Network Analysis and Modeling
Lecture 3
Prof. Aaron Clauset

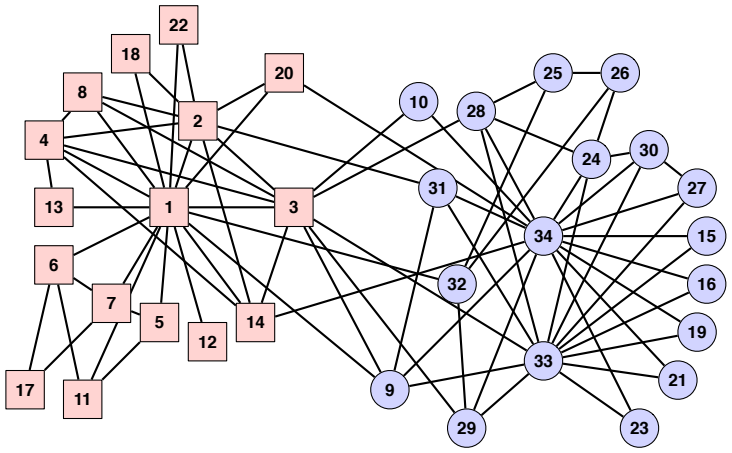
CSCI 5352
3 September 2013

1 Which vertices are important?

A common question when analyzing the structure of a network is which vertices are more or less important? This is not yet a well-defined question, and thus how we answer it depends greatly on what we mean by *important*. There are several general classes of answers. One is to define importance in terms of structural features in the network, e.g., high-degree vertices or being in the "middle" of the network, while another is to define importance in terms of some kind of dynamical process, e.g., vertices where random walkers tend to accumulate.

Measures of vertex importance are often called *centrality* measures, in which more central vertices are more important, and less central ones less important. Here, we will cover a few of the more commonly used forms of centrality measures. However, every centrality measure comes with a set of assumptions, and thus it is crucial to consider whether or not those assumptions fit with the system or question of interest. In most cases, a centrality measure can be calculated, but in some of those cases, its values may be meaningless or misleading.

As a running example for these centrality measures, we will apply each to a single network: the popular Zachary's karate club network,[1] which represents the social network of friendships between 34 members of a karate club at a US university in the 1970s. During the course of Zachary's study, the club split into two factions, centered around two leaders in the club.

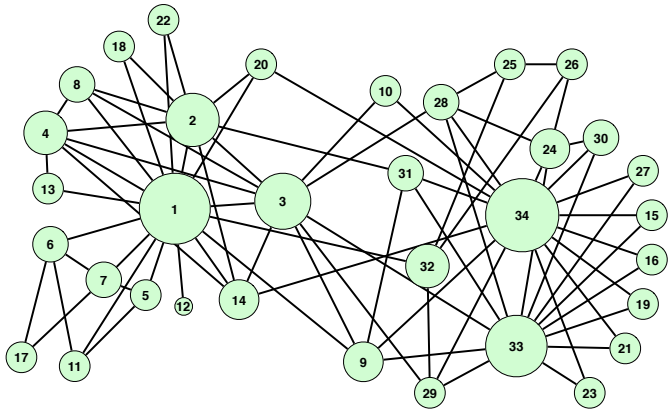


[1] From W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research* **33**, 452–473 (1977).

1.1 Centrality by degree

The simplest measure of importance is the degree of a vertex k_i, i.e., the number of edges that terminate or originate at i.[2] The idea is that vertices with larger degrees exert greater effect on the network, and thus identifying the most connected vertices is a useful way to identify these important vertices. In many situations, these highly-connected vertices play special roles in both large-scale organization of the network and in the dynamics of network processes. We will return to these ideas later in greater detail.

This figure shows the karate club network in which the area of each vertex's circle is proportional to its degree in the network, and the table lists the degree k and normalized degree k/m for each vertex.



group 1	1	2	3	4	5	6	7	8	11	12	13	14	17	18	20	22		
k	16	9	10	6	3	4	4	4	3	1	2	5	2	2	3	2		
k/m	0.10	0.06	0.06	0.04	0.02	0.03	0.03	0.03	0.02	0.01	0.01	0.03	0.01	0.01	0.02	0.01		
group 2	9	10	15	16	19	21	23	24	25	26	27	28	29	30	31	32	33	34
k	5	2	2	2	2	2	2	5	3	3	2	4	3	4	4	6	12	17
k/m	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.02	0.02	0.01	0.03	0.02	0.03	0.03	0.04	0.08	0.11

1.2 Centrality from eigenvectors

The degree of a vertex captures only a local measure of importance, and a natural generalization of degree centrality is to increase the importance of vertices who are connected to other high-degree vertices. That is, not all neighbors are equal and we may want a vertex's importance to be larger if it is connected to other important vertices. *Eigenvector centrality* accounts for these differences by assigning a vertex an importance score that is proportional to the importance scores of its neighbors.

There are several ways to formalize this recursive notion of importance, and each approach produces slightly different final scores. However, they are all forms of eigenvector centrality because they can be calculated as the principal eigenvector[3] for a particular eigenvalue problem—the differences lay in how we set up that problem. Here we will cover eigenvector centrality (as defined by Bonacich)

[2]The degree of a vertex is sometimes called *degree centrality* in the sociology literature, in order to emphasize its use as a centrality measure.

[3]The eigenvector associated with the largest (most positive) eigenvalue.

and PageRank. Other popular versions are the Katz centrality and hub/authority scores, which we will not cover here.

Eigenvector centrality
Taking the recursive idea about importance at face value, we may write down the following equation:

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)} \ , \tag{1}$$

where A_{ij} is an element of the adjacency matrix (and thus selects contributions to i's importance based on whether i and j are connected), and with the initial condition $x_i^{(0)} = 1$ for all i.

This formulation is a model in which each vertex "votes" for the importance of its neighbors by transferring some of its importance to them. By iterating the equation, with the iteration number indexed by t, importance flows across the network. However, this equation by itself will not produce useful importance estimates because the values x_i increase with t. But, absolute values are not of interest themselves, and relative values may be derived by normalizing at any (or every) step.[4]

Applying this method to the karate club for different choices of t yields the following table. Notice that by the $t = 15$th iteration, the vector x has essentially stopped changing, indicating convergence on a fixed point. (Convergence here is particularly fast in part because the network has a small diameter.)

vertex	$x^{(1)}$	$x^{(5)}$	$x^{(10)}$	$x^{(15)}$	$x^{(20)}$	degree, k
1	0.103	0.076	0.071	0.071	0.071	16
2	0.058	0.055	0.053	0.053	0.053	9
3	0.064	0.065	0.064	0.064	0.064	10
4	0.038	0.043	0.042	0.042	0.042	6
5	0.019	0.015	0.015	0.015	0.015	3
6	0.026	0.016	0.016	0.016	0.016	4
7	0.026	0.016	0.016	0.016	0.016	4
8	0.026	0.034	0.034	0.034	0.034	4
9	0.032	0.044	0.046	0.046	0.046	5
10	0.013	0.020	0.021	0.021	0.021	2
11	0.019	0.015	0.015	0.015	0.015	3
12	0.006	0.010	0.011	0.011	0.011	1
13	0.013	0.017	0.017	0.017	0.017	2
14	0.032	0.044	0.046	0.045	0.045	5
15	0.013	0.019	0.021	0.020	0.020	2
16	0.013	0.019	0.021	0.020	0.020	2
17	0.013	0.005	0.005	0.005	0.005	2
18	0.013	0.018	0.019	0.019	0.019	2
19	0.013	0.019	0.021	0.020	0.020	2
20	0.019	0.028	0.030	0.030	0.030	3
21	0.013	0.019	0.021	0.020	0.020	2
22	0.013	0.018	0.019	0.019	0.019	2
23	0.013	0.019	0.021	0.020	0.020	2
24	0.032	0.029	0.030	0.030	0.030	5
25	0.019	0.012	0.011	0.011	0.011	3
26	0.019	0.013	0.012	0.012	0.012	3
27	0.013	0.015	0.015	0.015	0.015	2
28	0.026	0.026	0.027	0.027	0.027	4
29	0.019	0.026	0.026	0.026	0.026	3
30	0.026	0.026	0.027	0.027	0.027	4
31	0.026	0.034	0.035	0.035	0.035	4
32	0.038	0.037	0.039	0.038	0.038	6
33	0.077	0.066	0.062	0.062	0.062	12
34	0.109	0.082	0.074	0.075	0.075	17

Illustrating the close relationship between degree and eigenvector centrality, the centrality scores here are larger among the high-degree vertices, e.g., 1, 34, 33, 3 and 2.

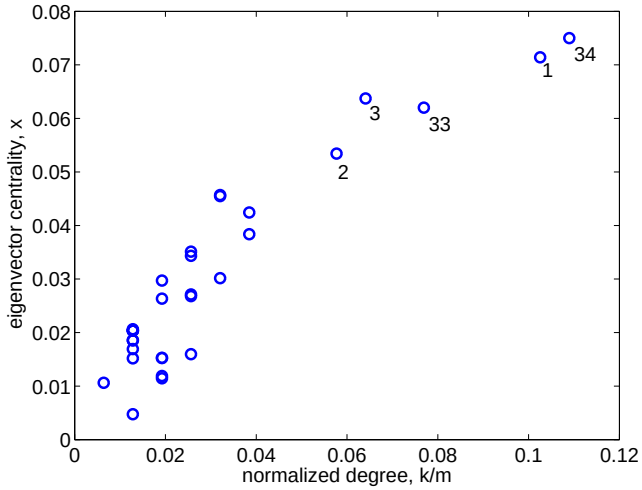
[4]Large values of t will tend to produce overflow errors in most matrix computations, and thus normalizing is a necessary component of a complete calculation.

The Perron-Frobenius theorem from linear algebra guarantees that when the network is an undirected, connected component, iterating Eq. (1) will always converge on a fixed point equivalent to the principal eigenvector of the adjacency matrix.[5] Thus, we can sidestep the iteration completely and formulate the calculation as an eigenvector problem of the form

$$\mathbf{A}\mathbf{x} = \lambda_1 \mathbf{x} \ , \tag{2}$$

where $\mathbf{A}$ is the adjacency matrix, $\mathbf{x}$ is a vector containing the eigenvector centralities, and λ_1 is the largest eigenvalue of $\mathbf{A}$.[6] Computing eigenvector centralities can be done in most modern mathematical computing software, or using common linear algebra libraries via matrix inversion techniques (which take $O(n^3)$ time, but can be as fast as $n^{2.373}$ or even $n^2 \ln n$ depending on some technical details.). Doing so with the karate club network yields exactly the same values we found above, via the iterative approach.

To more clearly illustrate the relationship between eigenvector centrality and degree, we can compare the scores given to vertices under each. This figure shows the very strong correlation between



eigenvector centrality and (normalized) degree centrality in the karate club. In fact, the Pearson correlation coefficient between $\mathbf{x}$ and k/m is $r^2 = 0.84$, indicating that knowing the value of one provides a great deal of information about the value of the other. There are, of course, differences, as the scatter plot shows, and these are related to the way eigenvector centrality allows a vertex's importance to be partly a function of the importance of its neighbors (and its neighbors' neighbors), which is information not included in the degree of a vertex.

[5] The Perron-Frobenius theorem provides the conditions under which this formulation holds: a real and irreducible square matrix with non-negative entries will have a unique largest real eigenvalue and that the corresponding eigenvector has non-negative components. What we are doing by iterating Eq. (1) is the "matrix power method" of computing the principal eigenvector.

[6] Originally given in P. Bonacich, "Power and Centrality: A Family of Measures." *American Journal of Sociology* **92**(5), 1170–1182 (1987).

PageRank

PageRank is another kind of eigenvector centrality,[7] but which has some nicer features than the Bonacich (and Katz) definitions. In particular, the classic eigenvector centrality performs poorly when applied to directed networks. In general, centralities will be zero for all vertices not within a strongly connected component, even if those vertices have high in-degree. Moreover, in an directed acyclic graph, there are no strongly connected components larger than a single vertex, and thus only vertices with no out-going edges ($k_{\text{out}} = 0$) will have non-zero centrality. These are not desirable behaviors for a centrality score.

PageRank solves these problems by adding two features to our vertex voting model. First, it assigns every vertex a small amount of centrality regardless of its network position. This eliminates the problems caused by vertices with zero in-degree—who have no other way of gaining any centrality—and allows them to contribute to the centrality of vertices they link to. As a result, vertices with high in-degree will tend to have higher centrality as a result of being linked to, regardless of whether those neighbors themselves have any links to them. Second, it divides the centrality contribution of a vertex by its out-degree. This eliminates the problematic situation in which a large number of vertex centralities are increased merely because they are pointed to by a single high-centrality vertex.

Mathematically, the addition of these features modifies Eq. (1) to become

$$x_i = \alpha \sum_{j=1}^{n} A_{ij} \frac{x_j}{k_j^{\text{out}}} + \beta \ , \tag{3}$$

where α and β are positive constants. The first term represents the contribution from the classic (Bonacich) eigenvector centrality, while the second is the "free" or uniform centrality that every vertex receives. The value of β is a simple scaling constant and thus by convention we will choose $\beta = 1$; as a result, α alone scales the relative contributions of the eigenvector and uniform centrality components. Further, we must choose a resolution method for the case of $k^{\text{out}} = 0$, which would result in a divide-by-zero in the calculation. This problem is solved by artificially simply setting $k^{\text{out}} = 1$ for each such vertex.

As a matrix formulation, PageRank is equivalent to

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{D}^{-1} \mathbf{x} + \beta \mathbf{1} \tag{4}$$

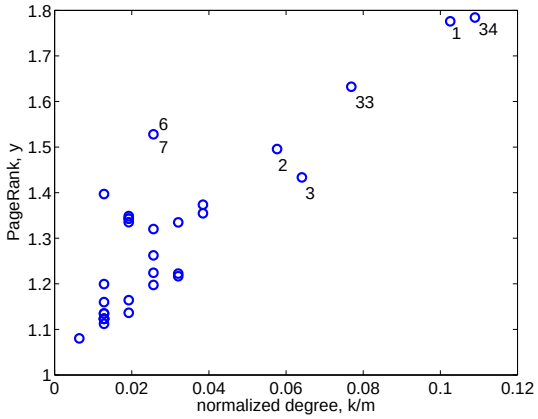
$$= \mathbf{D}(\mathbf{D} - \alpha \mathbf{A})^{-1} \mathbf{1} \ , \tag{5}$$

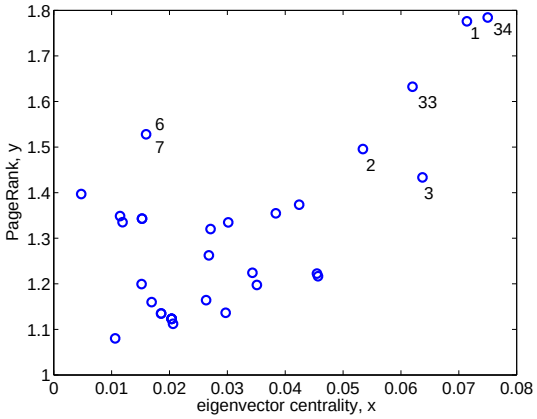
where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \max(k_i^{\text{out}}, 1)$, as described above, and where we have set $\beta = 1$.

[7] PageRank is usually attributed to S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine." *Computer Networks and ISDN Systems* **30**, 107–117 (1998). However, as is often the case with good ideas, it has been reinvented a number of times, and PageRank is, arguably, one of these reinventions. The idea of using eigenvectors in a manner very similar to PageRank goes back as least as far as G. Pinski and F. Narin, "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics." *Information Processing & Management* **12**(5): 297–312 (1976), but may even go back further than that.

How close are PageRank scores to degree and eigenvector centrality? This question depends on the choice of the free parameter α. When $\alpha = 1$, PageRank on an undirected network is mathematically equivalent to degree centrality, but not to eigenvector centrality (because PageRank normalizes voting by out-degree). In the limit of $\alpha \to 0$, only the "uniform" term remains and the contribution from the adjacency matrix goes to zero. In this limit, every centrality score converges on the constant β, which is not useful. A common choice is $\alpha = 0.85$ (see below), but in general there is little principled guidance about how to choose it.

Applied to the karate club network with $\alpha = 0.85$, PageRank is quite close to degree centrality and moderately close to eigenvector centrality, with $r^2 = 0.73$ for PageRank and normalized degree and $r^2 = 0.38$ for PageRank and eigenvector centrality. These figures illustrate the relationships. Perhaps most important, however, the overall ordering of the most important vertices is fairly





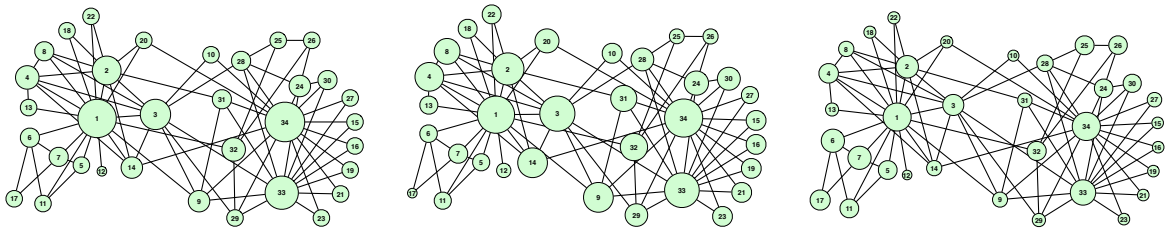
stable, with most of the disagreement between PageRank and eigenvector centrality being on the ordering of lower-importance vertices. If we examine the five most-important vertices under each measure, we see strong agreement on the two most-important vertices. All agree that vertex 33 is either 3rd or 4th, but PageRank chooses vertices 6 and 7 over vertices 2 and 3, for the remaining two slots, illustrating the point that these different measures, while related, are not equivalent.

	degree k/m		eigenvector		PageRank	
1^{st} largest	34	(0.1090)	34	(0.0750)	34	(1.7843)
2^{nd} largest	1	(0.1026)	1	(0.0714)	1	(1.7758)
3^{rd} largest	33	(0.0769)	3	(0.0637)	33	(1.6324)
4^{th} largest	3	(0.0641)	33	(0.0620)	6	(1.5280)
5^{th} largest	2	(0.0577)	2	(0.0534)	7	(1.5280)

And, here are pretty figures showing the full results visually, on the network itself.

PageRank, redux

Google is famous for using PageRank to estimate the importance of pages on the World Wide Web, which is a directed graph. They do not, however, use a large matrix calculation to estimate $\mathbf{x}$, as the size of the Web graph is $n \approx 10^{10}$. Instead, they use a streamlined version of the matrix power



degree eigenvector PageRank

method, in which they directly simulate the voting process. This presents us with an alternative interpretation of the underlying model for PageRank, which is related to random walks on networks.

Note that in the PageRank formulation, we normalize the contribution to i's importance from j by j's out-degree. Rewriting the summand as $x_j \times A_{ij} \Big/ k_j^{\text{out}}$, we observe that we are, in fact, working with a stochastic adjacency matrix, in which each row sums to 1. This is just another name for the transition matrix in a Markov process, which describes the probability that a random walker will move from state j to state i.

That is, PageRank is a first-order Markov model of a random walker on the network structure, in which the probability that the walker will be in state i at the next step depends only on the current state j and the probability of the transition $j \to i$. When a walker visits some vertex j, it chooses a new state uniformly at random from among the neighbors of j. The interpretation of the constant term α in the above formulation is a "teleportation" probability, i.e., with probability α, the random walker follows the Markov processes; otherwise, it chooses a uniformly random vertex to move to.

When α is large, the Markov process dominates, and the random walker tends to walk along the network's edges. When the walker enters a part of the network with few out-going edges, the teleportation probability allows the walk to restart somewhere else. On the World Wide Web, this process is crucial, as the strongly connected component of the web graph is only a modest portion of the entire graph, and Google would not be useful if its "web crawlers" were constantly getting stuck in obscure corners of the graph.

The streamlined matrix power method Google used to calculate PageRank essentially directly simulates these random walkers, having each vertex repeatedly vote for its neighbors in proportion to its current centrality divided by its out-degree.

2 At home

1. Read Chapter 7.1–7.5 (pages 168–181) in *Networks*
2. Next time: geometric centralities