Web-graphs. Lecture 6.

Degree centrality

- Network centrality is some method of nodes ranking: which are the most important or central vertices in a network?
- The answer depends on the meaning of the "importance". Each definition of importance leads to special centrality type.
- The simplest centrality measure in a network is just the **degree** of a vertex, the number of edges connected to it $\longrightarrow$ **Degree centrality**
- For directed networks $\longrightarrow$ **In-degree and out-degree centrality**
- Although degree centrality is a simple centrality measure, it can be very illuminating.
- Social networks: individuals who have connections to many others might have more influence, more access to information or more prestige than those who have fewer connections.
- Citation network: counts in the evaluation of scientific papers. The number of citations a paper receives from other papers, which is simply its in-degree in the citation network, gives a crude measure of whether the paper has been influential or not and is widely used as a metric for judging the impact of scientific research.

Eigenvector centrality

- We can think of degree centrality as awarding one "centrality point" for every network neighbor a vertex has.
- But neighbors are different
- Natural assumption: **vertex's importance in a network is increased by having connections to other vertices that are themselves important.**
- Instead of awarding vertices just one point for each neighbor, **eigenvector centrality gives each vertex a score proportional to the sum of the scores of its neighbors**

Eigenvector centrality

- Initial guess on the node i's centrality: $x_i^{(0)} = 1,\ \forall i$
- We start iterative procedure:

$$x_i^{(k)} = \sum_{j=1}^{N} A_{ij} x_j^{(k-1)},$$

where N in the number of nodes, $\mathbf{A} = \{A_{ij}\}_{i,j=1,\ldots,N}$ is the adjacency matrix. In matrix notation:

$$\mathbf{x}^{(k)} = \mathbf{A}\mathbf{x}^{(k-1)} = \mathbf{A}^k \mathbf{x}^{(0)},$$

where $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_N^{(k)})'$.

- Let us write $x^{(0)}$ as a linear combination of the eigenvectors $\mathbf{v}_i$ of the adjacency matrix $\mathbf{A}$ thus:

$$\mathbf{x}^{(0)} = \sum_i c_i \mathbf{v}_i,$$

for some appropriate choice of constants c_i. Then

$$\mathbf{x}^{(k)} = \mathbf{A}^k \sum_i c_i \mathbf{v}_i = \sum_i c_i \kappa_i^k \mathbf{v}_i = \kappa_1^k \sum_i c_i \left(\frac{\kappa_i}{\kappa_1}\right)^k \mathbf{v}_i,$$

where κ_i is the ith eigenvector of $\mathbf{A}$ and κ_1 **is the maximal one**.

Eigenvector centrality

- While $k \to \infty$ $\mathbf{x}^{(k)} \to c_1 \kappa_1^k \mathbf{v}_1$
- **Centrality = ranking, we are not interested in normalization constants**
- In other words, the limiting vector of centralities is simply proportional to the leading eigenvector of the adjacency matrix.
- Equivalently we could say that the centrality $\mathbf{x}$ satisfies

$$\mathbf{A}\mathbf{x} = \kappa_1 \mathbf{x}$$

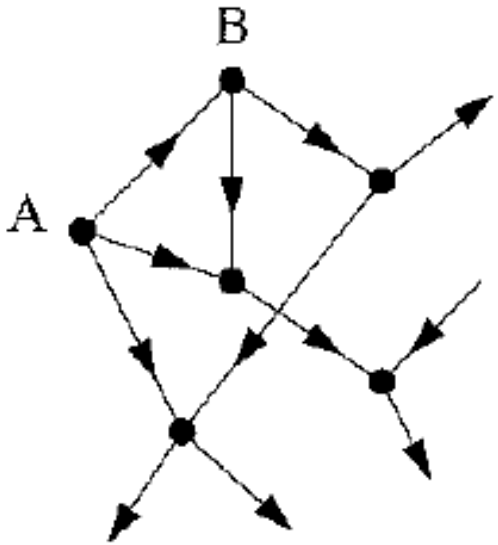
- As promised the centrality x_i of vertex i is proportional to the sum of the centralities of i's neighbors:

$$x_i = \kappa_1^{-1} \sum_j A_{ij} x_j$$

- which gives the eigenvector centrality the nice property that it can be large either because a vertex has many neighbors or because it has important neighbors (or both)

Eigenvector centrality

- In theory eigenvector centrality can be calculated for either undirected or directed networks.
- It works best for the undirected case.
- In the directed case other complications arise.
  - We should be careful with definition of the rank spread. Does the importance spreads by incoming or outgoing links?
  - Vertex A in the figure is connected to the rest of the network, but has only outgoing edges and no incoming ones. Such a vertex will always have centrality zero. This might not seem to be a problem: perhaps a vertex that no one points to should have centrality zero. But then consider vertex B, which has one ingoing edge, but that edge originates at vertex A, and hence B also has centrality zero, because the one term in its sum is zero. **Only nodes in SCC or its Out component can have nonzero eigenvector centralities.**



Katz centrality

- Katz centrality is the variation of eigenvector centrality which addresses the problems for directed networks.
- Idea: give each vertex a small amount of centrality "for free", regardless of its position in the network or the centrality of its neighbors.
- Formally

$$x_i = \alpha \sum_j A_{ij} x_j + \beta,$$

  where α and β are positive constants
- By adding this second term, even vertices with zero in-degree still get centrality β, and once they have a non-zero centrality, then the vertices they point to derive some advantage from being pointed to.
- In matrix notation

$$\mathbf{x} = \alpha \mathbf{A}\mathbf{x} + \beta \mathbf{1},$$

  where $\mathbf{1} = (1, \ldots, 1)'$.
- Resolving the last equation on $\mathbf{x}$ we obtain

$$\mathbf{x} = \beta(\mathbf{I} - \alpha \mathbf{A})^{-1}\mathbf{1}$$

- We can put $\beta = 1$ (or any other normalization)

Katz centrality

- The Katz centrality differs from ordinary eigenvector centrality in the important respect of having a free parameter α, which governs the balance between the eigenvector term and the constant term
- α cannot be arbitrary large, as its value preserves matrix inversion
- Critical point: $\det(\mathbf{I} - \alpha\mathbf{A}) = 0 \iff \det(\mathbf{A} - \alpha^{-1}\mathbf{I}) = 0$
- We see that it is simply the characteristic equation whose roots α^{-1} are equal to the eigenvalues of the adjacency matrix $\mathbf{A}$
- Critical α: $\alpha = \frac{1}{\kappa_1}$
- Katz centrality converges when we choose $\alpha < \frac{1}{\kappa_1}$

Katz centrality

- In practice, for large networks it is more efficient to calculate Katz centrality iteratively. For example, let $x_i^{(0)} = 0, \ \forall i$

$$\mathbf{x}^{(k)} = \alpha \mathbf{A} \mathbf{x}^{(k-1)} + \beta \mathbf{1}$$

- Katz centrality can be used for undirected networks too
- The idea of adding a constant term to the centrality so that each vertex gets some weight just by virtue of existing is a natural one. It allows a vertex that has many neighbors to have high centrality regardless of whether those neighbors themselves have high centrality, and this could be desirable in some applications.
- A possible extension of the Katz centrality is to consider cases in which the additive constant term is not the same for all vertices

$$x_i = \alpha \sum_j A_{ij} x_j + \beta_i,$$

  where β_i is some intrinsic, non-network contribution to the centrality for each node.
- For example, in a social network the importance of an individual might depend on non-network factors such as their age or income and if we had information about these factors we could incorporate it into the values of the β_i.

Katz centrality

- The Katz centrality has one feature that can be undesirable. If a vertex with high Katz centrality points to many others then those others also get high centrality.
- A high-centrality vertex pointing to one million others gives all one million of them **the same high centrality**.
- Example
  - *Yahoo!* web directory might contain a link to my web page, but it also has links to millions of other pages.
  - *Yahoo!* is an important website, and would have high centrality by any sensible measure
  - Should I therefore be considered very important by association?
  - Most people would say not: the high centrality of *Yahoo!* will get diluted and its contribution to the centrality of my page should be small because my page is only one of millions.

PageRank

- We can modify the Katz centrality by introduction of an assumption that the centrality node derive from its network neighbors is proportional to their centrality *divided by their out-degree* $\longrightarrow$ **PageRank centrality** – the trade name given it by the Google web search corporation, which uses it as a central part of their web ranking technology.
- Formal expression

$$x_i = \alpha \sum_{j:d_j^{out}>0} A_{ij} \frac{x_j}{d_j^{out}} + \beta,$$

  where d_j^{out} is out-degree of the node j.
- For nodes with $d_j^{out} = 0$ we will assume that $d_j^{out} = 1$ (or any other positive value) and define $\mathbf{D} = \text{diag}(d_1^{out}, \ldots, d_N^{out})$

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{D}^{-1} \mathbf{x} + \beta \mathbf{1}$$

- Resolving the last equation we obtain

$$\mathbf{x} = (\mathbf{I} - \alpha \mathbf{A} \mathbf{D}^{-1})^{-1} \mathbf{1}$$

- As with the Katz centrality, the formula for PageRank contains one free parameter α, whose value must be chosen somehow before the algorithm can be used.

PageRank

- By analogy with Katz centrality, the critical value for α is determined by the largest eigenvalue of $\mathbf{AD}^{-1}$.
- For an undirected network this largest eigenvalue turns out to be 1 and the corresponding eigenvector is $(d_1, \ldots, d_N)$, where d_i is node i's degree. Therefore, $\alpha < 1$.
- For directed networks in general the leading eigenvalue will be different from 1, although in practical cases it is usually still roughly of order 1.
- The traditional choice is $\alpha = 0.85$
- As with the Katz centrality we can generalize PageRank to the case where the additive constant term is different for different vertices:

$$x_i = \alpha \sum_{j:d_j^{out}>0} A_{ij} \frac{x_j}{d_j^{out}} + \beta_i,$$

PageRank: definition using random walks

- Let us assume that some person randomly surfs through the web-pages
- With probability α on each step he make transition using random link from the current page
- With probability $1 - \alpha$ he goes to any random web-page
- $1 - \alpha$ probability of teleportation
- PageRank centrality $=$ probability to appear in the web-page π_i

$$\pi_i = \alpha \sum_{j \to i} \frac{\pi_j}{d_j^{out}} + \frac{1-\alpha}{N} + \frac{\alpha}{N} \sum_{j : d_j^{out} = 0} \pi_j$$

Hubs & Authorities

- In the case of **directed networks**, there is another method to measure centrality.
- **Citation network**: a paper such as a review article may cite other articles that are authoritative sources for information on a particular subject. The review itself may contain relatively little information on the subject, but it tells us where to find the information, and this on its own makes the review useful, even if there are small number of other papers which cite the review.
- **Web-graphs:** are many examples of web pages that consist primarily of links to other pages on a given topic or topics and such a page of links could be very useful even if it does not itself contain explicit information on the topic in question.
- Thus there are really two types of important node in these networks: **authorities** are nodes that contain useful information on a topic of interest; **hubs** are nodes that tell us where the best authorities are to be found.
- An authority may also be a hub, and vice versa: review articles often contain useful discussions of the topic at hand as well as citations to other discussions.
- Hubs and authorities only exist in directed networks.

Hubs & Authorities

- Hyperlink-induced topic search (HITS)
- Each node has two types of centrality:
  - authority centrality – x_i
  - hybs centrality – y_i
- Idea: nodes with high authority centrality are the ones that connected with the nodes having high hubs centrality and vice versa.
- Formally

$$x_i = \alpha \sum_j A_{ij} y_j$$

$$y_i = \beta \sum_j A_{ji} x_j$$

- In the matrix form

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{y}$$

$$\mathbf{y} = \beta \mathbf{A}^T \mathbf{x}$$

Hubs & Authorities

- Combining the two expressions we obtain

$$\mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda\mathbf{x}$$
$$\mathbf{A}^T\mathbf{A}\mathbf{y} = \lambda\mathbf{y}$$

  where $\lambda = (\alpha\beta)^{-1}$
- Thus the authority and hub centralities are respectively given by eigenvectors of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$
- Let us proof that $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ have the same eigenvalues.
  - Let $\mathbf{x}$ is eigenvector of $\mathbf{A}\mathbf{A}^T \longrightarrow \mathbf{A}\mathbf{A}^T\mathbf{x} = \lambda\mathbf{x}$
  - $\mathbf{A}^T\mathbf{A}(\mathbf{A}^T\mathbf{x}) = \lambda\mathbf{A}^T\mathbf{x}$
  - $\mathbf{A}^T\mathbf{x}$ is eigenvector of $\mathbf{A}^T\mathbf{A}$ with the same eigenvalue λ
- $\mathbf{y} = \mathbf{A}^T\mathbf{x}$

Hubs & Authorities

- Introduction of the pair of hub and authority centralities instead of one eigenvector centrality is another way to resolve the problems that ordinary eigenvector centrality has with directed networks, that vertices outside of strongly connected components or their out-components always have centrality zero.
- In the hubs and authorities approach vertices not cited by any others have authority centrality zero (which is reasonable), but they can still have non-zero hub centrality. And the vertices that they cite can then have non-zero authority centrality by virtue of being cited.
- This is perhaps a more elegant solution to the problems of eigenvector centrality in directed networks than the more ad hoc method of introducing an additive constant term.
- We can still introduce such a constant term into the HITS algorithm if we wish, or employ any of the other variations considered previously, such as normalizing vertex centralities by the degrees of the vertices that point to them.

Closeness centrality

- Closeness centrality measures the mean distance from a vertex to other vertices.
- $d(u, v)$ is the shortest path length between u and v
- Node v's closeness centrality is defined as the value reciprocal to the sum of shortest paths to other vertices in the connected component:

$$c(v) = \frac{1}{\sum_{u \in V, u \neq v} d(u, v)}$$