Learning from Networks

Graph Analytics: Node-Level

Fabio Vandin

October 23rd, 2024

Pagerank Centrality

Intuition: a node is central if an agent navigating the graph *frequently visits* the node.

What is a model for an agent visiting the graph?

An agent performing a *random walk*.

To make things more interesting, let assume that G is a *directed* graph.

Random Walks

What is a random walk?

Consider a starting node i.

Let $\mathcal{N}(j) = \{k : (j,k) \in E\}$ be the neighbors of node j in G.

Algorithm `RandomWalk`(i)
Input: node $i \in V$ of graph $G = (V, E)$
Output: random walk starting from i
$v^{(0)} \leftarrow i$;
for $t \leftarrow 1, 2, \ldots$ **do**
    $v^{(t)} \leftarrow$ random vertex chosen uniformly at random from $\mathcal{N}\left(v^{(t-1)}\right)$;
return;

PageRank Centrality and Random Walks

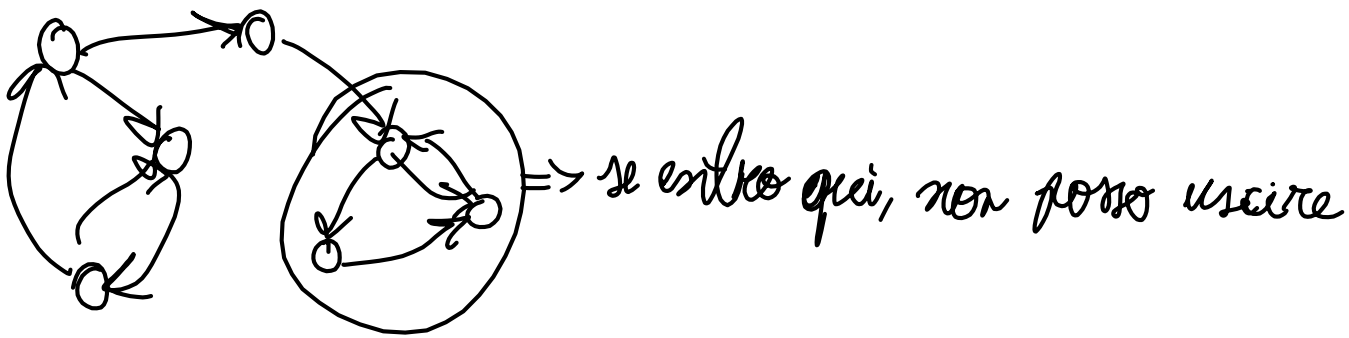
Definition

The *PageRank* centrality $PR(j)$ of a node j in a directed graph G is the fraction of times that a random walk on G visits node j.

Is the above well defined? NO

Does $PR(j)$ depends on the starting node (i) of the random walk?

What if the random walks gets stuck somewhere in G?



Random Walks and Markov Chains

Definition

A *Markov chain* is a collection of random variables $X_0, X_1, X_2, \ldots$ with the property:

$$\Pr[X_t = a_t | X_{t-1} = a_{t-1}, X_{t-2} = a_{t-2}, X_{t-3} = a_{t-3}, \ldots X_0 = a_0]$$

$$= \Pr[X_t = a_t | X_{t-1} = a_{t-1}]$$

The above is the *Markov property*.

Let $P_{i,j} = \Pr[X_t = j | X_{t-1} = i]$ be the *transition probability* from state i to state j.

The Markov property implies that a Markov chain is uniquely defined by the (one-step) *transition matrix* $\mathbf{P}$ with entries $\mathbf{P}_{i,j}$ equal to the transition probabilities.

Observation

Let $\mathbf{p}^{(t)} \in \mathbb{R}^d$ be a vector with $\mathbf{p}_i^{(t)} = \Pr[X_t = i]$. Then $\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)}\mathbf{P}$

The following establishes the connection between random walks and Markov chains.

Observation

A random walk is a Markov chain with transition matrix $\mathbf{P}_{i,j} = \frac{1}{|\mathcal{N}(i)|}$ if $j \in \mathcal{N}(i)$, and $\mathbf{P}_{i,j} = 0$ otherwise.

The questions we had before on random walks can then be translated in questions about the corresponding Markov chain...

by first observing that the PageRank $PR(j)$ corresponds to the probability $\mathbf{p}_j^{(t)} = \Pr[X_t = j]$ for $t \to \infty$

Does $\mathbf{p}^{(t)}$ for $t \to \infty$ exist? Does it depend on the starting node i?

Stationary Distribution of a Markov Chain

Definition

The stationary distribution π of a Markov chain with transition matrix $\mathbf{P}$ is a probability distribution (on the states of the Markov chain) such that:

$$\pi = \pi \mathbf{P}$$

When does a Markov chain have a unique stationary distribution?

Theorem

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

1. the chain has a unique stationary distribution π
2. for all j and i, the limit $\lim_{t\to\infty} \mathbf{P}_{j,i}^{(t)}$ exists and is independent of j
3. $\pi_i = \lim_{t\to\infty} \mathbf{P}_{j,i}^{(t)}$

- *finite*: X_t takes values from a finite set SÍ
- *irreducible*: its graph representation is *strongly connected* (every vertex i is reachable from any vertex j) NON necessariamente
- *ergodic*: for every node: 1) once it is visited it will eventually be visited again, and 2) the time of visit of the node is not *periodic* (e.g., the node is not visited only on even times)

NON necessariamente

Are the above properties satisfied by real world graphs?

The actual PageRank random walk

To make the corresponding Markov chain irreducible and ergodic, the random walk is modified as follows in PageRank.

Algorithm `PageRankRandomWalk`(i, α)
Input: node $i \in V$ of graph $G = (V, E)$; $\alpha \in (0, 1)$
Output: random walk starting from i
$v^{(0)} \leftarrow i$;
for $t \leftarrow 1, 2, \ldots$ **do**
  with probability α: $v^{(t)} \leftarrow$ random vertex chosen uniformly at random from $\mathcal{N}\left(v^{(t-1)}\right)$;
  otherwise: $v^{(t)} \leftarrow$ random vertex chosen uniformly at random from V;
return;

Idea: the random walker at some points jumps to a random page...

α is called the *damping factor*

da walk avremo $\pi_i = PR(i)$

PageRank Centrality: Computation

How do we compute the PageRank for all vertices in a graph G?

Let $\mathbf{p}_j^{(t)}$ be the probability that `PageRankRandomWalk`(i, α) is in node j at time t. Then we have:

$$\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)} \left(\alpha \mathbf{P} + (1-\alpha)\frac{1}{n}\mathbf{E} \right)$$

transition matrix

where $\mathbf{E}$ is a matrix with all entries equal to 1.

Let $\mathbf{p}$ be the vector where $\mathbf{p}_i = PR(i)$ is the PageRank of node i. Since $\mathbf{p}$ is the stationary distribution of the modified PageRank random walk we have

$$\mathbf{p} = \mathbf{p} \left(\alpha \mathbf{P} + (1-\alpha)\frac{1}{n}\mathbf{E} \right)$$

eigenvector

Therefore $\mathbf{p}$ can be computed with several methods:

1. find the eigenvector corresponding to eigenvalue 1 for the matrix $\alpha\mathbf{P} + (1-\alpha)\frac{1}{n}\mathbf{E}$
2. use the power method to compute $\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)}\left(\alpha\mathbf{P} + (1-\alpha)\frac{1}{n}\mathbf{E}\right)$ for a large enough t

   approssimazione $\uparrow$ $=$
3. ran several random walks starting from various nodes...
4. ...

Other Centrality Measures

There are several other centrality measures often implemented in graph analytics libraries:

- PageRank centrality
- harmonic centrality
- ...

Some useful libraries:

- https://graph-tool.skewed.de/
- https://networkx.org/
- https://networkit.github.io/
- https://snap.stanford.edu/

Comparison on Zachary's Karate Club Data

See jupyter notebook.