Welcome to San Diego

AN08-MS10

Probabilistic Methods

for Complex Graphs

10:30AM-12:30PM

July 7, 2008

Diameter of Random Spanning Trees in a Given Graph

Fan Chung Graham, Paul Horn
University of California, San Diego

Linyuan Lu*
University of South Carolina

Outline

- Motivation
- Laplacian eigenvalues
- Random walks
- Spanning trees
- Results
- Methods

Motivation

Liben-Nowell and Kleinberg (PNAS 2008) studied Internet chain-letter data.

PNAS

Tracing information flow on a global scale using Internet chain-letter data

David Liben-Nowell*[†] and Jon Kleinberg[‡‡]

*Department of Computer Science, Carleton College, Northfield, MN 55057; and [‡]Department of Computer Science, Cornell University, Ithaca, NY 14853

Edited by Ronald L. Graham, University of California at San Diego, La Jolla, CA, and approved January 25, 2008 (received for review September 6, 2007)

Although information, news, and opinions continuously circulate in the worldwide social network, the actual mechanics of how any single piece of information spreads on a global scale have been a mystery. Here, we trace such information-spreading processes at a person-by-person level using methods to reconstruct the propagation of massively circulated Internet chain letters. We find that rather than fanning out widely, reaching many people in very few steps according to "small-world" principles, the progress of these chain letters proceeds in a narrow but very deep tree-like pattern, continuing for several hundred steps. This suggests a new and more complex picture for the spread of information through a social network. We describe a probabilistic model based on network clustering and asynchronous response times that produces trees with this characteristic structure on social-network data.

social networks | algorithms | epidemics | diffusion in networks

information transmission in the local dynamics of communication within highly clustered social networks.

Reconstructing the Spread of Internet Chain Letters

To reconstruct instances in which specific pieces of information spread through large, globally distributed populations, we analyzed the dissemination of petitions that circulated widely in chain-letter form on the Internet over the past several years. The petitions instruct each recipient to append his or her name to a copy of the letter and then forward it to friends. Each copy will thus contain a list of people, representing a particular sequence of forwardings of the message; and hence different copies will contain different but overlapping lists of people, reflecting the paths they followed to their respective current recipients. This forwarding process is a readily recognizable mechanism by which jokes and news clippings can also achieve wide circulation through the global e-mail network; the explicit lists of names in the petition format, however, make it

Motivation

Liben-Nowell and Kleinberg (PNAS 2008) studied Internet chain-letter data.

Tracing information flow on a global scale using Internet chain-letter data

David Liben-Nowell*† and Jon Kleinberg‡‡

*Department of Computer Science, Carleton College, Northfield, MN 55057; and ‡Department of Computer Science, Cornell University, Ithaca, NY 14853

Edited by Ronald L. Graham, University of California at San Diego, La Jolla, CA, and approved January 25, 2008 (received for review September 6, 2007)

Although information, news, and opinions continuously circulate in the worldwide social network, the actual mechanics of how any single piece of information spreads on a global scale have been a mystery. Here, we trace such information-spreading processes at a person-by-person level using methods to reconstruct the propagation of massively circulated Internet chain letters. We find that rather than fanning out widely, reaching many people in very few steps according to "small-world" principles, the progress of these chain letters proceeds in a narrow but very deep tree-like pattern, continuing for several hundred steps. This suggests a new and more complex picture for the spread of information through a social network. We describe a probabilistic model based on network clustering and asynchronous response times that produces trees with this characteristic structure on social-network data.

social networks | algorithms | epidemics | diffusion in networks

information transmission in the local dynamics of communication within highly clustered social networks.

Reconstructing the Spread of Internet Chain Letters

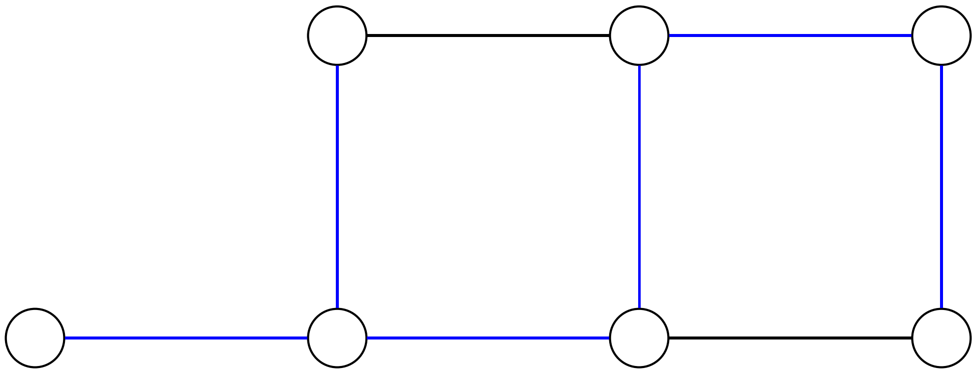
To reconstruct instances in which specific pieces of information spread through large, globally distributed populations, we analyzed the dissemination of petitions that circulated widely in chain-letter form on the Internet over the past several years. The petitions instruct each recipient to append his or her name to a copy of the letter and then forward it to friends. Each copy will thus contain a list of people, representing a particular sequence of forwardings of the message; and hence different copies will contain different but overlapping lists of people, reflecting the paths they followed to their respective current recipients. This forwarding process is a readily recognizable mechanism by which jokes and news clippings can also achieve wide circulation through the global e-mail network; the explicit lists of names in the petition format, however, make it

A typical spanning tree often has relatively large diameter.

Spanning trees

A subgraph T is a spanning tree of a connected graph G if

- $V(T) = V(G)$;
- T is a tree.



Enumerating spanning trees

- G: a connected graph on n vertices.
- A: adjacency matrix of G.
- D: the diagonal matrix of degrees.

Kirchoff's Matrix-Tree Theorem (1847):
The number of spanning trees in a graph G is the absolution value of the determinant of any $(n-1) \times (n-1)$ sub-matrix of $D - A$.

Enumerating spanning trees

- G: a connected graph on n vertices.
- A: adjacency matrix of G.
- D: the diagonal matrix of degrees.

Kirchoff's Matrix-Tree Theorem (1847):
The number of spanning trees in a graph G is the absolution value of the determinant of any $(n-1) \times (n-1)$ sub-matrix of $D - A$.

Cayley's Formula:
The number of spanning trees of K_n is n^{n-2}.

Diameter of Spanning trees

Rényi and Szekeres (1967): The diameter of a random spanning tree in the complete graph K_n is of order $\sqrt{n}$.

Diameter of Spanning trees

Rényi and Szekeres (1967): The diameter of a random spanning tree in the complete graph K_n is of order $\sqrt{n}$.

Aldous (1990): Let $diam(T)$ be the diameter of a random spanning tree in a regular graph with spectral bound σ. Then

$$\frac{c(1-\sigma)\sqrt{n}}{\log n} \leq \mathrm{E}(diam(T)) \leq \frac{c\sqrt{n}}{\sqrt{1-\sigma}} \log n.$$

Spectral bound σ

- Laplacian: $\mathcal{L} = I - D^{-1/2} A D^{-1/2}$
- Laplacian spectrum:

$$0 = \lambda_0 \le \lambda_1 \le \cdots \le \lambda_{n-1} \le 2.$$

- Spectrum bound σ:

$$\sigma = \max_{1 \le i \le n-1} \{|\lambda_i - 1|\}.$$

- $\sigma \le 1$. "$=$" holds if and only if G is disconnected or bipartite.

Main question

What is the diameter of random spanning trees of a given graph G?

Notations

For a given graph G, let

- n: the number of vertices.
- d_i: the degree of i-th vertex.
- $\mathrm{vol}(G) = \sum_{i=1}^{n} d_i$: the sum of degrees.
- $d = \frac{\mathrm{vol}(G)}{n}$: the average degree.
- $\tilde{d} = \frac{\sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{n} d_i}$: the second order average degree.
- δ: the minimum degree.
- σ: the spectral bound.
- $diam(T)$: the diameter of random spanning trees.

Notations

For a given graph G, let

- n: the number of vertices.
- d_i: the degree of i-th vertex.
- $\mathrm{vol}(G) = \sum_{i=1}^{n} d_i$: the sum of degrees.
- $d = \frac{\mathrm{vol}(G)}{n}$: the average degree.
- $\tilde{d} = \frac{\sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{n} d_i}$: the second order average degree.
- δ: the minimum degree.
- σ: the spectral bound.
- $diam(T)$: the diameter of random spanning trees.

We have

$$\delta \le d \le \tilde{d}.$$

Main result

Chung, Horn, Lu (2008)

If $d \gg \frac{\log^2 n}{\log^2 \sigma}$, then with probability $1 - \epsilon$, we have

$$(1-\epsilon)\sqrt{\frac{\epsilon n d}{\tilde{d}}} \leq diam(T) \leq \frac{c}{\epsilon}\sqrt{\frac{nd}{\delta \log(1/\sigma)}} \log n.$$

Main result

Chung, Horn, Lu (2008)

If $d \gg \frac{\log^2 n}{\log^2 \sigma}$, then with probability $1-\epsilon$, we have

$$(1-\epsilon)\sqrt{\frac{\epsilon n d}{\tilde{d}}} \leq diam(T) \leq \frac{c}{\epsilon}\sqrt{\frac{nd}{\delta \log(1/\sigma)}} \log n.$$

If $\tilde{d} \leq C\delta$, then

$$\Omega(\sqrt{n}) \leq \mathrm{E}(diam(T)) \leq O(\sqrt{n}\log n).$$

Main result

Chung, Horn, Lu (2008)

If $d \gg \frac{\log^2 n}{\log^2 \sigma}$, then with probability $1-\epsilon$, we have

$$(1-\epsilon)\sqrt{\frac{\epsilon n d}{\tilde{d}}} \leq diam(T) \leq \frac{c}{\epsilon}\sqrt{\frac{nd}{\delta \log(1/\sigma)}} \log n.$$

If $\tilde{d} \leq C\delta$, then

$$\Omega(\sqrt{n}) \leq \mathrm{E}(diam(T)) \leq O(\sqrt{n}\log n).$$

Applying to d-regular graphs, our results improve Aldous's result by a $\log n$-factor.

Next...

- Random walks

Next...

- Random walks
- Groundskeeper algorithm

Next...

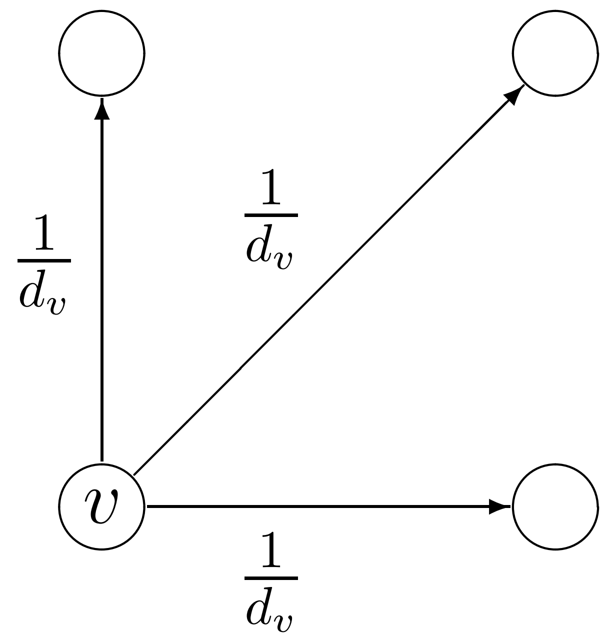
- Random walks
- Groundskeeper algorithm
- Proof for lower bound

Random walks

Random walks on a graph G:

$$P = D^{-1}A,$$

$$\beta_{t+1} = \beta_t P.$$

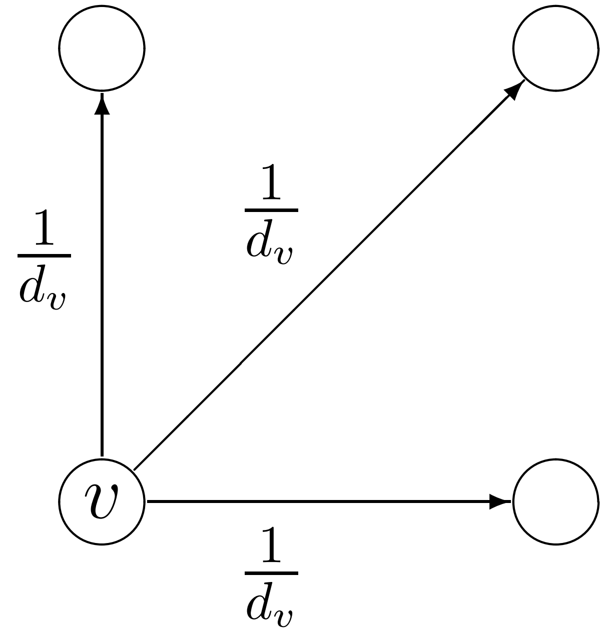


Random walks

Random walks on a graph G:

$$P = D^{-1}A,$$

$$\beta_{t+1} = \beta_t P.$$



$$P \sim D^{-1/2}AD^{-1/2} = I - \mathcal{L}.$$

The spectral bound σ measures the mixing rate of random walks.

Stationary distribution π

$\beta = (\beta_1, \ldots, \beta_n)$ is a probability distribution if

- $\beta_i \geq 0$, for $1 \leq i \leq n$.
- $\sum_{i=1}^{n} \beta_i = 1$.

Stationary distribution π

$\beta = (\beta_1, \ldots, \beta_n)$ is a probability distribution if

- $\beta_i \geq 0$, for $1 \leq i \leq n$.
- $\sum_{i=1}^{n} \beta_i = 1$.

P maps probability distributions to probability distributions.

Stationary distribution π

$\beta = (\beta_1, \ldots, \beta_n)$ is a probability distribution if

- $\beta_i \geq 0$, for $1 \leq i \leq n$.
- $\sum_{i=1}^{n} \beta_i = 1$.

P maps probability distributions to probability distributions.

This mapping has a unique fixed point:

$$\pi = \frac{1}{\mathrm{vol}(G)}(d_1, d_2, \ldots, d_n).$$

$$\pi P = \pi.$$

Mixing rate

Lemma For any integer $t > 0$, any $\alpha \in \mathbb{R}^n$, and any two probability distributions β and γ, we have

$$\langle (\beta - \gamma) P^t, \alpha D^{-1} \rangle \leq \sigma^t \|(\beta - \gamma) D^{-1/2})\| \|\alpha D^{-1/2}\|.$$

In particular,

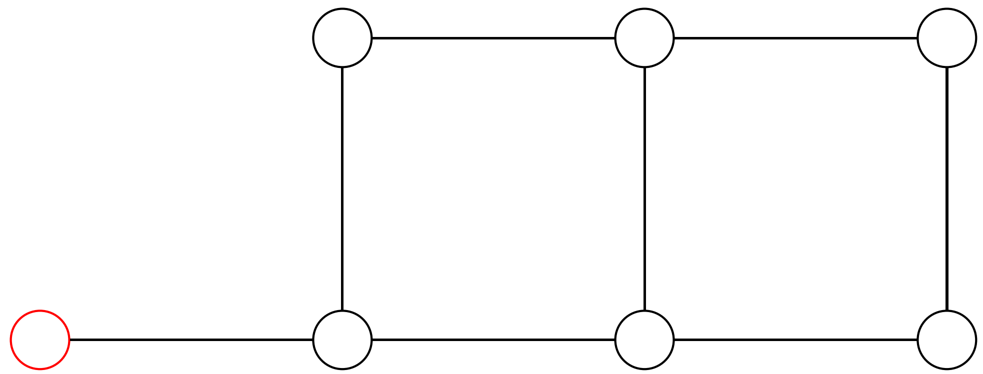
$$\|(\beta - \gamma) P^t D^{-1/2}\| \leq \sigma^t \|(\beta - \gamma) D^{-1/2}\|.$$

Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.

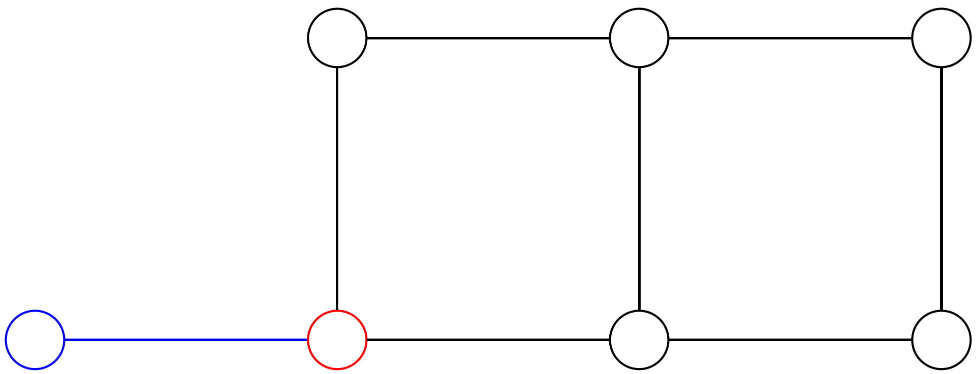
Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



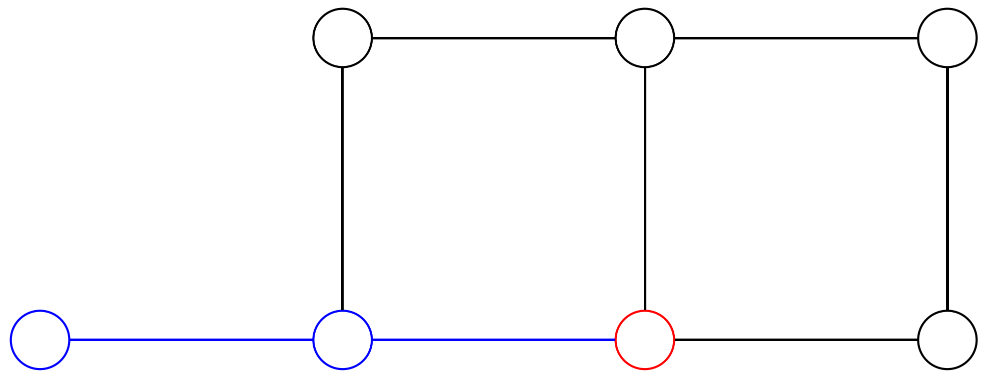
Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



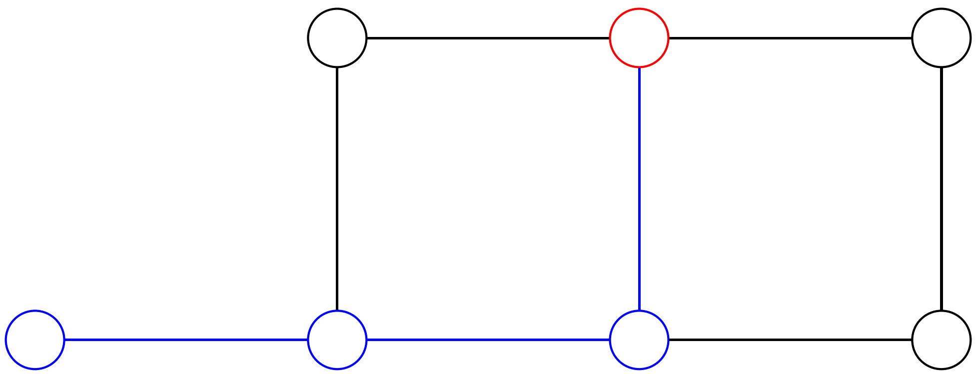
Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



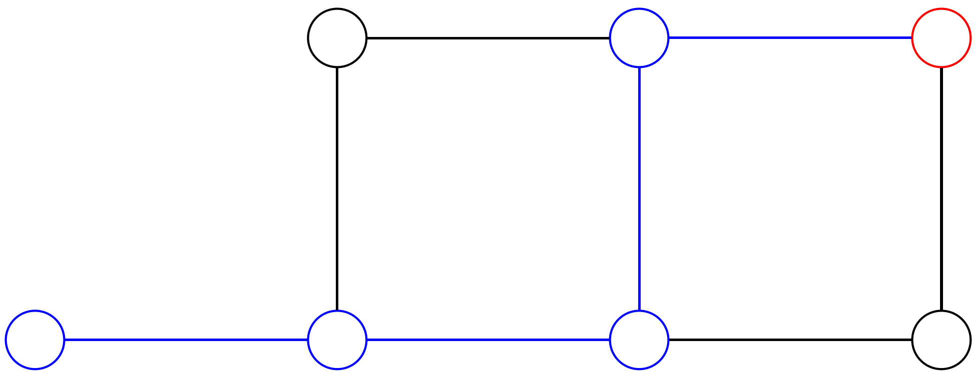
Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



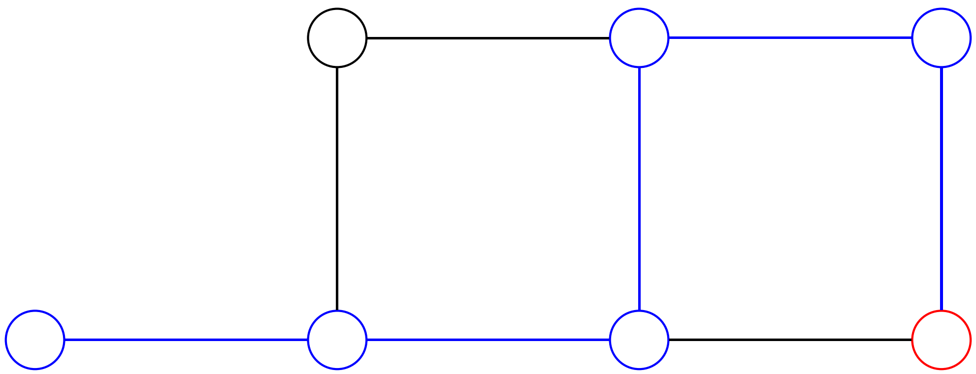
Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



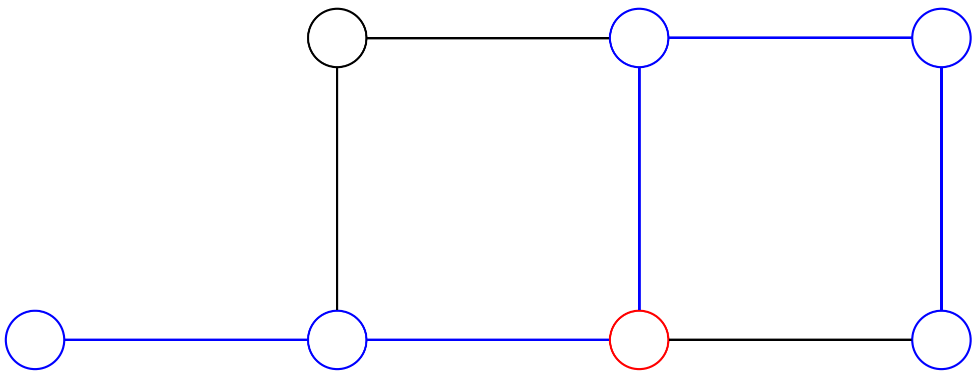
Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



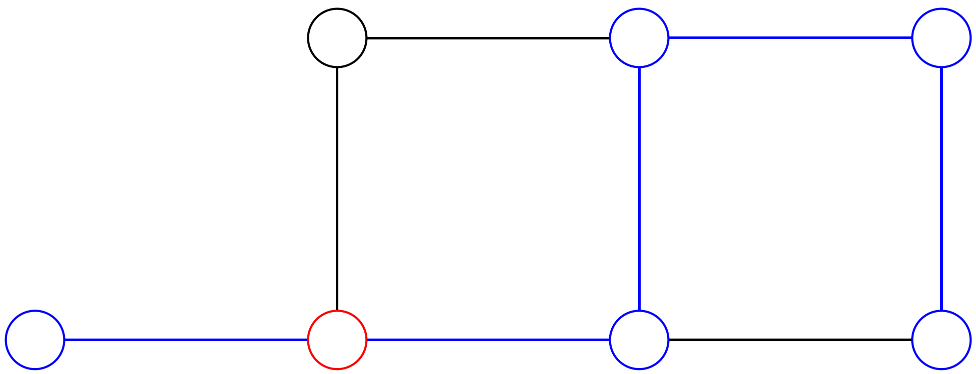
Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



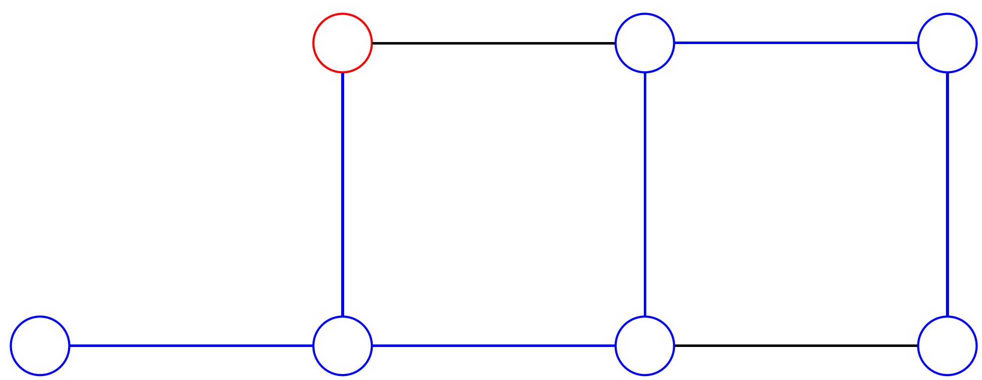
Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



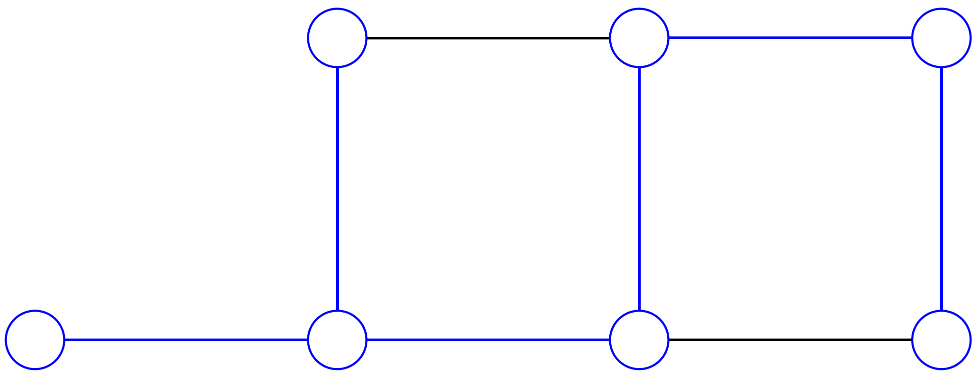
Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



Groundskeeper Algorithm

Starting a random walk at any vertex. The first time a vertex is visited through an edge f, we add the edge f to our spanning tree. Once the graph is covered, the resulting set of edges forms a spanning tree.



Φ: {random walks} $\rightarrow$ {random spanning trees}

Groundskeeper Algorithm

Aldous (1990), Broder (1989) The image of Φ is uniformly distributed over all spanning trees. It is independent of the choice of initial vertex v.

Groundskeeper Algorithm

Aldous (1990), Broder (1989) The image of Φ is uniformly distributed over all spanning trees. It is independent of the choice of initial vertex v.

We pick up a random initial vertex with stationary distribution π.

Difficulty

A random walk $\{v_t\}$ contains a circuit of length k if

$$v_{i+k} = v_i \quad \text{for some } i.$$

Difficulty

A random walk $\{v_t\}$ contains a circuit of length k if

$$v_{i+k} = v_i \quad \text{for some } i.$$

We try to analyze the length from v_0 to v_t in a random spanning tree. Here is the difficulty in analyzing this length.

- Long circuits
- Many short circuits

g-truncated random walks

We stop random walks when a circuit of length at least g is formed.

g-truncated random walks

We stop random walks when a circuit of length at least g is formed.

Lemma. For a fixed integer g, the probability that g-truncated random walks stop before time t is at most

$$\frac{t^2 \tilde{d}}{2nd} + t\frac{\sigma^g}{1-\sigma}.$$

g-truncated random walks

We stop random walks when a circuit of length at least g is formed.

Lemma. For a fixed integer g, the probability that g-truncated random walks stop before time t is at most

$$\frac{t^2\tilde{d}}{2nd} + t\frac{\sigma^g}{1-\sigma}.$$

Let $t = (1-\epsilon)\sqrt{\epsilon\frac{d}{\tilde{d}}n}$ and $g = \lceil \frac{\log\left(\frac{\epsilon(1-\sigma)\sqrt{\delta}}{4t\sqrt{\tilde{d}}}\right)}{\log(\sigma)} \rceil$. The g-truncated random walks will survive up to time t with probability at least $1-\frac{3\epsilon}{4}$.

Random variable X

Consider g-truncated random walks. For $i \le t$, let

$$X_i = \begin{cases} 1 & v_i \neq v_j \text{ for all } j < i \\ -k & v_i = v_{i-k} \text{ for some } k. \end{cases}$$

and $X = \sum_{i=1}^{t} X_i$.

Random variable X

Consider g-truncated random walks. For $i \leq t$, let

$$X_i = \begin{cases} 1 & v_i \neq v_j \text{ for all } j < i \\ -k & v_i = v_{i-k} \text{ for some } k. \end{cases}$$

and $X = \sum_{i=1}^{t} X_i$.

Observation.

$$X \leq d_T(v_0, v_t) \leq diam(T).$$

Exposed Martingale

Let $\mathcal{F}_i$ be the σ-algebra that $v_0, \ldots, v_i$ is revealed.
$\{\mathrm{E}(X \mid \mathcal{F}_i)\}_{0 \leq i \leq t}$ forms a martingale.

Exposed Martingale

Let $\mathcal{F}_i$ be the σ-algebra that $v_0, \ldots, v_i$ is revealed. $\{\mathrm{E}(X \mid \mathcal{F}_i)\}_{0 \leq i \leq t}$ forms a martingale. We have

$$|\mathrm{E}(X_j \mid \mathcal{F}_i) - \mathrm{E}(X_j \mid \mathcal{F}_{i-1})| \leq \begin{cases} 0 & \text{if } j < i; \\ 2(g-2)\frac{\sqrt{nd}}{\sqrt{\delta}}\sigma^{j-g+2-i} & \text{if } j \geq i + 2g + 2; \\ g - 1 & \text{otherwise.} \end{cases}$$

Exposed Martingale

Let $\mathcal{F}_i$ be the σ-algebra that $v_0, \ldots, v_i$ is revealed. $\{\mathrm{E}(X \mid \mathcal{F}_i)\}_{0 \leq i \leq t}$ forms a martingale. We have

$$|\mathrm{E}(X_j \mid \mathcal{F}_i) - \mathrm{E}(X_j \mid \mathcal{F}_{i-1})| \leq \begin{cases} 0 & \text{if } j < i; \\ 2(g-2)\frac{\sqrt{nd}}{\sqrt{\delta}}\sigma^{j-g+2-i} & \text{if } j \geq i + 2g + 2; \\ g - 1 & \text{otherwise.} \end{cases}$$

Summing up, we have the following Lipschitz Condition:

$$|\mathrm{E}(X \mid \mathcal{F}_i) - \mathrm{E}(X \mid \mathcal{F}_{i-1})| \leq 3g^2$$

Estimate $\mathrm{E}(X)$

$$\mathrm{E}(X) = \sum_{i=1}^{t} \mathrm{E}(X_i)$$

Estimate $\mathrm{E}(X)$

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{i=1}^{t} \mathrm{E}(X_i) \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n} \mathrm{E}(X_i \mid v_{i-1} = j) \Pr(V_{i-1} = j)
\end{aligned}$$

Estimate $\mathrm{E}(X)$

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{i=1}^{t} \mathrm{E}(X_i) \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n} \mathrm{E}(X_i \mid v_{i-1} = j) \Pr(V_{i-1} = j) \\
&\geq \sum_{i=1}^{t} \sum_{j=1}^{n} \left((1 - \frac{g-1}{d_j}) + \sum_{k=1}^{g-2} \frac{-k}{d_j} \right) \frac{d_j}{nd}
\end{aligned}
$$

Estimate $\mathrm{E}(X)$

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{i=1}^{t} \mathrm{E}(X_i) \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n} \mathrm{E}(X_i \mid v_{i-1} = j) \mathrm{Pr}(V_{i-1} = j) \\
&\geq \sum_{i=1}^{t} \sum_{j=1}^{n} \left((1 - \frac{g-1}{d_j}) + \sum_{k=1}^{g-2} \frac{-k}{d_j} \right) \frac{d_j}{nd} \\
&= (1 - \frac{g(g-1)}{2d}) t.
\end{aligned}$$

Put together

By applying Azuma's inequality, we have

$$\Pr(X - \mathrm{E}(X) < -\alpha) < e^{-\frac{\alpha^2}{18g^4t}}$$

Put together

By applying Azuma's inequality, we have

$$\Pr(X - \mathrm{E}(X) < -\alpha) < e^{-\frac{\alpha^2}{18g^4t}}$$

By choosing $\alpha = \sqrt{18g^4t\log\frac{4}{\epsilon}}$, we have

$$\Pr\left(X < (1 - \frac{g(g-1)}{2d})t - \sqrt{18g^4t\log\frac{4}{\epsilon}}\right) < \frac{\epsilon}{4}.$$

Put together

By applying Azuma's inequality, we have

$$\Pr(X - \mathrm{E}(X) < -\alpha) < e^{-\frac{\alpha^2}{18g^4t}}$$

By choosing $\alpha = \sqrt{18g^4t\log\frac{4}{\epsilon}}$, we have

$$\Pr\left(X < (1 - \frac{g(g-1)}{2d})t - \sqrt{18g^4t\log\frac{4}{\epsilon}}\right) < \frac{\epsilon}{4}.$$

Recall $t = (1-\epsilon)\sqrt{\epsilon\frac{d}{\tilde{d}}n}$, we have

$$(1 - \frac{g(g-1)}{2d})t - \sqrt{18g^4t\log\frac{4}{\epsilon}} = (1-\epsilon-o(1))\sqrt{\epsilon\frac{nd}{\tilde{d}}}.\square$$

Open questions

Recall we prove

$$(1-\epsilon)\sqrt{\frac{\epsilon n d}{\tilde{d}}} \leq diam(T) \leq \frac{c}{\epsilon}\sqrt{\frac{nd}{\delta \log(1/\sigma)}} \log n.$$

Open questions:

- In the upper bound, can we replace the minimum degree δ by the average degree d?
- Can we remove the multiplicative $\frac{1}{\sqrt{\log(1/\sigma)}} \log n$-factor?