $C(r)$ in $G$. Let then $g_1 = g \setminus \{z\} \cup \{x\}$, and let $C_1(r)$ be the connected component of $r$ in $g_1 \cap \bar{g}$. Now $g, g_1$ are adjacent in $\mathcal{G}_k(A)$, $g \cap g_1 = g \setminus z$ is connected by construction, and $|C_1(r)| = |C(r)| + 1$. Repeat the construction until $|C_\ell(r)| = k$, and $g_\ell = \bar{g}$. □

**Lemma 2.** *$d_g \leq k(k-1)\Delta$ for all $g \in \mathcal{V}_k$, and $|\mathcal{V}_k(v)| \leq 4d_v(4\Delta)^{k-2}$ for all $v \in G$.*

**Proof.** Any neighbor of $g$ can be built by replacing one of the $k$ nodes of $g$ with one of the at most $(k-1)\Delta$ neighbors of the other $k-1$ nodes, so $d_g \leq k(k-1)\Delta$. For $|\mathcal{V}_k(v)|$, an upper bound is the number of trees on $k$ nodes rooted at $v$, since any $g \in \mathcal{V}_k(v)$ is spanned by some such tree. Now, the number of non-isomorphic unlabeled rooted trees on $k$ nodes is at most $2^{2(k-1)} = 4^{k-1}$, since with $2(k-1)$ bits we can encode any such tree by the direction of its edges in a DFS visit. Moreover, any given tree on $k$ nodes has at most $d_v\Delta^{k-2}$ copies rooted at $v$ by a simple counting argument. Thus, $|\mathcal{V}_k(v)| \leq 4^{k-1}d_v\Delta^{k-2} = 4d_v(4\Delta)^{k-2}$. □

Let us now delve into the proof of Theorem 2. The idea is to take the partition $(\mathcal{U}, \overline{\mathcal{U}})$ of $\mathcal{V}_k$ that realizes $\Phi(\mathcal{U}) = \Phi(\mathcal{G}_k)$, build from it a certain partition $(U, \overline{U})$ of $G$, and show that $\Phi(\mathcal{U})$ is bounded from below by $\Phi(U)/4k^3(4\Delta)^{k-1}$. To this end we must properly decompose the volumes and the cuts of $\mathcal{U}$ and $U$ and relate them to each other. Before moving on note that, if $\Phi(\mathcal{G}_k) \geq \frac{1}{8k(4\Delta)^{k-1}}$, then $\Phi(\mathcal{G}_k) \geq \frac{\Phi(G)}{4k^3(4\Delta)^{k-1}}$ (the theorem's claim) since $1 \geq \Phi(G)$ and $8k < 4k^3$. From now on we thus assume $\Phi(\mathcal{G}_k) < \frac{1}{8k(4\Delta)^{k-1}}$.

*A cut for $G$.* Let $\mathcal{U} \subset \mathcal{V}_k$ satisfy $\mathrm{vol}(\mathcal{U}) \leq \frac{1}{2}\mathrm{vol}(\mathcal{V}_k)$ and $\Phi(\mathcal{G}_k) = \Phi(\mathcal{U}) = \frac{c(\mathcal{U})}{\mathrm{vol}(\mathcal{U})}$. Such $\mathcal{U}$ exists by Definition 2. We partition $G$ by taking every node whose graphlet set lies in $\mathcal{U}$ by at least a half. Formally:

$$U = \{v \in G : |\mathcal{V}_k(v) \cap \mathcal{U}| \geq \frac{1}{2}|\mathcal{V}_k(v)|\} \tag{1}$$

We now decompose $\mathrm{vol}(\mathcal{U})$ according to $U$. By definition, $\mathrm{vol}(\mathcal{U}) = \sum_{g\in\mathcal{U}} d_g$ where $d_g$ is the degree of $g$ in $\mathcal{G}_k$. Let $\mathcal{N}(g)$ be the set of neighbors of $g$ in $\mathcal{G}_k$. For any $g \in \mathcal{G}_k$ and any $v \in G$, let $d_g(v) = |\mathcal{V}_k(v) \cap \mathcal{N}(g)|$ if $v \in g$ and $d_g(v) = 0$ otherwise. Since for each $\bar{g} \in \mathcal{N}(g)$ we have $|g \cap \bar{g}| = k-1$, then $\bar{g}$ appears in $d_g(v)$ for exactly $(k-1)$ nodes $v$. Therefore $d_g = \frac{1}{k-1}\sum_{v\in V} d_g(v)$. We now generalize this expression to arbitrary subsets of $V$ and $\mathcal{V}_k$.

**Definition 3.** For all $X \subseteq V$ and $\mathcal{X} \subseteq \mathcal{V}_k$ we let $\mathrm{vol}_X(\mathcal{X}) = \frac{1}{k-1}\sum_{v\in X}\sum_{g\in\mathcal{X}} d_g(v)$.

It follows immediately that $\mathrm{vol}(\mathcal{X}) = \mathrm{vol}_X(\mathcal{X}) + \mathrm{vol}_{\overline{X}}(\mathcal{X})$, where $\overline{X} = V \setminus X$, for any $X \subseteq V$ and $\mathcal{X} \subseteq \mathcal{V}_k$. For $X = U$ and $\mathcal{X} = \mathcal{U}$, we obtain:

$$\Phi(\mathcal{G}_k) = \frac{c(\mathcal{U})}{\mathrm{vol}_U(\mathcal{U}) + \mathrm{vol}_{\overline{U}}(\mathcal{U})} \tag{2}$$

We shall then bound $\mathrm{vol}_U(\mathcal{U})$, $\mathrm{vol}_{\overline{U}}(\mathcal{U})$ in terms of $\mathrm{vol}(U)$, $c(U)$. We start by showing that $c(\mathcal{U})$ dominates $\mathrm{vol}_{\overline{U}}(\mathcal{U})$, so we can focus on $\mathrm{vol}_U(\mathcal{U})$ and $c(\mathcal{U})$.

**Lemma 3.** $\mathrm{vol}_{\overline{U}}(\mathcal{U}) < c(\mathcal{U}) \cdot 4(k-2)(4\Delta)^{k-1}$.

**Proof.** Recall that $\mathrm{vol}_{\overline{U}}(\mathcal{U}) = \frac{1}{k-1}\sum_{v\in\overline{U}}\sum_{g\in\mathcal{U}} d_g(v)$. Since $d_g(v) \neq 0$ only for $g \in \mathcal{U} \cap \mathcal{V}_k(v)$, we can restrict the outer sum to $\nu = \{v \in \overline{U} : \mathcal{U} \cap \mathcal{V}_k(v) \neq \emptyset\}$ and the inner sum to $\mathcal{U} \cap \mathcal{V}_k(v)$. Hence $\mathrm{vol}_{\overline{U}}(\mathcal{U}) = \frac{1}{k-1}\sum_{v\in\nu}\sum_{g\in\mathcal{U}\cap\mathcal{V}_k(v)} d_g(v)$. Now, for $v \in \overline{\mathcal{U}}$, we have $|\mathcal{U} \cap \mathcal{V}_k(v)| < \frac{1}{2}|\mathcal{V}_k(v)|$ by construction, while Lemma 2 gives $|\mathcal{V}_k(v)| \leq 4d_v(4\Delta)^{k-2}$ and $d_g(v) \leq k\Delta$. Thus:

$$\begin{aligned}\mathrm{vol}_{\overline{U}}(\mathcal{U}) &= \frac{1}{k-1}\sum_{v\in\nu}\sum_{g\in\mathcal{U}\cap\mathcal{V}_k(v)} d_g(v) \\ &< \sum_{v\in\nu}\frac{d_v}{2(k-1)}4(4\Delta)^{k-1}\end{aligned} \tag{3}$$

Let us now consider each $v \in \nu$ in turn. Note that $\overline{\mathcal{U}} \cap \mathcal{V}_k(v) \neq \emptyset$ since $\nu \subseteq \overline{U}$, and $\mathcal{U} \cap \mathcal{V}_k(v) \neq \emptyset$ by definition of $\nu$. By Lemma 1, there is an edge $g\bar{g} \in \mathcal{G}_k(v)$ with $g \in \mathcal{U} \cap \mathcal{V}_k(v)$ and $\bar{g} \in \overline{\mathcal{U}} \cap \mathcal{V}_k(v)$. Let $N(g \cap \bar{g}) = \cup_{u\in g\cap\bar{g}} N(u)$. Consider the graph $\mathcal{G}_k(g \cap \bar{g})$; observe that it is a clique of size at least $\frac{d_v}{2(k-2)} + 1$. We let $f(v) = C(T) \cap \mathcal{G}_k(g \cap \bar{g})$, and note that $|f(v)| \geq \frac{d_v}{2(k-2)}$. Finally, if $h\bar{h} \in f(v)$ then $v \in h\bar{h}$; and since $|h \cap \bar{h}| = k-1$, then $|\{u : h\bar{h} \in f(u)\}| \leq k-1$, i.e. $h\bar{h}$ is counted at most $k-1$ times. Summing over $\nu$ yields $\sum_{v\in\nu}\frac{d_v}{2(k-1)}\frac{1}{(k-2)} \leq c(\mathcal{U})$; comparing to (3) concludes the proof. □

We can now prove:

**Lemma 4.** $\Phi(\mathcal{G}_k) > \frac{1}{2}\frac{c(\mathcal{U})}{\mathrm{vol}_U(\mathcal{U})}$.

**Proof.** By chaining (2) and Lemma 3, we obtain:

$$\frac{1}{8k(4\Delta)^{k-1}} > \Phi(\mathcal{G}_k) = \frac{c(\mathcal{U})}{\mathrm{vol}_U(\mathcal{U}) + \mathrm{vol}_{\overline{U}}(\mathcal{U})} > \frac{c(\mathcal{U})}{\mathrm{vol}_U(\mathcal{U}) + 4k(4\Delta)^{k-1}c(\mathcal{U})} \tag{4}$$

This implies $4k(4\Delta)^{k-1}c(\mathcal{U}) < \mathrm{vol}_U(\mathcal{U})$, which plugged in the right-hand side proves the thesis. □

By Lemma 4 we can now focus on bounding $\frac{c(\mathcal{U})}{\mathrm{vol}_U(\mathcal{U})}$. We indeed bound $c(\mathcal{U})$ and $\mathrm{vol}_U(\mathcal{U})$ in the following two lemmata.

**Lemma 5.** $c(\mathcal{U}) \geq k^{-2}c(U)$.

**Proof.** Consider any edge $uv \in C(U)$ with $u \in U$, $v \in \overline{U}$. First, $\mathcal{G}_k(\{u,v\})$ is connected by Lemma 1. So if $\mathcal{G}_k(\{u,v\}) \cap C(\mathcal{U}) \neq \emptyset$ then there is an edge $g\bar{g} \in C(\mathcal{U})$ with $u \in g$ and $v \in \bar{g}$. Suppose instead that $\mathcal{G}_k(\{u,v\}) \cap C(\mathcal{U}) = \emptyset$, so $\mathcal{V}_k(\{u,v\}) \subseteq \mathcal{U}$ or $\mathcal{V}_k(\{u,v\}) \subseteq \overline{\mathcal{U}}$. Assume $\mathcal{V}_k(\{u,v\}) \subseteq \mathcal{U}$ (a symmetric argument applies to the other case). Fix any $\bar{g} \in \mathcal{V}_k(v) \cap \overline{\mathcal{U}}$ and let $g = \bar{g} \setminus \{z\} \cup \{u\}$ where $z \neq u$ is any leaf in a spanning tree of $\bar{g}$ rooted at $v$. Then $uv \in g$, and so $g \in \mathcal{V}_k(\{u,v\}) \subseteq \mathcal{U}$. Therefore again $g\bar{g} \in C(\mathcal{U})$ with $u \in g$ and $v \in \bar{g}$. In conclusion, every edge $uv \in C(U)$ can be mapped to an edge $g\bar{g} \in C(\mathcal{U})$ where $u \in g$ and $v \in \bar{g}$. Since for any $g\bar{g}$ we have at most $k^2$ such pairs $uv$, it follows that $g\bar{g}$ can be the image of at most $k^2$ distinct edges of $C(U)$. □

**Lemma 6.** $\mathrm{vol}_U(\mathcal{U}) \leq k(4\Delta)^{k-1}\,\mathrm{vol}(U)$.

**Proof.** For any $v$ we have $d_g(v) \leq d_g$. Moreover, Lemma 2 gives $|\mathcal{V}_k(v)| \leq 4d_v(4\Delta)^{k-2}$ and $d_g \leq k(k-1)\Delta$. Via the definition of $\mathrm{vol}_U(\mathcal{U})$ we then obtain:

$$\mathrm{vol}_U(\mathcal{U}) = \frac{1}{k-1}\sum_{v\in U}\sum_{g\in\mathcal{U}\cap\mathcal{V}_k(v)} d_g(v) \leq \frac{1}{k-1}\sum_{v\in U} 4d_v(4\Delta)^{k-2}k(k-1)\Delta \tag{5}$$

and the last term equals $k(4\Delta)^{k-1}\,\mathrm{vol}(U)$. □

It remains to wrap up the results proven above, taking care of two possible cases. Suppose first that $\mathrm{vol}(U) \leq \frac{1}{2}\mathrm{vol}(G)$ (the "good" case). This means $c(U)/\mathrm{vol}(U)$ is a valid upper bound to $\Phi(G)$. Then by chaining Lemma 4, Lemma 5 and Lemma 6 we obtain:

$$\Phi(\mathcal{G}_k) \geq \frac{1}{2}\frac{c(\mathcal{U})}{\mathrm{vol}_U(\mathcal{U})} \geq \frac{1}{2}\frac{k^{-2}c(U)}{k(4\Delta)^{k-1}\,\mathrm{vol}(U)} > \frac{1}{4k^3(4\Delta)^{k-1}}\Phi(G) \tag{6}$$

Suppose now $\mathrm{vol}(U) > \frac{1}{2}\mathrm{vol}(G)$ (the "bad" case). We then use $c(U)/\mathrm{vol}(\overline{U})$ as an upper bound to $\Phi(G)$, but we must bound $\mathrm{vol}(\overline{U})$ in terms of $\mathrm{vol}_U(\mathcal{U})$. Recall that $\Phi(\mathcal{G}_k) < \frac{1}{8k(4\Delta)^{k-1}}$. On one side, this implies $\mathrm{vol}(\mathcal{U}) \geq 8k(4\Delta)^{k-1}c(\mathcal{U})$. On the other side, it implies $\mathrm{vol}_U(\overline{\mathcal{U}}) < 4k(4\Delta)^{k-1}c(\mathcal{U})$; this follows from Lemma 3, by noting that it holds with $U$ and $\overline{\mathcal{U}}$ in place of $\overline{U}$ and $\mathcal{U}$. However, $\mathrm{vol}(\mathcal{U}) \leq \mathrm{vol}(\overline{\mathcal{U}}) = \mathrm{vol}_U(\overline{\mathcal{U}}) + \mathrm{vol}_{\overline{U}}(\overline{\mathcal{U}})$, thus $\mathrm{vol}_{\overline{U}}(\overline{\mathcal{U}}) \geq \mathrm{vol}(\mathcal{U}) - \mathrm{vol}_U(\overline{\mathcal{U}}) > 4k(4\Delta)^{k-1}c(\mathcal{U})$ or, equivalently, $c(\mathcal{U}) < \frac{1}{4k(4\Delta)^{k-1}}\mathrm{vol}_{\overline{U}}(\overline{\mathcal{U}})$.

We turn to $\mathrm{vol}_U(\mathcal{U})$. First, $\mathrm{vol}_U(\mathcal{U}) \leq \mathrm{vol}(\mathcal{U}) \leq \mathrm{vol}(\overline{\mathcal{U}}) = \mathrm{vol}_U(\overline{\mathcal{U}}) + \mathrm{vol}_{\overline{U}}(\overline{\mathcal{U}})$. Lemma 3 and the bound on $c(\mathcal{U})$ give $\mathrm{vol}_U(\overline{\mathcal{U}}) < \mathrm{vol}_{\overline{U}}(\overline{\mathcal{U}})$. Then by Lemma 6:

$$\mathrm{vol}_U(\mathcal{U}) \leq \mathrm{vol}_U(\overline{\mathcal{U}}) + \mathrm{vol}_{\overline{U}}(\overline{\mathcal{U}}) < 2\,\mathrm{vol}_{\overline{U}}(\overline{\mathcal{U}}) \leq 2k(4\Delta)^{k-1}\,\mathrm{vol}(\overline{U}) \tag{7}$$

By invoking (7), Lemma 4, and Lemma 5, we finally obtain:

$$\Phi(\mathcal{G}_k) > \frac{1}{2}\frac{c(\mathcal{U})}{\mathrm{vol}_U(\mathcal{U})} > \frac{k^{-2}c(U)}{4k(4\Delta)^{k-1}\,\mathrm{vol}(\overline{U})} \geq \frac{1}{4k^3(4\Delta)^{k-1}}\Phi(G) \tag{8}$$

concluding the proof of Theorem 2.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

M. Bressan and S. Haddadan are supported in part by the ERC Starting Grant DMAP 680153, by a Google Focused Award "Algorithms and Learning for AI", and by the MIUR grant "Dipartimenti di Eccellenza 2018-2022" awarded to the Department of Computer Science of the Sapienza University of Rome.

## References

[1] M. Bressan, F. Chierichetti, R. Kumar, S. Leucci, A. Panconesi, Counting graphlets: space vs time, in: ACM WSDM 2017, pp. 557–566..
[2] M. Bressan, F. Chierichetti, R. Kumar, S. Leucci, A. Panconesi, Motif counting beyond five nodes, ACM Trans. Knowl. Discov. Data 12 (4) (2018) 48.
[3] M. Bressan, S. Leucci, A. Panconesi, MOTIVO: fast motif counting via succinct color coding and adaptive sampling, Proc. VLDB Endow. 12 (11) (2019) 1651–1663.
[4] M.A. Bhuiyan, M. Rahman, M. Rahman, M. Al Hasan, GUISE: uniform sampling of graphlets for large graph analysis, in: IEEE ICDM 2012, pp. 91–100.
[5] P. Wang, J.C.S. Lui, B. Ribeiro, D. Towsley, J. Zhao, X. Guan, Efficiently estimating motif statistics of large networks, ACM Trans. Knowl. Discov. Data 9 (2) (2014) 8.
[6] X. Chen, Y. Li, P. Wang, J.C.S. Lui, A general framework for estimating graphlet statistics via random walk, Proc. VLDB Endow. 10 (3) (2016) 253–264.
[7] G. Han, H. Sethu, Waddling random walk: fast and accurate mining of motif statistics in large graphs, in: IEEE ICDM 2016, pp. 181–190.
[8] A. Mohaisen, A. Yun, Y. Kim, Measuring the mixing time of social graphs, in: ACM SIGCOMM 2010, pp. 383–389.
[9] D.A. Levin, Y. Peres, E.L. Wilmer, Markov Chains and Mixing Times, American Mathematical Society, 2009.