ESTIMATING GRAPH PARAMETERS VIA RANDOM WALKS WITH RESTARTS

Anna Ben-Hamou, Roberto I Oliveira, Yuval Peres

ESTIMATING GRAPH PARAMETERS VIA RANDOM WALKS WITH RESTARTS

ANNA BEN-HAMOU, ROBERTO I. OLIVEIRA, AND YUVAL PERES

Abstract. In this paper we discuss the problem of estimating graph parameters from a random walk with restarts. In this setting, an algorithm observes the trajectory of a random walk over an unknown graph G, starting from a vertex x. The algorithm also sees the degrees along the trajectory. The only other power that the algorithm has is to request that the random walk be reset to its initial state x at any given time, based on what it has seen so far. Our main results are as follows. For *regular graphs* G, one can estimate the number of vertices n_G and the ℓ^2 mixing time of G from x in $\widetilde{O}(\sqrt{n_G}\,(t^G_{\text{unif}})^{3/4})$ steps, where t^G_{unif} is the uniform mixing time on G. The algorithm is based on the number of intersections of random walk paths X, Y, *i.e.* the number of times (t, s) such that $X_t = Y_s$. Our method improves on previous methods by various authors which only consider collisions (*i.e.* times t with $X_t = Y_t$). We also show that the time complexity of our algorithm is optimal (up to log factors) for 3-regular graphs with prescribed mixing times. For *general graphs*, we adapt the intersections algorithm to compute the number of edges m_G and the ℓ^2 mixing time from the starting vertex x in $\widetilde{O}(\sqrt{m_G}\,(t^G_{\text{unif}})^{3/4})$ steps. Under mild additional assumptions (which hold *e.g.* for sparse graphs) the number of vertices can also be estimated by this time. Finally, we show that these algorithms, which may take sublinear time, have a *fundamental limitation*: it is not possible to devise a sublinear *stopping time* at which one can be reasonably sure that our parameters are well estimated. On the other hand, we show that, given either m_G or the mixing time of G, we can compute the "other parameter" with a self-stopping algorithm.

1. Introduction

What can one learn about a graph from random walk trajectories on it? The (trivial) answer is that, given enough time and resources, we can learn everything. If the graph is connected, and one is willing to wait for long enough, eventually all edges of the graph are crossed by the walk. Given some more time, one can even be nearly sure that no other edges exist and it is safe to stop exploring the graph.

This paper is inspired by a more interesting question: what can one learn from the random walk way before the graph is fully covered? Our motivation is the analysis of large networks that can contain millions (or even billions) of nodes and edges. Direct manipulation or full observations of such huge graphs are typically impractical. Random-walk-based methods, which are local and lightweight, are often used in dealing with this kind of graph (see Das Sarma et al. [8] and the references therein). Our problem, then, is

to determine the least number of random walk steps that are needed to compute interesting graph parameters via random walks.

Our main contribution is to analyze this problem in an algorithmic model that we call *random walks with restarts* (RWR). We assume our algorithm has black-box access to a random walk on a graph G starting from a vertex x. At each time step t, the algorithm sees the current vertex and its degree. It then decides whether it wants to jump to a neighboring vertex, or to "reset" the walker back to x.

The algorithm produces an estimate $\widehat{\gamma}_t$ of a parameter $\gamma = \gamma(G)$ of interest after t time steps, solely by looking at the traces of the random walk and the vertex degrees along the way. The goal is to achieve

$$\forall t \geq t_0 \,:\, \mathbb{P}_x^G\left(\left|\frac{\widehat{\gamma}_t}{\gamma(G)} - 1\right| \leq \frac{1}{2}\right) \geq 1 - \varepsilon,$$

with t_0 as small as possible.

In general, the time complexity parameter t_0 will depend on the error parameter ε and on unknown characteristics of the graph. This leads us to consider the possibility of "self-stopping" algorithms that decide on their own when to stop exploring G.

Section 2 defines the RWR model and self stopping algorithms more precisely. For now, we point out that our model is *one of the most restrictive models for random walk algorithms that actually make sense in real life.* In practical settings, if we can simulate a random walk over a graph G, we can most likely restart it at will, and also compute degrees along the way.

1.1. **What we do.** In a nutshell, this paper gives nearly optimal algorithms for estimating the *number of vertices, number of edges, and mixing time of G from the starting point x* in the RWR model. For regular graphs G, one can estimate these parameters with about $\sqrt{n}\, t_{\text{unif}}^{3/4}$ random walk steps, where n is the number of vertices and t_{unif} is the uniform mixing time of G. For general graphs, our algorithms use about $\sqrt{m}\, t_{\text{unif}}^{3/4}$ steps, where m is the number of edges (this requires minor assumptions on degrees in the case of estimating n). For estimation of n in regular graphs, these complexity bounds on the number of random walk steps are then shown to be optimal up to a factor of order at most $(\log n)^{3/4}$.

Let us describe our results in more detail, postponing the definition of the model to Section 2. In Section 3, we review results by Peres et al. [17] on intersections of two independent random walks X, Y on a regular graph. By definition, intersections are pairs of times (t, s) with $X_t = Y_s$. Using intersection counts gives us a simple algorithm for estimating numbers of vertices n_G of a regular graph G in $O\left(\sqrt{n_G}\,(t_{\text{unif}}^G)^{3/4}\right)$ steps, where t_{unif}^G is its uniform mixing time.

In Section 4, we prove that this algorithm is optimal up to a factor of $(\ln n_G)^{3/4}$. More specifically, for any pre-specified function $\mathbf{t} : \mathbb{N} \to \mathbb{N}$, we construct an infinite sequence of 3-regular graphs G with uniform mixing time $t_{\text{unif}}^G = O(\mathbf{t}(n_G))$. We then show that

any RWR algorithm that finds the number of vertices of these graphs requires at least $\Omega\left(\sqrt{n_G}\left(\frac{t^G_{\text{unif}}}{\ln n_G}\right)^{3/4}\right)$ time steps.

Our next step is to consider arbitrary graphs G. In Section 5, we adapt the intersections algorithm to show that the number of edges m_G of G can be estimated in time $O(\sqrt{m_G}\,(t^G_{\text{unif}})^{3/4})$. Under simple assumptions – for instance, if G is sparse –, the same bounds apply to estimating the number of vertices n_G.

Up to this point all algorithms we described are essentially optimal for our model. They are also space-efficient. They just need to store a single real number and maintain a list of visits to each vertex, which is only read or changed during visits. Another desirable trait of our algorithms is that they run in sub-linear time when the mixing time is small (less than $o(m_G^{2/3})$). This property of (relatively) fast mixing is expected to hold in social networks [12] and other large graphs.

However, our algorithms also suffer from a serious drawback: they are not self-stopping. As it turns out, this is unavoidable. We argue in Section 6 that self-stopping algorithms for the number of vertices must cover nearly all edges of the graph. This is true even if our graph is guaranteed to be 3-regular and have polylog mixing time. We deduce that, while it may be possible to know the size of a graph after sub-linear time, knowing that we already know the size may take much longer.

We complement these results by showing that if either m_G or the mixing time is known, the other parameter can be estimated with few steps via a self-stopping algorithm. In Section 7, we show how one can use an upper-bound τ on the mixing time to compute the number of edges via a self-stopping algorithm with time complexity $O\left(\sqrt{m_G}\,\tau^{3/4}\log\log m_G\right)$ (or $O(\sqrt{n_G}\,\tau^{3/4}\log\log n_G)$ steps if G is regular). Section 8 then presents a result for estimating $t_x(\delta)$, the ℓ_2-mixing time from x, with time complexity $O(\sqrt{m_G}\,(t^G_{\text{unif}})^{3/4}\log\log m_G)$, assuming a good estimate for the number of edges is available. A corollary is that both the mixing time from x and the number of edges m can be approximated by a self-stopping algorithm with time complexity $O\left(\sqrt{m_G}\,\tau^{3/4}\log\log m_G\right)$, assuming an upper-bound τ on the uniform mixing time is available.

1.2. **Background.** Our result relates to the a large body of work on inferring graph (or Markov chain) parameters from random walks. We give here a brief overview of these papers, with a focus on results most closely resembling ours.

In some cases, one has to estimate parameters from a single path of the random walk. One possibility is to use return times to the initial vertex to estimate n_G or m_G, as proposed by Cooper et al. [7] and Benjamini et al. [3]. Other parameters, such as the spectral gap, may be quite challenging to estimate (see Hsu et al. [9] and Levin and Peres [13]). In any case, all of these algorithms require time that is at least of the order of the number of vertices, whereas our own algorithms are sublinear in certain cases.

Another line of work, which is similar to our random walks with restarts, is to consider several random walks on the same graph. Typically, estimators in this case rely on

collisions of random walks at their endpoints. Assume for instance that G is regular and an upper bound for the mixing time is known. In order to find the number of vertices of G, one can then run k random walks from point x for a time larger then t_{mix}. The endpoints form an independent sample with nearly uniform distribution over the vertex set. There are then multiple methods for estimating the number of vertices, most of which take advantage of the birthday paradox (see [6] for a review). As the first collision in an I.I.D. drawn from the uniform distribution occurs at time of order $\sqrt{n}$, the running time of such procedures on regular graphs is typically $O(t_{\text{mix}}\sqrt{n})$. Less brutal strategies allow to improve this upper-bound up to $O(t_{\text{rel}}\sqrt{n})$, where t_{rel} is the relaxation time of the walk. Similar methods may be designed to estimate the number of edges and vertices in non-regular graphs Katzir et al. [11] and to estimate mixing times Benjamini and Morris [2]. Our results show that random walk *intersections*, which take whole trajectories into account, give strictly more information than collisions, and lead to nearly optimal time dependence on the mixing time.

1.3. **Future directions.** Our results are just a first step towards understanding estimation via random walks. It would be interesting to understand what other graph parameters can be computed efficiently in our model. Extensions of our results to oriented graphs and other models of access to the graph (including distributed access as in [8]) would also be worthwhile.

2. Notation and definitions

Let $G = (V, E)$ be a finite connected graph on n_G vertices and m_G edges. For $u \in V$, we let $N(u)$ be the set of neighbors of u in G, and $\deg(u) = |N(u)|$ be the degree of u. Let also

$$\mathcal{V} = \bigcup_{t=0}^{\infty} V^t ,$$

the set of finite length sequences of elements of V.

2.1. **Patterns of sequences.** For $t \geq 0$ and for a sequence of vertices $u_0^t = (u_0, \dots, u_t) \in V^{t+1}$, let $r(u_i)$ be the index of the first occurrence of u_i in u_0^t and define the *pattern* of u_0^t as the length-$(t+1)$ sequence $(r(u_0), \dots, r(u_t))$, each vertex being replaced by its rank of occurrence in u_0^t. For instance, the pattern of $(g, a, a, c, g, d, a, b, d)$ is $(1, 2, 2, 3, 1, 4, 2, 5, 4)$. Note that the pattern is invariant under vertex-relabelling. Also let Φ be the map defined on $\mathcal{V}$ by: for all $t \geq 0$, for all $u_0^t \in V^{t+1}$,

$$\Phi(u_0^t) = \Big((r(u_i), \deg(u_i)) \Big)_{i=0}^{t} .$$

In other words, for each finite length sequence of vertices u_0^t, the function Φ captures the pattern and the sequence of degrees, and takes values in

$$\mathcal{S} = \bigcup_{t \geq 0} (\mathbb{N} \times \mathbb{N})^t .$$

From now on, the term *pattern* will actually refer to the function Φ.

2.2. **Random walk with restarts and estimators.** Fix $x \in V$ and a map RESTART : $\mathcal{S} \to \{0,1\}$, and generate a sample as follows: initially $X_0 = x$ and for all $s \geq 0$, conditionally on $X_0^s = (X_0, \ldots, X_s)$, the distribution of X_{s+1} is given by: for all $y \in V$,

$$\mathbb{P}\left(X_{s+1} = y \,|\, X_0^s\right) \;=\; \begin{cases} \mathbb{1}_{\{y=x\}} & \text{if } \textsc{restart}\left(\Phi(X_0^s)\right) = 1 \;, \\ \frac{1}{\deg(X_s)} \mathbb{1}_{\{y \in N(X_s)\}} & \text{if } \textsc{restart}\left(\Phi(X_0^s)\right) = 0 \;, \end{cases}$$

The sample $X_0^t = (X_0, \ldots, X_t)$ will be called a random walk with restarts (RWR) at x, and we denote by $\mathbb{P}_x^G$ the corresponding probability measure over trajectories. To avoid periodicity issues, it will be convenient to consider the *lazy* version of a RWR: if $\textsc{restart}\left(\Phi(X_0^s)\right) = 0$, the walk stays at its current position with probability $1/2$, and moves to a uniformly chosen neighbor with probability $1/2$.

An *estimator* is a pair (RESTART, EST) with RESTART : $\mathcal{S} \to \{0,1\}$ and EST : $\mathcal{S} \to \mathbb{R}$, which returns the value $\textsc{est}\left(\Phi(X_0^t)\right)$ for a RWR X_0^t characterized by RESTART. More precisely, letting $\gamma(G)$ be some parameter of interest (*e.g.* $\gamma(G) = n_G$ or $\gamma(G) = m_G$), the goal is to produce a map RESTART : $\mathcal{S} \to \{0,1\}$ and an estimator EST : $\mathcal{S} \to \mathbb{R}$ such that, for all graph $G = (V, E)$ for all starting point $x \in V$, for all $t \geq t(\varepsilon, G)$

$$\mathbb{P}_x^G \left(\left| \frac{\textsc{est}\left(\Phi(X_0^t)\right)}{\gamma(G)} - 1 \right| > \frac{1}{2} \right) \leq \varepsilon \,, \tag{2.1}$$

for $t(\varepsilon, G)$ as small as possible.

2.3. **Lower bounds.** The lower bound problem can be formalized as follows: we say that $t(G)$ is a lower bound for the estimation of $\gamma(G)$ if there exists $\delta > 0$ such that for all function RESTART, for all estimator EST, there exists an infinite sequence of graphs G and $x \in V(G)$ such that for all $t \leq \delta t(G)$,

$$\mathbb{P}_x^G \left(\left| \frac{\textsc{est}\left(\Phi(X_0^t)\right)}{\gamma(G)} - 1 \right| > \frac{1}{2} \right) \geq \frac{1}{4} \,. \tag{2.2}$$

To obtain more refined lower bound, one may further require that all graphs in the infinite sequence belong to some specified class.

2.4. **Self-stopping algorithms.** The time $t(\varepsilon, G)$ above which inequality (2.1) holds usually depends on unknown parameters of the graph, possibly on $\gamma(G)$ itself. This prompts the search for *self-stopping* algorithms.

In addition to the functions RESTART and EST, self-stopping algorithms also rely on a function STOP : $\mathcal{S} \to \{0,1\}$. Defining, for a RWR X_0^t (for a given function RESTART),

$$\tau = \inf\{t \geq 0,\, \textsc{stop}\left(\Phi(X_0^t)\right) = 1\} \,,$$

then the self-stopping algorithm defined by RESTART, STOP and EST returns the value $\textsc{est}\left(\Phi(X_0^\tau)\right)$. One then has to control the deviations of $\textsc{est}\left(\Phi(X_0^\tau)\right)$ with respect to $\gamma(G)$ and the expectation of the stopping time τ.

3. Intersections and regular graphs

3.1. **Definitions and preliminary results.** Let $G = (V, E)$ be a finite connected regular graph with n vertices. Let X and Y be two independent lazy random walks started at the same vertex $x \in V$. Define

$$I_t \;=\; \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} \mathbb{1}_{\{X_i = Y_j\}}\,,$$

i.e. I_t is the number of intersections between the trajectories of X and Y up to time $t-1$. Let also

$$g_t(x,u) \;=\; \sum_{i=0}^{t-1} \mathbb{P}_x(X_i = u)$$

be the expected number of visits to vertex u before time t (also known as the Green's function). It is not hard to see that

$$\mathbb{E}_x I_t \;=\; \sum_{u \in V} g_t(x,u)^2\,. \tag{3.1}$$

Denote by t_{unif} the uniform mixing time of the chain, *i.e.*

$$t_{\mathrm{unif}} \;=\; \inf\left\{ t \ge 0,\; \max_{x,y \in V} \left| \frac{\mathbb{P}_x(X_t = y)}{\pi(y)} - 1 \right| \le \frac{1}{4} \right\}\,,$$

where π is the stationary distribution of the chain. Also, letting $1 = \lambda_1 > \lambda_2 \ge \ldots \ge \lambda_n \ge 0$ be the eigenvalues of the transition matrix of the walk, in decreasing order (the fact that all eigenvalues are positive is by laziness of the walk), the *relaxation time* is defined as

$$t_{\mathrm{rel}} \;=\; \frac{1}{1-\lambda_2}\,.$$

For regular graphs, the following inequalities were established by Peres et al. [17].

Lemma 1 ([17])**.** *Assume that $G = (V, E)$ is a finite connected d-regular graph. Let*

$$t_\star \;=\; \min\left\{ t_{\mathrm{unif}}, t_{\mathrm{rel}}\left(\log(t_{\mathrm{rel}}) + 1\right) \right\}\,.$$

Then for all $x \in V$,

$$\frac{t^2}{n} \;\le\; \mathbb{E}_x I_t \;\le\; C t_\star^{3/2} + \frac{5t^2}{4n}\,,$$

for some universal constant $C \ge 0$. Moreover

$$\mathbb{E}_x I_t^2 \;\lesssim\; \left(t_\star^{3/2} + \frac{t^2}{n} \right) \mathbb{E}_x I_t\,.$$

Here and throughout the paper, for two functions f, g, the notation $f(n) \lesssim g(n)$ means that there exists an absolute constant $C > 0$ such that $f(n) \le C g(n)$ for all $n \ge 1$.

Let us note that a critical element in the proof of Lemma 1 is Aldous and Fill [1, Proposition 6.16], which establishes that on regular graphs, for all $t \le 5n^2$,

$$g_t(x,x) \;\le\; 5\sqrt{t}\,. \tag{3.2}$$

3.2. **A simple estimator for the number of vertices.** Lemma 1 suggests the following simple estimator for the number of vertices in a regular graph: consider $2K$ independent lazy random walks $X^{(1)}, Y^{(1)}, \ldots, X^{(K)}, Y^{(K)}$ all started at the same vertex $x \in V$. For each k between 1 and K, let $I_t^{(k)}$ be the number of intersections of $X^{(k)}$ and $Y^{(k)}$ between 0 and $t-1$, and define

$$\widehat{n}_t \;=\; \frac{t^2}{\frac{1}{K}\sum_{k=1}^{K} I_t^{(k)}}\,. \tag{3.3}$$

This estimator clearly falls into the RWR model, the function RESTART being simply given by

$$\textsc{restart}\left(\Phi(X_0^T)\right) = \begin{cases} 1 & \text{if } T+1 \equiv 0 \pmod{t}\ , \\ 0 & \text{otherwise.} \end{cases}$$

For $t \gtrsim t_\star^{3/4}\sqrt{n}$ and K large enough, this estimator starts returning a faithful value for n. Indeed, using Lemma 1, for a large enough constant A and $t \geq A t_\star^{3/4}\sqrt{n}$, we have $\frac{t^2}{n} \leq \mathbb{E}_x I_t \leq \frac{3t^2}{2n}$ and $\mathrm{Var}_x I_t \lesssim (\mathbb{E}_x I_t)^2$. Hence, by Chebyshev's Inequality

$$\mathbb{P}_x^G\left(\left|\frac{\widehat{n}_t}{n} - 1\right| > \frac{1}{2}\right) \leq \mathbb{P}_x^G\left(\frac{1}{K}\sum_{k=1}^{K} I_t^{(k)} > \frac{2t^2}{n}\right) + \mathbb{P}_x^G\left(\frac{1}{K}\sum_{k=1}^{K} I_t^{(k)} < \frac{2t^2}{3n}\right) \;=\; O\left(\frac{1}{K}\right).$$

The case of the cycle on n vertices gives an example where this bound is tight. Indeed, in this case, $t_{\mathrm{unif}} \asymp t_{\mathrm{rel}} \asymp n^2$, and thus $t_\star^{3/4}\sqrt{n} \asymp n^2$. And any procedure based on random walks requires at least order n^2 steps to distinguish between a cycle of size n and a cycle of size $2n$. Section 4 is devoted to the elaboration of a more refined lower bound.

4. Lower bounds for regular graphs

For a given function $\mathbf{t} : \mathbb{N} \to \mathbb{N}$, let us denote by $\mathcal{C}(\mathbf{t})$ the class of connected 3-regular graphs with uniform mixing time t_{unif}^G smaller than $\mathbf{t}(n_G)$. Note that for the class $\mathcal{C}(\mathbf{t})$ to be non-empty, one has at least to assume $\mathbf{t}(n) \geq \frac{\log(3n/4)}{\log(3)}$, which is a general lower bound for the mixing time of 3-regular graphs (see [14, Chapter 7]).

Proposition 2. *There exists $\delta > 0$ such that for any function $\mathbf{t} : \mathbb{N} \to \mathbb{N}$ with $\mathbf{t}(\cdot) \geq 5\log(\cdot)$, for any functions* RESTART $: \mathcal{S} \to \{0,1\}$ *and* EST $: \mathcal{S} \to \mathbb{N}$, *there exists an infinite sequence of 3-regular graphs $G \in \mathcal{C}(\mathbf{t})$ and $x \in V(G)$ such that, for all $t \leq \delta\left(\frac{t_{\mathrm{unif}}^G}{\log n_G}\right)^{3/4}\sqrt{n_G}$,*

$$\mathbb{P}_x^G\left(\left|\frac{\textsc{est}\left(\Phi(X_0^t)\right)}{n_G} - 1\right| > \frac{1}{2}\right) \;\geq\; \frac{1}{4}\,,$$

where X_0^t is a RWR *characterized by* RESTART.

Before proving Proposition 2, we first establish the following lemma.

Lemma 3. *There exists $k_0 \geq 1$ such that for all even $k \geq k_0$, there exists a connected 3-regular graph $\mathcal{E}_k$ with $|V(\mathcal{E}_k)| = k$ and $x \in V(\mathcal{E}_k)$ satisfying*

- $t_{\mathrm{unif}}(\mathcal{E}_k) \leq 5 \log k$,
- *denoting $\mathbf{G}_t$ the subgraph spanned by the edges visited by a random walk on $\mathcal{E}_k$ with restarts at x (for any function* RESTART*), then, if $t \leq \sqrt{k}/3$,*

$$\mathbb{P}_x^{\mathcal{E}_k}(\mathbf{G}_t \text{ is a tree}) \geq \frac{3}{4}. \tag{4.1}$$

Proof of Lemma 3. To establish Lemma 3, it is sufficient to show that, with positive probability, a uniform random 3-regular graphs satisfy those properties. First, by the results of [5], we know that, with probability tending to 1 with k, a uniform random 3-regular graph is an expander, and thus the mixing time of the simple random walk on such a graph is of order $\log k$. Lubetzky et al. [15] actually determine the precise order of the mixing time: with probability tending to 1, a uniform random d-regular graphs on k vertices has mixing time equivalent to $\frac{d}{d-2}\frac{\log k}{\log(d-1)}$. For $d = 3$, we see that the first property in Lemma 3 easily follows. Now, to establish (4.1), we use a common method to generate a uniform 3-regular random graph, known as the *configuration model* (see [4]). One initially considers k isolated vertices, each vertex v being endowed with 3 *half-edges* $(v, 1)$, $(v, 2)$, $(v, 3)$. A random matching on the half-edges is then chosen uniformly, and each pair of half-edges is interpreted as an edge between the corresponding vertices. It is well-known that G_k is simple with probability bounded away from 0 (see for instance [10]), and that, conditionally on being simple, its distribution is uniform over simple 3-regular graphs. One nice feature of this model is that it allows to generate sequentially and simultaneously the graph and the random walk (with restarts), as follows. Let RESTART be any function from $\mathcal{S}$ to $\{0, 1\}$. Initially, all half-edges are unpaired and $X_0 = x$. Then, at each step $s \geq 0$,

- either $\textsc{restart}(\Phi(X_0^s)) = 1$ and we set $X_{s+1} = x$,
- or $\textsc{restart}(\Phi(X_0^s)) = 0$ and we then choose with probability $1/3$ a half-edge $(X_s, *)$ attached to X_s. If $(X_s, *)$ has already been paired to some half-edge $(v, *)$, we let $X_{s+1} = v$. Otherwise, we choose uniformly at random an unpaired half-edge $(u, *)$, match $(X_s, *)$ and $(u, *)$, and let $X_{s+1} = u$.

With this procedure, it is not hard to see that the first cycle is formed at time s with probability smaller than $\frac{3s}{3k-3s}$ (by time s, we have exposed at most $3s$ half-edges). Hence, the (annealed) probability that $\mathbf{G}_t$ contains a cycle is smaller than $\frac{3t^2}{3k-3t}$. For $t = \sqrt{k}/3$, this probability is smaller than $1/8$. If $\mathbb{P}_x^{G_k}$ denotes the (quenched) probability associated with the random walk on G_k with restarts at x, then, by Markov's Inequality,

$$\mathbb{P}\left(\left\{\mathbb{P}_x^{G_k}(\mathbf{G}_t \text{ is not a tree}) > \frac{1}{4}\right\}\right) \leq \frac{1}{2}.$$

This entails that, with positive probability, a uniform 3-regular random graph satisfies (4.1). ∎

In particular, it means that for any function RESTART, one can find a coupling (X, Y) where X (*resp.* Y) is a random walk on $\mathcal{E}_k$ with restarts at x (*resp.* on $\mathcal{E}_{4k}$ with restarts at y), such that for all $t \leq \sqrt{k}/3$,

$$\mathbb{P}_{x,y}\left(\Phi(X_0^t) = \Phi(Y_0^t)\right) \geq \frac{3}{4}\,, \tag{4.2}$$

since, with probability $3/4$, none of the walks is able to distinguish its base-graph from a 3-regular infinite tree.

Let us now turn to the proof of Proposition 2.

Proof of Proposition 2. For $k \geq 1$, consider a 3-regular graph $\mathcal{E}_k$ satisfying the properties of Lemma 3. Now, in place of each edge $e \in E(\mathcal{E}_k)$, put a path of length $m \geq 1$. To make the graph 3-regular, we add edges between pairs of interior vertices at distance 2 within the same path (assuming $m-1$ is even). Let $G_{k,m}$ be the resulting graph (see Figure 1). Note that

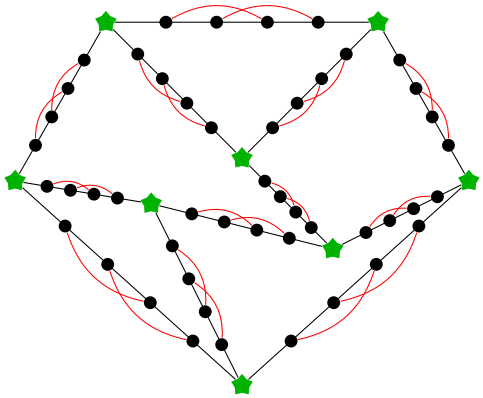

FIGURE 1. The graph $G_{k,m}$ ($k = 8$, $m = 5$). The green star-shaped vertices are the original vertices of $\mathcal{E}_k$.

$$n(G_{k,m}) = k + (m-1)\frac{3k}{2}\,, \tag{4.3}$$

and, as the time needed to cross one path is of order m^2, the mixing time on $G_{m,k}$ is

$$t_{\text{unif}}(G_{k,m}) = \Theta\left(m^2 \log k\right)\,. \tag{4.4}$$

For any target function $\mathbf{t}$ (with $\mathbf{t}(\cdot) \geq 5\log(\cdot)$), and for an infinity of integers n, the parameters k and m may be adjusted so that $n(G_{k,m}) = n$ and $t_{\text{unif}}(G_{k,m}) \leq \mathbf{t}(n)$. We now consider the graphs $G_{k,m}$ on n vertices and $G_{4k,m}$ on $4n$ vertices. Combining equation (4.2) and the m^2-slow down induced by paths, we get that we can find $\delta > 0$, starting points $(x, y) \in V(G_{k,m}) \times V(G_{4k,m})$, and a coupling (X, Y) of random walks

with restarts at x and y (for the same function RESTART) such that

$$\mathbb{P}_{x,y}\left(\mathbf{A}_t\right) \geq \frac{2}{3}, \quad \text{with} \quad \mathbf{A}_t = \left\{\Phi(X_0^t) = \Phi(Y_0^t)\right\} \quad \text{and} \quad t = \delta m^2\sqrt{k}\,. \tag{4.5}$$

Let $\textsc{est} : \mathcal{S} \to \mathbb{N}$ be an estimator, and define

$$B_t^X = \left\{\frac{1}{2} \leq \frac{\textsc{est}\left(\Phi(X_0^t)\right)}{n} \leq \frac{3}{2}\right\}, \quad \text{and} \quad B_t^Y = \left\{\frac{1}{2} \leq \frac{\textsc{est}\left(\Phi(Y_0^t)\right)}{4n} \leq \frac{3}{2}\right\}.$$

Assume that it holds simultaneously that $\mathbb{P}_x\left(B_t^X\right) \geq 3/4$ and $\mathbb{P}_y\left(B_t^Y\right) \geq 3/4$. Then, by (4.5),

$$\mathbb{P}_{x,y}\left(B_t^X \mid \mathbf{A}_t\right) = \frac{\mathbb{P}_{x,y}\left(B_t^X \cap \mathbf{A}_t\right)}{\mathbb{P}_{x,y}(\mathbf{A}_t)} \geq 1 - \frac{1 - \mathbb{P}_x\left(B_t^X\right)}{\mathbb{P}_{x,y}(\mathbf{A}_t)} \geq \frac{5}{8},$$

and similarly, $\mathbb{P}_{x,y}\left(B_t^Y \mid \mathbf{A}_t\right) \geq \frac{5}{8}$, so that

$$\mathbb{P}_{x,y}\left(B_t^X \cap B_t^Y \mid \mathbf{A}_t\right) \quad \geq \quad \frac{1}{4}.$$

However, on the event $\mathbf{A}_t$, the events B_t^X and B_t^Y can not occur simultaneously, implying a contradiction. Hence, we either have $\mathbb{P}_x\left(|\frac{\textsc{est}\left(\Phi(X_0^t)\right)}{n} - 1| > \frac{1}{2}\right) \geq \frac{1}{4}$ or $\mathbb{P}_y\left(|\frac{\textsc{est}\left(\Phi(Y_0^t)\right)}{4n} - 1| > \frac{1}{2}\right) \geq \frac{1}{4}$. The proof is then concluded by noticing that, thanks to (4.3) and (4.4),

$$m^2\sqrt{k} \quad \gtrsim \quad \left(\frac{t_{\text{unif}}}{\log n}\right)^{3/4}\sqrt{n}\,.$$

■

5. Computing parameters of general graphs

5.1. **Intersections on general graphs.** To deal with non-regular graphs, it will be convenient to consider a weighted version of the number of intersections, which we call the *weighted number of intersections*, defined as

$$\mathcal{I}_t \quad = \quad \sum_{i=0}^{t-1}\sum_{j=0}^{t-1} \frac{1}{\deg(X_i)}\mathbb{1}_{\{X_i = Y_j\}}\,,$$

where X and Y are two independent lazy simple random walks on G.

As in the previous section, t_{unif} and t_{rel} are respectively the uniform mixing time and the relaxation time of the walk, and $t_\star = \min\left\{t_{\text{unif}}, t_{\text{rel}}\left(\log(t_{\text{rel}}) + 1\right)\right\}$. The following Lemma is an analogue of Lemma 1 for non-regular graphs..

Lemma 4. *For all $x \in V$,*

$$\frac{t^2}{2m} \;\leq\; \mathbb{E}_x\mathcal{I}_t \;\leq\; 12t_\star^{3/2} + \frac{5t^2}{8m}\,.$$

Moreover

$$\mathbb{E}_x \mathcal{I}_t^2 \quad \lesssim \quad \left(t_\star^{3/2} + \frac{t^2}{2m} \right) \mathbb{E}_x \mathcal{I}_t \,.$$

Let us first state the following generalization to general graphs of inequality (3.2), as established by Lyons [16, Lemma 3.4].

Lemma 5 (Lyons [16])**.** *For a lazy random walk X on G, for all $t \geq 0$,*

$$\mathbb{P}_x(X_t = x) \quad \leq \quad \pi(x) + \frac{4 \deg(x)}{\sqrt{t+1}} \,.$$

Note that Lemma 5 implies that $g_t(x,x) \leq t\pi(x) + 8\deg(x)\sqrt{t}$ for all $t \geq 0$, which in turn yields that

$$\forall t \leq 16m^2, \quad g_t(x,x) \leq 10 \deg(x) \sqrt{t} \,. \tag{5.1}$$

Let us also recall the following alternative bound on the return probabilities, which follows for instance from [14, Inequality 12.11].

Claim 6. *For all $x \in V$ and $t \geq 0$,*

$$\mathbb{P}_x(X_t = x) \quad \leq \quad \pi(x) + \lambda_2^t \,.$$

We are now ready to prove Lemma 4.

Proof of Lemma 4. Using Jensen's Inequality,

$$\begin{aligned} \mathbb{E}_x \mathcal{I}_t \quad &= \quad \sum_u \frac{g_t(x,u)^2}{\deg(u)} \quad = \quad 2m \sum_u \pi(u) \left(\frac{g_t(x,u)}{\deg(u)} \right)^2 \\ &\geq \quad 2m \left(\sum_u \pi(u) \frac{g_t(x,u)}{\deg(u)} \right)^2 \quad = \quad \frac{t^2}{2m} \,, \end{aligned}$$

establishing the lower bound on the first moment. For the upper bound, the weighting in the definition of $\mathcal{I}_t$ allows us to use reversibility and obtain that

$$\begin{aligned} \mathbb{E}_x \mathcal{I}_t \quad &= \quad \frac{1}{\deg(x)} \sum_{i,j=0}^{t-1} \sum_u \mathbb{P}_x(X_i = u) \mathbb{P}_u(X_j = x) \\ &= \quad \frac{1}{\deg(x)} \sum_{i,j=0}^{t-1} \mathbb{P}_x(X_{i+j} = x) \,. \end{aligned} \tag{5.2}$$

Dividing the sum according to whether $i + j \geq t_\star$, we obtain

$$\mathbb{E}_x \mathcal{I}_t \quad \leq \quad \frac{1}{\deg(x)} \sum_{k=0}^{t_\star - 1} (k+1) \mathbb{P}_x(X_k = x) + \frac{1}{\deg(x)} \sum_{\substack{i,j=0 \\ i+j \geq t_\star}}^{t-1} \mathbb{P}_x(X_{i+j} = x) \,. \tag{5.3}$$

For the first term in the right-hand side of (5.3), we note that as $t_\star \leq t_{\mathrm{unif}}$ and as, on all connected graphs $t_{\mathrm{unif}} \leq 8mn$ (combining for instance [1, Corollary 6.8] and [14, Theorem 10.14]), we may resort to inequality (5.1) and get

$$\frac{1}{\deg(x)} \sum_{k=0}^{t_\star - 1} (k+1) \mathbb{P}_x(X_k = x) \quad \leq \quad \frac{t_\star g_{t_\star}(x,x)}{\deg(x)} \quad \leq \quad 10 t_\star^{3/2} \,.$$

For the second term in the right-hand side of (5.3), we consider two cases. The first case is when $t_\star = t_{\mathrm{unif}}$. Then, we use that for all i, j such that $i + j \geq t_{\mathrm{unif}}$, we have $\mathbb{P}_x(X_{i+j} = x) \leq \frac{5 \deg(x)}{8m}$, which gives

$$\frac{1}{\deg(x)} \sum_{\substack{i,j=0 \\ i+j \geq t_\star}}^{t-1} \mathbb{P}_x(X_{i+j} = x) \quad \leq \quad \frac{5t^2}{8m} \,.$$

In the second case, $t_\star = t_{\mathrm{rel}} \left(\log(t_{\mathrm{rel}}) + 1 \right)$. Then, we use Claim 6, and obtain

$$\begin{aligned}
\frac{1}{\deg(x)} \sum_{\substack{i,j=0 \\ i+j \geq t_\star}}^{t-1} \mathbb{P}_x(X_{i+j} = x) \quad &\leq \quad \frac{t^2}{2m} + \sum_{k=t_\star}^{+\infty} (k+1) \lambda_2^k \\
&\leq \quad \frac{t^2}{2m} + \lambda_2^{t_\star} \left(\frac{t_\star}{1-\lambda_2} + \frac{1}{(1-\lambda_2)^2} \right) \\
&\leq \quad \frac{t^2}{2m} + 2t_\star \,.
\end{aligned}$$

Altogether, we have checked that for all $x \in V$ and $t \geq 0$,

$$\mathbb{E}_x \mathcal{I}_t \quad \leq \quad 12 t_\star^{3/2} + \frac{5t^2}{8m} \,,$$

as desired.

Moving on to the bound on the second moment, we have

$$\begin{aligned}
\mathbb{E}_x \mathcal{I}_t^2 \quad &= \quad \sum_{u,v} \frac{1}{\deg(u) \deg(v)} \left(\sum_{0 \leq i,k \leq t-1} \mathbb{P}_x(X_i = u, X_k = v) \right)^2 \\
&\lesssim \quad \sum_{u,v} \frac{g_t(x,u)^2 g_t(u,v)^2}{\deg(u) \deg(v)} \\
&= \quad \sum_u \frac{g_t(x,u)^2}{\deg(u)} \mathbb{E}_u \mathcal{I}_t \\
&\lesssim \quad \left(t_\star^{3/2} + \frac{t^2}{2m} \right) \mathbb{E}_x \mathcal{I}_t \,,
\end{aligned}$$

by the previously established upper-bound on the first moment. ∎

5.2. **A simple estimator for the number of edges.** Lemma 4 suggests the following simple estimator for the number of edges, namely:

$$\widehat{m}_t \quad = \quad \frac{t^2}{\frac{2}{K}\sum_{k=1}^{K} \mathcal{I}_t^{(k)}}\,, \tag{5.4}$$

where $\left\{\mathcal{I}_t^{(k)}\right\}_{k=1}^{K}$ are independent copies of $\mathcal{I}_t$, the weighted number of intersections between to independent random walks started at some $x \in V$. Using Lemma 4, for a large enough constant A and $t \geq A t_\star^{3/4}\sqrt{m}$, we have $\frac{t^2}{2m} \leq \mathbb{E}_x \mathcal{I}_t \leq \frac{3t^2}{4m}$ and $\mathrm{Var}_x \mathcal{I}_t \lesssim \left(\mathbb{E}_x \mathcal{I}_t\right)^2$. Hence, by Chebyshev's Inequality

$$\mathbb{P}_x^G\left(\left|\frac{\widehat{m}_t}{m} - 1\right| > \frac{1}{2}\right) = O\left(\frac{1}{K}\right)\,.$$

Remark 1. *Note that once we have a good estimate for the number of edges, it is quite easy to deduce an estimate for the number of vertices. Indeed, what remains to estimate is just the mean degree. Consider the function $f : x \in V \mapsto f(x) = \frac{1}{\deg(x)}$, and note that $\mathbb{E}_\pi f = \frac{n}{2m}$. Applying [14, Proposition 12.19] to the function f, we know that for $r \geq t_{\mathrm{mix}}(\varepsilon/2)$ and $t \geq \frac{4\,\mathrm{Var}_\pi f}{\eta^2 \varepsilon (\mathbb{E}_\pi f)^2} t_{\mathrm{rel}}$, for all $x \in V$,*

$$\mathbb{P}_x\left(\left|\frac{1}{t}\sum_{s=0}^{t-1} f(X_{r+s}) - \mathbb{E}_\pi f\right| > \eta \mathbb{E}_\pi f\right) \quad \leq \quad \varepsilon\,.$$

Observing that $\mathrm{Var}_\pi f \leq \mathbb{E}_\pi f^2 = \frac{1}{2m}\sum_{u\in V} \deg(u)^{-1}$, we see that estimating the mean degree can be done in time of order $t_{\mathrm{mix}} + t_{\mathrm{rel}}\frac{m}{n^2}\sum_{u\in V}\deg(u)^{-1}$. Under some weak assumption on the degrees' heterogeneity, or assuming that the graph is sparse ($m = O(n)$), the dominant term in this sum is t_{mix}, which is smaller than $t_\star^{3/4}\sqrt{m}$ (see [17, Claim 4.4]).

The estimators $\widehat{n}_t$ and $\widehat{m}_t$ defined at (3.3) and (5.4) thus start "being right" after times $t_\star^{3/4}\sqrt{n}$ and $t_\star^{3/4}\sqrt{m}$ respectively. In practice however, a user would want to know when to stop the random walks. This prompts the search for self-stopping procedures. The next section, however, is devoted to establishing a negative result: there is no sublinear self-stopping algorithm for the estimation of the number of vertices.

6. No self-stopping algorithms in general

In this section, we show that one can not hope for a general sublinear self-stopping algorithm, even with the restriction that the graphs have polylog mixing time.

Let $\mathcal{C}$ be the class of graphs G such that $t_{\mathrm{unif}}^G \leq (\log n_G)^3$.

Proposition 7. *There exists $\delta > 0$, such that, for all functions* RESTART, STOP *and* EST, *there is an infinite sequence of graphs $G \in \mathcal{C}$ and $x \in V(G)$ such that*

$$\mathbb{P}_x^G\left(\{\tau \geq \delta n_G\} \cup \left\{\left|\frac{\mathrm{EST}\left(\Phi(X_0^\tau)\right)}{n_G} - 1\right| > \frac{1}{2}\right\}\right) \quad \geq \quad \frac{1}{4}\,,$$

where X_0^t *is a* RWR *characterized by* RESTART *and* $\tau = \inf\{t \geq 0,\ \textsc{stop}\,(\Phi(X_0^\tau)) = 1\}$.

Proof of Proposition 7. Consider a 3-regular expander G on $n_G = n$ vertices and another graph H obtained from G as follows: let $G^{(1)}, \dots, G^{(2^n)}$ be 2^n identical copies of G. For all $i \in \{1, \dots, 2^n\}$, choose three distinct vertices (u_1^i, u_2^i, u_3^i) uniformly at random in $V(G^{(i)})$. Now let $\widetilde{G}$ be some other 3-regular expander on 2^n vertices, labelled from 1 to 2^n. One may mark each edge of $\widetilde{G}$ by a label in $\{1,2,3\}$ in such a way that no pair of edges incident to the same vertex have the same label. Now if there is an edge between vertices i and j in $\widetilde{G}$ and if this edge has label $k \in \{1,2,3\}$, then we create an edge between vertices u_k^i and u_k^j. Let H be the resulting graph ($n_H = n2^n$). Note that, as $\widetilde{G}$ is an expander, and as the random walk on H needs order n steps to go from some u_1^i to either u_2^i or u_3^i, we have $t_{\text{unif}}^H \lesssim n \log(2^n)$, so that both G and H belong to the class $\mathcal{C}$. It is not hard to check that one can find $y \in V(G^{(1)})$ and $\delta > 0$, such that for any function RESTART,

$$\mathbb{P}_y^H\left(\bigcap_{s=0}^{\delta n}\left\{Y_s \notin \{u_1^1, u_2^1, u_3^1\}\right\}\right) \geq \frac{2}{3}\,.$$

Therefore, there exist starting points $(x, y) \in V(G) \times V(H)$, and a coupling (X, Y) of random walks with restarts at x and y (for the same function RESTART) such that

$$\mathbb{P}_{x,y}\left(\mathbf{A}_t\right) \geq \frac{2}{3}\,, \quad \text{with} \quad \mathbf{A}_t = \left\{\Phi(X_0^t) = \Phi(Y_0^t)\right\} \quad \text{and} \quad t = \delta n\,. \tag{6.1}$$

Let $\textsc{est} : \mathcal{S} \to \mathbb{N}$ be an estimator and $\textsc{stop} : \mathcal{S} \to \{0,1\}$. Define

$$B_G^X = \{\tau^X < \delta n_G\} \cap \left\{\left|\frac{\textsc{est}\left(\Phi(X_0^{\tau^X})\right)}{n_G} - 1\right| \leq \frac{1}{2}\right\},$$

and

$$B_H^Y = \{\tau^Y < \delta n_H\} \cap \left\{\left|\frac{\textsc{est}\left(\Phi(Y_0^{\tau^Y})\right)}{n_H} - 1\right| \leq \frac{1}{2}\right\},$$

where $\tau^X = \inf\{s \geq 0,\ \textsc{stop}\,(\Phi(X_0^s)) = 1\}$ and $\tau^Y = \inf\{s \geq 0,\ \textsc{stop}\,(\Phi(Y_0^s)) = 1\}$. Assume that we both have $\mathbb{P}_x\left(B_G^X\right) \geq 3/4$ and $\mathbb{P}_y\left(B_H^Y\right) \geq 3/4$. Then, by (6.1),

$$\mathbb{P}_{x,y}\left(B_G^X \,|\, \mathbf{A}_t\right) = \frac{\mathbb{P}_{x,y}\left(B_G^X \cap \mathbf{A}_t\right)}{\mathbb{P}_{x,y}(\mathbf{A}_t)} \geq 1 - \frac{1 - \mathbb{P}_x\left(B_G^X\right)}{\mathbb{P}_{x,y}(\mathbf{A}_t)} \geq \frac{5}{8}\,,$$

and similarly, $\mathbb{P}_{x,y}\left(B_H^Y \,|\, \mathbf{A}_t\right) \geq \frac{5}{8}$, so that

$$\mathbb{P}_{x,y}\left(B_G^X \cap B_H^Y \,|\, \mathbf{A}_t\right) \quad \geq \quad \frac{1}{4}\,.$$

However, on the event $\mathbf{A}_t$, we have $\{\tau^X < \delta n_G\} \cap \{\tau^Y < \delta n_H\} = \{\tau^X < \delta n_G\} \cap \{\tau^Y = \tau^X\}$, so that $\textsc{est}\left(\Phi(X_0^{\tau^X})\right) = \textsc{est}\left(\Phi(Y_0^{\tau^Y})\right)$ and the events B_G^X and B_H^Y can not occur simultaneously, implying a contradiction. ∎

7. A self-stopping algorithm for the number of edges

Let $G = (V, E)$ be a finite connected graph. Initially, the only information we have about G is an upper-bound τ on the uniform mixing time $t_{\text{unif}}(1/4)$, and access to a random walk with restarts at a fixed vertex $x \in V$. The goal is to estimate $m = |E|$.

The algorithm is as follows (and borrows several ideas from [2]). For $q = 0, 1, \ldots,$ iterate the following procedure until stopped:

- let $\widehat{m} = 2^q$ be the current guess for the number of edges and let $t = t_q = \tau^{3/4}\sqrt{2\widehat{m}}$.
- let $R = R_q = \lceil 8\log(2/\varepsilon) + 8C\log q \rceil$ (for a constant C to be specified later) and repeat the following experiment R times.
  - let $X^{(1)}, Y^{(1)}, \ldots, X^{(K)}, Y^{(K)}$ be $2K$ independent random walks started from x (for a fixed integer $K \geq 1$ to be specified later) and define
$$\mathcal{W}_t = \frac{1}{K}\sum_{\ell=1}^{K} \mathcal{I}_t^{(\ell)}, \quad \text{where} \quad \mathcal{I}_t^{(\ell)} = \sum_{i,j=0}^{t-1} \frac{1}{\deg(X_i^{(\ell)})} \mathbb{1}_{\left\{X_i^{(\ell)} = Y_j^{(\ell)}\right\}}.$$
  - If $\mathcal{W}_t \geq 8\tau^{3/2}$, call the experiment a success.
- If the number of successes is larger than $R/2$, then stop and estimate m by $\widehat{m} = 2^q$; otherwise, go from q to $q+1$.

This algorithm satisfies the two following properties.

Fact 1. *The probability that the algorithm stops at a value of q such that $2^q \leq m/2$ is smaller than ε.*

Fact 2. *The expected running time of the algorithm is $O\left(\sqrt{m}\tau^{3/4}\log\log m\right)$.*

Proof of Fact 1. Note that, by Lemma 4, it always holds that

$$\frac{\widehat{m}}{m}\tau^{3/2} \;\leq\; \mathbb{E}_x\mathcal{W}_t \;\leq\; 14\frac{\widehat{m}}{m}\tau^{3/2}. \tag{7.1}$$

Hence, assuming that q is such that $\widehat{m} = 2^q \leq m/2$, the expectation of $\mathcal{W}_t$ is smaller than $7\tau^{3/2}$. By Chebyshev's Inequality,

$$\mathbb{P}_x\left(\mathcal{W}_t \geq 8\tau^{3/2}\right) \;\leq\; \mathbb{P}_x\left(\mathcal{W}_t - \mathbb{E}_x\mathcal{W}_t \geq \tau^{3/2}\right) \;\lesssim\; \frac{\mathrm{Var}_x\mathcal{I}_t}{K\tau^3}.$$

Now by Lemma 4, $\mathrm{Var}_x\mathcal{I}_t \lesssim \left(\tau^{3/2} + \frac{t^2}{m}\right)^2 \lesssim \tau^3$. Hence, we may choose K large enough such that $\mathbb{P}_x\left(\mathcal{W}_t \geq 8\tau^{3/2}\right) \leq 1/4$. Using Hoeffding's Inequality, the probability that there are more than $R/2$ successes is smaller than $\exp(-R/8) = \frac{\varepsilon}{2}q^{-C}$. Choosing C large enough and taking a union bound, we obtain that the probability for the algorithm to return an estimate smaller than $m/2$ is smaller than ε. ∎

Proof of Fact 2. Let q be such that $2^q \geq m$. By equation (7.1), the expectation of $\mathcal{W}_t$ is larger than $14\tau^{3/2}$. Hence

$$\mathbb{P}_x\left(\mathcal{W}_t \leq 8\tau^{3/2}\right) \;\leq\; \mathbb{P}_x\left(\mathcal{W}_t \leq \frac{4}{7}\mathbb{E}_x\mathcal{W}_t\right) \;\lesssim\; \frac{\mathrm{Var}_x\mathcal{I}_t}{K\left(\mathbb{E}_x\mathcal{I}_t\right)^2}.$$

Again, Lemma 4 entails that the constant K may be chosen such that the above probability is smaller than $1/4$. If $q^\star = \inf\{q \geq 0,\ 2^q \geq m\}$, then, for all $q > q^\star$, the probability that the algorithm stops at step q is smaller than $(1/4)^{q-q^\star}$. Now when the algorithm stops at step q, the running time is smaller, up to constant factors, than

$$\sum_{i=0}^{q} R_i t_i \lesssim R_q t_q \,,$$

so that the expected running time is smaller up to constant factors, than

$$R_{q^\star} t_{q^\star} + \sum_{q>q^\star} \left(\frac{1}{4}\right)^{q-q^\star} R_q t_q \quad = \quad O\left(\sqrt{m}\tau^{3/4} \log\log m\right) .$$

■

Remark 2. *On d-regular graphs, one may simplify this algorithm using the unweighted number of intersections and resorting to Lemma 1, and obtain an algorithm which estimates n, the number of vertices, in expected time $O\left(\tau^{3/4}\sqrt{n}\log\log n\right)$, without any further dependence in d than the one which might comes from τ.*

8. Algorithms for the mixing time

The number of intersections may also be used to estimate the mixing time from a given vertex $x \in V$. Assume that the number of edges m in $G = (V, E)$ is known. Let

$$d_x(t) \quad = \quad \sqrt{\sum_y \pi(y) \left(\frac{\mathbb{P}_x(X_t = y)}{\pi(y)} - 1\right)^2}$$

be the ℓ_2-distance between $\mathbb{P}_x(X_t \in \cdot)$ and $\pi(\cdot)$. Our goal now is to estimate

$$t_x(\delta) \quad = \quad \inf\left\{t \geq 0,\ d_x(t)^2 \leq \delta\right\} ,$$

for $0 < \delta < 1$. By reversibility, $d_x(t)^2 = \frac{\mathbb{P}_x(X_{2t}=x)}{\pi(x)} - 1$, so that, using (5.2),

$$\mathbb{E}_x \mathcal{I}_t \quad = \quad \sum_{i,j=0}^{t-1} \frac{d_x\left(\frac{i+j}{2}\right)^2 + 1}{2m} . \tag{8.1}$$

Equation (8.1) suggests the following self-stopping algorithm. For $q = 0, 1, \ldots,$ iterate the following procedure until stopped:

- Let $t = t_q = 2^q$ be the current guess for the mixing time $t_x(\delta)$ and let
$$K \ = \ K_q \quad = \quad \left\lceil \frac{C_1}{\delta^2} \max\left\{1, \frac{\sqrt{m}}{t^{1/4}}\right\} \right\rceil ,$$
for a constant $C_1 > 0$ to be specified later.
- Let $R = R_q = \lceil 8\log(2/\varepsilon) + 8C_2 \log q \rceil$ (for a constant C_2 to be specified later) and repeat the following experiment R times.

– Let $X^{(1)}, \dots, X^{(K)}$ be K independent random walks started from x and define

$$\mathcal{J}_t = \frac{1}{\binom{K}{2}} \sum_{1 \le \ell < k \le K} \mathcal{J}_t^{(\ell,k)}, \quad \text{where} \quad \mathcal{J}_t^{(\ell,k)} = \sum_{i,j=t}^{2t-1} \frac{1}{\deg(X_i^{(\ell)})} \mathbb{1}_{\left\{X_i^{(\ell)} = X_j^{(k)}\right\}}.$$

– If $\mathcal{J}_t \le \left(1 + \frac{\delta}{2}\right) \frac{t^2}{2m}$, call the experiment a success.

- If the number of successes is larger than $R/2$, then stop and estimate $t_x(\delta)$ by $t = 2^q$; otherwise, go from q to $q+1$.

This algorithm satisfies the two following properties.

Fact 3. *The probability that the algorithm stops at a value of q such that $2^q \le t_x(\delta)/2$ is smaller than ε.*

Fact 4. *The expected running time of the algorithm is*

$$O\left(\frac{\sqrt{m}}{\delta^2} t_x(\delta/4)^{3/4} \log\log t_x(\delta/4)\right).$$

Before analysing this self-stopping algorithm, we prove the following useful lemma.

Lemma 8. *Let X, Y, Z be three independent random walks started at x and let*

$$\mathcal{J}_t^{(X,Y)} = \sum_{i,j=t}^{2t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}}.$$

Define similarly $\mathcal{J}_t^{(X,Z)}$ to be the weighted number of intersections of X and Z between time t and $2t$. Then for all $t \ge 0$,

$$\mathrm{Var}_x \mathcal{J}_t^{(X,Y)} \lesssim \max_{u \in V} \{\mathbb{E}_u \mathcal{I}_t\} \, \mathbb{E}_x \mathcal{J}_t^{(X,Y)}, \tag{8.2}$$

and

$$\mathrm{Cov}_x\left(\mathcal{J}_t^{(X,Y)}, \mathcal{J}_t^{(X,Z)}\right) \lesssim \sqrt{\max_{u \in V} \{\mathbb{E}_u \mathcal{I}_t\}} \left(\mathbb{E}_x \mathcal{J}_t^{(X,Y)}\right)^{3/2}. \tag{8.3}$$

Proof of Lemma 8. Let us define

$$g_{t \to 2t}(x,u) = \sum_{i=t}^{2t-1} \mathbb{P}_x(X_i = u) = g_{2t}(x,u) - g_t(x,u).$$

One easily checks that

$$\mathbb{E}_x \mathcal{J}_t^{(X,Y)} = \sum_u \frac{g_{t \to 2t}(x,u)^2}{\deg(u)},$$

and that

$$\mathbb{E}_x\left(\left(\mathcal{J}_t^{(X,Y)}\right)^2\right) \lesssim \sum_{u,v} \frac{g_{t \to 2t}(x,u)^2 g_t(u,v)^2}{\deg(u)\deg(v)} = \sum_u \frac{g_{t \to 2t}(x,u)^2}{\deg(u)} \mathbb{E}_u \mathcal{I}_t.$$

Taking the maximum over $u \in V$ of $\mathbb{E}_u \mathcal{I}_t$ establishes inequality (8.2). Moving on to the covariance, we have

$$\mathbb{E}_x \left(\mathcal{J}_t^{(X,Y)} \mathcal{J}_t^{(X,Z)} \right) \lesssim \sum_{u,v} \frac{g_{t\to 2t}(x,u)^2 g_t(u,v) g_{t\to 2t}(x,v)}{\deg(u)\deg(v)},$$

and, by Cauchy-Schwarz Inequality,

$$\mathbb{E}_x \left(\mathcal{J}_t^{(X,Y)} \mathcal{J}_t^{(X,Z)} \right) \lesssim \mathbb{E}_x \mathcal{J}_t^{(X,Y)} \sqrt{\sum_u \frac{g_{t\to 2t}(x,u)^2}{\deg(u)} \mathbb{E}_u \mathcal{I}_t}.$$

Again, maximizing $\mathbb{E}_u \mathcal{I}_t$ over $u \in V$ establishes inequality (8.3). ■

Remark 3. *By the results of Section 5, we know that for all $t \le 16m^2$,*

$$\max_{u\in V} \{\mathbb{E}_u \mathcal{I}_t\} \lesssim t^{3/2},$$

whereas for $t \ge 16m^2$, as $t_{\text{unif}} \le 8m^2$,

$$\max_{u\in V} \{\mathbb{E}_u \mathcal{I}_t\} \lesssim \frac{t^2}{2m}.$$

We are now ready to prove Facts 3 and 4.

Proof of Fact 3. Assume that q is such that $t = 2^q \le t_x(\delta)/2$. By equation (8.1), the expectation of $\mathcal{J}_t$ is larger than $(1+\delta)\frac{t^2}{2m}$. By Chebyshev's Inequality,

$$\mathbb{P}_x \left(\mathcal{J}_t \le \left(1 + \frac{\delta}{2}\right) \frac{t^2}{2m} \right) \lesssim \frac{\mathrm{Var}_x \mathcal{J}_t}{\delta^2 (\mathbb{E}_x \mathcal{J}_t)^2}.$$

We have

$$\mathrm{Var}_x \mathcal{J}_t \lesssim \frac{\mathrm{Var}_x \mathcal{J}_t^{(X,Y)}}{K^2} + \frac{\mathrm{Cov}_x \left(\mathcal{J}_t^{(X,Y)}, \mathcal{J}_t^{(X,Z)} \right)}{K}.$$

Now combining Lemma 8 and Remark 3, we see that the constant C_1 in the definition of K can be made large enough so that the above probability is smaller than $1/4$. Using Hoeffding's Inequality, the probability that there are more than $R/2$ successes is then smaller than $\exp(-R/8) = \frac{\varepsilon}{2} q^{-C_2}$. Choosing C_2 large enough and taking a union bound, we obtain that the probability for the algorithm to return an estimate smaller than $t_x(\delta)/2$ is smaller than ε. ■

Proof of Fact 4. Let q be such that $t = 2^q \ge t_x(\delta/4)$. Then $\mathbb{E}_x \mathcal{J}_t \le (1+\delta/4)\frac{t^2}{2m}$ and by Chebyshev's Inequality

$$\mathbb{P}_x \left(\mathcal{J}_t > \left(1 + \frac{\delta}{2}\right) \frac{t^2}{2m} \right) \lesssim \frac{\mathrm{Var}_x \mathcal{J}_t}{\delta^2 (\mathbb{E}_x \mathcal{J}_t)^2}.$$

As in the proof of Fact 3, combining Lemma 8 and Remark 3, and taking C_1 large enough, the above probability is smaller than $1/4$. Hence, if $q^\star = \inf\{q \ge 0,\ 2^q \ge t_x(\delta/4)\}$, then

for all $q > q^\star$, the probability that the algorithm stops at q is smaller than $(1/4)^{q-q^\star}$. Now, the running time up to some step q is smaller, up to constant factors, than

$$\sum_{i=0}^{q} R_i K_i t_i \quad \lesssim \quad \frac{\sqrt{m}}{\delta^2}(t_q)^{3/4} \log q \,.$$

Altogether, the expected running time is less, up to constant factor, than

$$\frac{\sqrt{m}}{\delta^2}(t_{q^\star})^{3/4} \log q^\star + \sum_{q>q^\star} (1/4)^{q-q^\star} \frac{\sqrt{m}}{\delta^2}(t_q)^{3/4} \log q \,,$$

which is $O\left(\frac{\sqrt{m}}{\delta^2} t_x(\delta/4)^{3/4} \log\log t_x(\delta/4)\right)$. ■

Remark 4. *On d-regular graphs, one may simplify this algorithm using the unweighted number of intersections, and obtain an algorithm that estimates $t_x(\delta)$ in expected time*

$$O\left(\frac{\sqrt{n}}{\delta^2} t_x(\delta/4)^{3/4} \log\log t_x(\delta/4)\right) .$$

We assume, for simplicity, that the true value of m is known. However, our estimation scheme can easily be extended to the case where only a good approximation of m is available. In Section 7, we showed how an upper-bound on the uniform mixing could be used to devise a self-stopping algorithm returning a faithful estimate for the number of edges, entailing the following.

Corollary 9. *An upper-bound τ on the uniform mixing time can be used to precisely estimate both the number of edges and the mixing time from x, via a self-stopping algorithm with time complexity $O\left(\sqrt{m}\tau^{3/4} \log\log m\right)$.*

Acknowledgement. The question of estimating the mixing time with random walks trajectories was posed by Gábor Lugosi, during the *Eleventh annual workshop in Probability and Combinatorics*, Barbados, April 1-8, 2016. We are grateful to him for bringing this problem to our attention, and we thank the Bellairs Institute where this work was initiated.

References

[1] D. Aldous and J. Fill. Reversible Markov chains and random walks on graphs, 2002. URL https://www.stat.berkeley.edu/~aldous/RWG/book.html.

[2] I. Benjamini and B. Morris. The birthday problem and markov chain monte carlo. *arXiv preprint math/0701390*, 2007.

[3] I. Benjamini, G. Kozma, L. Lovász, D. Romik, and G. Tardos. Waiting for a bat to fly by (in polynomial time). *Combinatorics, Probability and Computing*, 15(5): 673–683, 2006. doi: 10.1017/S0963548306007590.

[4] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

[5] A. Broder and E. Shamir. On the second eigenvalue of random regular graphs. In *Foundations of Computer Science, 1987., 28th Annual Symposium on*, pages 286–294. IEEE, 1987.

[6] J. Bunge and M. Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[7] C. Cooper, T. Radzik, and Y. Siantos. Estimating network parameters using random walks. *Social Network Analysis and Mining*, 4(1):168, 2014.

[8] A. Das Sarma, D. Nanongkai, G. Pandurangan, and P. Tetali. Distributed random walks. *J. ACM*, 60(1):2:1–2:31, feb 2013. ISSN 0004-5411. doi: 10.1145/2432622.2432624. URL http://doi.acm.org/10.1145/2432622.2432624.

[9] D. J. Hsu, A. Kontorovich, and C. Szepesvári. Mixing time estimation in reversible markov chains from a single sample path. In *Advances in neural information processing systems*, pages 1459–1467, 2015.

[10] S. Janson. The probability that a random multigraph is simple. *Combinatorics, Probability and Computing*, 18(1-2):205–225, 2009.

[11] L. Katzir, E. Liberty, O. Somekh, and I. A. Cosma. Estimating sizes of social networks via biased sampling. *Internet Mathematics*, 10(3-4):335–359, 2014.

[12] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009. doi: 10.1080/15427951.2009.10129177. URL http://dx.doi.org/10.1080/15427951.2009.10129177.

[13] D. A. Levin and Y. Peres. Estimating the spectral gap of a reversible markov chain from a short trajectory. *arXiv preprint arXiv:1612.05330*, 2016.

[14] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

[15] E. Lubetzky, A. Sly, et al. Cutoff phenomena for random walks on random regular graphs. *Duke Mathematical Journal*, 153(3):475–510, 2010.

[16] R. Lyons. Asymptotic enumeration of spanning trees. *Combinatorics, Probability and Computing*, 14(04):491–522, 2005.

[17] Y. Peres, T. Sauerwald, P. Sousi, A. Stauffer, et al. Intersection and mixing times for reversible chains. *Electronic Journal of Probability*, 22, 2017.

A. Ben-Hamou
IMPA, Estrada Dona Castorina, 110, Rio de Janeiro 22460-320, Brazil.
E-mail address: benhamou@impa.br

R. Oliveira
IMPA, Estrada Dona Castorina, 110, Rio de Janeiro 22460-320, Brazil.
E-mail address: rimfo@impa.br

Y. Peres
Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA.
E-mail address: peres@microsoft.com