Math. Stat. Learn. 1 (2018), 375–399
DOI 10.4171/MSL/9

Mathematical Statistics and Learning

Estimating graph parameters with random walks

Anna Ben-Hamou, Roberto I. Oliveira and Yuval Peres

Abstract. An algorithm observes the trajectories of random walks over an unknown graph G, starting from the same vertex x, as well as the degrees along the trajectories. For all finite connected graphs, one can estimate the number of edges m up to a bounded factor in $O(t_{\text{rel}}{}^{3/4}\sqrt{m/d})$ steps, where t_{rel} is the relaxation time of the lazy random walk on G and d is the minimum degree in G. Alternatively, m can be estimated in $O(t_{\text{unif}} + t_{\text{rel}}{}^{5/6}\sqrt{n})$, where n is the number of vertices and t_{unif} is the uniform mixing time on G. The number of vertices n can then be estimated up to a bounded factor in an additional $O(t_{\text{unif}}\frac{m}{n})$ steps. Our algorithms are based on counting the number of intersections of random walk paths X, Y, i.e. the number of pairs (t, s) such that $X_t = Y_s$. This improves on previous estimates which only consider collisions (i.e. times t with $X_t = Y_t$). We also show that the complexity of our algorithms is optimal, even when restricting to graphs with a prescribed relaxation time. Finally, we show that, given either m or the mixing time of G, we can compute the "other parameter" with a self-stopping algorithm.

Mathematics Subject Classification (2010). 60J10, 05C81, 05C85, 62M05.

Keywords. Graph inference, intersections of random walks.

1. Introduction

What can one learn from the random walk on a graph long 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.*[1]

We assume our algorithm has black-box access to K random walks of length t on a graph G starting from the same fixed vertex x. It then produces an estimate $\widehat{\gamma}_t$ of a parameter $\gamma = \gamma(G)$ of interest, solely by looking at the traces of the random walks

[1]This paper is an extended and improved version of the SODA conference proceeding [2].

and the vertex degrees along the way. The goal is to achieve

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

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.

1.1. What we do. Let us describe our results in more detail, postponing the precise definition of the model to Section 2. In Section 3, we build on recent results of Peres et al. [19] and Oliveira and Peres [18] to derive bounds on the number of intersections between two independent random walks X, Y, i.e. the number of pairs of times (t, s) with $X_t = Y_s$. Using new bounds from [18], we show in particular that if X and Y are two independent lazy random walks on G, and if τ_{I} denotes the time of the first intersection between X and Y, i.e.

$$\tau_{\mathrm{I}} = \inf\left\{t \geq 0,\, \{X_0, \ldots, X_t\} \cap \{Y_0, \ldots, Y_t\} \neq \emptyset\right\},$$

then

$$\max_{x,y \in V} \mathbb{E}_{x,y}\, \tau_{\mathrm{I}} \lesssim {t_{\mathrm{rel}}}^{3/4} \sqrt{m/d}$$

where m is the number of edges in G and d is the minimum degree.

In Section 4, we focus on the particular case of regular graphs. Using intersection counts gives us a simple algorithm for estimating numbers of vertices n of a regular graph G in $O({t_{\mathrm{rel}}}^{3/4}\sqrt{n}\,)$ random walk steps. Moreover, we prove that this algorithm is optimal. More specifically, for any n and $1 \lesssim \mathbf{t}(n) \lesssim n^2$, we construct a graph G with about n vertices and relaxation time about $\mathbf{t}(n)$. We then show that any RW algorithm that finds the number of vertices of this graph requires at least $\Omega(\mathbf{t}(n)^{3/4}\sqrt{n}\,)$ time steps.

We then consider arbitrary graphs G in Section 5. In Section 5.1, we show that the number of edges m of G can be estimated in time of order $({t_{\mathrm{rel}}}^{3/4}\sqrt{m/d}\,) \wedge (t_{\mathrm{unif}} + {t_{\mathrm{rel}}}^{5/6}\sqrt{n}\,)$, where t_{unif} is the uniform mixing time on G, and we prove in Section 5.2 that the bound ${t_{\mathrm{rel}}}^{5/6}\sqrt{n}$ is tight for the estimation of the number of edges on graphs with any prescribed relaxation time. We then show in Section 5.3 that the bound ${t_{\mathrm{unif}}}^{5/6}\sqrt{n}$, which suffices to estimate the number of edges, may not be sufficient to estimate the number of vertices. However, provided a good estimate for the number of edges is known, the number of vertices follows from estimation of the mean degree, which can be done in times of order $(m/n)t_{\mathrm{unif}}$. Altogether, the number of vertices in general graphs can be estimated with random walks in time of order

$$\left({t_{\mathrm{rel}}}^{3/4}\sqrt{m/d}\,\right) \wedge \left({t_{\mathrm{rel}}}^{5/6}\sqrt{n}\,\right) + t_{\mathrm{unif}}\frac{m}{n},$$

and this is optimal.

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 $o(n^{3/5})$). This property of (relatively) fast mixing is expected to hold in social networks [14] 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 it is not possible to devise a sublinear stopping time at which one can be reasonably sure that our parameters are well estimated. 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 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 $O(\tau^{3/4}\sqrt{m}\log\log m)$ (or $O(\tau^{3/4}\sqrt{n}\log\log n)$ 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 $O(t_x(\delta/4)^{3/4}\sqrt{m}\log\log t_x(\delta/4))$, 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 $O(\tau^{3/4}\sqrt{m}\log\log m)$, assuming an upper-bound τ on the uniform mixing time is available.

1.2. Background. Our result relates to the 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 or m, as proposed by Cooper et al. [7] and Benjamini et al. [4]. Other parameters, such as the spectral gap, may be quite challenging to estimate (see Hsu et al. [10] and Levin and Peres [15]). 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 closer to ours, is to consider several, say k, random walks started from the same vertex x. Typically, estimators in this case rely on collisions of random walks at their endpoints. If each random walk has length greater than the mixing time t_{unif}, then the k-sample formed by their endpoints is an independent sample with nearly stationary distribution over the vertex set. In the case where G is regular, the problem comes down to estimating the size of a finite set through independent uniform samples from that set. It is well known that counting

collisions and resorting to the *birthday paradox* allow one to estimate n with order $\sqrt{n}$ samples. The time complexity, measured by the total number of random walk steps, is then of order $t_{\text{unif}}\sqrt{n}$ (the same kind of method was also used by Benjamini and Morris [3] to estimate the mixing time of regular graphs). If the graph is not regular, the stationary distribution is no longer uniform, and estimation of the support size can be more challenging (see [6] and [20] on support size estimation, and [1] on the related question of testing closeness between distributions). Katzir et al. [13] showed, through a variant of collision counting, that taking $k = O(\sqrt{n} + m/n)$ suffices to estimate n (if one is willing to use more information about the graph, the bound may be improved to $k = O(\|\pi\|_2^{-1} + m/n)$, where $\|\pi\|_2$ is the Euclidean norm of the stationary distribution π). Kanade et al. [12] established a corresponding lower bound for k in this setting. This yields a time complexity of $t_{\text{unif}}(\sqrt{n} + m/n)$. Kanade et al. [12] asked whether the factor t_{unif} in those bounds was really necessary or whether more efficient estimators could be designed. Indeed, in those methods, each unit of information already costs t_{unif} steps. Can we improve the performance by using the information held by the whole trajectories of walks ? We show that this is indeed the case, and that considering intersections of random walks' paths (instead of collisions at their endpoints) gives strictly more information, and leads to optimal time complexity.

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 vertices and m edges. For $u \in V$, we let $\deg(u) = |\{v \in V, \{u, v\} \in E\}|$ be the degree of u.

Random walks and estimators. Our estimators take as input trajectories of random walks, along with the degrees of visited vertices. However, they do not rely on a particular vertex labeling. To make this more precise, we introduce the *profile* of a sequence of vertices. For $t \geq 1$ and for a sequence of vertices $\mathbf{u}_t = (u_0, \ldots, u_{t-1}) \in V^t$, let $r(\mathbf{u}_t)$ be the length-t sequence where each vertex is replaced by the index of its first occurrence in $\mathbf{u}_t$. For instance, the image of the sequence $(g, a, a, c, g, d, a, b, d)$ by r is $(1, 2, 2, 3, 1, 4, 2, 5, 4)$. Note that r is invariant under vertex-relabeling. The profile Φ of $\mathbf{u}_t$ is then defined as

$$\Phi(\mathbf{u}_t) = \left(r(\mathbf{u}_t), (\deg(u_i))_{i=0}^{t-1}\right).$$

In other words, for each finite length sequence of vertices $\mathbf{u}_t$, the function Φ captures the ranks of occurrence and the degrees, and takes values in

$$\mathcal{S} = \bigcup_{t \geq 1} \mathbb{N}^{2t}\,.$$

Now let $x \in V$ be some fixed vertex. An *estimator* is a function EST: $\mathcal{S} \to \mathbb{R}$, which takes as input the profile of the trajectories of K independent lazy random walks (LRW) of length t, all started at x. More precisely, for integers K, $t \geq 1$, let $X^{(1)}, \ldots, X^{(K)}$ be K independent LRW on G started at x, and define $\mathbf{X}_t^{(i)} = (X_0^{(i)}, \ldots, X_{t-1}^{(i)})$, the trajectory of $X^{(i)}$ up to time $t-1$, for $i = 1, \ldots, K$. Letting $\gamma(G)$ be some parameter of interest (e.g. $\gamma(G) = n$ or $\gamma(G) = m$), the goal is to produce a map EST: $\mathcal{S} \to \mathbb{R}$, returning the value

$$\widehat{\gamma}_{K,t} = \text{EST}\big(\Phi\big(\mathbf{X}_t^{(1)}, \ldots, \mathbf{X}_t^{(K)}\big)\big)\,,$$

such that, for all connected graph $G = (V, E)$, for all $x \in V$, for all $t \geq t(\varepsilon, G)$ and $K \geq K(\varepsilon, G)$,

$$\mathbb{P}_x\left(\left|\frac{\widehat{\gamma}_{K,t}}{\gamma(G)} - 1\right| > \frac{1}{2}\right) \leq \varepsilon\,, \tag{2.1}$$

for $t(\varepsilon, G) \times K(\varepsilon, G)$ as small as possible. The product $t(\varepsilon, G) \times K(\varepsilon, G)$ corresponds to the total number of random walk steps and will often be referred to as the time complexity of the estimator. Let us point out right away that, in our estimation procedures, the critical quantity will be $t(\varepsilon, G)$, the random walks' length, rather than $K(\varepsilon, G)$, the number of random walks, which will simply be chosen according to the desired precision ε.

Convergence of random walks. To study the large-time behavior of our estimators, it is natural to take advantage of the convergence of LRW to its stationary distribution π, given by $\pi(u) = \deg(u)/2m$. Denote by t_{unif} the uniform mixing time defined as

$$t_{\text{unif}} = \inf\left\{t \geq 0,\ \max_{x,y \in V}\left|\frac{P^t(x,y)}{\pi(y)} - 1\right| \leq \frac{1}{4}\right\},$$

Also, letting $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ be the eigenvalues of P in decreasing order (the fact that all eigenvalues are non-negative is by laziness of the walk), the relaxation time is defined as

$$t_{\text{rel}} = \left\lceil \frac{1}{1-\lambda_2} \right\rceil.$$

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, i.e. algorithms which automatically stop at some random time, according to what has been seen so far. One then needs

to control both the error probability for the returned value, and the expected value of the stopping time (see Sections 6, 7 and 8).

3. Intersections of random walks

We start by some results on intersections of random walks' trajectories.

For X and Y two independent LRW on a finite connected graph $G = (V, E)$, the number of intersections between X and Y up to time $t-1$ is defined as

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

For non-regular graphs, a more relevant quantity is the *weighted* number of intersections, defined as

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

When X and Y start at x and y respectively, the probability law will be denoted $\mathbb{P}_{x,y}$ and the corresponding expectation $\mathbb{E}_{x,y}$. When $x = y$, we just write $\mathbb{P}_x$ and $\mathbb{E}_x$. Let P be the transition matrix of X and

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

be the expected number of visits to vertex u before time t (also known as Green's function). We have

$$\mathbb{E}_x \mathcal{I}_t = \sum_{u \in V} \frac{g_t(x,u)^2}{\deg(u)}. \tag{3.1}$$

The expected number of intersections is intimately related to return probabilities. Indeed, by reversibility, $\deg(x) g_t(x,u) = \deg(u) g_t(u,v)$ and we get

$$\mathbb{E}_x \mathcal{I}_t = \sum_{i,j=0}^{t-1} \frac{P^{i+j}(x,x)}{\deg(x)}. \tag{3.2}$$

We also define $\mathcal{J}_t$ to be the weighted number of intersections counted from the mixing time t_{unif}, i.e.

$$\mathcal{J}_t = \sum_{i,j=t_{\text{unif}}}^{t_{\text{unif}}+t-1} \frac{1}{\deg(X_i)} \mathbb{I}_{\{X_i = Y_j\}}.$$

Proposition 1. *For all finite connected graph $G = (V, E)$ with m edges, minimum degree d and relaxation time t_{rel}, for all $x \in V$,*

$$\frac{t^2}{2m} \le \mathbb{E}_x \mathcal{I}_t \le \frac{t^2}{2m} + \frac{16 t_{\mathrm{rel}}{}^{3/2}}{d}, \tag{3.3}$$

and

$$\mathbb{E}_x \mathcal{I}_t^2 \le 4\big(\max_{a \in V} \mathbb{E}_a \mathcal{I}_t\big)\mathbb{E}_x \mathcal{I}_t . \tag{3.4}$$

Proposition 2. *For all finite connected graph $G = (V, E)$ with m edges, n vertices and relaxation time t_{rel}, for all $x \in V$,*

$$\Big(\frac{3}{4}\Big)^2 \frac{t^2}{2m} \le \mathbb{E}_x \mathcal{J}_t \le \Big(\frac{5}{4}\Big)^2 \frac{t^2}{2m}, \tag{3.5}$$

and

$$\mathbb{E}_x \mathcal{J}_t^2 \lesssim \frac{t^2}{m^2}\big(t^2 + n t_{\mathrm{rel}}{}^{5/3}\big) . \tag{3.6}$$

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 $n \ge 1$.

Before proving Propositions 1 and 2, let us state three useful results. The following bound on Green's function was established by [18].

Lemma 3 ([18, Lemma 2]). *Let X be a LRW on G. For all $x \in V$, for all $1 \le t \le \frac{36m^2}{d}$,*

$$g_t(x, x) \le \frac{6 \deg(x)\sqrt{t}}{d} .$$

By [18, Proposition 1], we have

$$t_{\mathrm{rel}} \le \frac{12 m n}{d} . \tag{3.7}$$

In particular, the bound of Lemma 3 is valid up to t_{rel}. The following powerful result on the sum of return probabilities was established by Lyons and Oveis Gharan [17].

Lemma 4 ([17]). *For a lazy random walk X on G, for all $t \ge 0$,*

$$\sum_{u \in V} P^t(u, u) \le 1 + \frac{13n}{(t+1)^{1/3}} .$$

Finally, we also need the following lemma.

Lemma 5. *For any $f \in \mathbb{R}^{\mathcal{X}}$, if P is reversible, irreducible and has non-negative spectrum, then*

$$\sum_{s=0}^{+\infty} (s+1)\big(\langle f, P^s f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2\big) \le \frac{t_{\mathrm{rel}}}{(1 - 1/\mathrm{e})^2} \sum_{s=0}^{t_{\mathrm{rel}}-1} \big[\langle f, P^s f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2\big] .$$

Proof of Lemma 5. Partitioning $\mathbb{N}$ in blocks of length t_{rel}, we may write

$$\sum_{s=0}^{+\infty}(s+1)\big(\langle f, P^s f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2\big) = \sum_{k=0}^{+\infty}\sum_{s=0}^{t_{\mathrm{rel}}-1}(t_{\mathrm{rel}}k+s+1)\big(\langle f, P^{t_{\mathrm{rel}}k+s} f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2\big).$$

The terms in the above sums can be written in the form:

$$\langle f, P^r f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2 = \sum_{j=2}^{n} \lambda_j^r \, \langle f, \Psi_j\rangle_\pi^2$$

where $\lambda_1 = 1 > \lambda_2 \geq \lambda_2 \geq \cdots \leq \lambda_n \geq 0$ are the eigenvalues of P and $(\Psi_1 = \mathbf{1}, \Psi_2, \ldots, \Psi_n)$ is an orthonormal basis of eigenvectors for the inner product $\langle \cdot, \cdot \rangle_\pi$. By definition of t_{rel}, we have $\lambda_j^{t_{\mathrm{rel}}k} \leq \mathrm{e}^{-k}$ for all $j \geq 2$. Therefore,

$$\langle f, P^{t_{\mathrm{rel}}k+s} f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2 \leq \mathrm{e}^{-k}\big[\langle f, P^s f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2\big].$$

Summing the bounds, we obtain

$$\begin{aligned}
&\sum_{s=0}^{+\infty}(s+1)\left(\langle f, P^s f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2\right) \\
&\qquad \leq \sum_{k=0}^{+\infty}\sum_{s=0}^{t_{\mathrm{rel}}-1}(t_{\mathrm{rel}}k+s+1)\,\mathrm{e}^{-k}\left(\langle f, P^s f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2\right) \\
&\qquad \leq \sum_{k=0}^{+\infty} t_{\mathrm{rel}}(k+1)\,\mathrm{e}^{-k}\sum_{s=0}^{t_{\mathrm{rel}}-1}\left(\langle f, P^s f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2\right) \\
&\qquad \leq \frac{t_{\mathrm{rel}}}{(1-1/\mathrm{e})^2}\sum_{s=0}^{t_{\mathrm{rel}}-1}\left[\langle f, P^s f\rangle_\pi - \langle f, \mathbf{1}\rangle_\pi^2\right].
\end{aligned}$$

□

Proof of Proposition 1. By (3.2), we have

$$\mathbb{E}_x \mathcal{I}_t = \sum_{i,j=0}^{t-1}\frac{P^{i+j}(x,x)}{\deg(x)} = \frac{t^2}{m} + \frac{1}{2m}\sum_{i,j=0}^{t-1}\left(\frac{P^{i+j}(x,x)}{\pi(x)} - 1\right).$$

All summands in the right-hand side are non-negative (this can be seen, for instance, by the spectral decomposition

$$P^r(x,x) = \pi(x) + \sum_{j=2}^{n}\lambda_j^r \Psi_j(x)^2\pi(x)$$

and by non-negativity of the eigenvalues). Moreover, by Lemma 5 applied to the function $f = \frac{\mathbb{I}_{\{\cdot=x\}}}{\pi(x)}$,

$$\begin{aligned}\sum_{i,j=0}^{t-1}\left(\frac{P^{i+j}(x,x)}{\pi(x)}-1\right) &\le \sum_{s=0}^{+\infty}(s+1)\left(\frac{P^s(x,x)}{\pi(x)}-1\right)\\ &\le \frac{t_{\mathrm{rel}}}{(1-1/\mathrm{e})^2}\sum_{s=0}^{t_{\mathrm{rel}}-1}\left(\frac{P^s(x,x)}{\pi(x)}-1\right) \qquad (3.8)\\ &\le \frac{t_{\mathrm{rel}}}{(1-1/\mathrm{e})^2}\frac{g_{t_{\mathrm{rel}}}(x,x)}{\pi(x)}\,.\end{aligned}$$

Resorting to Lemma 3, we obtain

$$\mathbb{E}_x \mathcal{I}_t \le \frac{t^2}{2m} + \frac{6{t_{\mathrm{rel}}}^{3/2}}{(1-1/\mathrm{e})^2 d} \le \frac{t^2}{2m} + \frac{16{t_{\mathrm{rel}}}^{3/2}}{d}\,,$$

concluding the proof of the first moment bounds. Moving on to the second moment, we have

$$\begin{aligned}\mathbb{E}_x \mathcal{I}_t^2 &= \sum_{u,v}\frac{1}{\deg(u)\deg(v)}\Bigg(\sum_{i,k=0}^{t-1}\mathbb{P}_x(X_i=u,X_k=v)\Bigg)^2\\ &\le \sum_{u,v}\frac{1}{\deg(u)\deg(v)}\big(g_t(x,u)g_t(u,v)+g_t(x,v)g_t(v,u)\big)^2\\ &\le 4\sum_{u,v}\frac{g_t(x,u)^2 g_t(u,v)^2}{\deg(u)\deg(v)}\\ &= 4\sum_{u}\frac{g_t(x,u)^2}{\deg(u)}\mathbb{E}_u\mathcal{I}_t \;\le\; 4\big(\max_{u\in V}\mathbb{E}_u\mathcal{I}_t\big)\mathbb{E}_x\mathcal{I}_t\,,\end{aligned}$$

and (3.4) follows from the upper-bound in (3.3). □

Proof of Proposition 2. The bounds on the expectation of $\mathcal{J}_t$ are straightforward. Indeed

$$\mathbb{E}_x\mathcal{J}_t = \sum_{y,z} P^{t_{\mathrm{unif}}}(x,y)P^{t_{\mathrm{unif}}}(x,z)\mathbb{E}_{y,z}\mathcal{I}_t\,,$$

so that, by definition of t_{unif} and the fact that $\sum_{y,z}\pi(y)\pi(z)\mathbb{E}_{y,z}\mathcal{I}_t = t^2/2m$,

$$\left(\frac{3}{4}\right)^2\frac{t^2}{2m} \le \mathbb{E}_x\mathcal{J}_t \le \left(\frac{5}{4}\right)^2\frac{t^2}{2m}\cdot$$

Moving on to (3.6), again by definition of t_{unif}, we have

$$\begin{aligned}
\mathbb{E}_x \mathcal{J}_t^2 &\lesssim \sum_{y,z} \pi(y)\pi(z)\mathbb{E}_{y,z}\mathcal{I}_t^2 \\
&\lesssim \sum_{y,z}\sum_{u,v} \frac{\pi(y)\pi(z)}{\deg(u)\deg(v)} \sum_{i,j,k,\ell} \mathbb{P}_y(X_i = u, X_k = v)\mathbb{P}_z(Y_j = u, Y_\ell = v) \\
&\lesssim \sum_{u,v} \frac{1}{\deg(u)\deg(v)} \Big(\sum_y \pi(y) \sum_{i,k} \mathbb{P}_y(X_i = u, X_k = v) \Big)^2 \\
&\lesssim \sum_{u,v} \frac{1}{\deg(u)\deg(v)} \Big(\sum_y \pi(y) g_t(y,u) g_t(u,v) \Big)^2 .
\end{aligned}$$

Using that $\sum_y \pi(y) g_t(y,u) = \sum_y \pi(u) g_t(u,y) = t\pi(u)$, we have

$$\mathbb{E}_x \mathcal{J}_t^2 \lesssim \frac{t^2}{m^2} \sum_{u,v} \frac{\deg(u)}{\deg(v)} g_t(u,v)^2 = \frac{t^2}{m^2} \sum_{i,j=0}^{t-1} \sum_u P^{i+j}(u,u) ,$$

where the last equality comes from reversibility. Now, by Inequality (3.8),

$$\begin{aligned}
\sum_{i,j=0}^{t-1} \sum_u P^{i+j}(u,u) &= t^2 + \sum_u \pi(u) \sum_{i,j=0}^{t-1} \left(\frac{P^{i+j}(u,u)}{\pi(u)} - 1 \right) \\
&\le t^2 + \frac{t_{\text{rel}}}{(1-1/\mathrm{e})^2} \sum_{s=0}^{t_{\text{rel}}-1} \Big(\sum_u P^s(u,u) - 1 \Big) .
\end{aligned}$$

Finally, resorting to Lemma 4, we obtain

$$\mathbb{E}_x \mathcal{J}_t^2 \lesssim \frac{t^2}{m^2} \big(t^2 + n {t_{\text{rel}}}^{5/3} \big) ,$$

concluding the proof of Proposition 2. □

Remark 1. Proposition 1 entails bounds on $\mathbb{E}_{x,y}\mathcal{I}_t$. Indeed, for $x \neq y$, we may use the bound

$$\left| \frac{P^t(x,y)}{\pi(y)} - 1 \right| \le \sqrt{\frac{P^t(x,x)}{\pi(x)} - 1} \sqrt{\frac{P^t(y,y)}{\pi(y)} - 1} ,$$

which follows, for instance, from the Cauchy–Schwarz inequality in the spectral decomposition $P^t(x,y) = \pi(y)(1 + \sum_{j=2}^n \lambda_j^t \Psi_j(x)\Psi_j(y))$. This entails

$$\left| \mathbb{E}_{x,y}\mathcal{I}_t - \frac{t^2}{2m} \right| \le \sqrt{\mathbb{E}_{x,x}\mathcal{I}_t - \frac{t^2}{2m}} \sqrt{\mathbb{E}_{y,y}\mathcal{I}_t - \frac{t^2}{2m}} .$$

Moreover, one may check easily that $\max_{x,y} \mathbb{E}_{x,y} \mathcal{I}_t^2 \leq \max_x \mathcal{I}_t^2$. From those bounds, one may derive the following new bound on the first intersection time: for $t \gtrsim {t_{\mathrm{rel}}}^{3/4} \sqrt{m/d}$, by the second-moment method, $\mathbb{P}_{x,y}(\mathcal{I}_t > 0) \geq 1/8$. Since this holds uniformly in x and y, one may perform independent experiments to conclude that

$$\max_{x,y} \mathbb{E}_{x,y} \tau_{\mathrm{I}} \lesssim {t_{\mathrm{rel}}}^{3/4} \sqrt{m/d} \,.$$

4. Estimating the number of vertices on regular graphs

4.1. A simple estimator for the number of vertices. Specifying to regular graphs with degree $d \geq 1$ and considering the unweighted number of intersections I_t, Proposition 1 entails

$$\frac{t^2}{n} \leq \mathbb{E}_x I_t \leq \frac{t^2}{n} + 16 {t_{\mathrm{rel}}}^{3/2} \,,$$

and

$$\mathbb{E}_x I_t^2 \lesssim \Big(\frac{t^2}{n} + {t_{\mathrm{rel}}}^{3/2} \Big)^2 .$$

This 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

$$\widehat{n}_t = \frac{t^2}{\frac{1}{K} \sum_{k=1}^K I_t^{(k)}} \cdot \tag{4.1}$$

For $t \geq 2\sqrt{6} {t_{\mathrm{rel}}}^{3/4} \sqrt{n}$, we have $\frac{t^2}{n} \leq \mathbb{E}_x I_t \leq \frac{5t^2}{3n}$ and $\mathrm{Var}_x I_t \lesssim t^4/n^2$. Hence, by Chebyshev's inequality

$$\mathbb{P}_x \Big(\Big| \frac{\widehat{n}_t}{n} - 1 \Big| > \frac{1}{2} \Big) \leq \mathbb{P}_x \Big(\Big| \frac{1}{K} \sum_{k=1}^K I_t^{(k)} - \mathbb{E}_x \mathcal{I}_t \Big| > \frac{t^2}{3n} \Big) = O\Big(\frac{1}{K} \Big) .$$

4.2. Lower bounds for regular graphs. The case of the cycle on n vertices gives an example where the bound ${t_{\mathrm{rel}}}^{3/4} \sqrt{n}$ is tight. Indeed, in this case, $t_{\mathrm{rel}} \asymp n^2$, and thus ${t_{\mathrm{rel}}}^{3/4} \sqrt{n} \asymp n^2$. 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 $2n$.

This section is devoted to a stronger version of tightness. Namely, we exhibit graphs achieving the bound ${t_{\mathrm{rel}}}^{3/4} \sqrt{n}$ for any n, *and* for any relaxation time t_{rel}.

Proposition 6. *There exist absolute constants $\delta, \Lambda > 0$ such that the following holds. For all integers $n \geq \Lambda$ and $\mathbf{t}(n)$ with $\Lambda \leq \mathbf{t}(n) \leq \Lambda n^2$, for all map* EST*: $\mathcal{S} \to \mathbb{R}$, there exists a 3-regular graph $G = (V, E)$ such that:*

- $|V| \in [n, 14n]$*;*
- $t_{\mathrm{rel}} \leq \mathbf{t}(n)$*;*
- *for more than $9/10^{\text{th}}$ of the vertices $x \in V$, for all $t, K \geq 1$ with $tK \leq \delta \mathbf{t}(n)^{3/4}\sqrt{n}$,*

$$\mathbb{P}_x\left(\left|\frac{\widehat{n}_t}{n} - 1\right| > \frac{1}{2}\right) \geq \frac{1}{4},$$

where $\widehat{n}_t = \text{EST}\big(\Phi\big(\mathbf{X}_t^{(1)}, \dots, \mathbf{X}_t^{(K)}\big)\big)$.

Before proving Proposition 6, we first establish the following lemma.

Lemma 7. *For $k \geq 2$ even, let G_k be a uniform random 3-regular graph on k vertices. Then:*

(1) *The probability that G_k is connected tends to 1 as $k \to \infty$;*

(2) *The relaxation time $t_{\mathrm{rel}}(G_k)$ tends to $(1 - 2\sqrt{2}/3)^{-1}$ in probability;*

(3) *For k large enough, for all $x \in V(G_k)$, letting $(X_s)_{s\geq 0}$ be the concatenation of independent* RWS *of length $t \geq 1$ on G_k started at x (i.e. $(X_s)_{s\geq 0} = (\mathbf{X}_t^{(1)}, \mathbf{X}_t^{(2)}, \dots)$), we have, as soon as $s \leq \sqrt{k}/20$,*

$$\mathbb{P}_x(\mathbf{G}_s \text{ is a tree}) \geq \frac{93}{95},$$

where $\mathbf{G}_s$ is the subgraph induced by the edges visited by $(X_0, \dots, X_{s-1})$.

Proof of Lemma 7. The first item is a well-known fact, valid for random graphs with given degrees, as soon as the minimum degree is larger or equal to 3. The second item is by Friedman's theorem [9], which states that a random d-regular graph is with high probability weakly Ramanujan, i.e. its relaxation time is asymptotic to $(1 - 2\sqrt{d-1}/d)^{-1}$. Now, to establish the third item, we use a common method to generate a uniform 3-regular random graph, known as the *configuration model* (see [5]). 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 half-edges is then chosen uniformly, and each pair of matched half-edges is interpreted as an edge between the corresponding vertices. The probability that this creates a simple graph tends to e^{-2} (see for instance [11]), and, conditionally on being simple, the graph is uniformly distributed 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 walks, as follows. Initially, all half-edges are unpaired and $X_0 = x$. Then, at each step $s \geq 1$,

- either s is a multiple of t and we set $X_s = x$ (hereby starting a new walk),

- or s is not a multiple of t and we then choose with probability $1/3$ a half-edge $(X_{s-1}, *)$ attached to X_{s-1}. If $(X_{s-1}, *)$ has already been paired to some half-edge $(v, *)$, we let $X_s = v$. Otherwise, we choose uniformly at random an unpaired half-edge $(u, *)$, match $(X_{s-1}, *)$ and $(u, *)$, and let $X_s = u$.

Observe that the edges spanned by (X_s) form a tree up to the first time s when $(X_{s-1}, *)$ is unpaired but is then matched to a half-edge attached to a visited vertex (creating a cycle in the induced graph). The probability that this event first occurs at time s is smaller than $\frac{3s}{3k-3s}$ (by time $s-1$, we have exposed at most $3s$ half-edges). Hence, the (annealed) probability that this event occurs before time s is smaller than $\frac{3s^2}{3k-3s}$. For $s = \sqrt{k}/20$, this probability is smaller than $1/380$. For k large enough, the probability for the configuration model to yield a simple graph is larger than $1/8$, hence, on G_k, we have $\mathbb{P}_x(\mathbf{G}_s \text{ is a tree }) \geq 1 - 8/380 = 93/95$. □

Lemma 7 entails the following: there exists $k_0 \geq 1$ such that for all even $k \geq k_0$, there exist connected 3-regular graphs $\mathcal{E}_k$ and $\mathcal{E}_{4k}$ on k and $4k$ vertices respectively, satisfying

$$\max\left\{t_{\text{rel}}(\mathcal{E}_k), t_{\text{rel}}(\mathcal{E}_{4k})\right\} \leq 18\,, \tag{4.2}$$

and, for more than $9/10^{\text{th}}$ of the pairs of vertices $(x, y) \in V(\mathcal{E}_k) \times V(\mathcal{E}_{4k})$, there is a coupling of (X_s) and (Y_s), where (X_s) (resp. (Y_s)) is the concatenation of independent RWs of length t on $\mathcal{E}_k$ (resp. $\mathcal{E}_{4k}$) started at x (resp. y) such that, if $s \leq \sqrt{k}/20$,

$$\mathbb{P}_{x,y}\big(\Phi(X_0^s) = \Phi(Y_0^s)\big) \geq \frac{3}{4}\cdot \tag{4.3}$$

Indeed, on uniform 3-regular random graphs G_k and G_{4k}, the two processes (X_s) and (Y_s) can be successfully coupled up to the first time s when $\mathbf{G}_s$ is not a tree. By Lemma 7, this has probability less than $2/95$ for $s \leq \sqrt{k}/20$. Letting $\mathbf{P}_{x,y}$ denote the (quenched) probability associated with the coupled random walks on G_k and G_{4k}, by Markov's inequality (applied twice),

$$\mathbb{P}\left(\left|\left\{(x, y),\, \mathbf{P}_{x,y}\big(\Phi(X_0^s) = \Phi(Y_0^s)\big) < \frac{3}{4}\right\}\right| > \frac{1}{10}(k \times 4k)\right) \leq \frac{80}{95}\cdot$$

Hence we can find graphs $\mathcal{E}_k$ and $\mathcal{E}_{4k}$ satisfying (4.3).

Proof of Proposition 6. For some constant $\Lambda > 0$ to be specified later, let $n \geq \Lambda$ and $\Lambda \leq \mathbf{t}(n) \leq \Lambda n^2$, and define

$$\ell = 4\left\lfloor \frac{1}{4}\sqrt{\frac{\mathbf{t}(n)}{\Lambda}} \right\rfloor + 1 \quad \text{and} \quad k = \left\lceil \frac{2n}{3\ell - 1} \right\rceil.$$

Now let $\mathcal{G}_{k,\ell}$ and $\mathcal{G}_{4k,\ell}$ be constructed as follows:

(1) take two 3-regular graphs $\mathcal{E}_k$ and $\mathcal{E}_{4k}$ satisfying (4.3) (by our assumptions on n and $\mathbf{t}(n)$, the constant Λ can be chosen large enough so that $k \geq k_0$);

(2) in each graph, in place of each edge, put a path of length ℓ.;

(3) to make those graphs 3-regular, add edges between pairs of interior vertices at distance 2 on the same path (this is possible because $\ell - 1$ is a multiple of 4).

See Figure 1.

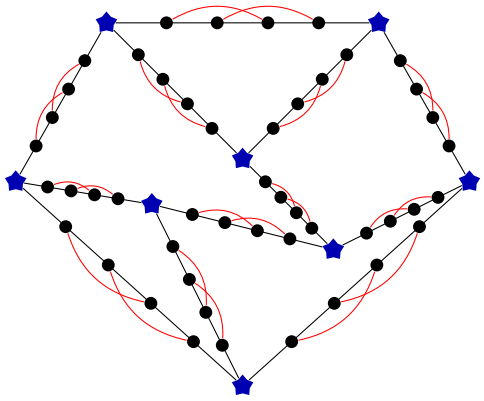

Figure 1. The graph $\mathcal{G}_{k,\ell}$ ($k = 8, \ell = 5$). The blue star-shaped vertices are the original vertices of $\mathcal{E}_k$.

Note that, using $\ell \leq n + 1$,

$$n \leq |V(\mathcal{G}_{k,\ell})| = \frac{k}{2}(3\ell - 1) \leq \frac{7n}{2},$$

and similarly $4n \leq |V(\mathcal{G}_{4k,m})| \leq 14n$. Moreover, choosing Λ large enough, we have

$$\max\left\{t_{\mathrm{rel}}(\mathcal{G}_{k,\ell}), t_{\mathrm{rel}}(\mathcal{G}_{k,\ell})\right\} \leq \frac{\Lambda}{4}\ell^2.$$

This can be seen by conductance arguments (the bottleneck ratio of $\mathcal{E}_k$ is bounded away from 0 by expansion, entailing that the one of $\mathcal{G}_{k,\ell}$ is up to constant factors larger than $1/\ell$, and by Cheeger's inequality, the relaxtion time is smaller than ℓ^2 up to constant factors). By definition of ℓ and the fact that $\Lambda \leq \mathbf{t}(n)$,

$$\max\left\{t_{\mathrm{rel}}(\mathcal{G}_{k,\ell}), t_{\mathrm{rel}}(\mathcal{G}_{4k,\ell})\right\} \leq \frac{\Lambda}{4}\left(\sqrt{\frac{\mathbf{t}(n)}{\Lambda}} + 1\right)^2 \leq \frac{\Lambda}{4}\left(2\sqrt{\frac{\mathbf{t}(n)}{\Lambda}}\right)^2 \leq \mathbf{t}(n).$$

Combining Equation (4.3) and the ℓ^2-slow down induced by paths, we obtain that for 9/10 of the starting points $(x, y) \in V(G_{k,\ell}) \times V(G_{4k,\ell})$, there is a coupling of random walks such that, letting

$$\mathbf{A}_t = \left\{\Phi\left(\mathbf{X}_t^{(1)}, \ldots, \mathbf{X}_t^{(K)}\right) = \Phi\left(\mathbf{Y}_t^{(1)}, \ldots, \mathbf{Y}_t^{(K)}\right)\right\},$$

we have

$$\mathbb{P}_{x,y}\left(\mathbf{A}_t\right) \geq \frac{3}{4}, \quad \text{with } Kt = \delta \ell^2 \sqrt{k}\,, \tag{4.4}$$

for some $\delta > 0$ small enough. Let EST: $\mathcal{S} \to \mathbb{N}$ be an estimator and let $\widehat{n}_t(X) = \textsc{est}\big(\Phi(\mathbf{X}_t^{(1)}, \ldots, \mathbf{X}_t^{(K)})\big)$ and $\widehat{n}_t(Y) = \textsc{est}\big(\Phi(\mathbf{Y}_t^{(1)}, \ldots, \mathbf{Y}_t^{(K)})\big)$. Define

$$B_t^X = \Big\{\frac{1}{2} \leq \frac{\widehat{n}_t(X)}{n} \leq \frac{3}{2}\Big\} \quad \text{and} \quad B_t^Y = \Big\{\frac{1}{2} \leq \frac{\widehat{n}_t(Y)}{4n} \leq \frac{3}{2}\Big\}.$$

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 (4.4),

$$\mathbb{P}_{x,y}\big(B_t^X \,\big|\, \mathbf{A}_t\big) = \frac{\mathbb{P}_{x,y}(B_t^X \cap \mathbf{A}_t)}{\mathbb{P}_{x,y}(\mathbf{A}_t)} \geq 1 - \frac{1 - \mathbb{P}_x(B_t^X)}{\mathbb{P}_{x,y}(\mathbf{A}_t)} \geq \frac{2}{3},$$

and similarly, $\mathbb{P}_{x,y}(B_t^Y \,|\, \mathbf{A}_t) \geq \frac{2}{3}$, so that $\mathbb{P}_{x,y}(B_t^X \cap B_t^Y \,|\, \mathbf{A}_t) \geq \frac{1}{3}$. However, on the event $\mathbf{A}_t$, the events B_t^X and B_t^Y can not occur simultaneously, implying a contradiction. We either have $\mathbb{P}_x\big(\big|\frac{\widehat{n}_t(X)}{n} - 1\big| > \frac{1}{2}\big) \geq \frac{1}{4}$ or $\mathbb{P}_y\big(\big|\frac{\widehat{n}_t(Y)}{4n} - 1\big| > \frac{1}{2}\big) \geq \frac{1}{4}$. The proof is then concluded by noticing that

$$\ell^2 \sqrt{k} \gtrsim \mathbf{t}(n)^{3/4} \sqrt{n}\,. \qquad \square$$

5. Computing parameters of general graphs

5.1. A simple estimator for the number of edges. In the non-regular case, Proposition 1 suggests the following simple estimator for the number of edges, namely:

$$\widehat{m}_t = \frac{t^2}{\frac{2}{K}\sum_{k=1}^K \mathcal{I}_t^{(k)}}, \tag{5.1}$$

where $\{\mathcal{I}_t^{(k)}\}_{k=1}^K$ are independent copies of $\mathcal{I}_t$, the weighted number of intersections between to independent random walks started at some $x \in V$. For $t \geq 4\sqrt{3} t_{\text{rel}}{}^{3/4} \sqrt{m/d}$, we have $\frac{t^2}{2m} \leq \mathbb{E}_x \mathcal{I}_t \leq \frac{5t^2}{6m}$ and $\mathrm{Var}_x \mathcal{I}_t \lesssim t^4/m^2$. Hence, by Chebyshev's inequality

$$\mathbb{P}_x\left(\Big|\frac{\widehat{m}_t}{m} - 1\Big| > \frac{1}{2}\right) = \mathbb{P}_x\left(\Big|\frac{1}{K}\sum_{k=1}^K \mathcal{I}_t^{(k)} - \mathbb{E}_x \mathcal{I}_t\Big| > \frac{t^2}{6m}\right) = O\Big(\frac{1}{K}\Big).$$

Alternatively, considering the other estimator

$$\widetilde{m}_t = \frac{t^2}{\frac{2}{K}\sum_{k=1}^K \mathcal{J}_t^{(k)}}, \tag{5.2}$$

where $\{\mathcal{J}_t^{(k)}\}_{k=1}^K$ are independent copies of $\mathcal{J}_t$, we obtain, by Proposition 2, that for $t \gtrsim {t_{\mathrm{rel}}}^{5/6}\sqrt{n}$,

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

Since intersections are counted from the uniform mixing time, the total time complexity of $\tilde{m}_t$ to reach error probability ε is $O(\varepsilon^{-1}(t_{\mathrm{unif}} + {t_{\mathrm{rel}}}^{5/6}\sqrt{n}))$.

5.2. Lower bounds for general graphs. The bound ${t_{\mathrm{rel}}}^{5/6}\sqrt{n}$ is achieved on a graph known as the barbell, formed by two cliques of size n joined by a path of length n. Indeed, the relaxation time of this graph has order n^3, so that ${t_{\mathrm{rel}}}^{5/6}\sqrt{n} \asymp n^3$, and any procedure based on random walks needs time n^3 to correctly estimate n, since this is the time needed by a random walk to go from one clique to the other.

As in Section 4.2, we now exhibit graphs achieving the bound ${t_{\mathrm{rel}}}^{5/6}\sqrt{n}$ for any n and any relaxation time t_{rel}. For two integers $k, q \geq 1$, consider the graph constructed as follows:

(1) Take a 3-regular graph $\mathcal{E}_k$ on k vertices, satisfying the properties of Lemma 7;

(2) Replace each node of $\mathcal{E}_k$ by a clique of size q;

(3) Replace each edge of $\mathcal{E}_k$ by a path of length q.

See Figure 2.

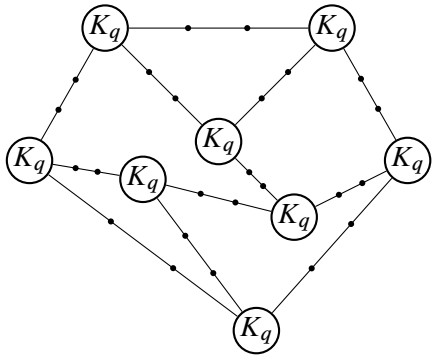

Figure 2.

Such a graph has a number of vertices n of order kq and relaxation time t_{rel} of order q^3. Parameters k and q may then be tuned so as to obtained (almost) any possible n and t_{rel}. Now, to estimate correctly the number of edges, one needs to get the correct order for k. By Lemma 7, a random walk on $\mathcal{E}_k$ needs order $\sqrt{k}$ steps to make a cycle and thus be able to distinguish $\mathcal{E}_k$ from an infinite 3-regular tree. Since adding cliques and paths of size q slows down the random walk by a factor of q^3 (the time to go from one clique to an other in the modified graph), the estimation of the number of edges on such a graph requires at least order $q^3\sqrt{k} \asymp {t_{\mathrm{rel}}}^{5/6}\sqrt{n}$ steps.

5.3. Estimating the number of vertices on general graphs. We first note that estimating the number of vertices might take much more time than estimating the number of edges. More precisely, we show that order ${t_{\text{rel}}}^{5/6}\sqrt{n}$ steps may not be enough to estimate n. Indeed, consider the graph formed by a clique of size ℓ with path of length q attached to each vertex of the clique, with $q \ll \ell$ (see Figure 3).

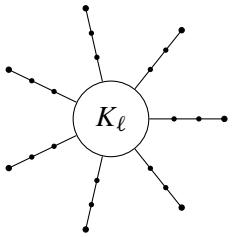

Figure 3.

The number n of vertices is of order ℓq, and, as $q \ll \ell$, the number m of edges if of order ℓ^2. Moreover, the relaxation time is of order q^2 (this can be seen by a coupling argument). Estimating m is relatively easy: starting from the end of one path, the walk has to traverse it to reach the clique, which takes time q^2, and then to wait for a collision in the clique, which, by the birthday paradox, takes time $\sqrt{\ell}$. Estimating n however takes more time: starting from the clique, the walk has to visit a positive fraction of at least one of the paths, and this takes time ℓq. As soon as $q \ll \ell^{3/7}$, we have $\ell q \gg \sqrt{\ell} q^{13/6} \asymp {t_{\text{rel}}}^{5/6}\sqrt{n}$.

Estimating n might thus require more time. However, once a good estimate for the number of edges is known, 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 [16, Proposition 12.19] to the function f, we know that for $r \geq t_{\text{mix}}(\varepsilon/2)$ and $t \geq \frac{16\,\text{Var}_\pi f}{\varepsilon(\mathbb{E}_\pi f)^2} t_{\text{rel}}$, for all $x \in V$,

$$\mathbb{P}_x\left(\left|\frac{1}{t}\sum_{s=0}^{t-1} f(X_{r+s}) - \mathbb{E}_\pi f\right| > \frac{\mathbb{E}_\pi f}{2}\right) \leq \varepsilon\,.$$

Observing that $\text{Var}_\pi f \leq \mathbb{E}_\pi f^2 = (2m)^{-1}\sum_u \deg(u)^{-1}$ and that $t_{\text{mix}}(\varepsilon/2) \lesssim \log(1/\varepsilon) t_{\text{unif}}$, the mean degree can be estimated with error probability less than ε in time of order

$$\log(1/\varepsilon) t_{\text{unif}} + \frac{t_{\text{rel}} m}{\varepsilon n^2}\sum_{u\in V}\deg(u)^{-1} \lesssim \varepsilon^{-1} t_{\text{unif}}\frac{m}{n}\,.$$

Note that this is optimal by the previous example of Figure 3, for which $\ell q \asymp t_{\text{unif}} m/n$.

Altogether, the number of vertices of a connected graph can be estimated by random walks in time

$$\varepsilon^{-1}\left(\left(t_{\text{rel}}{}^{3/4}\sqrt{m/d}\,\right)\wedge\left(t_{\text{rel}}{}^{5/6}\sqrt{n}\,\right)+t_{\text{unif}}\frac{m}{n}\right).$$

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 when restricting to graphs with polylog mixing time.

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

Consider the following process on a graph, called *random walk with restarts*: at each time step $t\ge 0$, based on $\Phi(X_0,\dots,X_t)$, the process decides whether it wants to make a random walk step from X_t, or to reset back to the starting point x. A self-stopping algorithm is based on the profile of a random walk with restarts, up to some stopping time τ. More precisely, it relies on a function $\textsc{stop}\colon\mathcal{S}\to\{0,1\}$. Defining

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

where $X_0^t=(X_0,\dots,X_t)$ is the trajectory of a random walk with restarts up to time t, then the self-stopping algorithm defined by $\textsc{stop}$ and $\textsc{est}$ returns the value $\textsc{est}(\Phi(X_0^\tau))$.

Proposition 8. *There exists $\delta>0$, such that, for all functions $\textsc{stop}$ and $\textsc{est}$, there is an infinite sequence of graphs $G\in\mathcal{C}$ and $x\in V(G)$ such that*

$$\mathbb{P}_x^G\left(\{\tau\ge\delta n_G\}\cup\left\{\left|\frac{\textsc{est}(\Phi(X_0^\tau))}{n_G}-1\right|>\frac{1}{2}\right\}\right)\ge\frac{1}{4},$$

where X is a $\textsc{rw}$ with restarts and $\tau=\inf\{t\ge 0,\ \textsc{stop}(\Phi(X_0^\tau))=1\}$.

Proof of Proposition 8. Consider a 3-regular expander G on n vertices and a graph $G^\star$ 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_i,v_i,w_i) uniformly at random in $V(G^{(i)})$. Now let F be some other 3-regular expander on 2^n vertices, labelled from 1 to 2^n. For all $1\le i\le 2^n$, if i has neighbors $j<k<\ell$ in F, put an edge between u_i and u_j, between v_i and v_k, between w_i and w_ℓ. Let $G^\star$ be the resulting graph (on $|V(G^\star)|=n2^n$ vertices). Note that, as F is an expander, and as the random walk on $G^\star$ needs order n steps to go from some u_i to either v_i or w_i, we have $t_{\text{unif}}(G^\star)\lesssim n\log(2^n)$, so that both G and $G^\star$ 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

$$\mathbb{P}_y^{G^\star}\left(\bigcap_{s=0}^{\delta n}\{Y_s\notin\{u_1,v_1,w_1\}\}\right)\ge\frac{2}{3}\cdot$$

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

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

Let EST: $\mathcal{S} \to \mathbb{N}$ be an estimator and STOP: $\mathcal{S} \to \{0, 1\}$. For $(Z, H) \in \{(X, G), (Y, G^\star)\}$, define

$$B_H^Z = \big\{\tau^Z < \delta|V(H)|\big\} \cap \left\{\left|\frac{\textsc{est}(\Phi(Z_0^{\tau^Z}))}{|V(H)|} - 1\right| \leq \frac{1}{2}\right\},$$

where $\tau^Z = \inf\{s \geq 0,\ \textsc{stop}(\Phi(Z_0^s)) = 1\}$. Assume that we both have $\mathbb{P}_x(B_G^X) \geq 3/4$ and $\mathbb{P}_y(B_{G^\star}^Y) \geq 3/4$. Then, by (6.1),

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

and similarly, $\mathbb{P}_{x,y}(B_{G^\star}^Y \,|\, \mathbf{A}_t) \geq \frac{5}{8}$, so that $\mathbb{P}_{x,y}(B_G^X \cap B_{G^\star}^Y \,|\, \mathbf{A}_t) \geq \frac{1}{4}$. However, on the event $\mathbf{A}_t$, we have

$$\big\{\tau^X < \delta|V(G)|\big\} \cap \big\{\tau^Y < \delta|V(G^\star)|\big\} = \big\{\tau^X < \delta n\big\} \cap \big\{\tau^Y = \tau^X\big\},$$

so that $\textsc{est}(\Phi(X_0^{\tau^X})) = \textsc{est}(\Phi(Y_0^{\tau^Y}))$ and the events B_G^X and $B_{G^\star}^Y$ 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 and let τ be an upper-bound on the relaxation time t_{rel}.

Algorithm 1. 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 $R = R_q = \lceil 8\log(4/\varepsilon) + 16\log(q+1)\rceil$ and repeat the following experiment R times.
  - let $X^{(1)}, Y^{(1)}, \ldots, X^{(K)}, Y^{(K)}$ be $2K$ independent LRW started from x (for a fixed integer $K \geq 1$ to be specified later) and define

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

  - If $\mathcal{Q}_t \geq 18\tau^{3/2}$, 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 $q+1$.

Proposition 9. *Algorithm 1 satisfies the two following properties:*

(1) *The probability that the algorithm stops at a value of q such that $2^q < m$ or $2^q > 38m$ is smaller than ε.*

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

Proof of Proposition 9. (1) By Equation (3.3) in Proposition 1 and since $d \geq 1$, it always holds that

$$\frac{\tau^{3/2}\widehat{m}}{m} \leq \mathbb{E}_x \mathcal{Q}_t \leq \frac{\tau^{3/2}\widehat{m}}{m} + 16\tau^{3/2}. \tag{7.1}$$

Assume that q is such that $\widehat{m} = 2^q < m$. Then the expectation of $\mathcal{Q}_t$ is smaller than $17\tau^{3/2}$. By Chebyshev's inequality,

$$\mathbb{P}_x\big(\mathcal{Q}_t \geq 18\tau^{3/2}\big) \leq \mathbb{P}_x\big(\mathcal{Q}_t - \mathbb{E}_x\mathcal{Q}_t \geq \tau^{3/2}\big) \lesssim \frac{\mathrm{Var}_x \mathcal{I}_t}{K\tau^3}.$$

Now by Equation (3.4) and since $t < \tau^{3/4}\sqrt{2m}$, we have $\mathrm{Var}_x \mathcal{I}_t \lesssim \tau^3$. Hence, we may choose K large enough such that $\mathbb{P}_x(\mathcal{Q}_t \geq 20\tau^{3/2}) \leq 1/4$. Using Hoeffding's inequality, the probability that there are more than $R/2$ successes at this step is smaller than $\exp(-R/8) = \frac{\varepsilon}{4}(q+1)^{-2}$. Taking a union bound, the probability for the algorithm to stop at a value of q such that $2^q < m$ is smaller than $\varepsilon/2$.

Let now q be such that $\widehat{m} = 2^q > 19m$. By Equation (7.1), the expectation of $\mathcal{Q}_t$ is larger than $19\tau^{3/2}$. Hence

$$\mathbb{P}_x\big(\mathcal{Q}_t < 18\tau^{3/2}\big) \leq \mathbb{P}_x\Big(\mathcal{Q}_t < \frac{18}{19}\mathbb{E}_x\mathcal{Q}_t\Big) \lesssim \frac{\mathrm{Var}_x \mathcal{I}_t}{K(\mathbb{E}_x \mathcal{I}_t)^2}.$$

Again, Equation (3.4) entails that the constant K may be chosen such that the above probability is smaller than $1/4$. And by Hoeffding's inequality, the probability that there are less than $R/2$ successes is smaller than $\exp(-R/8) \leq \varepsilon/4$. Clearly, the probability to stop at a step q with $2^q > 38m$ is smaller than the probability not to have stopped at $q^\star = \inf\{q \geq 0,\ 2^q > 19m\}$, which is smaller than $\varepsilon/4$.

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

$$R_{q^\star} t_{q^\star} + \sum_{q > q^\star} \Big(\frac{\varepsilon}{4}\Big)^{q-q^\star} R_q t_q = O\big(\sqrt{m}\tau^{3/4}\log\log m\big). \qquad \square$$

Remark 2. If the graph is d-regular of if the minimum degree d is known, Proposition 1 allows to design an algorithm which estimates m (or rather n in the case of regular graphs) in expected time $O(\sqrt{m/d}\,\tau^{3/4}\log\log(m/d))$.

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

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

be the $\ell_2(\pi)$-distance between $\mathbb{P}_x(X_t \in \cdot)/\pi(\cdot)$ and 1. Our goal now is to estimate

$$t_x(\delta) = \inf \left\{ t \geq 0,\ d_x(t)^2 \leq \delta \right\}, \quad 0 < \delta < 1 .$$

Before describing a self-stopping algorithm to estimate $t_x(\delta)$, we prove the following useful lemma.

Lemma 10. *Let X, Y, Z be three independent random walks started at x and let $\mathcal{L}_t^{(X,Y)} = \mathcal{I}_{2t}^{(X,Y)} - \mathcal{I}_t^{(X,Y)}$ be the weighted number of intersections of X and Y between t and $2t$. Define $\mathcal{L}_t^{(X,Z)}$ similarly. For all $t \geq 0$,*

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

$$\mathrm{Var}_x \mathcal{L}_t^{(X,Y)} \lesssim \mathbb{E}_x \mathcal{L}_t^{(X,Y)} \max_u \mathbb{E}_u \mathcal{I}_t , \tag{8.2}$$

and

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

Proof of Lemma 10. By reversibility, $d_x(t)^2 = \frac{\mathbb{P}_x(X_{2t}=x)}{\pi(x)} - 1$, and

$$\mathbb{E}_x \mathcal{L}_t^{(X,Y)} = \frac{1}{\deg(x)} \sum_{i,j=t}^{2t-1} \mathbb{P}_x(X_{i+j} = x) = \sum_{i,j=t}^{2t-1} \frac{d_x\left(\frac{i+j}{2}\right)^2 + 1}{2m} .$$

Moving on to (8.2), defining $g_{t\to 2t}(x,u) = g_{2t}(x,u) - g_t(x,u)$, one easily checks that

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

and that

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

which implies

$$\mathbb{E}_x\left(\left(\mathcal{L}_t^{(X,Y)}\right)^2\right) \lesssim \mathbb{E}_x \mathcal{L}_t^{(X,Y)} \max_u \mathbb{E}_u \mathcal{I}_t .$$

Finally, to establish (8.3), observe that

$$\mathbb{E}_x\big(\mathscr{L}_t^{(X,Y)}\mathscr{L}_t^{(X,Z)}\big) \lesssim \sum_{u,v} \frac{g_{t\to 2t}(x,u)^2 g_t(u,v) g_{t\to 2t}(x,v)}{\deg(u)\deg(v)}\,,$$

and that, by the Cauchy–Schwarz inequality,

$$\mathbb{E}_x\big(\mathscr{L}_t^{(X,Y)}\mathscr{L}_t^{(X,Z)}\big) \le \big(\mathbb{E}_x\mathscr{L}_t^{(X,Y)}\big)^{3/2}\sqrt{\max_u \mathbb{E}_u \mathscr{I}_t}\,. \qquad \square$$

Algorithm 2. For $q = 0, 1, \dots$, 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 $K = K_q = \lceil C\delta^{-2}\lceil\sqrt{m}t^{-1/4}\rceil\rceil$, for a constant $C > 0$ to be specified later.
- Let $R = R_q = \lceil 8\log(4/\varepsilon) + 16\log(q+1)\rceil$ and repeat the following experiment R times.
  - Let $X^{(1)}, \dots, X^{(K)}$ be K independent LRW started from x and define

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

  - If $\mathscr{L}_t \le \big(1+\frac{\delta}{2}\big)\frac{t^2}{2m}$, 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 $q+1$.

Proposition 11. *Algorithm 2 satisfies the two following properties:*

(1) *The probability that the algorithm stops at a value of q such that $2^q < t_x(\delta)/2$ or $2^q > 2t_x(\delta/4)$ is smaller than ε.*

(2) *The expected running time of the algorithm is $O(\delta^{-2}\sqrt{m}t_x(\delta/4)^{3/4}\log\log t_x(\delta/4))$.*

Proof of Proposition 11. (1) Assume that q is such that $t = 2^q < t_x(\delta)/2$. By Equation (8.1), the expectation of $\mathscr{L}_t$ is larger than $(1+\delta)\frac{t^2}{2m}$. By Chebyshev's inequality,

$$\mathbb{P}_x\Big(\mathscr{L}_t \le \Big(1+\frac{\delta}{2}\Big)\frac{t^2}{2m}\Big) \lesssim \frac{\mathrm{Var}_x\mathscr{L}_t}{\delta^2(\mathbb{E}_x\mathscr{L}_t)^2}\,. \tag{8.4}$$

Since for all ℓ, ℓ', k, k' pairwise distinct, $\mathrm{Cov}_x\big(\mathscr{L}_t^{(\ell,k)}, \mathscr{L}_t^{(\ell',k')}\big) = 0$, we have

$$\mathrm{Var}_x\mathscr{L}_t \lesssim \frac{\mathrm{Var}_x\mathscr{L}_t^{(X,Y)}}{K^2} + \frac{\mathrm{Cov}_x\big(\mathscr{L}_t^{(X,Y)}, \mathscr{L}_t^{(X,Z)}\big)}{K}\,,$$

so that, by Lemma 10 and using that $\mathbb{E}_x\mathscr{L}_t \ge t^2/2m$, we get

$$\frac{\mathrm{Var}_x\mathscr{L}_t}{\delta^2(\mathbb{E}_x\mathscr{L}_t)^2} \lesssim \kappa + \sqrt{\kappa}\,,$$

where

$$\kappa = \frac{\max_u \mathbb{E}_u \mathcal{I}_t}{C^2 \lceil \sqrt{m}/t^{1/4} \rceil^2 (t^2/m)} .$$

Now, if $t \leq \frac{36m^2}{d}$, then applying Lemma 3 directly in (3.2) yields $\max_u \mathbb{E}_u \mathcal{I}_t \lesssim t^{3/2}$. On the other hand, if $t > \frac{36m^2}{d}$, then by (3.7), $t \gtrsim t_{\text{rel}}{}^{3/4} \sqrt{m/d}$, which by Proposition 1 yields $\max_u \mathbb{E}_u \mathcal{I}_t \lesssim t^2/m$. Hence, in both cases, we have $\kappa \lesssim 1/C^2$, and the constant C can be made large enough so that the right-hand side in (8.4) is smaller than $1/4$. Using Hoeffding's inequality, the probability that there are more than $R/2$ successes is then smaller than $\exp(-R/8) = \frac{\varepsilon}{4}(q+1)^{-2}$. 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 $\varepsilon/2$.

Now let q be such that $t = 2^q > t_x(\delta/4)$. Then $\mathbb{E}_x \mathcal{L}_t \leq (1+\delta/4)\frac{t^2}{2m}$ and by Chebyshev's inequality

$$\mathbb{P}_x\Big(\mathcal{L}_t > \Big(1+\frac{\delta}{2}\Big)\frac{t^2}{2m}\Big) \lesssim \frac{\mathrm{Var}_x \mathcal{L}_t}{\delta^2 (t^2/m)^2} .$$

By the same arguments as above, the constant C can be chosen large enough so that the above probability is smaller than $1/4$. By Hoeffding's inequality, the probability that there are less than $R/2$ successes is smaller than $\exp(-R/8)\,\frac{\varepsilon}{4}$. Clearly, the probability to stop at a value q such that $2^q > 2t_x(\delta/4)$ is smaller than the probability not to have stopped at $q^\star = \inf\{q \geq 0,\ 2^q > t_x(\delta/4)\}$, which is smaller than $\varepsilon/4$.

(2) By the above, for all $q > q^\star$, the probability that the algorithm stops at step q is smaller than $(\varepsilon/4)^{q-q^\star}$. Moreover, the running time up to step q is smaller, up to constant factors, than $\sum_{i=0}^{q} R_i K_i t_i \lesssim \delta^{-2}\sqrt{m}(t_q)^{3/4}\log(q+1)$. Altogether, the expected running time is less, up to constant factor, than

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

which is $O\big(\delta^{-2}\sqrt{m}t_x(\delta/4)^{3/4}\log\log t_x(\delta/4)\big)$. □

Remark 3. If the graph is d-regular of if the minimum degree d is known, Proposition 1 actually allows to design an algorithm which estimates $t_x(\delta)$ in expected time $O(\delta^{-2}\sqrt{m/d}\,t_x(\delta/4)^{3/4}\log\log t_x(\delta/4))$.

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. Combining Proposition 9 and 11 then entails the following corollary.

Corollary 12. *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 $O(\sqrt{m}\,\tau^{3/4}\log\log m)$.*

Acknowledgements. 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. RO was funded by a *Bolsa de Produtividade em Pesquisa* from CNPq, Brazil and a *Cientista do Nosso Estado* grant from FAPERJ, Rio de Janeiro, Brazil. His work in this article is part of the activities of FAPESP Center for Neuromathematics (grant # 2013/ 07699-0 , FAPESP - S. Paulo Research Foundation).

References

[1] J. Acharya, C. Daskalakis, and G. C. Kamath, Optimal testing for properties of distributions, in *Advances in Neural Information Processing Systems. 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), 3591–3599, Curran Associates, Inc., 2015.

[2] A. Ben-Hamou, R. I. Oliveira, and Y. Peres, Estimating graph parameters via random walks with restarts, in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1702–1714, SIAM, Philadelphia, PA, 2018. Zbl 1403.05139 MR 3775899

[3] I. Benjamini and B. Morris, The birthday problem and Markov chain Monte Carlo, 2007. arXiv:math/0701390

[4] I. Benjamini, G. Kozma, L. Lovász, D. Romik, and G. Tardos, Waiting for a bat to fly by (in polynomial time), *Combin. Probab. Comput.*, **15** (2006), no. 5, 673–683. Zbl 1107.05057 MR 2248320

[5] B. Bollobás, A probabilistic proof of an asymptotic formula for the number of labelled regular graphs, *European J. Combin.*, **1** (1980), no. 4, 311–316. Zbl 0457.05038 MR 595929

[6] J. Bunge and M. Fitzpatrick, Estimating the number of species: a review, *J. Amer. Statist. Assoc.*, **88** (1993), no. 421, 1993, 364–373.

[7] C. Cooper, T. Radzik, and Y. Siantos, Estimating network parameters using random walks, *Soc. Netw. Anal. Min.*, **4** (2014), no. 1, Art. 168.

[8] A. Das Sarma, D. Nanongkai, G. Pandurangan, and P. Tetali, Distributed random walks, *J. ACM*, **60** (2013), no. 1, Art. 2, 31pp. Zbl 1281.68225 MR 3033219

[9] J. Friedman, A proof of Alon's second eigenvalue conjecture and related problems, *Mem. Amer. Math. Soc.*, **195** (2008), no. 910, viii+100pp. Zbl 1177.05070 MR 2437174

[10] 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. 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), 1459–1467, Curran Associates, Inc., 2015.

[11] S. Janson, The probability that a random multigraph is simple, *Combin. Probab. Comput.*, **18** (2009), no. 1-2, 205–225. Zbl 1216.05145 MR 2497380

[12] V. Kanade, F. Mallmann-Trenn, and V. Verdugo, How large is your graph?, in *31 International Symposium on Distributed Computing*, Art. No. 34, 16pp., LIPIcs. Leibniz Int. Proc. Inform., 91, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2017. Zbl 1380.68316 MR 3746297

[13] L. Katzir, E. Liberty, O. Somekh, and I. A. Cosma, Estimating sizes of social networks via biased sampling, *Internet Math.*, **10** (2014), no. 3-4, 335–359. MR 3259270

[14] 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 Math.*, **6** (2009), no. 1, 29–123. Zbl 1205.91144 MR 2736090

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

[16] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. With a chapter by James G. Propp and David B. Wilson, American Mathematical Society, Providence, RI, 2009. Zbl 1160.60001 MR 2466937

[17] R. Lyons and S. Oveis Gharan, Sharp bounds on random walk eigenvalues via spectral embedding, *Int. Math. Res. Not. IMRN*, (2018), no. 24, 7555–7605. MR 3892273

[18] R. I. Oliveira and Y. Peres, Random walks on graphs: new bounds on hitting, meeting, coalescing and returning, in *2019 Proceedings of the Sixteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, 119–126, SIAM, Philadelphia, PA, 2019. MR 3909448

[19] Y. Peres, T. Sauerwald, P. Sousi, and A. Stauffer, Intersection and mixing times for reversible chains, *Electron. J. Probab.*, **22** (2017), Paper No. 12, 16pp. Zbl 1357.60076 MR 3613705

[20] G. Valiant and P. Valiant, Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts, in *STOC'11 – Proceedings of the 43rd ACM Symposium on Theory of Computing*, 685–694, ACM, New York, 2011. Zbl 1288.68186 MR 2932019

Received 17 August, 2018

A. Ben-Hamou, Sorbonne Université, LPSM,
4, place Jussieu, 75005 Paris, France
E-mail: anna.ben-hamou@upmc.fr

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

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