Estimating graph parameters with random walks

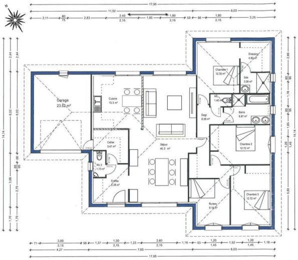
Anna Ben-Hamou (Sorbonne Université)

Joint work with Roberto Oliveira (IMPA) and Yuval Peres (Microsoft)

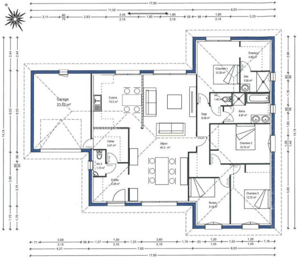
Approximation Algorithms and Networks

Collège de France

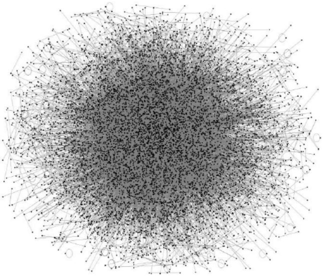
June 7th, 2018



Can we estimate the number of rooms in a house by randomly walking through adjacent rooms?



Can we estimate the number of rooms in a house by randomly walking through adjacent rooms?



Can we estimate the size of a possibly huge network using random walks?

Problem setting

Let $G = (V, E)$ be a finite connected graph.

$$n = |V| \qquad m = |E|$$

We start at a given $x \in V$.

Problem setting

Let $G = (V, E)$ be a finite connected graph.

$$n = |V| \qquad m = |E|$$

We start at a given $x \in V$.

Estimating with random walks

- $X^{(1)}, \ldots, X^{(K)}$: K independent RWs of length t, all started at x.
- we observe the label and the degree of visited vertices.

Problem setting

Let $G = (V, E)$ be a finite connected graph.

$$n = |V| \qquad m = |E|$$

We start at a given $x \in V$.

Estimating with random walks

- $X^{(1)}, \dots, X^{(K)}$: K independent RWs of length t, all started at x.
- we observe the label and the degree of visited vertices.

Goal: design an estimator $\widehat{n}_t = \widehat{n}_t(X^{(1)}, \dots, X^{(K)})$ such that for all $G = (V, E)$ connected, for all $x \in V$, for all $t \geq t(\varepsilon, G)$ and $K \geq K(\varepsilon, G)$,

$$\mathbb{P}_x^G \left(\left| \frac{\widehat{n}_t}{n_G} - 1 \right| > \frac{1}{2} \right) \leq \varepsilon \,,$$

with $K(\varepsilon, G) \times t(\varepsilon, G)$ as small as possible.

Other parameters of interest: number of edges, mixing time...

Convergence of random walks

$X = (X_t)_{t \geq 0}$ lazy RW on G with transition matrix P:

$$\forall \, (x, y) \in V^2 \, , \; P^t(x, y) \underset{t \to \infty}{\longrightarrow} \pi(y) = \frac{\deg(y)}{2m}$$

Convergence of random walks

$X = (X_t)_{t\geq 0}$ lazy RW on G with transition matrix P:

$$\forall\,(x,y)\in V^2\,,\ P^t(x,y) \underset{t\to\infty}{\longrightarrow} \pi(y) = \frac{\deg(y)}{2m}$$

Reversibility:

$$\pi(x)P(x,y) = \pi(y)P(y,x)\,.$$

Convergence of random walks

$X = (X_t)_{t \geq 0}$ lazy RW on G with transition matrix P:

$$\forall (x,y) \in V^2 , \; P^t(x,y) \underset{t\to\infty}{\longrightarrow} \pi(y) = \frac{\deg(y)}{2m}$$

Reversibility:

$$\pi(x)P(x,y) = \pi(y)P(y,x) \, .$$

Speed of convergence measured by uniform mixing time:

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

Convergence of random walks

$X = (X_t)_{t \geq 0}$ lazy RW on G with transition matrix P:

$$\forall\,(x,y) \in V^2\,,\; P^t(x,y) \underset{t\to\infty}{\longrightarrow} \pi(y) = \frac{\deg(y)}{2m}$$

Reversibility:

$$\pi(x)P(x,y) = \pi(y)P(y,x)\,.$$

Speed of convergence measured by uniform mixing time:

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

Or by relaxation time:

$$t_{\mathrm{rel}} = \frac{1}{1-\lambda_2} \quad \text{where} \quad \lambda_2 = \max\{\lambda \in \mathrm{Sp}(P),\; \lambda \neq 1\}$$

Reducing to i.i.d. samples

- Consider K lazy RWs started from $x \in V$, with length $t \geq t_{\mathrm{unif}}$.
- The sample $(X_t^{(1)}, \ldots, X_t^{(K)})$ is (almost) i.i.d. with law π.

Reducing to i.i.d. samples

- Consider K lazy RWs started from $x \in V$, with length $t \geq t_{\text{unif}}$.
- The sample $(X_t^{(1)}, \ldots, X_t^{(K)})$ is (almost) I.I.D. with law π.

On regular graphs ($\pi = 1/n$)

Count the number of collisions:

$$C_K = \sum_{i<j} \mathbb{1}_{\{X_t^{(i)} = X_t^{(j)}\}} .$$

Birthday paradox: $K \asymp \sqrt{n}$

Reducing to i.i.d. samples

- Consider K lazy RWs started from $x \in V$, with length $t \geq t_{\text{unif}}$.
- The sample $(X_t^{(1)}, \ldots, X_t^{(K)})$ is (almost) I.I.D. with law π.

On regular graphs ($\pi = 1/n$)

Count the number of collisions:

$$C_K = \sum_{i<j} \mathbb{1}_{\{X_t^{(i)} = X_t^{(j)}\}} .$$

Birthday paradox: $K \asymp \sqrt{n} \implies$ Time complexity $O\left(t_{\text{unif}}\sqrt{n}\right)$.

Reducing to i.i.d. samples

- Consider K lazy RWs started from $x \in V$, with length $t \geq t_{\text{unif}}$.
- The sample $(X_t^{(1)}, \dots, X_t^{(K)})$ is (almost) i.i.d. with law π.

On regular graphs ($\pi = 1/n$)

Count the number of collisions:

$$C_K = \sum_{i<j} \mathbb{1}_{\{X_t^{(i)} = X_t^{(j)}\}}.$$

Birthday paradox: $K \asymp \sqrt{n} \implies$ Time complexity $O\left(t_{\text{unif}}\sqrt{n}\right)$.

On non-regular graphs

Katzir, Liberty, Somekh, Cosma (2014); Kanade, Mallmann-Trenn, Verdugo (2017)

$$K \asymp \sqrt{n} + \frac{m}{n}$$

Reducing to i.i.d. samples

- Consider K lazy RWs started from $x \in V$, with length $t \geq t_{\text{unif}}$.
- The sample $(X_t^{(1)}, \ldots, X_t^{(K)})$ is (almost) I.I.D. with law π.

On regular graphs ($\pi = 1/n$)

Count the number of collisions:

$$C_K = \sum_{i<j} \mathbb{1}_{\{X_t^{(i)} = X_t^{(j)}\}}.$$

Birthday paradox: $K \asymp \sqrt{n} \implies$ Time complexity $O\left(t_{\text{unif}}\sqrt{n}\right)$.

On non-regular graphs

Katzir, Liberty, Somekh, Cosma (2014); Kanade, Mallmann-Trenn, Verdugo (2017)

$$K \asymp \sqrt{n} + \frac{m}{n} \implies \text{Time complexity } O\left(t_{\text{unif}}\left(\sqrt{n} + \frac{m}{n}\right)\right)$$

Reducing to i.i.d. samples

- Consider K lazy RWs started from $x \in V$, with length $t \geq t_{\text{unif}}$.
- The sample $(X_t^{(1)}, \ldots, X_t^{(K)})$ is (almost) I.I.D. with law π.

On regular graphs ($\pi = 1/n$)

Count the number of collisions:

$$C_K = \sum_{i<j} \mathbb{1}_{\{X_t^{(i)} = X_t^{(j)}\}} .$$

Birthday paradox: $K \asymp \sqrt{n} \implies$ Time complexity $O\left(t_{\text{unif}}\sqrt{n}\right)$.

On non-regular graphs

Katzir, Liberty, Somekh, Cosma (2014); Kanade, Mallmann-Trenn, Verdugo (2017)

$$K \asymp \sqrt{n} + \frac{m}{n} \implies \text{Time complexity } O\left(t_{\text{unif}}\left(\sqrt{n} + \frac{m}{n}\right)\right)$$

If one is willing to use more info on the graph: $K \asymp \left(\sum \pi(u)^2\right)^{-1/2} + \frac{m}{n}$.

Reducing to i.i.d. samples

- Consider K lazy RWs started from $x \in V$, with length $t \geq t_{\text{unif}}$.
- The sample $(X_t^{(1)}, \ldots, X_t^{(K)})$ is (almost) i.i.d. with law π.

On regular graphs ($\pi = 1/n$)

Count the number of collisions:

$$C_K = \sum_{i<j} \mathbb{1}_{\{X_t^{(i)} = X_t^{(j)}\}} .$$

Birthday paradox: $K \asymp \sqrt{n} \implies$ Time complexity $O\left(t_{\text{unif}} \sqrt{n}\right)$.

On non-regular graphs

Katzir, Liberty, Somekh, Cosma (2014); Kanade, Mallmann-Trenn, Verdugo (2017)

$$K \asymp \sqrt{n} + \frac{m}{n} \implies \text{Time complexity } O\left(t_{\text{unif}} \left(\sqrt{n} + \frac{m}{n}\right)\right)$$

If one is willing to use more info on the graph: $K \asymp \left(\sum \pi(u)^2\right)^{-1/2} + \frac{m}{n}$.

- Can we do better using the whole trajectories of walks?
- Is the factor t_{unif} necessary?

Regular graphs

Let X and Y be two indep. lazy RWs started at $x \in V$.

$$I_t = \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} \mathbb{1}_{\{X_i = Y_j\}} \qquad \left(\mathbb{E}_{\pi,\pi} I_t = \frac{t^2}{n} \right).$$

Regular graphs

Let X and Y be two indep. lazy RWs started at $x \in V$.

$$I_t = \sum_{i=0}^{t-1} \sum_{j=0}^{t-1} \mathbb{1}_{\{X_i = Y_j\}} \qquad \left(\mathbb{E}_{\pi,\pi} I_t = \frac{t^2}{n} \right).$$

Peres, Sauerwald, Sousi, Stauffer (2017)

$$\begin{aligned} \frac{t^2}{n} \ \leq \ \mathbb{E}_{x,x} I_t &= \sum_{i,j=0}^{t-1} \sum_{u \in V} P^i(x,u) P^j(x,u) \\ &= \sum_{i,j=0}^{t-1} \sum_{u \in V} P^i(x,u) P^j(u,x) \\ &= \sum_{i,j=0}^{t-1} P^{i+j}(x,x) \\ &\lesssim \sum_{i+j < t_{\mathrm{unif}}} P^{i+j}(x,x) + \frac{t^2}{n} \cdot \end{aligned}$$

Regular graphs

Estimate on return probabilities

$$P^t(x,x) \le \frac{1}{n} + \frac{5}{\sqrt{t}} \qquad \text{Aldous and Fill}$$

Regular graphs

Estimate on return probabilities

$$P^t(x,x) \leq \frac{1}{n} + \frac{5}{\sqrt{t}} \qquad \text{Aldous and Fill}$$

Bound on first moment

$$\frac{t^2}{n} \;\leq\; \mathbb{E}_{x,x} I_t \;\lesssim\; t_{\text{unif}}^{3/2} + \frac{t^2}{n}\,,$$

$$\mathbb{E}_{x,x} I_t \asymp \frac{t^2}{n} \quad \text{for} \quad t \gtrsim t_{\text{unif}}^{3/4} \sqrt{n}.$$

Regular graphs

Estimate on return probabilities

$$P^t(x,x) \le \frac{1}{n} + \frac{5}{\sqrt{t}} \qquad \text{Aldous and Fill}$$

Bound on first moment

$$\frac{t^2}{n} \le \mathbb{E}_{x,x} I_t \lesssim t_{\text{unif}}^{3/2} + \frac{t^2}{n},$$

$$\mathbb{E}_{x,x} I_t \asymp \frac{t^2}{n} \quad \text{for} \quad t \gtrsim t_{\text{unif}}^{3/4}\sqrt{n}.$$

Bound on second moment

$$\mathbb{E}_{x,x}\left[I_t^2\right] \lesssim \mathbb{E}_{x,x}[I_t] \max_{u \in V} \mathbb{E}_{u,u}[I_t].$$

Regular graphs

Consider K pairs of RWs $\left\{(X^{(k)}, Y^{(k)})\right\}_{k=1}^{K}$ and the estimator

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

Regular graphs

Consider K pairs of RWs $\{(X^{(k)}, Y^{(k)})\}_{k=1}^{K}$ and the estimator

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

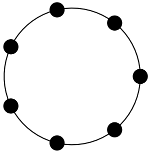
Theorem (B., Oliveira and Peres)

For all $G = (V, E)$ connected regular, for all $x \in V$,

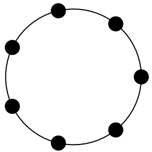
$$\forall \, t \gtrsim t_{\mathrm{rel}}^{3/4}\sqrt{n}\, , \quad \mathbb{P}_x\left(\left|\frac{\widehat{n}_t}{n} - 1\right| > \frac{1}{2}\right) = O\left(\frac{1}{K}\right) \, .$$

Lower bound

Bound $t_{\mathrm{rel}}^{3/4}\sqrt{n}$ is achieved by the cycle.



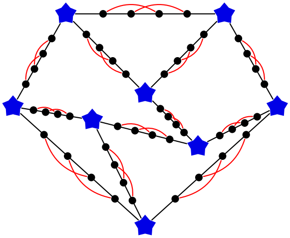
Lower bound



Bound $t_{\mathrm{rel}}^{3/4}\sqrt{n}$ is achieved by the cycle.

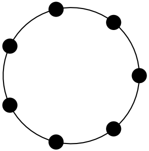
Lower bound for each possible mixing time

1. Start with a 3-regular expander $\mathcal{E}_k$ of size k;
2. Replace each edge of $\mathcal{E}_k$ by a path of length $\ell \geq 1$;
3. Make the graph 3-regular by adding short edges.



$$\begin{aligned} n &\asymp k\ell \\ t_{\mathrm{rel}} &\asymp \ell^2 \end{aligned}$$

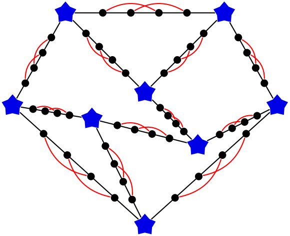
Lower bound



Bound $t_{\mathrm{rel}}^{3/4}\sqrt{n}$ is achieved by the cycle.

Lower bound for each possible mixing time

1. Start with a 3-regular expander $\mathcal{E}_k$ of size k;
2. Replace each edge of $\mathcal{E}_k$ by a path of length $\ell \geq 1$;
3. Make the graph 3-regular by adding short edges.



$$n \asymp k\ell$$
$$t_{\mathrm{rel}} \asymp \ell^2$$

No RW is able to distinguish $G_{k,\ell}$ and $G_{2k,\ell}$ before time

$$\ell^2\sqrt{k} \gtrsim t_{\mathrm{rel}}^{3/4}\sqrt{n}$$

Non-regular graphs

Let X and Y be two indep. lazy RWs started at $x \in V$.

$$\mathcal{I}_t = \sum_{i,j=0}^{t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}} \qquad \mathcal{J}_t = \sum_{i,j=t_{\mathrm{unif}}}^{t_{\mathrm{unif}}+t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}}$$

Non-regular graphs

Let X and Y be two indep. lazy RWs started at $x \in V$.

$$\mathcal{I}_t = \sum_{i,j=0}^{t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}} \qquad \mathcal{J}_t = \sum_{i,j=t_{\text{unif}}}^{t_{\text{unif}}+t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}}$$

First moment

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

Non-regular graphs

Let X and Y be two indep. lazy RWs started at $x \in V$.

$$\mathcal{I}_t = \sum_{i,j=0}^{t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}} \qquad \mathcal{J}_t = \sum_{i,j=t_{\text{unif}}}^{t_{\text{unif}}+t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}}$$

First moment

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

Bound on return probabilities

$$P^t(x,x) \leq \pi(x) + \frac{4\deg(x)}{\sqrt{t}} \qquad \text{Lyons, 2005}$$

Non-regular graphs

Let X and Y be two indep. lazy RWs started at $x \in V$.

$$\mathcal{I}_t = \sum_{i,j=0}^{t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}} \qquad \mathcal{J}_t = \sum_{i,j=t_{\text{unif}}}^{t_{\text{unif}}+t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}}$$

First moment

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

Bound on return probabilities

$$P^t(x,x) \leq \pi(x) + \frac{4\deg(x)}{\sqrt{t}} \qquad \text{Lyons, 2005}$$

Bound on the sum

$$\sum_{x \in V} P^t(x,x) \leq 1 + \frac{13n}{t^{1/3}} \qquad \text{Lyons and Oveis Gharan, 2017}$$

Non-regular graphs

Consider K pairs of RWs $\left\{\left(X^{(k)}, Y^{(k)}\right)\right\}_{k=1}^{K}$ and the estimator

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

Non-regular graphs

Consider K pairs of RWs $\left\{(X^{(k)}, Y^{(k)})\right\}_{k=1}^{K}$ and the estimator

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

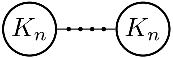
Theorem (B., Oliveira and Peres)
For all $G = (V, E)$ connected, for all $x \in V$,

$$\forall\, t \geq t_{\text{rel}}^{5/6}\sqrt{n}\,,\ \mathbb{P}_x\left(\left|\frac{\widehat{m}_t}{m} - 1\right| > \frac{1}{2}\right) = O\left(\frac{1}{K}\right).$$

Alternatively: $t \geq t_{\text{rel}}^{3/4}\sqrt{m}$.

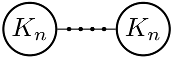
Lower bound

Bound $t_{\text{rel}}^{5/6}\sqrt{n}$ attained by the barbell.



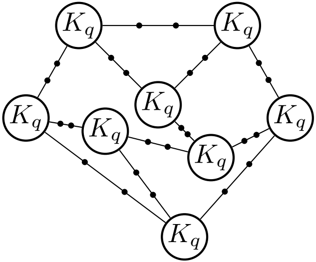
Lower bound

Bound $t_{\text{rel}}^{5/6}\sqrt{n}$ attained by the barbell.



Lower bound for each possible mixing time

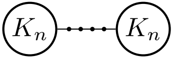
1. Start with a 3-regular expander $\mathcal{E}_k$ of size k;
2. Replace each node of $\mathcal{E}_k$ by a clique K_q of size q;
3. Replace each edge of $\mathcal{E}_k$ by a path of length q.



$$
\begin{aligned}
n &\asymp kq \\
t_{\text{rel}} &\asymp q^3
\end{aligned}
$$

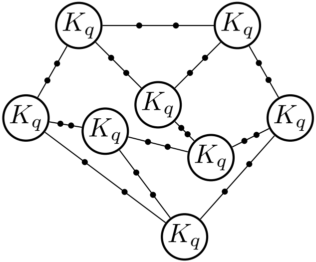
Lower bound

Bound $t_{\text{rel}}^{5/6}\sqrt{n}$ attained by the barbell.



Lower bound for each possible mixing time

1. Start with a 3-regular expander $\mathcal{E}_k$ of size k;
2. Replace each node of $\mathcal{E}_k$ by a clique K_q of size q;
3. Replace each edge of $\mathcal{E}_k$ by a path of length q.



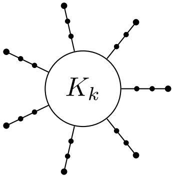
$$n \asymp kq$$
$$t_{\text{rel}} \asymp q^3$$

No RW is able to distinguish $G_{k,q}$ and $G_{2k,q}$ before time

$$q^3\sqrt{k} \gtrsim t_{\text{rel}}^{5/6}\sqrt{n}$$

From edges to vertices

Time $t_{\text{unif}}^{5/6}\sqrt{n}$ is not enough to estimate the number of vertices.



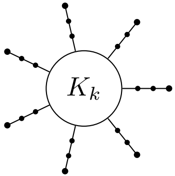
1. Take a complete graph of size k;
2. Add paths of length q to each vertex, with $q << k$.

$n \asymp kq \qquad m \asymp k^2 \qquad t_{\text{unif}} \asymp q^2$

$T(n) \asymp qk \gg t_{\text{unif}}^{5/6}\sqrt{n}$ for q small enough.

From edges to vertices

Time $t_{\text{unif}}^{5/6}\sqrt{n}$ is not enough to estimate the number of vertices.



1. Take a complete graph of size k;
2. Add paths of length q to each vertex, with $q << k$.

$n \asymp kq \qquad m \asymp k^2 \qquad t_{\text{unif}} \asymp q^2$

$T(n) \asymp qk \gg t_{\text{unif}}^{5/6}\sqrt{n}$ for q small enough.

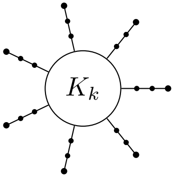
However, once a good estimate for m is available, it suffices to estimate the mean degree, which can be done in

$$O\left(t_{\text{unif}}\frac{m}{n}\right)$$

By previous example, this is sharp: $T(n) \asymp qk \asymp t_{\text{unif}}\frac{m}{n}$.

From edges to vertices

Time $t_{\mathrm{unif}}^{5/6}\sqrt{n}$ is not enough to estimate the number of vertices.



1. Take a complete graph of size k;
2. Add paths of length q to each vertex, with $q << k$.

$n \asymp kq \qquad m \asymp k^2 \qquad t_{\mathrm{unif}} \asymp q^2$

$T(n) \asymp qk \gg t_{\mathrm{unif}}^{5/6}\sqrt{n}$ for q small enough.

However, once a good estimate for m is available, it suffices to estimate the mean degree, which can be done in

$$O\left(t_{\mathrm{unif}}\frac{m}{n}\right)$$

By previous example, this is sharp: $T(n) \asymp qk \asymp t_{\mathrm{unif}}\frac{m}{n}$.

All in all, the number of vertices can be estimated in time

$$t_{\mathrm{rel}}^{5/6}\sqrt{n} + t_{\mathrm{unif}}\frac{m}{n}$$

A self-stopping algorithm for the number of edges

Assume that an upper bound T on t_{unif} is available.

For all $\varepsilon > 0$, one may design a self-stopping algorithm such that

- with probability $1 - \varepsilon$, the returned value $\widehat{m}$ satisfies $\left|\frac{\widehat{m}}{m} - 1\right| \leq \frac{1}{2}$;
- the expected running time is $O\left(\sqrt{m}T^{3/4}\log\log m\right)$.

A self-stopping algorithm for the mixing time

Assume that m is known (or that we have a good approximation).

We want to estimate $t_x(\delta) = \inf\{t \geq 0,\ d_x(t) \leq \delta\}$, where

$$d_x(t) = \sum_y \pi(y) \left(\frac{P^t(x,y)}{\pi(y)} - 1 \right)^2 .$$

A self-stopping algorithm for the mixing time

Assume that m is known (or that we have a good approximation).

We want to estimate $t_x(\delta) = \inf\{t \geq 0,\ d_x(t) \leq \delta\}$, where

$$d_x(t) = \sum_y \pi(y) \left(\frac{P^t(x,y)}{\pi(y)} - 1 \right)^2 .$$

Connection with intersections

If $\mathcal{I}_t = \sum_{i,j=0}^{t-1} \frac{1}{\deg(X_i)} \mathbb{1}_{\{X_i = Y_j\}}$, then $\mathbb{E}_x \mathcal{I}_t = \sum_{i,j=0}^{t-1} \frac{d_x\left(\frac{i+j}{2}\right)^2 + 1}{2m}$.

One can design a self-stopping algorithm such that

- with probability $1 - \varepsilon$, the returned value $\widehat{t_x(\delta)}$ satisfies $\frac{t_x(\delta)}{2} \leq \widehat{t_x(\delta)} \leq t_x(\delta/4)$.
- the expected running time is $O\left(\frac{\sqrt{m}}{\delta} t_x(\delta/4)^{3/4} \log\log t_x(\delta/4) \right)$.