Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

DPRank centrality: Finding important vertices based on random walks with a new defined transition matrix

Min Liu [a], Zhen Xiong [a], Yue Ma [a], Peng Zhang [a], Jianliang Wu [b], Xingqin Qi [a,*]

[a] School of Mathematics and Statistics, Shandong University (Weihai), Weihai, 264209, China
[b] School of Mathematics, Shandong University, Jinan, 250100, China

HIGHLIGHTS

- A new vertex centrality method named DPRank is presented by introducing a new transition matrix.
- In DPRank the probability of the vertex i jumps to its neighbor j is proportional to the degree (or out-degree) of the vertex j, instead of the reciprocal of the degree of vertex i as in PageRank.
- DPRank centrality method takes a bigger environment around a node into account.

ARTICLE INFO

Article history:
Received 26 October 2016
Received in revised form 21 October 2017
Accepted 24 October 2017
Available online xxxx

Keywords:
Centrality
Random walk
Transition matrix
Network

ABSTRACT

The vertices centrality, as an indicator, aims to find important vertices within a network (undirected or directed). It is a crucial issue in social network analysis to find important vertices, which has significant applications in diverse domains. PageRank is the most known algorithm to rank vertices in a directed network, where a random walker always selects next arriving node from its neighborhood uniformly. But in the real world, a selection or transition is more likely to have "tendentiousness". Thus in this paper, we propose a new nodes centrality mechanism taking "tendentiousness" into consideration. The main idea is that, instead of selecting next node uniformly from its neighbors, a "far-sighted" random walker prefers to move to a neighbor with greater degree (or out-degree for directed network, respectively), so that the information can be spread rapidly and will not be trapped by dangling nodes (without outgoing arcs). This new centrality method is thus called Degree-Preferential PageRank centrality, short for DPRank centrality. One can see that, DPRank centrality method gives more accurate evaluation of a node's ability by taking not only the immediate local environment around it but also the bigger environment (i.e., its neighbor's neighbors) into consideration. This new DPRank centrality method performs very well when applying it on several data sets including directed and undirected networks. It gives a new perspective of evaluating a node importance, and is expected to have a promising application in the future.

1. Introduction

Social networks permeate social and economic lives and play a central role in measuring relationships between individuals, groups, organizations even countries, with vertices representing individuals or organizations and edges representing relationships between vertices. In network analysis, the *centrality* of a node describes its relative importance in a network. Measuring the centrality of vertices has been a crucial issue in social network analysis and has significant use in diverse domains, including disease spread, transmission of information and logistic distribution. Applications include identifying the most influential person(s) in a social network, the key infrastructure nodes in urban networks, or the super-spreaders of diseases.

Several centrality methods have been proposed, for example: degree centrality, closeness centrality [1], betweenness centrality [2,3] and so on. *Degree centrality* is defined as the number of edges incident upon a vertex. Its simplicity and low computing complexity are advantages. However, degree centrality has some limitations, such as: the measure does not take the global structure of the graph into consideration. *Closeness centrality* [1] is defined as the inverse of the sum of shortest distances to all other vertices from a focal vertex i, i.e., $\frac{n-1}{\sum_j d_{ij}}$, where d_{ij} is the shortest distance between the vertex i and the vertex j reachable from the vertex i. Closeness can be treated as a measure of how efficiently it exchanges information with others in a graph. A main limitation of closeness is the lack of applicability to graphs with disconnected

* Corresponding author.
E-mail address: qixingqin@sdu.edu.cn (X. Qi).

https://doi.org/10.1016/j.future.2017.10.036
0167-739X/

components: two nodes that belong to different components do not have a finite distance between them. Thus, for disconnected graphs, a popular approach is to calculate the closeness centrality in terms of the inverse of the harmonic mean distances between the nodes, i.e., $\frac{1}{n-1}\sum_{j\neq i}\frac{1}{d_{ij}}$, where $\frac{1}{\infty}=0$. *Betweenness centrality* [2] of a vertex i is equal to the number of all shortest paths that pass through the vertex i. It was introduced as a measure for quantifying the control of a human on the communication among other humans in a social network by Linton Freeman. The betweenness centrality of a vertex v can be represented as: $\sum_{s\neq v\neq t\in V}\frac{\sigma_{st}(v)}{\sigma_{st}}$, where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v. The betweenness centrality may be normalized by dividing it by the total number of pairs of vertices not including v, which for directed graphs is $(n-1)(n-2)$ and for undirected graphs is $\frac{(n-1)(n-2)}{2}$. Although the betweenness centrality takes the global graph structure into consideration and can be applied to graphs with disconnected components, it has limitations. For example, a great proportion of nodes in a graph generally do not lie in a shortest path between any two other nodes, and therefore receive the same betweenness centrality score 0, thus we cannot judge which one is more central than others. In 2016, Lü et al. [4] gave an intensive survey for vital nodes identification in complex network, which gives a systematic review for this problem.

There are a few algorithms have been proposed for directed graphs. PageRank centrality method [5] is the most popular and famous one. Starting from a vertex, a walker randomly and uniformly selects next node he will move to among its neighbors if exists, and repeats the process. This process can be specified by a matrix P called *transition matrix*, whose element P_{ij} denotes the probability of transiting from the vertex i to the vertex j in a given step.

Any probability distribution on a graph G of n nodes can be represented by a row vector $\overrightarrow{\pi}=(\pi_1,\pi_2,\ldots,\pi_n)$ with $\sum_{i=1}^{n}\pi_i=1$, where the ith entry captures the distribution residing at node i. If the probability of a walker to stay on i does not change at time $t\to\infty$, this random walk is said to have *stationary distribution*. In other words, the stationary distribution does not change over time and describes the probability that a walker stays at a specific node after a sufficiently long time. The stationary distribution $\overrightarrow{\pi}$ is then specified by $\overrightarrow{\pi}=\overrightarrow{\pi}\cdot P$, i.e.,

$$P^T\cdot\overrightarrow{\pi}^T=\overrightarrow{\pi}^T \tag{1}$$

where $\overrightarrow{\pi}^T$ is the eigenvector of P^T corresponding to the eigenvalue 1. The stationary distribution $\overrightarrow{\pi}$ is usually used as a metric of importance of vertices.

Definition 1. A non-negative square matrix P is called **primitive** if there is an integer k such that all the entries of P^k are positive. It is called **irreducible** if for any i,j, there is a $k=k(i,j)$ such that $(P^k)_{ij}>0$.

Theorem 1 (*Perron–Frobenius Theorem [6]*)**.** *Suppose P is an irreducible non-negative square matrix, then*

(1) *The spectral radius $\rho=\rho(P)$ of P is a positive* ***real*** *number and it is an eigenvalue of the matrix P, which is called the Perron–Frobenius eigenvalue.*

(2) *The Perron–Frobenius eigenvalue ρ is simple. Both right and left eigenspaces associated with ρ are one-dimensional, and P has a right (or left) eigenvector corresponding to eigenvalue ρ whose components are all* ***positive****.*

Based on Theorem 1, the connectedness of an undirected graph or a strongly connected digraph guarantees that the transition matrix P is irreducible and stochastic, meaning that the stationary state of a random walk for such graphs always exists.[1] But a general directed graph is not always strongly connected, so its transition matrix P is not irreducible, and hence the existence of the stationary distribution is not guaranteed. Page and Brin [5] introduced a special transition matrix which is primitive[2] (and hence irreducible), and successfully used it as an important tool to rank web pages. The (i,j)-entry of the transition matrix can be interpreted as the probability of a walker jumping from i to j . If $k_i^{out}\neq 0$, the probability or the (i,j)-entry is $\alpha\frac{A_{i,j}}{k_i^{out}}+(1-\alpha)\frac{1}{n}$, and if $k_i^{out}=0$, the probability or the (i,j)-entry is $\frac{1}{n}$. In other words, the PR value of the vertex i at t step is:

$$PR_i(t)=\alpha\cdot\sum_{j=1}^{n}A_{ji}\frac{PR_j(t-1)}{k_j^{out}}+(1-\alpha)\frac{1}{n} \tag{2}$$

where A is the adjacency matrix and k_j^{out} is the number of arcs going out from the vertex j. The higher the value of damping factors α, the more accurately the topological structure will be preserved.

To improve the accuracy of PageRank, many variants have been presented. For example, LeaderRank [7] and Pro-PageRank [8], which will be illustrated detailedly in next section. Liu et al. [9] developed an improved algorithm with the residence time of website added in the original algorithm mechanism. Atish Das Sarma et al. [10] presented fast random walk-based distributed algorithms for computing PageRank in general graphs and proved strong bounds on the round complexity. Luo et al. [11] proposed Time-Weighted PageRank, extending PageRank by introducing a time decaying factor. An improved PageRank algorithm based on time feedback and topic similarity was proposed by Yang et al. [12] in 2016.

In this paper, we propose a new node centrality mechanism. Compared with PageRank and its variants where a random walker selects next arriving node uniformly from its neighbors, the random walker in our strategy is "far-sighted" and prefers to select the neighbor with greater degree (or greater out-degree for directed graphs, similarly hereinafter) as the arriving node of next step, so that the information can be spread rapidly and not trapped by dangling nodes (without outgoing arcs). In other words, the probability that the vertex i moves to its neighbor j is proportional to the degree (or out-degree) of the vertex j. This new centrality method is thus called Degree-Preferential PageRank centrality, or simply DPRank centrality. We will present the DPRank centrality method detailedly in Section 2, and test it on several data sets (including directed or undirected networks) to show its utilities in Section 3, then make further quantitative analysis in Section 4; Conclusions are made in Section 5.

2. Degree-preferential PageRank centrality

Let $G=(V,E)$ be an undirected graph with the vertex set $V(G)=\{v_1,v_2,\ldots,v_n\}$, and edge set $E=\{(v_i,v_j)\}$. Its adjacency matrix A is an $n\times n$ matrix where $A_{ij}=1$ if there is an edge between the vertex v_i and the vertex v_j, otherwise $A_{ij}=0$. Similarly, if $G=(V,E)$ is a directed graph, its adjacency matrix A is an $n\times n$ matrix where $A_{ij}=1$ if there is an arc from v_i to v_j, otherwise $A_{ij}=0$. Clearly, the adjacency matrix of an undirected graph is symmetric, while it is asymmetric for a directed graph.

In an undirected graph, the *degree* of the vertex v_i is the number of edges incident to it, which is denoted by k_i. The *neighborhood* of v_i is the set of vertices adjacent to it, which is denoted by $N(v_i)$. For a directed graph, the *out-neighborhood* of the vertex v_i is the

[1] Actually $\pi_i=\frac{k_i}{2M}$, where M is the sum of edges in the graph.

[2] For a primitive matrix, we can use the power method to calculate its largest eigenvalue and corresponding eigenvector.

Please cite this article in press as: M. Liu, et al., DPRank centrality: Finding important vertices based on random walks with a new defined transition matrix, Future Generation Computer Systems (2017), https://doi.org/10.1016/j.future.2017.10.036.