Application of random walks in designing an s-t connectivity algorithm

Abhinav Chaudhary
ES18BTECH11026

Karan Bhukar
CS18BTECH11021

Raj Patil
CS18BTECH11039

Tanishq Bakliwal
ES18BTECH11020

Vedant Singh
CS18BTECH11047

1 Overview

We aim to concisely present the application of random walks on undirected graphs to test for connectivity between two vertices. We first set out to outline the prerequisites to understand the same and finally state the randomized algorithm vis-à-vis its deterministic counterparts.

2 Prerequisites

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

Definition 2.1. *A random walk on G is a Markov chain defined by the sequence of moves of a particle between vertices of G. In this process, the place of the particle at a given time step is the state of the system. If the particle is at vertex i and if i has d(i) outgoing edges , then the probability that the particle follows edge(i,j) and moves to a neighbor j is 1/d(i).*

Definition 2.2. *A stationary distribution (also called an equilibrium distribution) of a Markov chain and is a probability distribution $\bar{\pi}$ (over the states of the chain) such that*

$$\bar{\pi} = \bar{\pi}\boldsymbol{P}$$

where $\boldsymbol{P}$ is the transition matrix corresponding to the Markov chain with $\boldsymbol{P}_{i,j}$ representing the transition probability from state i to state j

Theorem 2.3. *Any finite, irreducible, and ergodic Markov chain has the following properties:*

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

From this theorem we get 2 interpretations of $\bar{\pi}$. The first is that if we ran for long enough initial state of the chain is almost forgotten, and eventually the probability of the chain being in state i will converge to π_i. The other interpretation is that π_i is equal to the inverse of $h_{i,i}$, which is the expected number of steps/time taken by the chain whose initial state is i to return to state i. This can be understood intuitively as well, since if the expected steps to return back to i is $h_{i,i}$, we expect the chain to be in state i for $1/h_{i,i}$ of the time, and thus, in the limit, we have $\pi_i = 1/h_{i,i}$.

We will not be getting into the proof of this theorem and will be using this as a black box.

Lemma 2.4. *A random walk on an undirected and connected graph* G *is aperiodic if and only if* G *is not bipartite.*

Proof. As we know, a graph is bipartite if and only if it does not have any cycles with odd number of edges. So if a graph is not bipartite, this implies that there exists at least one cycle in it with odd number of edges. Now any vertex in this cycle will have an odd length path back to itself. Also any vertex in the graph has a path back to itself of length 2 since the graph is undirected and connected. So the states in the Markov chain corresponding to the vertices in the odd length cycle are aperiodic as gcd of 2 and an odd number is 1. So there is atleast one aperiodic state in the Markov chain. Thus the random walk is aperiodic if G is not bipartite.
For the other direction, consider the vertex corresponding to an aperiodic state. Let t be the time step at which the chain is in this aperiodic state, so X_t is aperiodic. Now $Pr(X_{t+2} = j|X_t = j) \neq 0$ as there exists a path of length 2 back to any vertex. So there must exist an odd integer s such that $Pr(X_{t+s} = j|X_t = j) \neq 0$. This corresponds to a path of odd length back to the corresponding vertex in the graph which is possible only if there exists an odd length cycle in the graph containing the vertex. This implies that G can not be bipartite if the the random walk is aperiodic which concludes the proof. □

From here on, we will be assuming that G is not bipartite.

Theorem 2.5. *A random walk on G converges to a stationary distribution $\bar{\pi}$, where*

$$\pi_v = \frac{d(v)}{2|E|}$$

Proof. As we have assumed that G is not bipartite, it will be aperiodic and thus ergodic. Now being finite, irreducible and ergodic, from Theorem 2.3, we know it will have a unique stationary distribution $\bar{\pi} = (\pi_0, \pi_1, ..., \pi_n)$.
Now as we know $\sum_{v \in V} d(v) = 2|E|$. Also by definition $\sum_{v \in V} \pi_v = 1$. Therefore we can say

$$\sum_{v \in V} \pi_v = \sum_{v \in V} \frac{d(v)}{2|E|} = 1$$

Now, let $\mathbf{P}$ represent the transition probability matrix of the Markov chain and let $N(u)$ represent the set of neighbours of vertex v.
Using the relation $\bar{\pi} = \bar{\pi} P$, we get

$$\pi_v = \sum_{u \in N(v)} \frac{d(u)}{2|E|} \frac{1}{d(u)} = \frac{d(v)}{2|E|}$$

as $\mathbf{P}$ has value $\frac{1}{d(u)}$ for the $(u, v)^{th}$ element for any $u \in N(v)$ and 0 if $u \notin N(v)$. □

As we know, $h_{u,v}$ represent the expected number of steps taken to reach u from v. From theorem 2.3, as $\pi_i = 1/h_{i,i}$, we get the the following corollary.

Corollary 2.6. *For any vertex in G:*

$$h_{u,u} = \frac{2|E|}{d(u)}$$

Lemma 2.7. *If $(u, v) \in E$, $h_{v,u} < 2|E|$*

Proof. Let $N(u)$ represent the set of neighbours of vertex u in G. Now as we have seen in Corollary 2.6:

$$h_{u,u} = \frac{2|E|}{d(u)}$$

Now if we consider a vertex $w \in N(u)$, we can say that at the beginning of any random walk from u, it will jump to w with probability $\frac{1}{d(u)}$. Therefore we can write

$$h_{u,u} = \frac{1}{d(u)} \sum_{w \in N(u)} (1 + h_{w,u})$$

On equating these two equations, we get,

$$2|E| = \sum_{w \in N(u)} (1 + h_{w,u})$$

Now as $h_{w,u} >= 0$ for any $w \in N(u)$, we can conclude that for any edge $(u, v) \in E$, $h_{v,u} < 2|E|$ □

Definition 2.8. *The cover time of a graph $G = (V, E)$ is the maximum over all vertices $v \in V$ of the expected time to visit all the nodes in the graph by a random walk starting from v*

Lemma 2.9. *The cover time of G = (V,E) is bounded above by $4|V| \cdot |E|$*

Proof. Considering a spanning tree of G i.e. $G' = (V', E')$ where $V' = V$ and $|E'| = |V| - 1$ which is an acyclic graph covering all the vertices of G. Now, the consider the vertex traversal order encountered when doing a depth first search of the spanning tree. Note that we traverse each edge twice and end back at the root vertex in the end. This corresponds to a cyclic Eulerian tour of an augmented graph of the spanning tree (decomposing each undirected edge into two directed edges in the opposite direction) with the traversal order identifiable by the $2|V| - 2$ vertices visited (each edge was traversed twice), say $v_0, v_1, v_2 \ldots, v_{2|V|-2}$ with $v_{2|V|-2} = v_0$. Observe that the cover time of G will be bounded above by the expected number of steps for the aforementioned traversal quantifiable as:

$$\mathbb{E}\left[\sum_{i=0}^{2|V|-3} H_{v_i, v_{i+1}}\right]$$

here H is a random variable corresponding to the number of steps required to reach from a vertex to the next. Now, by linearity of expectation we have:

$$\sum_{i=0}^{2|V|-3} h_{v_i, v_{i+1}} < (2|V| - 2) \cdot 2|E| < 4|V| \cdot |E|$$

$$\because \mathbb{E}\left[H_{v_i, v_{i+1}}\right] = h_{v_i, v_{i+1}}$$

where we leverage lemma 2.7 for the intermediate inequality □

3 Application : s-t Connectivity algorithm

Drawing parallels between the concerned undirected graph and a Markov chain as established in the previous section, we can now devise a randomized algorithm.

Given an undirected graph $G = (V, E)$ and two vertices s, t in G. Let $n \triangleq |V|$ and $m \triangleq |E|$. Our objective is to find if there exists a path connecting s and t. Deterministic algorithms like depth or breadth first search require $\Omega(n)$ memory as they need to keep track of the visited vertices. Here, we only need $O(log(n))$ bits of extra memory as we store only the index of the current vertex and a counter of the number of steps, which is bounded by $4n^3$ and therefore also only requires $O(log(n))$ bits. Note that this is the space requirement over the basics which include storing the graph and a randomization mechanism.

s-t connectivity algorithm

1. `Start a random walk from` s
2. `If the walk reaches` t `within` $4n^3$ `steps,`
   `output that a path exists.`
   `Otherwise, report the contrary.`

3.1 Analysis

For convenience, we will be assuming that G has no bipartite connected components, so that we can directly use the results from Theorem 2.5. However, the analysis can be generalized with some adjustments that we will not be delving into.

Theorem 3.1. *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. If no such path exists from s to t, then the algorithm will always return the correct answer as it will not find a path within $4n^3$ steps. This algorithm only errs if there exists a path which it has not found within $4n^3$ steps. Let Z denote the number of steps taken to find the path, and let T_{exp} denote the expected time to reach from s to t. Assuming that a path exists, by the definition of cover time, T_{exp} will be bounded by the cover time of their shared component. As we know from Lemma 2.9, the cover time is bounded by $4nm$, also m can be at most $\binom{n}{2}$. Therefore $T_{exp} <= 2n^3$.

Now applying Markov's inequality,

$$P(Z > 4n^3) <= \frac{T_{exp}}{4n^3} <= \frac{1}{2}$$

$\square$

Therefore we get a Monte Carlo algorithm with a one sided error, and the probability of a wrong answer is bounded by 1/2.