Notes on Lawler's 1983 paper "Expected hitting times for a random walk on a connected graph"

September 2023

Notes

0.1 Definitions, vocabulary, nomenclature

- $G = (V, E)$
- Let S_n be a random walk of n steps on graph G. S_0 denotes the initial state of the random walk.
- Let $\Pr_x$ and $\mathbb{E}_x$ be the probability and expectation of events associated with random walks starting at node x i.e $S_0 = x$. Unless otherwise specified all of our random walks will start at node x. Importantly, starting at x counts as a visit to x.
- To "hit" y means to visit node y for the first time.
- Let τ^y be the number of steps till the first visit to node y i.e the hitting time to node y from x.
- $\mathbb{E}_x[\tau^y]$ is the expected hitting time to node y.
- Let $\pi(x, y)$ be the transition function, i.e it denotes the probability of jumping from x to y in a single step. For connected graphs:

$$\pi(x, y) = \begin{cases} 0, & (x, y) \notin E \\ \frac{1}{\deg(x)} & (x, y) \in E \end{cases}$$

- Let $p(x, y)$ be the probability that a random walk starting at x hits y without ever hitting itself again. When necessary, we will make use of the notation $p(x, y; G)$ to make explicit which graph we are doing a random walk on. Mathematically, $p(x, y) = \Pr_x[\exists j \text{ s.t } S_j = y, S_i \neq x, 1 \leq i < j]$. Note that this is for random walks of all lengths that have this property.

  An example:
  Suppose $x = 1$ and $y = 4$.
  Let A be a random walk sequence of $[1, 2, 3, 2, 3, 4]$ and B be a random walk sequence of $[1, 2, 1, 2, 3, 4]$.
  A is a random walk where x does not hit itself before hitting y, but B is not.

  Notice that $p(x, x) = 1$.

- Let $p(x, z, y)$ be the probability that a random walk starting at x hits z without ever hitting y. Mathematically, $p(x, z, y) = \Pr_x[\exists j \geq 0 \text{ s.t } S_j = z, S_i \neq y, 0 \leq i < j]$.

  An example:
  Suppose $x = 1$, $z = 4$ and $y = 3$.

Let A be a random walk sequence of $[1, 2, 4]$ and B be a random walk sequence of $[1, 2, 3, 4]$.
A is a random walk where x does not hit y before hitting z, but B is not.

Notice that $p(x, y, y) = 1$ and $p(x, x, y) = 0$.

- Let $\delta_x(S_i)$ be an indicator random variable s.t

$$\delta_x = \begin{cases} 1 \text{ if } S_i = x, \\ 0 \text{ otherwise} \end{cases}$$

  Then, $\sum\limits_{j=0}^{\tau^y} \delta_x(S_j)$ is the number of times that the random walk hit visited node x before hitting y for the first time.

- $R(x, y) = \mathbb{E}_x\left[\sum\limits_{j=0}^{\tau^y} \delta_x(S_j)\right]$ is the expected number of visits to x before hitting y. Notice that $R(x, x) = 1$.

Lemma 0.1.

$$R(x, y) = \frac{1}{p(x, y)}$$

Proof. When you start a random walk at x, there are two possible outcomes:

- your walk hits y without "looping" back to x, w.p $p(x, y)$.
- your walk "loops" back to x, w.p $1 - p(x, y)$.

You can think of this as flipping a coin, where you succeed w.p $p(x, y)$ and fail w.p $1 - p(x, y)$.

Let X be the number of times you have to flip this coin before observing the success event. This is a geometric random variable, so $\mathbb{E}[X] = \frac{1}{p(x,y)}$. The expected number of failures therefore is $\frac{1}{p(x,y)} - 1$. This is the expected number of times our walk loops back to its starting point. $R(x, y)$ is the expectation of the number of times our random walk visits x, which is $1 +$ the expected number of failures. We add this one because starting the walk at x also counts as visiting x. The expected number of visits to x is therefore $1 + \frac{1}{p(x,y)} - 1 = \frac{1}{p(x,y)}$. $\square$

Also,

$$\mathbb{E}_x[\tau^y] = \mathbb{E}_x\Bigg[\underbrace{\sum_{z\in G} \underbrace{\sum_{j=1}^{\tau^y} \delta_z(S_j)}_{\text{number of visits to } z \text{ on rw}}}_{\text{number of visits to all nodes}}\Bigg] \qquad \text{(This term does not count the start node as a visit.)}$$

$$= \mathbb{E}_x\left[\sum_{z\in G}\sum_{j=0}^{\tau^y} \delta_z(S_j) - 1\right] \qquad \text{(This term does counts the start node as well.)}$$

$$= \sum_{z\in G} \mathbb{E}_x\left[\sum_{j=0}^{\tau^y} \delta_z(S_j)\right] - 1$$

$$= \left(\sum_{z\in G} p(x,z,y)\cdot \mathbb{E}_z\left[\sum_{j=0}^{\tau^y} \delta_z(S_j)\right]\right) - 1$$

$$= \left(\sum_{z\in G} p(x,z,y)\cdot R(z,y)\right) - 1$$

$$= \left(\sum_{z\in G} \frac{p(x,z,y)}{p(z,y)}\right) - 1 \qquad \text{(From Lemma 0.1)}$$

The major theorem of the paper is the following, and we will spend the most of our effort proving this:

Theorem 0.1. Let G be a connected graph with N vertices. Then for every $x, y \in G$,

$$\deg(x)\cdot p(x,y) \geq \frac{1}{N-1}$$

The above theorem allows us to prove the more interesting:

Theorem 0.2. Let G be a connected graph with N vertices. Then for every $x, y \in G$,

$$\mathbb{E}_x[\tau^y] \leq N(N-1)^2$$

Proof. Assume that Theorem 0.2 is true. Then we get:

$$\deg(x)\cdot p(x,y) \geq \frac{1}{N-1}$$

$$p(x,y) \geq \frac{1}{(N-1)^2}$$

$$(N-1)^2 \geq \frac{1}{p(x,y)}$$

Using this for our above expression of the expected hitting time, we get:

$$\begin{aligned}
\mathbb{E}_x[\tau^y] &= \left(\sum_{z\in G} \frac{p(x,z,y)}{p(z,y)}\right) - 1 \\
&\leq \left(\sum_{z\in G} p(x,z,y)(N-1)^2\right) - 1 \\
&= (N-1)^2 \underbrace{\big(\sum_{z\in G} p(x,z,y)\big)}_{\leq N} -1 \\
&\leq N(N-1)^2 - 1
\end{aligned}$$

□

This gives us the upper bound on the expected hitting time.

Notice that this theorem is for any two nodes though. More importantly, this theorem says that the expected hitting time between the farthest away nodes is upper bounded by $O(n^3)$. We are interested in the expected hitting time between a randomly chosen node and the most central node of the graph in terms of random-walk distance.

This is a Lemma that I don't fully understand yet, but will have to take for granted to prove Theorem x.

Lemma 0.2. If G is a graph, and $x, y \in G$, then:

$$\deg(x) \cdot p(x,y) = \deg(y) \cdot p(y,x)$$

Proof. We know that for irreducible Markov Chains, there exists an invariant probability measure ϕ on G s.t

$$\sum_{x\in G} \phi(x) = 1$$

and for every $x \in G$,

$$\sum_{y\in G} \phi(y) \cdot \pi(y,x) = \phi(x)$$

The function $\phi(x)$ is simply the probability of being at the node x in the stationary distribution. I don't prove this but it is known that

$$\phi(x) = \frac{\deg(x)}{2|E|}$$

We combine this fact with another well-known fact about Markov Chains

$$\begin{aligned}
\frac{p(y,x)}{p(x,y)} &= \frac{\phi(x)}{\phi(y)} \\
&= \frac{\deg(x)}{\deg(y)}
\end{aligned}$$

Rearranging the above term completes the proof. □

We now prove Theorem 0.2

The proof is by induction on the number of vertices N.

We start with the base case where G has only two vertices, i.e $N = 2$. The result is immediate:

$$\deg(x) \cdot p(x, y) \geq \frac{1}{N-1}$$
$$1 \cdot 1 \geq \frac{1}{2-1}$$

which is true.

The inductive case is when $N \geq 3$.

We first look at the case of $(x, y) \in E$, which happens to be relatively straightforward:

$$\deg(x) \cdot p(x, y) \geq deg(x) \cdot \pi(x, y) = deg(x) \cdot \frac{1}{deg(x)} = 1 \geq \frac{1}{N-1}$$

for $N \geq 3$.

We can therefore now focus on the case of when $(x, y) \notin E$.

For such graphs, we define three cases of what can happen when you delete node x from the graph:

1. Case 1: $\deg(x) = 1$.
2. Case 2: $\deg(x) > 1$, and deleting x results in a disconnected graph.
3. Case 3: $\deg(x) > 1$, and deleting x results in a connected graph.

For Case 3, we will formulate our analysis to be that for such graphs, $\exists z$ s.t deleting (x, z) from G still leaves a connected graph. This is a peculiar way to analyze Case 3, because this analysis will also work for some Case 2 graphs. An explainer is provided in Figure 3 and 4. The crucial thing to understand here is that although this analysis also applies for some Case 2 graphs, it applies for *all* Case 3 graphs.

Case 1: $\deg(x) = 1$

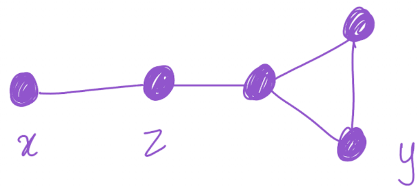

Figure 1: An example of Case 1

We are trying to prove that

$$\deg(x) \cdot p(x, y) \geq \frac{1}{N-1}$$

$\deg(x) = 1$ by definition, so it remains that we lower bound $p(x, y)$. Let G_1 be the graph obtained after deleting x from G.

By our inductive hypothesis:

$$\deg(z; G_1) \cdot p(z, y; G_1) \geq \frac{1}{N-2}$$

We want to compute $p(x, y; G)$, which we observe for this case to be

$$p(x, y; G) = p(z, y, x; G)$$

For this case, the probability that a random walk starting at x hits y without hitting x is the same as the probability of starting at z and never hitting x before hitting y.

Let us compute two terms that will come up when we solve for $p(z, y, x; G)$:

The probability of *not* choosing x in the random walk from z in the first step is:

$$\Pr[S_1 \neq x] = \frac{\deg(z; G) - 1}{\deg(z; G)} = \frac{\deg(z; G_1)}{\deg(z; G_1) + 1} \tag{1}$$

For this particular case, let us compute a particular event that we will use later:

$$\begin{aligned} p(z, y, x; G | S_1 \neq x) = & \underbrace{p(z, y; G_1)}_{\text{hit } y \text{ without hitting } z \text{ in } G_1} \\ & + \underbrace{(1 - p(z, y; G_1)) \cdot p(z, y, x; G)}_{\text{hit } z \text{ before hitting } y \text{, and we start at the same place}} \end{aligned} \tag{2}$$

Combining this, we get:

$$
\begin{aligned}
p(z,y,x;G) &= \underbrace{p(z,y,x;G|S_1=x)}_{=0} \cdot \Pr[S_1=x] + p(z,y,x;G|S_1\neq x)\cdot \Pr[S_1\neq x] \\
&= p(z,y,x;G|S_1\neq x)\cdot \Pr[S_1\neq x] \\
&= \frac{\deg(z;G_1)}{\deg(z;G_1)+1}\cdot\Bigg(p(z,y;G_1)+(1-p(z,y;G_1))\cdot p(z,y,x;G)\Bigg) && \text{using (1) and (2)} \\
&= \frac{\deg(z;G_1)p(z,y;G_1)}{1+\deg(z;G_1)p(z,y;G_1)} && \text{(just some algebra)} \\
&\geq \frac{\frac{1}{N-2}}{1+\frac{1}{N-2}} \\
&= \frac{1}{N-1}
\end{aligned}
$$

Case 2: $\deg(x) > 1$, and deleting x results in a disconnected graph.

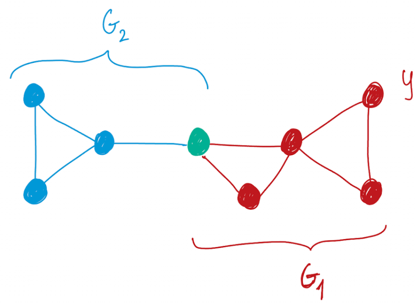

Figure 2: An example of Case 2

This is the case where the removal of x separates the graph G. Let $G = G_1 \cup G_2$ and $G_1 \cap G_2 = \{x\}$, and by definition of the case there are no edges shared between G_1 and G_2.

Let us suppose that $y \in G_1$. Then,

$$\begin{aligned}
p(x, y; G) &= p(x, y; G | S_1 \in G_1) \cdot \Pr[S_1 \in G_1] \\
&\quad + p(x, y; G | S_1 \neq G_1) \cdot \underbrace{\Pr[S_1 \neq G_1]}_{=0} \\
&= p(x, y; G | S_1 \in G_1) \cdot \Pr[S_1 \in G_1] \\
&= p(x, y; G | S_1 \in G_1) \cdot \frac{\deg(x; G_1)}{\deg(x; G)} \\
&= \frac{\deg(x; G_1)}{\deg(x; G)} \cdot p(x, y; G_1) \\
p(x, y; G) \cdot \deg(x; G) &= p(x, y; G_1) \cdot \deg(x; G_1) \\
&\geq \frac{1}{|G_1| - 1} && \text{(by our inductive hypothesis)} \\
&\geq \frac{1}{N - 1}
\end{aligned}$$

which concludes our proof for Case 2.

Case 3: $\deg(x) > 1$, and deleting x results in a connected graph. We will formulate our analysis by first noticing that for such graphs,

$\exists z \in G$ s.t deleting (x, z) from G still leaves the new graph H connected.

This formulation of Case 3 confused me for a while. Why would we define our case as that of deleting a *vertex*, but then do analysis on the case of deleting an edge?

This is why I think it works.

- Case 3 is when deleting x from G leaves H still connected.
- This is equivalent to saying that deleting all of x's edges (and x) leaves H connected.
- If deleting *all* of x's edges leaves H connected, then deleting just one of them definitely also leaves it connected. [1]
- i.e, $\exists z \in G$ s.t deleting (x, z) from G still leaves H connected.

The reason I got confused about this is because there are graphs where this formulation works even for Case 2. The key thing to understand here is that although this formulation is valid for some Case 2 graphs, it is valid for all Case 3 graphs, therefore proving assertions based off of this formulation still proves Case 3. The following figures show examples

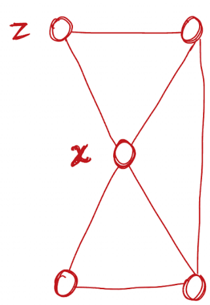

Figure 3: A canonical Case 3 graph

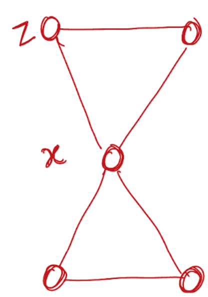

Figure 4: An example of a Case 2 graph that the formulation is also valid for

[1] Say thats not true. Deleting one of x's edges leaves H disconnected. "Disconnectedness" is not something you can fix by deleting more edges, so by the point you delete all edges you cannot have H connected.

So we have established exists an edge $(x, z) \in E$ s.t the graph H gotten by deleting the edge (x, z) is still connected. Importantly, notice that if we prove our theorem for this case, then subsequent deletion of nodes (and their edges) will result in our graph to eventually land in Case 1 or Case 2, which we have taken care of!

To prove the theorem for this case, it will suffice to show

$$\deg(x; G) \cdot p(x, y; G) \geq \deg(x; H) \cdot p(x, y; H)$$

Why? Because if this is true, then we can keep removing edges until we land in either Case 1 or Case 2.
By Lemma 0.2 it suffices to show

$$\begin{aligned} \deg(y; G) \cdot p(y, x; G) &\geq \deg(y; H) \cdot p(y, x; H) \\ p(y, x; G) &\geq p(y, x; H) \qquad \text{(because } \deg(y; G) = \deg(y; H)\text{)} \end{aligned}$$

Next we define two probabilities:

- β_x: the Pr that a random walk starting at y ends at x without hitting y or z first.
- β_z: the Pr that a random walk starting at y ends at z without hitting x or x first.

Note: If you want the formalisms for the above terms, check the original paper. It is not clear to me why in the above definitions it does not matter whether we are random-walking in G or H, but if we take that to be true then:

$$p(y, x; H) = \beta_x + \beta_z \cdot p(z, x, y; H) \tag{3}$$

$$p(y, x; G) = \beta_x + \beta_z \cdot p(z, x, y; G) \tag{4}$$

Looking at the above two equations, if we are able to prove that

$$p(z, x, y; H) \leq p(z, x, y; G)$$

then we are done, so that is what we will aim to do next.

We define the following probabilities:

- γ_x: the Pr of a random walk starting at z and hitting x without first hitting y or z
- γ_y: the Pr of a random walk starting at z and hitting y without first hitting x or z
- γ_z: the Pr of a random walk starting at z and hitting z after the first step without first hitting x or y

Notice that $\gamma_x + \gamma_y + \gamma_z = 1$. Then,

$$\begin{aligned} p(z,x,y;H) &= \gamma_x + \gamma_z \cdot p(z,x,y:H) \\ &= \frac{\gamma_x}{1-\gamma_z} && \text{(simply rearranging)} \\ &= \frac{\gamma_x}{\gamma_x+\gamma_y} \end{aligned}$$

Also,

$$\begin{aligned} p(z,x,y;G) &= \underbrace{p(z,x,y;G|S_1=x)}_{=1} \cdot \underbrace{\Pr[S_1=x]}_{=\frac{1}{\deg(z)}} + \underbrace{p(z,x,y;G|S_1\neq x)}_{=\gamma_x+\gamma_z\cdot p(z,x,y;G)} \cdot \underbrace{\Pr[S_1\neq x]}_{=\frac{\deg(z;G)-1}{\deg(z;G)}} \\ &= \frac{1}{\deg(z)} + (\gamma_x + \gamma_z \cdot p(z,x,y;G)) \cdot \frac{\deg(z;G)-1}{\deg(z;G)} \\ &= \frac{1+(\deg(z;G)-1)\cdot\gamma_x}{1+(\deg(z;G)-1)\cdot(1-\gamma_z)} && \text{(just rearranging)} \\ &\geq \frac{(\deg(z;G)-1)\cdot\gamma_x}{(\deg(z;G)-1)\cdot(\gamma_x+\gamma_y)} \\ &= \frac{\gamma_x}{\gamma_x+\gamma_y} \\ &= p(z,x,y;H) \end{aligned}$$

And we are done!