Network Science: Principles and Applications

CS 695 - Fall 2016

Amarda Shehu, Fei Li

Department of Computer Science
George Mason University

Random walks and Markov chain

Definition (Random walk)

Let $G = (V, E)$ be a connected, *undirected graph*. A *random walk* on G, starting from vertex $s \in V$, is a random process to follow an outgoing edge chosen uniformly at random. A Markov chain is similar, except the outgoing edge is chosen according to an arbitrary distribution.

Algorithm 1 Random walk

```
$u := s$.
for $i = 1$ to $T$ do
    choose a neighbor $v$ of $u$, uniformly at random;
    $u := v$;
end for
```

1. 1D line
2. **2D grid/graph**
3. 3D grid/graph
4. . . .

Exercise

Let G be the complete graph K_n on n vertices. Let u and v be two vertices in G. Prove that:

1. The expected number of steps in a simple random walk that begins at u and ends upon first reaching v is $n-1$.
2. The expected number of steps to visit all vertices in G starting from u is $(n-1)H_{n-1}$, where $H_{n-1} = \sum_{j=1}^{n-1} 1/j$ is the Harmonic number.

Example (Betting game)

A player bets \$1, and either loses it or wins an addition dollar with probability $1/2$.
Think of this as a random walk on a line graph, where each node represents the amount of wealth at any point of time.
We can learn about the probability distribution of the amount of money at a given time. We can also ask about the probability of the player running out of money before winning a certain amount, and if that happens, what is the expected amount of time before that happens.

Random walk on a line

Lemma

Consider an infinite random walk on the integer line, starting from 0*. The expected number of times that such a walk visits* 0 *is unbounded.*

Lemma

Consider a random walk on the integer line, starting from 0*. If he walks at* -1 *or* 3*, then this random walk terminates. What is the probability that this walk terminates?*

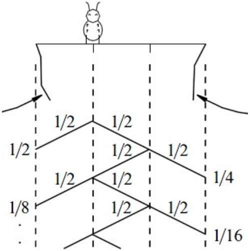

Figure : from "Mathematics for Computer Science" by Tom Leighton et. al, MIT 2010

Problem

Consider a random walk on the integer line, starting from n. The random walk terminates when he stops at 0 or w. What is the probability that this random walk terminates at w?

Proof.

Let R_n be the probability that this walker reaches w. We have

$$\begin{aligned}
R_0 &= 0 \\
R_w &= 1 \\
R_n &= \frac{1}{2}R_{n-1} + \frac{1}{2}R_{n+1} \\
R_{n+1} &= 2R_n - R_{n-1}
\end{aligned}$$

Assume $R_n = a \cdot n + b$. Note $R_0 = a \cdot 0 + b = b = 0$ and $R_w = a \cdot w + b = 1$. We have $a = 1/w$ and $b = 0$ and $R_n = n/w$. □

Problem

Consider a random walk on the integer line, starting from n. The random walk terminates when he stops at 0 *or* w*. What is the expected number of steps for this random walker reaches* 0 *or* w*?*

Proof.

Let X_n be the expected steps. We have

$$\begin{aligned} X_0 &= 0 \\ X_w &= 0 \end{aligned}$$

If he starts somewhere in the middle ($0 < n < w$), we can again break down the analysis into two cases based on his first step:

1. If his first step is to the left, then he lands at position $n-1$ and can expect to move for another X_{n-1} steps.
2. If his first step is to the right, then he lands at position $n+1$ and can expect to move for another X_{n+1} steps.

□

So $X_n = 1 + \frac{1}{2}X_{n-1} + \frac{1}{2}X_{n+1}$. We thus have $X_n = w \cdot n - n^2$.

Summary (random walk on a line)

A gambler goes to Las Vegas with \$$n$ in her pocket. Her plan is to make only \$1 bets and somehow she has found a casino that she will win or lose \$1 on each bet with probability 1/2. She will play until she is broke or she has won \$$m$.

1. The gambler goes home *broke* with probability $\frac{n}{w} = \frac{m}{n+m}$.
2. The gambler goes home a *winner* with probability $\frac{w-n}{w} = \frac{n}{n+w}$.
3. The gambler goes home with probability $\frac{n}{n+m} + \frac{m}{n+m} = 1$.
4. The number of bets before the gambler goes home is expected to be $n(w-n) = n \cdot m$.
5. If the gambler gets greedy and keeps playing until she goes broke, then the gambler eventually goes broke with probability 1, and the number of bets before the gambler goes broke is expected to be infinite.

Quantities associate with random walks

Definition (Stationary distribution, hitting time $H_{u,v}$, cover time G_u, cover time $C(G)$ of G)

Stationary distribution: What is the distribution if we run the random walk for an infinite number of steps? (See the next page.)

Hitting time $H_{u,v}$: the (expected) hitting time from u to v, i.e., the expected number of steps taken by a random walking starting from u to visit v for the first time.

(*Commute time* $C_{u,v}$: the expected time to get from u to v, and back to u.)

Cover time C_u: the (expected) time from u, i.e., the expected number of steps taken by a random walk starting from u to visit *all vertices* of G.

Cover time $C(G) := \max_{u \in V} C_u$.

Stationary distribution

Definition (Transition matrix)

A random walk (or Markov chain), is most conveniently represented by its *transition matrix* P. P is a square matrix denoting the probability of transitioning from any vertex in the graph to any other vertex. Formally,

$$P_{u,v} = \mathrm{Pr}[\text{going from } u \text{ to } v\text{, given that we are at } u].$$

For a random walk, $P_{u,v} = 1/d_u$ if $(u, v) \in E$, and 0 otherwise (where d_u is the degree of u).

Definition (Stationary distribution)

If we have a distribution π over the nodes, we can obtain the distribution after one step by computing $\pi' = P^T \cdot \pi$ $(= (\pi^T \cdot P)^T)$ (P^T is P's transpose). A *stationary distribution* π_s is a distribution with the property that $P^T \cdot \pi_s = \pi_s$.

Remark

Stationary distributions are not always unique, but under certain conditions, they are. It also is the case that under certain conditions, $\lim_{t \to \infty} (P^T)^t \pi = \pi_s$ for all starting distributions π.

Lemma

Consider $G = (V, E)$ *with* $|E| = m$. $P_{v,u} = \frac{d_v}{2m} = 1$ *if* $(v, u) \in E$ *and* 0 *otherwise.* P *is a stationary distribution.*

Proof.

$$
\begin{aligned}
(P^T \cdot \pi)_u &= \sum_v P_{v,u} \cdot \pi_v \\
&= \sum_{v:(v,u)\in E} \frac{d_v}{2m} \cdot \frac{1}{d_v} \\
&= \sum_{v:(v,u)\in E} \frac{1}{2m} \\
&= \pi_u
\end{aligned}
$$

□

Commute time

Lemma

$\forall(u,v) : (u,v) \in E$, we have $C_{u,v} \leq 2m$.

Proof.

(Sketch.) If we view the process as a random walk on sequence of edges, we can bound the commute time by the expected amount of time between consecutive occurrences of the edge $u \rightarrow v$.
The expected length of the gap between consecutive occurrences if we run for t steps is simply t divided by the actual number of times we see the edge $u \rightarrow v$. We also know that since the stationary distribution is uniform, we expect to see the edge $t/(2m)$ times. As t goes to infinity, the actual number of times we see $u \rightarrow v$ approaches its expectation $t/(2m)$ with probability 1 (due to the law of large numbers). We can then approximate the actual number seen by the expected number seen, and thus we expect the length of the gap to be $t/(t/(2m)) = 2m$. □

Cover time

If we had a bound on the commute time for all pairs (u, v) (call this bound x), we could get a bound (in expectation) on the cover time by running the random walk for $x \cdot n$ steps. Unfortunately, the bound given by the previous lemma is only valid for pairs (u, v) where there is an edge between u and v. However, we can still come up with a different method for bounding the cover time.

Lemma

$C(G) \leq 2m(n-1)$.

Proof.

Let T be an arbitrary spanning tree of G. For each edge (u, v) in T, add the edge (v, u). We can then bound the cover time of G with the expected time needed to complete an Euler tour of T. Since each node in T has even degree (due to the doubling of the edges), we know that an Euler tour must exist. If we list the vertices visited as $v_0, v_1, v_2, \dots, v_k = v_0$, we have $C(G) \leq h(v_0, v_1) + h(v_1, v_2) + \cdots + h(v_{k-1}, v_0) = \sum_{(u,v) \in T} (h(u, v) + h(v, u)) = \sum_{(u,v) \in T} C_{u,v} \leq 2m(n - 1)$, since each of the $(n - 1)$ edges that was in T originally shows up in both directions. □

Markov chains

Definition (Irreducible chain)

A Markov chain is irreducible if all states belong to one communicating class.

Definition (Recurrent state)

A *recurrent* state i is *positive recurrent* if $h_{i,i} < \infty$. Otherwise, it is *null recurrent*.

Definition (Periodic, aperiodic state/chain)

A state j in a discrete time Markov chain is *periodic* if there exists an integer $\Delta > 1$ such that $\Pr(X_{t+s} = j | X_t = j) = 0$ unless s is divisible by Δ. A discrete time Markov chain is periodic if any state in the chain is periodic. A state or chain that is not periodic is aperiodic.

Definition (Ergodic state/chain)

An aperiodic, positive recurrent state is an *ergodic* state. A Markov chain is ergodic if all its states are ergodic.

Theorem

Any finite, irreducible, and ergodic Markov chain has the following properties:

1. *the chain has a unique stationary distribution* $\bar{\pi} = (\pi_0, \pi_1, \ldots, \pi_n)$*;*
2. *for all* j *and* i*, the limit* $\lim_{t \to \infty} P^t_{j,i}$ *exists and it is independent of* j*;*
3. $\pi_i = \lim_{t \to \infty} P^t_{j,i} = \frac{1}{h_{i,i}}$.

Application: an *s*-*t* connectivity algorithm

Algorithm 2 *s*-*t* Connectivity Algorithm

```
Start a random walk from $s$.
if the walk reaches $t$ within $2n^3$ steps then
    return there is a path;
else
    return there is no path.
end if
```

Theorem

The s-t connectivity algorithm returns the correct answer with probability $1/2$ *and it only errs by returning that there is no path from s to t when there is such a path.*

Proof.

1. The expected time to reach t from s (if there is a path) is bounded from the cover time of their shared component, which is $2 \cdot n \cdot m < n^3$.
2. By Markov's inequality, the probability that a walk takes more than $4n^3$ steps to reach t from s is at most $1/2$.

□

Application: the Internet

View the Web as a Markov chain. Take a random walk on the web viewed as a directed graph with an edge corresponding to each hypertext link and rank pages according to their stationary probability.

1. Assume there is a vertex with no out edges.
   When the walk encounters this vertex, the walk disappears.
2. Assume a vertex or a strongly connected component with no in edges is never reached.
   Introduce a random restart condition.

Application: PageRank

The intuition behind PageRank starts with simple voting based on in-links, and refines it using the Principle of Repeated Improvement. In particular, the Principle is applied here by having nodes repeatedly pass endorsements across their out-going links, with the weight of a node's endorsement based on the current estimate of its PageRank: nodes that are currently viewed as more important get to make stronger endorsements.

Algorithm 3 PageRank by Jon Kleinberg

```
In a network with $n$ nodes, assign all nodes the same initial PageRank, set to be $1/n$.
Choose a number of steps $k$.
Perform a sequence of $k$ updates to the PageRank values, using the following rule for each update:
    • *Basic PageRank Update Rule*:
      Each page divides its current PageRank equally across its out-going links, and passes these equal shares to the pages it points to. (If a page has no out-going links, it passes all its current PageRank to itself.)
      Each page updates its new PageRank to be the sum of the shares it receives.
```

Notice that the total PageRank in the network will remain constant as we apply these steps.

Consider someone who is randomly browsing a network of Web pages. They start by choosing a page at random, picking each page with equal probability. They then follow links for a sequence of k steps: in each step, they pick a random out-going link from their current page, and follow it to where it leads. (If their current page has no out-going links, they just stay where they are.)

Claim

The probability of being at a page X after k steps of this random walk is precisely the PageRank of X after k applications of the *Basic PageRank Update Rule*.