Applications to network analysis: Eigenvector centrality indices Lecture notes

Dario Fasino, University of Udine (Italy)

Lecture notes for the second part of the course "Nonnegative and spectral matrix theory with applications to network analysis", held within the Rome-Moscow school on Matrix Methods and Applied Linear Algebra, August–September 2014. The author is partially supported by Istituto Nazionale di Alta Matematica. This document is intended for internal circulation only.

1 A brief introduction to the analysis of complex networks

Complex networks is a common name for various real networks which are usually presented by graphs with a large number of vertices. Here belong Internet graphs, phone graphs, e-mail graphs, social networks and many other. The term *network analysis* refers to a wealth of mathematical techniques aiming at describing the structure, function, and evolution, of complex networks.

- One of the main tasks in network analysis is the localization of nodes that, in some sense, are the "most important" in a given graph. The main tool to quantify the relevance of nodes in a graph is through the computation of suitably defined *centrality indices*.

  Many centrality indices have been invented during time. Each one of them refers to a particular definition of "importance" or "relevance" that is most useful in a given context.

- *Graph partitioning* is the problem of dividing the vertices of a graph into a given number of disjoint subsets of given sizes such that the total weight of edges between such sets is minimized. The best known example of a graph partitioning problem is the problem of dividing a graph into two subsets of comparable size, such that the number of edges between them is minimized.

- *Community detection* differs from graph partitioning in that the number and size of the subsets into which the network is divided are generally not apriori specified. Instead it is assumed that the graph is intrinsically structured into communities or groups of vertices which are more or less evidently delimited, the aim being to reveal the presence and the consistency of such groups.

1.1 Notations and definitions

Two graphs $G = (V, E)$ and $G' = (V, E')$ are called *isomorphic* if there exists a permutation matrix P such that $A_{G'} = PA_GP^T$.

Let $A = A_G$. The *in-degree* and the *out-degree* of node i are respectively the numbers

$$d_i^{\text{in}} = \sum_{j=1}^{n} A_{ij}, \qquad d_i^{\text{out}} = \sum_{j=1}^{n} A_{ji}.$$

They represent the number (or overall weight) of edges that arrive to or depart from node i, respectively. If G is not oriented the two numbers are the same and their common value is the *degree* d_i.

Let $V = \{1, \dots, n\}$ and let $\mathcal{G}_n$ be the set of all graphs whose node set is V. A *graph invariant* is any function $f : \mathcal{G}_n \mapsto \mathbb{R}$ which is invariant under graph isomorphisms: If $A'_G = PA_GP^T$ then

$f(G) = f(G')$. A *centrality index* is any function $c : \mathcal{G}_n \mapsto \mathbb{R}^n$ such that if $A'_G = PA_GP^T$ then $Pc(G) = c(G')$.

The degree vectors $d^{\text{in}} = A\mathbf{1}$ and $d^{\text{out}} = A^T\mathbf{1}$ are the most simple (somehow trivial) examples of a centrality index. Clearly, $\mathbf{1}^T d^{\text{out}} = \mathbf{1}^T d^{\text{out}}$, and the sum is equal to the total edge weight of G, which is a graph invariant called *volume*. Many interesting graph invariants and centrality indices are based on linear algebraic properties (in particular, eigenpairs) of A_G and variations thereof.

Remark 1.1. *Let $G \in \mathcal{G}_n$ and let $A = A_G$. Let $c : \mathcal{G}_n \mapsto \mathbb{R}^n$ be a centrality index. Suppose that there exists a nontrivial permutation matrix P such that $A = P^T AP$ (that is, the graph owns a nontrivial automorphism). Then for all $i = 1 \ldots, n$ the centrality index of node i is the same as that of node j, where $e_j = Pe_i$.*

2 Eigenvector centralities

The purpose of this section is to describe some of the most important centrality indices, whose definition is largely based on tools and concepts borrowed from linear algebra and matrix analysis. They are the Bonacich index, PageRank, and HITS scores. The common feature shared by these indices is that they are Perron eigenvectors of suitably defined nonnegative matrices.

2.1 The Bonacich index

One of the first centrality indices (besides degrees) was introduced by the american sociologist Bonacich [1]. Let G be a directed graph. For simplicity, assume it is strongly connected, and let $A = A_G$. The original idea is that *a node is important it is connected to other important nodes.* This sort of circular defintion can be formalized rigorously by assuming that the score of node i is proportional to the sum of scores of all nodes j such that $i \to j$:

$$\lambda b_i = \sum_{j:j\to i} b_j = \sum_{j=1}^{n} A_{ji} b_j,$$

where $A = A_G$. Hence, the vector of Bonacich indices fulfills the eigenvalue equation $A^T b = \lambda b$. Among the possible solutions of the previous equation, the Bonacich index is the one which corresponds to the Perron eigenpair of A^T. In fact, if G is strongly connected then the Bonacich score vector $b = (b_1, \ldots, b_n)^T$ immediately obtains a number of useful properties from Perron–Frobenius theory:

- It is uniquely defined, apart of a scaling factor; its entries are positive (every node gets a nonzero score).
- If we add a new edge $i \to j$ to G then the node whose Bonacich index receives the largest increase is i (by Dietzenbacher theorem), consistently to the idea that its influence has increased.

Remark 2.1. *Recall that the number $(A^k)_{ij}$ represents the number of distinct paths from node j to node i whose length is k. If A is primitive then we can compute the Bonacich index by applying the (normalized) power method to A^T. Hence, apart of a scaling factor, $b = \lim_{k\to\infty}(A^T)^k\mathbf{1}/\|(A^T)^k\mathbf{1}\|$. In particular, b_i is proportional to limit for $k \to \infty$ of the number of different paths whose length is k from node i to every node in the graph. Thus a node with a large Bonacich index is a node that originates many long paths.*

Exercise 2.2. A *star graph* with n nodes is the undirected graph whose adjacency matrix is

$$A = \begin{pmatrix} 0 & 1 \cdots 1 \\ 1 & 0 \cdots 0 \\ \vdots & \vdots \quad \vdots \\ 1 & 0 \cdots 0 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Prove that the spectrum of A is $\{-\rho, 0, \rho\}$ and a Perron eigenvector of A (which contains the Bonacich centrality index of the nodes) is $x = (\rho, 1, \ldots, 1)^T$ where $\rho = \sqrt{n-1}$.

2.2 PageRank

One of the best known centrality indices for arbitrary graphs is PageRank, whose fortune is due to its usage in the Google search engine [5].

The original formula by S. Brin and L. Page [2] defines the PageRank vector $\pi = (\pi_1, \dots, \pi_n)^T$ of a graph G as the solution of the following linear system:

$$\pi_i = (1-\alpha) + \alpha \sum_{j:j\to i} \frac{\pi_j}{d_j^{\text{out}}},$$

where $\alpha \in (0,1)$ is a fixed constant called the *damping factor*, originally set to $\alpha = 0.85$. In matrix form,

$$(I - \alpha M)\pi = (1-\alpha)\mathbf{1}, \tag{1}$$

where $M \geq O$ is the so-called *link matrix* which is defined as

$$M_{ij} = \begin{cases} A_{ij}/d_j^{\text{out}} & \text{if } d_j^{\text{out}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

and $A = A_G$.

For simplicity of exposition, hereafter I assume that all nodes in G have at least one outgoing edge, that is, $d^{\text{out}} > 0$. In this case, the sum of all entries in any column of M is 1. Unfortunately, M is seldom irreducible. Nevertheless, we can say something about $\rho(M)$:

Lemma 2.3. $\rho(M) = 1$.

Proof. Since $M \geq O$, there exists $x \geq 0$ such that $Mx = \rho(M)x$. Moreover, we can rewrite $\sum_i M_{ij} = 1$ as $M^T\mathbf{1} = \mathbf{1}$. Hence, $\mathbf{1}^T x = \mathbf{1}^T Mx = \rho(M)\mathbf{1}^T x$, and the proof is complete, by observing that $\mathbf{1}^T x > 0$. ∎

The surprise is that there exists a matrix $\Gamma > O$ (which is called *Google matrix*) such that, apart of a scaling factor, π is a Perron–Frobenius eigenvector of Γ.

Theorem 2.4. *Let $\Gamma \in \mathbb{R}^{n\times n}$ be the positive matrix defined as*

$$\Gamma = \alpha M + \frac{1-\alpha}{n}\mathbf{1}\mathbf{1}^T.$$

for $0 < \alpha < 1$. Then, $\rho(\Gamma) = 1$ and the vector π defined in (1) is a Perron eigenvector.

Proof. Observe that $\Gamma > O$ by construction; in particular, it is primitive, hence irreducible. Simple computations show that $\Gamma^T\mathbf{1} = \mathbf{1}$, so that $\rho(\Gamma) = 1$.

Finally, let x be a Perron eigenvector of Γ normalized so that $\mathbf{1}^T x = n$. Then,

$$x = \Gamma x = \alpha M x + \frac{1-\alpha}{n}\mathbf{1}\mathbf{1}^T x = \alpha M x + (1-\alpha)\mathbf{1}.$$

Rearranging terms, $(I - \alpha M)x = (1-\alpha)\mathbf{1}$, which is (1).

To complete the proof it remains to prove that $I - \alpha M$ is nonsingular. Using Lemma 2.3 we can derive that all eigenvalues of $I - \alpha M$ are contained in the circle $\{z \in \mathbb{C} : |1 - z| \leq \alpha\}$, which excludes 0 since $\alpha < 1$ by hypothesis. ∎

2.3 Hubs and Authorities

Exactly in the same year Brin and Page invented PageRank, J. Kleinberg introduced another algorithm to evaluate the relevance of documents in a large hypertext, such as the internet [3, 5]. This algorithm (HITS, *Hypertext Induced Topic Search*) quantifies the importance of nodes in a graph according to two centrality indices: the *hub score* and the *authority score*.

Very informally, the hub score of a node is a measure of how good it is as "access point" or "portal", while the authority score is a measure of how good a node is as "final document". Kleinberg original idea is that *a node is a good hub if it points to good authorities; and a node is a good authority if it is pointed by good hubs.*

This concept has been formalized by the following equations: Let h_i and a_i be the hub score and authority score of node i, respectively. Then,

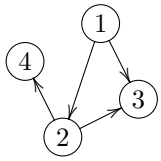
$$\lambda h_i = \sum_{j:i\to j} a_j \qquad \lambda a_i = \sum_{j:j\to i} h_j, \qquad i = 1\ldots, n, \tag{2}$$

where λ is a proportionality constant, to be defined. In matrix notations, $\lambda h = Aa$ and $\lambda a = A^T h$. The two equations can be uncoupled as follows:

$$\lambda^2 h = A^T A h, \qquad \lambda^2 a = A A^T a.$$

Let $M_{\text{hub}} = A^T A$ and $M_{\text{auth}} = AA^T$ be the hub matrix and the authority matrix, respectively. These two matrices are symmetric, nonnegative, positive semidefinite, and have exactly the same eigenvalues (why?). In particular, $\rho(M_{\text{hub}}) = \rho(M_{\text{auth}})$. The preferred solution to (2) corresponds to Perron eigenvectors, with $\lambda = \sqrt{\rho(M_{\text{hub}})}$. Unfortunately, M_{hub} and M_{auth} are usually not irreducible, even if the original graph is strongly connected.

As an exercise, let's compute hub and authority scores of the following graph:



The adjacency and hub matrix are

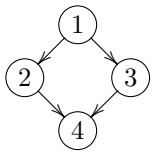
$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \qquad M_{\text{hub}} = A^T A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The eigenvalues of M_{hub} are $3, 1, 0, 0$. An eigenvector associated to $\rho(M_{\text{hub}}) = 3$ is $h = (1, 1, 0, 0)^T$. We can compute authority scores from the formula $\lambda a = Ah$. We obtain $a = (0, 1, 2, 1)^T/\sqrt{3}$. We conclude that nodes 1 and 2 are good hubs, nodes 3 and 4 are not (indeed, they have no outgoing links). The best authority node is 3, which is pointed by both best hubs. Node 1 is not an authority, because it has no ingoing links.

Exercise 2.5. Suppose that in a given graph G there are two nodes, say i and j such that for every $k \in V$ it holds $k \to i \Rightarrow k \to j$. Prove that $a_i \le a_j$.

HITS algorithm is essentially the power method (with normalization) applied to M_{hub} and M_{auth} starting from the initial vector $\mathbf{1}$ [3]. However, the largest eigenvalue of these matrices may be not simple, and this fact implies that hub scores and authority scores may be not uniquely defined (apart of a scaling factor), since the convergence of the power method may be affected by the choiche of the starting vector.

For example, consider



$$\rightsquigarrow \qquad M_{\text{hub}} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The eigenvalues of M_{hub} are $2, 2, 0, 0$. Any vector of the form $h = (\alpha\ \beta\ \beta\ 0)^T$ is an eigenvector corresponding to $\rho(M_{\text{hub}}) = 2$. If we apply the power method to M_{hub} starting from $(1\ 1\ 1\ 1)^T$ we obtain $h \sim (1\ 1\ 1\ 0)^T$, while if the starting vector is $(1\ 0\ 0\ 0)^T$ then we obtain $h \sim (1\ 0\ 0\ 0)^T$.

Various modifications have been devised in order to make hub-authority scores well defined under rather general hypotheses, see e.g., [4]. One of these tricks is described in the following exercise:

Exercise 2.6. Let $\widehat{A} = A + \varepsilon I$ where $\varepsilon > 0$ (note: this modification corresponds to adding a loop with weight ε to every node in the graph) and let $\widehat{M}_{\text{hub}} = \widehat{A}^T \widehat{A}$. Prove that if G is not disconnected[1] then $\widehat{M}_{\text{hub}}$ is irreducible.

Hint: $\widehat{M}_{\text{hub}} = \varepsilon(A + A^T) +$ *other nonnegative matrices.*

3 Exercises and problems

Exercises marked with a star ($\star$) are requested for the final evaluation.

1. ($\star$) For any given integers $p, q \geq 1$, let G be the graph having $1 + p + pq$ nodes, which is defined as follows:

   - There is a root node, which is connected to p star nodes;
   - every star node is connected to q peripheral nodes.

   Compute the Bonachich centrality index for the nodes of this graph.

   Use the preceding result to prove that the Bonacich index of a node is not an inccreasing function of its degree.

2. ($\star$) Let $G = (V, E)$ be an undirected, connected graph. Consider the following centrality indices for both nodes and edges: To any $i \in V$ and $ij \in E$ associate variables q_i and e_{ij} respectively, by means of these equations:

$$\lambda q_i = \sum_{j : i \sim j} e_{ij}, \qquad \lambda e_{ij} = q_i + q_j.$$

   where λ is a constant (depending on G) to be determined.

   Discuss existence, uniqueness, positivity... (open ended problem).

References

[1] Phillip Bonacich. Factoring and Weighting Approaches to Status Scores and Clique Identification. *Journal of Mathematical Sociology*, 2 (1972), 113–120.

[2] Sergey Brin, Larry Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine.* Proceedings of the 7th international conference on World Wide Web (WWW). Brisbane, Australia (1998), 107–117.

[3] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46 (1999), 604–632.

This is the seminal paper on HITS. Preliminary versions circulated in 1997, see Kleinberg homepage, `http://www.cs.cornell.edu/home/kleinber/`

[4] A. Farahat, T. Lofaro, J. Miller, G. C. Rae, L. A. Ward. Authority rankings from HITS, PageRank, and SALSA: existence, uniqueness, and effect of initialization. *SIAM J. Sci. Comput.* 27 (2006), 1181–1201.

[5] A. N. Langville, C. D. Meyer. A survey of eigenvector methods for Web information retrieval. *SIAM Rev.*, 47(2005), 135–161.

Expository paper on PageRank and HITS.

`Last update: September 6, 2014`

[1] A graph is *disconnected* if its vertex set can be partitioned into two subsets, $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \varnothing$, so that no edge belongs to $(V_1 \times V_2) \cup (V_2 \times V_1)$.