1 Rankings and importance in networks

When analyzing the structure of a network, a common question is

How important is this vertex?

There are two notable things about this question: (i) it implies a model f that defines "importance" according to some as-yet-unspecified assumptions, and (ii) it is a node-level question, and so we use node-level measures of a network's structure to answer it.

Answers often take the form of assigning each node i a score $\theta_i = f(i, G)$, which quantifies the importance of node i within the structure of network G. If θ_i is ordinal, then we might call the list of scores $\vec{\theta}$ a *ranking*; if θ_i is real-valued, we might instead call it an *embedding*. How we define f in practice serves to codify what we mean by "important." For instance, we might want to say that a node is more important if it is located higher up in a competitive hierarchy among all nodes, or if it serves as a conduit for relatively more information flowing across the network, or if it has more influence on the behavior of other nodes, e.g., under a cascade model.

Every definition of f takes as input a graph G and returns a vector $\vec{\theta}$ containing an importance score for every node. Hence, f is a function that projects a network onto a vector in an n-dimensional space, where the ith element of this vector gives the importance of node i. Notably, there are an infinite number of ways to specify a function f.

Importance functions can be divided into two main classes:

- **Structural importance**: θ_i characterizes i's role in the structure of the observed network G.

  These importance functions often utilize the same types of node-level summary statistics (node connectivity, motifs, or position) we saw earlier in the class; this class subsumes most of the popular "centrality" scores from the social sciences, such as betweenness centrality, and many competition-based rankings like Bradley-Terry-Luce.

- **Dynamical importance**: θ_i characterizes i's role in a dynamical process running on top of the observed network G.

  These importance functions measure how "influential" a node is in driving or shaping the evolution of the dynamical process.[1] For instance, a node u might be very important if a change to u's state variable x_u induces state-variable changes in many other nodes, e.g., in network epidemics or information cascades.

[1] The term "influence" is like the term "importance" in that it is not well defined, and the literature on network influence admits both no accepted rigorous definition of influence and no accepted way to rigorously measure it, and hence also little rigorous science. There are some good papers on influence in networks, but they are few and far between.

Conceptually, every choice of f is a dynamical importance, because importance is always defined under a model, even if that model is only implied. Indeed, each of the structural measures that we will explore below is motivated by some kind of dynamical model over a network, e.g., information flowing across edges, or something traveling across a network, or nodes competing with each other, etc. Unfortunately, calculating even very simple measures of dynamical importance can be computationally expensive, and especially so if they require simulation.

Structural importances, in contrast, are often computationally far less expensive. Their popularity and utility stems from being able to use a cheap structural importance to approximate the answer of an expensive dynamical importance. But, the quality of the approximation is an assumption that should be tested, i.e., how well does a particular structural importance function $f'(i, G)$ predict our desired dynamical importance function $f(i, G)$? Depending on the particular measures we choose, a structurally important vertex may not itself be dynamically important, and vice versa.[2]

Within both classes of importance functions, we can further subdivide them as being either

- (supervised) if the function f is parameterized, and we have a downstream prediction task, then we are in the "learning to rank" domain applied to networks, in which we choose the version of f that maximizes our performance on the downstream task.[3]
- (unsupervised) if f is not parameterized, and we instead calculate some function (stochastic or deterministic) purely from the graph structure G.

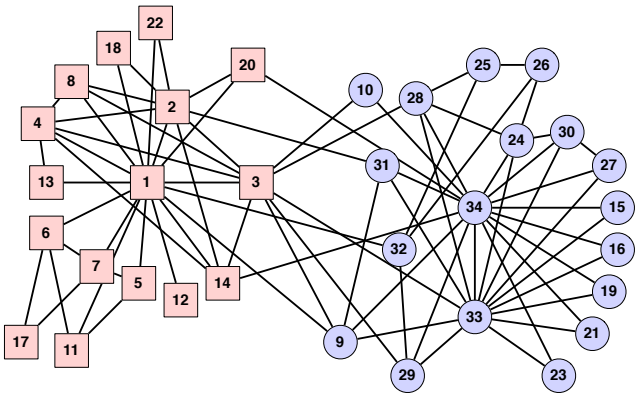
Below, we will survey some of the more commonly used unsupervised measures of structural importance, which serve as the foundation of more complicated measures of dynamical importance, and which are also useful for developing network intuition. Understanding how different measures produce different results on the same network will illustrate key network concepts and measures. Most of these measures were first developed in sociology, and hence are called "centrality" measures, in which more important means more "central" in the network.

As a running example for these centrality measures, we will apply each to a single network: the popular Zachary's karate club network,[4] 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 (nodes 1 and 34). The picture below shows the network and the social partition.

[2]For a discussion of subtleties, see Borgatti, "Centrality and network flow." *Social Networks* **27**, 55–71 (2005).

[3]For example, see Agarwal, "Learning to rank on graphs". *Machine Learning* **81**, 333–357 (2010).

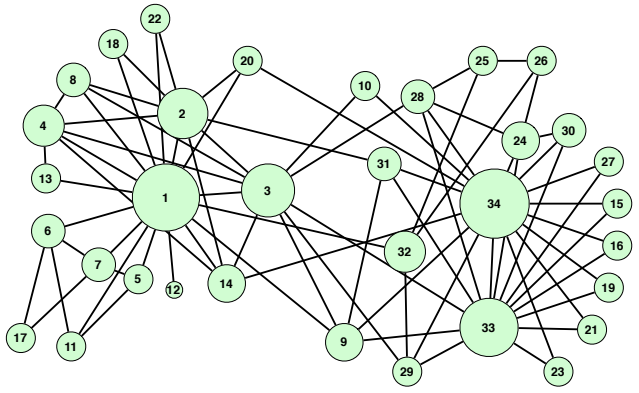
[4]From Zachary, "An information flow model for conflict and fission in small groups." *J. Anthropol. Res.* **33**, 452–473 (1977).



2 Connectivity centralities

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, a measure that is sometimes called *degree centrality* in sociology. The assumed model is one in which vertices exert influence in proportion to the number of other nodes they directly connect to, e.g., in a network SIR model or under the Independent Cascade model. Hence, node degree is a direct way to measure a node's relative importance. That is, influence (importance) is entirely local and direct. (Can you think of a type of influence in a networked system that is not local, or not direct?)

The figure below 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

2.1 Centrality from eigenvectors

Vertex degree counts all neighbors the same, but some neighbors are more important than others, i.e., how much influence a node exerts is not simply proportional to the number of neighbors, but can vary more or less depending on which neighbors it has. A natural generalization of degree centrality is to define importance in terms of the importance of a node's neighbors. That is, not all neighbors are equal and a node's importance may 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[5] 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) 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:

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

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.[6]

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

[6] 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.

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.) 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.

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

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.[7] Thus, we can sidestep the iteration completely and formulate the calculation as an eigenvector problem of the form

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

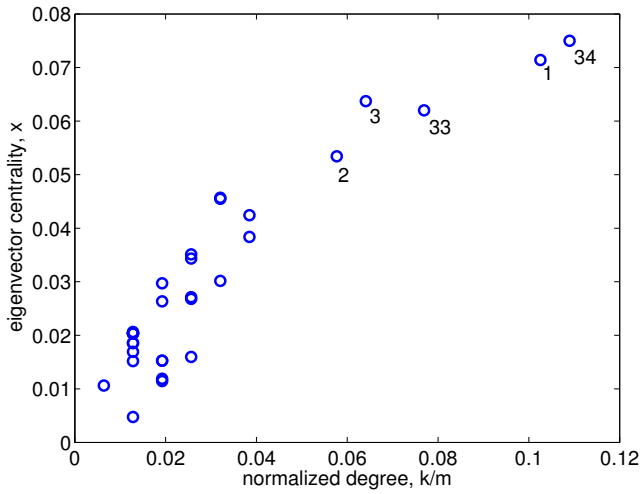
where $\mathbf{A}$ is the adjacency matrix, $\mathbf{x}$ is a vector containing the eigenvector centralities, and λ_1 is the largest eigenvalue of $\mathbf{A}$.[8] Computing eigenvector centralities can be done in most modern mathematical computing software, or using common linear algebra libraries via 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

[7] 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.

[8] Originally given in Bonacich, *J. Math. Soc.* **2**, 113 (1972), and Bonacich, *Social Meth.* **4**, 176 (1972).

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.

PageRank

PageRank is another kind of eigenvector centrality,[9] 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

[9]PageRank is usually attributed to Brin and 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 Pinski and 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.

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

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

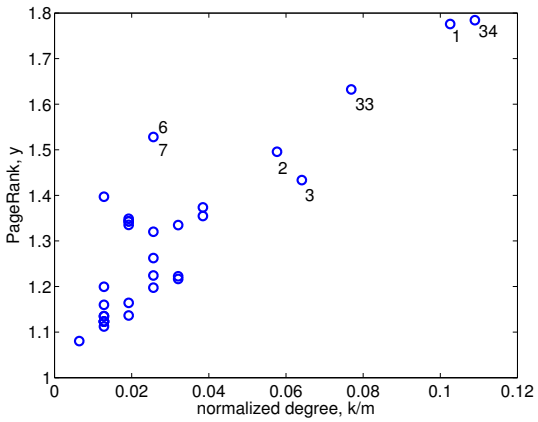
$$= \mathbf{D}(\mathbf{D} - \alpha \mathbf{A})^{-1} \mathbf{1} \ , \tag{5}$$

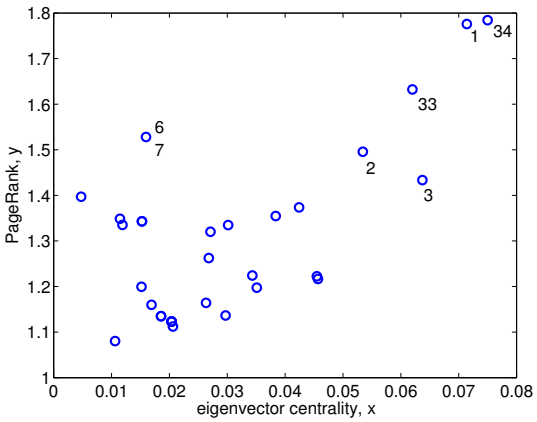
where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \max(k_i^{\text{out}}, 1)$, as described above, and where we have set $\beta = 1$.

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





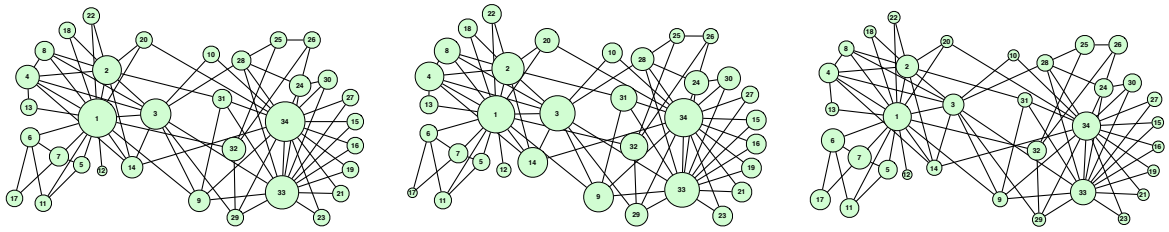
stable, with most of the disagreement between PageRank and eigenvector centrality being on the ordering of lower-importance vertices. If we examine the five most-important vertices under each measure, we see strong agreement on the two most-important vertices. All agree that vertex 33 is either 3rd or 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, below 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



degree eigenvector PageRank

the size of the Web graph is $n \approx 10^{10}$. Instead, they use a streamlined version of the matrix power 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.

3 Geometric centralities

Another class of centrality measures takes a geometric approach to identifying important vertices, relying on geodesic paths between pairs of vertices. Notably, geodesic distances are not metric—they do not obey the triangle inequality—which means applying our (Euclidean) intuition may provide incorrect interpretations of the results. In many cases, the most central vertices under these measures are completely different from those identified by degree-based measures. Here we will study closeness and betweenness centrality scores. As a final point, we note that there are a number of other centrality measures to be found in the literature—and some are even used to study real networks—but the ones we have covered here represent the most common ones.

3.1 Centrality by closeness

A literal interpretation of "centrality" takes inspiration from geometry: the most central point in a k-dimensional body has short paths—it is the point closest—to all other points in the body. If d_{ij} denotes the geodesic distance between vertices i and j, then the average distance from i to all other vertices[10] is given by

$$\ell_i = \frac{1}{n}\sum_{j=1}^{n} d_{ij} \quad . \tag{6}$$

This quantity is large for peripheral vertices, i.e., those far from most other vertices, and small for central vertices, i.e., those close to other vertices. This pattern runs in the opposite direction of most other centrality measures, which are large for central vertices and small for non-central vertices. Thus, the *closeness centrality* of vertex i is typically defined as its inverse:

$$C_i = \frac{1}{\ell_i} = \frac{n}{\sum_{j=1}^{n} d_{ij}} \quad . \tag{7}$$

There are two practical problems with this definition. First, most networks have small diameters (being roughly $\log n$), and thus the range of values that C_i assumes is fairly narrow. Small variations in network topology, perhaps generated by a few missing edges, will produce large changes in a relative ordering. Second, closeness cannot be calculated for a network that is not a single strongly connected component, e.g., an undirected network with only one component. A pair of nodes in distinct components have, by definition, a geodesic distance of $d_{ij} = \infty$, which results in $C_i = 0$.[11] As a result, closeness may only be used in specific contexts.

[10] Sometimes researchers use a summation that omits the path from i to i, which is a geodesic path of length zero, in which case we replace n by $n-1$ in Eq. (6). However, this choice simply rescales all mean distances by a factor of $n/(n-1)$, which cannot change the relative ordering. Eq. (6) is a more convenient mathematical form, and thus we employ it here.

[11] Some researchers have attempted to fix this latter problem by only averaging over distances to vertices in the same component as i, but this introduces a new problem, in which a vertex in a small component, which generally

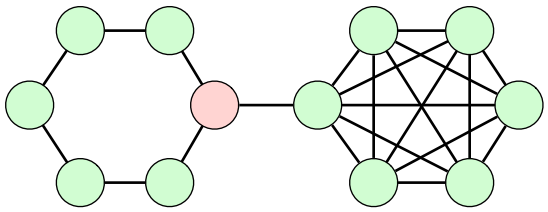
Harmonic centrality
An elegant solution to several of these problems, sometimes called the *harmonic centrality*, is to take the harmonic mean of the geodesic distances from i:

$$C_i = \frac{1}{n-1} \sum_{j(\neq i)} \frac{1}{d_{ij}} \ , \tag{8}$$

where $d_{ij} = \infty$ if there is no path between i and j and we exclude the term $d_{ii} = 0$ to prevent the sum from diverging for trivial reasons. This formulation naturally handles disconnected components, as the $d_{ij} = \infty$ terms contribute 0 to the sum; it also has several other nice mathematical properties.[12]

The calculation of harmonic centrality may be done efficiently using any standard single-source shortest-paths (SSSP) algorithm. For undirected graphs, a breadth first search forest is sufficient, while for directed or weighted graphs, Dijkstra's algorithm works well. In either case, only the actual distances (number of edges) need be retained, rather than the paths themselves.

For example, consider the following small network. The highlighted vertex has a path of length 0 to



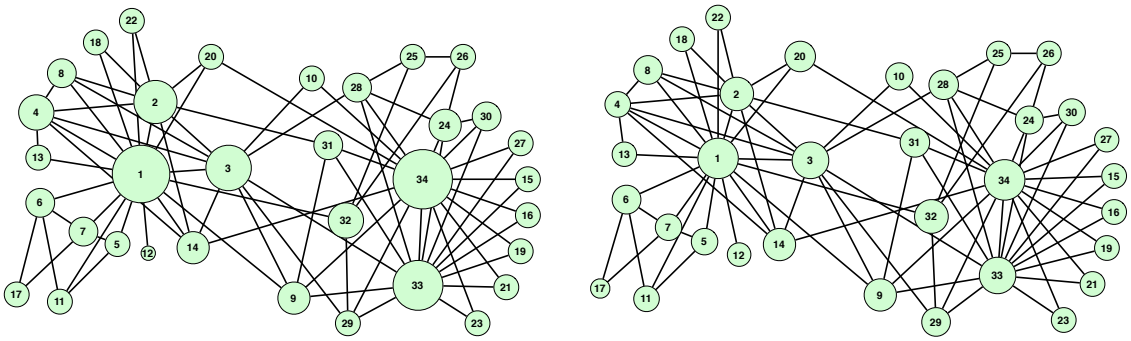
itself, paths of lengths $\{1, 2, 2, 2, 2, 2\}$ to the vertices in the clique on the right, and paths of lengths $\{1, 1, 2, 2, 3\}$ to the vertices in the cycle on the left. Its closeness centrality is thus $12/20 = 0.6$, which is the maximal score in the network, but one other vertex has the same closeness (which one?). Its harmonic centrality is $0.6212\ldots$, which is the second largest value (what is the largest?). The minimal scores are 0.316 (closeness) and 0.417 (harmonic), which illustrates the narrow range of variation of closeness (less than a factor of 2). (Do you see which vertex produces these scores?)

Applying the harmonic centrality calculation to the karate club network yields the figure on the next page (with circle size scaled to be proportional to the score). The small size of this network tends to compress the centrality scores into a narrow range. Comparing the harmonic scores to degrees, we observe several differences. For instance, the centrality of vertex 17, the only vertex in

are considered genuinely peripheral, can have a closeness score comparable to that of an important vertex in a large component.

[12] For a detailed explanation, see Boldi and Vigna, "Axioms for Centrality." Preprint, `arxiv:1308.2140` (2013).

group 1 that does not connect to the hub vertex 1, is lower than that of vertex 12, which has the lowest degree but connects to the high-degree vertex 1. And, vertex 3 has a harmonic centrality close to that of the main hubs 1 and 34, by virtue of it being "between" the two groups and thus having short paths to all members of each.



degree harmonic

Relationship to degree-based centralities
In fact, degree-based centrality measures are related to geodesic-based measures like closeness and harmonic centrality, although they do emphasize different aspects of network structure. For instance, the Katz centrality can be seen as a weighted sum of all paths of different lengths to a vertex i (weighted so that the summation converges), while PageRank can be viewed as a sum of random paths on the network that touch i (recall the Markov model or "random surfer" interpretation). Both of these scores are measuring different paths of different types than the geodesic-based measures, which assume that only the shortest-path is relevant, but they can be viewed as path-based measures nevertheless. As a result, we can expect these measures to be correlated for certain types of networks, as we will see below.

3.2 Centrality by betweenness

Our final measure of importance is also derived from geodesic paths, and relies on the notion that important vertices are the "bridges" over which information tends to flow. This idea is based, in part, on a seminal paper by Mark Granovetter called "The strength of weak ties"[13] in which it was shown that most job seekers (who participated in the study) found their ultimate employment through a weak tie, that is, through an acquaintance, rather than a strong tie or a close friend.

[13]See Granovetter, "The strength of weak ties" *Am. J. Sociology* **78**, 1360–1380 (1973).

The theoretical argument for this pattern was that the information residing at either end of a strong tie is nearly identical because these vertices frequently exchange what information they have. Thus, you and your friends are mostly aware of the same job opportunities, which, had you been qualified for them, you would not be still seeking a job. In contrast, weak ties synchronize their information more rarely, and thus serve as greater sources of novel information when such information is needed. That is, your acquaintances are more likely to know about jobs you have not already considered.

The implication is that vertices that serve as information bridges for many pairs of other vertices are important. Let us make the unrealistic assumption that each pair of vertices exchanges information as a constant rate, and that information is passed along geodesic paths on the network (i.e., information always follows the shortest path between two points). The number of these geodesic paths that cross some vertex i is thus a measure of its importance for synchronizing information across the network, and this is precisely what we call betweenness centrality. There are several different mathematical definitions of betweenness, and we will cover the main ones here.[14]

Our first definition of betweenness is to simply count the number of geodesics that pass through a particular vertex i:

$$\begin{aligned} b_i &= \sum_{jk} \#\{\text{geodesic paths } j \to \cdots \to i \to \cdots \to k\} \\ &= \sum_{jk} \sigma_{jk}(i) \ , \end{aligned} \tag{9}$$

where $\sigma_{jk}(i)$ denotes the number of paths from $j \to k$ that pass through i. Note that applied to an undirected network, this definition double counts each path, once for the $j \to k$ direction and once for the $k \to j$ direction. This behavior is not an issue, however, as multiplication by a constant does not alter the final ordering. Furthermore, this definition includes paths from j to $k = j$. This too simply adds a constant to each centrality score, which does not alter the final ordering.

A second definition of betweenness divides each count by the number of possible geodesics from $j \to k$:

$$\begin{aligned} b_i &= \sum_{jk} \frac{\#\{\text{geodesic paths } j \to \cdots \to i \to \cdots \to k\}}{\#\{\text{geodesic paths } j \to \cdots \to k\}} \\ &= \sum_{jk} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \ , \end{aligned} \tag{10}$$

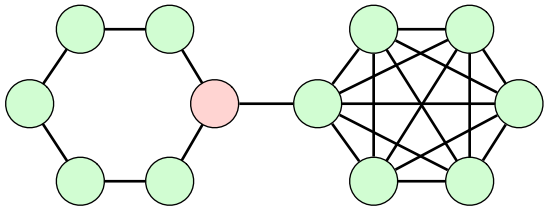
[14] Some of the variations observed in the literature, particularly the sociology literature, differ only in a multiplicative or additive constant to all scores. These constants cannot alter the relative ordering of vertices, and thus here we prefer the more mathematically simple forms.

where σ_{jk} counts all the geodesic paths $j \to k$, not just those that pass through i, and where we define $0/0 = 0$ for disconnected pairs of vertices. This version has the nice feature that if there are multiple geodesic paths from $j \to k$, some of which pass through i and others of which pass through ℓ, both i and ℓ get equal credit for each path.

Finally, a third definition normalizes Eq. (10) to fall on the unit interval $[0, 1]$ by dividing by the number of pairs in the network:

$$b_i = \frac{1}{n^2} \sum_{jk} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \quad . \tag{11}$$

For example, consider again our small network of a clique and a cycle. The highlighted vertex lies



on every geodesic between the left and right groups, of which there are 72. (Why 72?) It also lies on every geodesic path from it to other vertices in the left group, of which there are 6. Thus, the first definition of betweenness would yield $b_\circ = 78$. One of the vertices in the cycle has two geodesic paths to each of the vertices in the clique plus the highlighted vertex (and vice versa); however, both pass through the highlighted vertex, and so the corresponding term in Eq. (10) is 1, as before. All other pairs of vertices have a unique geodesic path, and thus the second definition of betweenness also yields $b_\circ = 78$. Finally, the third definition divides this value by n^2, yielding $b_\circ = 78/144 \approx 0.542$. In each of the three definitions, this score is the maximal value, making the highlighted vertex the most central. The minimal scores are achieved by only one vertex (which one?), and are $b_\bullet = 23$ (being $2n - 1$; why is this the lower bound?), $b_\bullet = 23$, and $b_\bullet = 23/144 \approx 0.160$ (about 3.4 times smaller than the maximum value).

To compute betweenness for an arbitrary network requires enumerating the geodesic paths between all pairs of vertices in the network; this can be done naïvely in $O(n^3)$ time and $O(n^2)$ space.[15] A rough approximation to betweenness may be calculated by solving the SSSP problem once for each vertex and then counting the number of times a vertex i appears in any of the resulting search trees.

[15] It can be done faster, however, using the accumulation algorithm described in Brandes, "A Faster Algorithm for Betweenness Centrality." *J. Math. Sociology* **25**(2), 163–177 (2001). This algorithm takes $O(n + m)$ space, and $O(nm)$ time for unweighted networks or $O(nm + n^2 \log n)$ time for weighted networks.

This procedure takes $O(n(n+m))$ time for unweighted networks, using a breadth-first search forest, and $O(n(m + n \log n))$ for weighted networks, using Dijkstra's algorithm with a Fibonacci heap. However, this approach makes errors whenever there are multiple geodesics between some j and k, a situation that is common in unweighted networks. In this event, full weight will be assigned to the vertices along only one of the geodesics rather than dividing that weight evenly across all of them.[16]

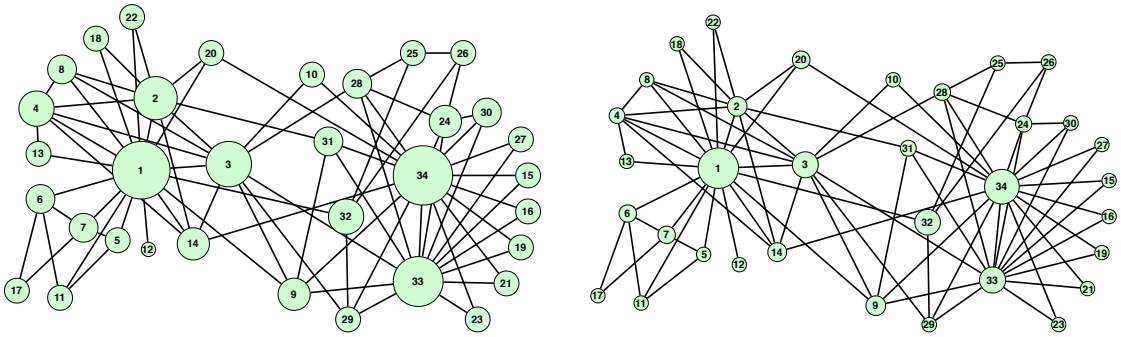
Applied to the karate club, the figures on the following page illustrate that betweenness assigns much smaller relative scores to a greater portion of the network than we saw in degree or harmonic centrality. The most between vertices are still the high-degree nodes (1, 33 and 34), mainly because these nodes are the "brokers" for many other nodes' access to the rest of the network. Vertices that lay between the two groups, like 3 and 32, also receive relatively high betweenness for similar reasons.

The following table compares the relative rankings derived from the different measures defined in this lecture, along with a ranking by degree, for the top few most central vertices. (The last column gives the betweenness estimated by using only a single SSSP tree from each vertex, in order to illustrate the differences this approximation produces.) A few details are worth pointing out. For instance, closeness scores are all very similar, differing only in their second decimal place, while other scores have greater variability, with betweenness having the broadest range. Although all measures generally agree on which vertices are among the most important (mainly 1, 3, 32 and 34), they disagree on the precise ordering of their importance. Consider applying these measures to a novel network: how would such disagreements complicate your interpretation of which vertices are most important? what if there were more disagreement about which vertices were in the top five?

	degree k/m		closeness Eq. (7)		harmonic Eq. (8)		betweenness Eq. (11)		betweenness* Eq. (11)	
1^{st} largest	34	(0.1090)	1	(0.5862)	34	(0.7045)	1	(0.4577)	1	(0.4939)
2^{nd} largest	1	(0.1026)	3	(0.5763)	1	(0.7020)	34	(0.3357)	34	(0.2708)
3^{rd} largest	33	(0.0769)	34	(0.5667)	3	(0.6364)	33	(0.1906)	32	(0.2638)
4^{th} largest	3	(0.0641)	32	(0.5574)	33	(0.6338)	3	(0.1892)	33	(0.2439)
5^{th} largest	2	(0.0577)	9	(0.5312)	32	(0.5859)	32	(0.1843)	3	(0.1912)

Finally, we return to the question of correlation between measures, as both harmonic and betweenness centrality are functions of geodesic paths, which may produce correlated rankings. On the

[16] We may sidestep such a tie-breaking problem by adding a small amount of noise to each edge weight. With high probability, this perturbed network will have a unique geodesic path between each pair of vertices, and each perturbation chooses that geodesic uniformly at random from the original set. (Alternatively, if the network is stored as an adjacency list, we may simply randomly permute the ordering of each vertex's adjacencies.) For up to moderate-sized networks, we can use this trick to enumerate all geodesic paths by repeating the following process: perturb the edge weights, run the SSSP algorithm from each vertex, take the union of the identified geodesics with those of the past step.



degree betweenness

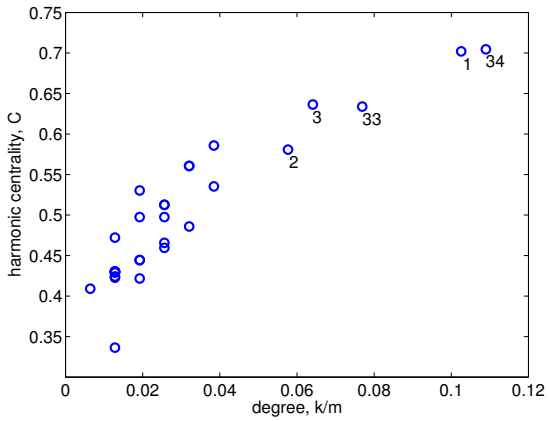
other hand, such a correlation is not a foregone conclusion. Consider a vertex v that has only a single connection to another, highly central vertex. This vertex would have a minimal betweenness score, as it lies on no geodesic paths that do not begin or terminate at v, but its path lengths are all very short, being a single step longer than the paths to the highly central node to which it connects. Thus, v would have high closeness or harmonic centrality, but low betweenness.

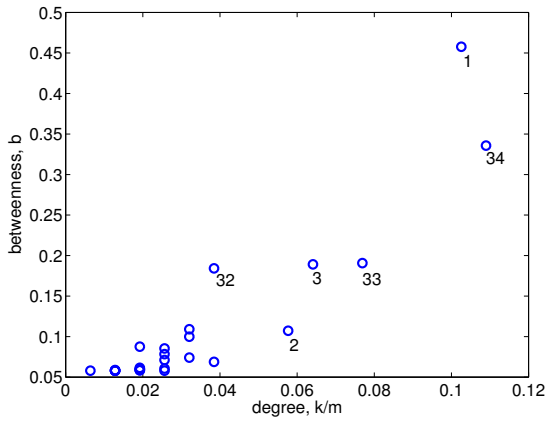
For the karate club network, we see no such disagreement: harmonic and betweenness (and degree) centrality are highly correlated, as the figures below illustrate. (Moreover, $r^2 = 0.69$ for harmonic and betweenness centralities, and they both have $r^2 = 0.83$ with degree centrality.) That is, all of our centrality measures more-or-less agree that the most important vertices in the karate club network are vertices like 1, 34, 33, 32 and 3, plus a few others. These particular vertices are distinguished along several different measures of importance, and we may conclude that they play special roles in structuring the karate club.[17]

3.3 Caveats about centrality

It is worth reiterating that each measure of centrality is fundamentally a proxy of some an underlying network process or processes. If the particular network process is irrelevant or unrealistic for a given network, then any measure of centrality based on that process will produce nonsense. For instance, betweenness centrality attempts to get at the common idea in network science that connecting disparate parts of the network is important. Betweenness formalizes this notion using a particular model—a model in which communication occurs only over geodesic paths, in which the routing of information is maximally efficient, in which each vertex has full information about the routes through the network, in which communications occur at regular intervals, etc.—and this

[17]The story behind this network reinforces this conclusion, as the highest degree vertices 1 and 34 were the president and leader of the karate club before it split into to factions, and each went on to form distinct clubs after the split.





model may have little to do with actual importance in actual network systems.[18]

This does not mean that centrality measures cannot or should not be used. Rather, they should be used mainly in an exploratory manner, to gain some insight into the general structure and pattern of a network and to generate hypotheses about what processes might have generated that structure. These measures may also serve the useful purpose of building our intuition about what kinds of structural patterns correlate with other types of structural patterns, a topic we will revisit when we study random graph models.

As a nice visualization, below is a fun image by Claudio Rocchini on the Wikipedia page for centralities, which shows how different centrality measures think different vertices are more or less important. Do you see why?

Finally, note that we did not cover any number of other importance functions, ranging from ranking functions for competitive interactions (e.g., Minimum Violation Rankings, Bradley-Terry-Luce models, Ello scores, SpringRank, and more), nor did we cover any dynamical importance scores, like "spreading centrality." Nor did we cover any of the supervised "learning to rank" methods.

[18] The development of new centrality measures is partly driven by researchers tinkering with the underlying model. For instance, if we don't like the assumption that all vertices send messages to each other at exactly the same rate, or if we don't like the assumption that each message from j to k follows a randomly chosen geodesic, we can apply a weight function to the summation, giving some paths more or less weight than others. The key question, however, is whether or now such a variation provides *more* useful insights for some system than existing measures.

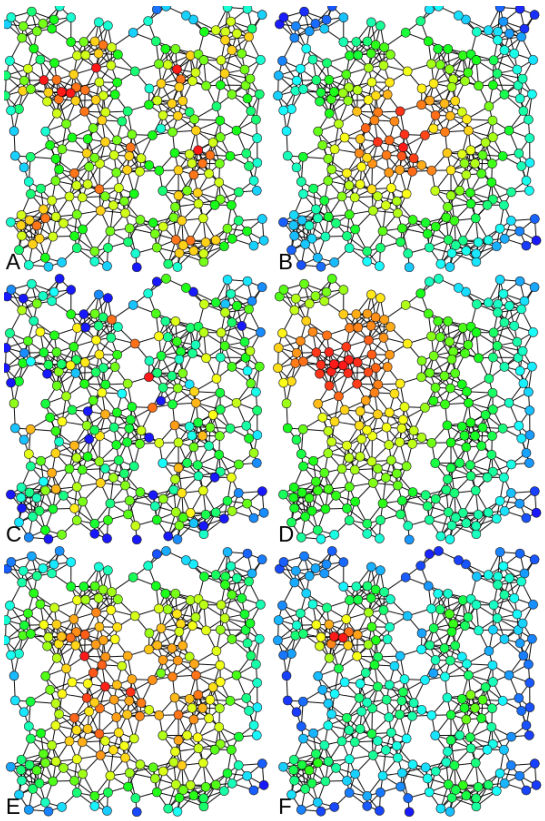


Figure 1: An example calculation by Claudio Rocchini (from the Wikipedia page on Centrality) of (A) degree centrality, (B) closeness centrality, (C) betweenness centrality, (D) eigenvector centrality, (E) katz centrality and (F) alpha centrality.

4 Supplemental Readings

1. Read Chapter 7.1–7.5 (pages 168–181) in *Networks* (degree centralities)
2. Read Chapter 7.6–7.8 (pages 181–198) in *Networks* (geometric centralities)