1 Geometric centrality

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.

1.1 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[1] is given by

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

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:

$$C_i = \frac{1}{\ell_i} = \frac{n}{\sum_{i=1}^{n} d_{ij}} \quad . \tag{2}$$

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 $C_i = 0$.[2] As a result, closeness may only be used in specific contexts.

[1] 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. (1). However, this choice simply rescales all mean distances by a factor of $n/(n-1)$, which cannot change the relative ordering. Eq. (1) is a more convenient mathematical form, and thus we employ it here.

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

Harmonic centrality

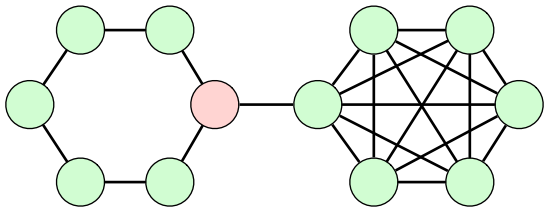
An elegant solution to several of these problems, sometimes called the *harmonic centrality*, is to take the harmonic mean of the geodesic distances from i:

$$C_i = \frac{1}{n-1} \sum_{j(\neq i)} \frac{1}{d_{ij}} \ , \tag{3}$$

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.[3]

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.

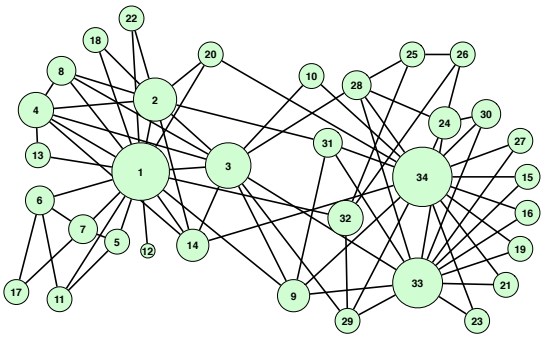
For example, consider the following small network. The highlighted vertex has a path of length 0 to



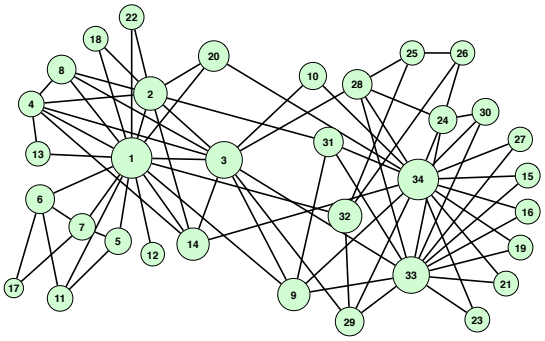
itself, paths of lengths $\{1, 2, 2, 2, 2, 2\}$ to the vertices in the clique on the right, and paths of lengths $\{1, 1, 2, 2, 3\}$ to the vertices in the cycle on the left. Its closeness centrality is thus $12/20 = 0.6$, which is the maximal score in the network, but one other vertex has the same closeness (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 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.

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



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.

1.2 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"[4] 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.

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

[4] In *American Journal of Sociology* **78**, 1360–1380 (1973).

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.[5]

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

$$\begin{aligned} b_i &= \sum_{jk} \#\{\text{geodesic paths } j \to \cdots \to i \to \cdots \to k\} \\ &= \sum_{jk} \sigma_{jk}(i) \ , \end{aligned} \tag{4}$$

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 $j \to k$:

$$\begin{aligned} b_i &= \sum_{jk} \frac{\#\{\text{geodesic paths } j \to \cdots \to i \to \cdots \to k\}}{\#\{\text{geodesic paths } j \to \cdots \to k\}} \\ &= \sum_{jk} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \ , \end{aligned} \tag{5}$$

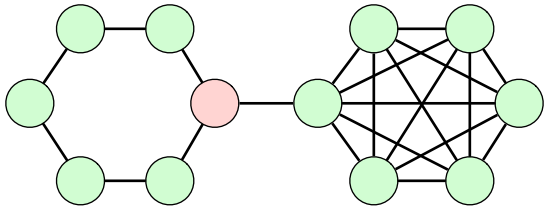
where σ_{jk} counts all the geodesic paths $j \to k$, not just those that pass through i, and where we define $0/0 = 0$ for disconnected pairs of vertices. This version has the nice feature that if there are multiple geodesic paths from $j \to k$, some of which pass through i and others of which pass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

Finally, a third definition normalizes Eq. (5) to fall on the unit interval $[0, 1]$ by dividing by the number of pairs in the network:

$$b_i = \frac{1}{n^2} \sum_{jk} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \ . \tag{6}$$

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. (5) is

[5] 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.



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.[6] 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. 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.[7]

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.

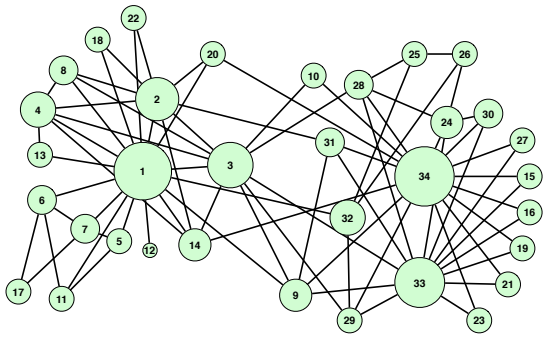
The following table compares the relative rankings derived from the different measures defined in this lecture, along with a ranking by degree, for the top few most central vertices. (The last column gives the betweenness estimated by using only a single SSSP tree from each vertex, in order

[6] It can be done faster, however, using the accumulation algorithm described in U. Brandes, "A Faster Algorithm for Betweenness Centrality." *Journal of Mathematical 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.

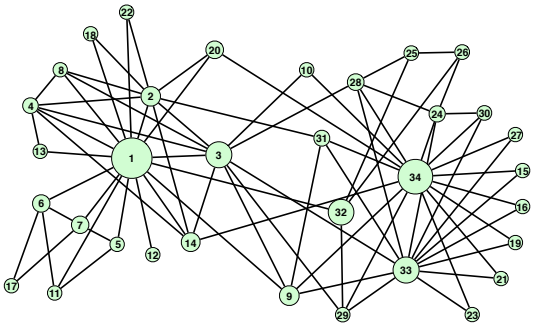
[7] We may sidestep such a tie-breaking problem by adding a small amount of noise to each edge weight. With high probability, this perturbed network will have a unique geodesic path between each pair of vertices, and each perturbation chooses that geodesic uniformly at random from the original set. (Alternatively, 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.

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. (2)		harmonic Eq. (3)		betweenness Eq. (6)		betweenness* Eq. (6)	
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)



degree

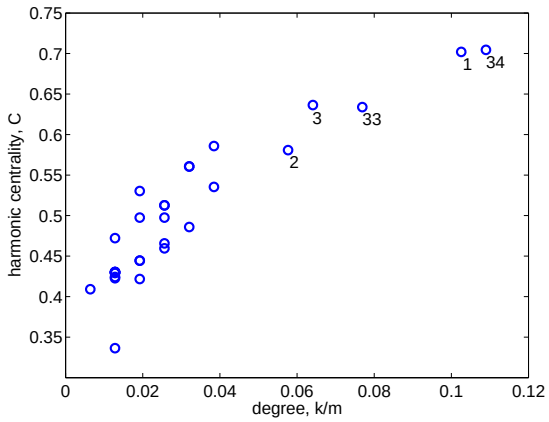


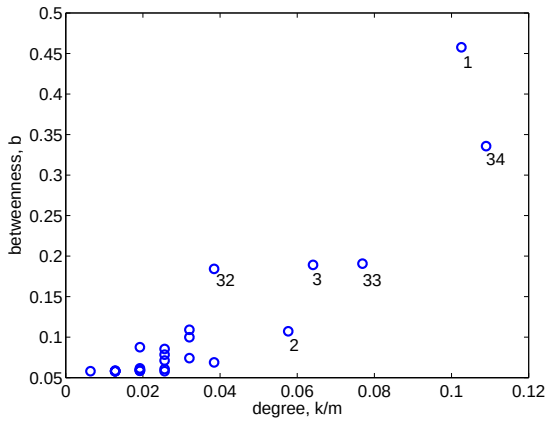
betweenness

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 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.[8]

[8]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.





1.3 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 model may have little to do with actual importance in actual network systems.[9]

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.

2 At home

1. Read Chapter 7.6–7.8 (pages 181–198) in *Networks*
2. Next time: transitivity, reciprocity and assortativity
3. Optional reading: P. Boldi and S. Vigna, "Axioms for Centrality." Preprint, `arxiv:1308.2140` (2013).

[9]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.