OUP • UNCORRECTED PROOF

CHAPTER 17

THREE PERSPECTIVES ON CENTRALITY

STEPHEN P. BORGATTI AND
MARTIN G. EVERETT

Social network analysis is blessed with a large number of measures of centrality. We know these are node-level measures and that they are in a very general sense about measuring node importance. But beyond that it isn't easy to define what counts as a centrality measure and what doesn't. Part of the problem, of course, is that people can call anything they want "centrality," and they do. They can also use much more fanciful names instead of centrality, such as influence, prominence, status, power, and so on, which they also do, even though it is an empirical question how a graph-theoretic concept like centrality relates to these variables. Finally, it is an evolving category, so things that today we might not call centrality still carry the name.

This chapter is about three perspectives on centrality, which we refer to as the "walk structure perspective," the "induced centrality perspective," and the "flow outcomes perspective." They are presented not as competing frameworks—the first is so general as to easily coexist with any other—but as conceptual tools that provide insight into centrality and perhaps into the enterprise of social network analysis generally. Before getting into these, however, a few prefatory notes are in order.

It is often stated that centrality measures are structural measures. The idea here is to distinguish a centrality measure from, say, a node's wealth or social class. The measures are constructed from the network (i.e., "endogenous"), and not from qualities of the actors that one would need to look beyond the pattern of ties to obtain (i.e., "exogenous"). This is a useful distinction but overstated. It is not that one cannot incorporate financial status into a centrality measure, but that a centrality measure must incorporate ties as a key feature. So, for example, in a study of organizations, we might define a variant of degree centrality (i.e., the number of ties a node has) by differentially weighting ties by the organizational rank of the person being connected to. Thus, my ties to high-ranking managers count more than my ties to ordinary workers. Such a measure incorporates both a structural aspect (the ties) and an exogenous attribute of the nodes (their rank in the organization). This is a legitimate and, indeed, useful measure of centrality.

It is also worth noting that centrality concepts are universally described as "measures." This causes all kinds of confusion. For most social scientists, the word *measure* indicates

0004829569.INDD 334 Dictionary: NOAD 6/15/2020 1:13:28 PM

OUP • UNCORRECTED PROOF

that centrality measures fall under methodology and are best described in the methods section of an empirical paper. Thus, the hypothesis in the main section of the paper is going to be something like "More central managers will be more likely to receive promotion in the next year." The specific centrality concept is left unstated. What these researchers don't understand is that "centrality" is just a family of concepts on par with "demographics" or "personality traits." Who writes a hypothesis like "Demographics will be associated with postsurgical survival"? No one would accept this because it is understood that "demographics" is just a category of constructs. ~~Similarly,~~ with personality, a theory would argue, with appropriate reasoning, that a specific personality characteristic, such as neuroticism, should contribute to survival. That trait would be part of the hypothesis, as in "Higher levels of neuroticism will be associated with reduced five-year survival rates." How neuroticism is actually measured is left for the methods section, because there are multiple ways to measure the same psychological construct. Similarly, a good network hypothesis would state that, say, closeness centrality will be associated with being perceived to be a good source of gossip. The methods section would then discuss exactly how closeness was being calculated (and there are indeed multiple ways).

Finally, it should be mentioned that there have been some attempts to define centrality measures, or at least to lay down characteristics of "well formedness." For example, Sabidussi (1966) suggested five properties that every centrality measure should observe. These were things like adding a tie to a node should always increase its centrality and adding a tie anywhere else in the network should never reduce a node's centrality. Unfortunately, the only measure to pass all five criteria was his own, a form of closeness centrality. Freeman (1978/1979) argued that any centrality measure worth its salt should achieve its highest value with the node at the center of a star graph. Specifically, if we consider all possible graphs of a given size (defined by number of nodes), no node in any graph should get a higher score than the node at the center of a star graph. This is an appealing criterion but does leave out measures—such as the PN measure of Everett and Borgatti (2014)—that intuitively seem as qualified to be called centrality measures as any other.

The one property that does make sense to require all purely structural centrality measures to obey is the isomorphism rule: if two nodes are isomorphic, they must receive the same score. In a sense, this is what we mean by "structural"—if the values are calculated entirely from the pattern of ties in the network, structurally identical nodes will necessarily receive the same scores. Of course, if the measure is a hybrid that deliberately incorporates an exogenous node attribute such as personality, then we simply modify the isomorphism rule correspondingly: if two nodes are isomorphic in the network and have the same personality, then they must receive the same score.[1]

Schoch and Brandes (2016) have examined a simple property that is shared by nearly all centrality measures. The open neighborhood of a node v is the subgraph induced by all the nodes adjacent to v and is denoted by $\mathrm{N}(v)$. That is, $\mathrm{N}(v)$ is the graph that consists of all nodes adjacent to v and all the edges in the original graph that connect these vertices but does not include v itself. If we include v, it is called the closed neighborhood and is denoted by $\mathrm{N}[v]$. A node u is said to dominate a node v if $N[u] \supseteq N(v)$, meaning that the open neighborhood of v is included in the closed neighborhood of u. A centrality measure $\mathbf{c}$ is said to be *neighborhood preserving* if whenever u dominates v, then $\mathrm{c}(u) \geq \mathrm{c}(v)$. This implies that, given a graph in which there are some nodes that satisfy neighborhood inclusion, then the centrality ranking is predetermined. Of the centrality measures commonly used in

0004829569.INDD 335 Dictionary: NOAD 6/15/2020 1:13:28 PM

OUP • UNCORRECTED PROOF

social networks, the only one that does not obey this property is Bonacich's (1987) beta centrality (also known as Bonacich power), but only when beta is negative. It should be noted that there exist graphs for which every pair of nodes satisfy neighborhood inclusion, and so the ranking of nodes by almost every centrality measure is completely determined regardless of which measure is used. On the other hand, it is possible to have graphs for which no pairs of nodes satisfy the inclusion property, in which case the centrality rankings across measures can be quite independent.

In reading this chapter, one theme that the reader might keep in mind is the tension between using standard measures that the field as a whole has a great deal of experience using and constructing new measures tailor-made for a specific application. The downside of the former is that no off-the-shelf measure may be quite right for a given application, while the downside of the latter is that the new measure has not been vetted by the community and could have serious flaws. Moreover, consumers of research utilizing new measures may wonder if the measure was somehow ginned up specifically because it would yield significant results for the data in question—a kind of overfitting problem.

The Walk Structure Perspective

In social network analysis, our central object is the graph—a set of nodes together with a set of dyadic ties[2] that connect them. The definition makes it sound like the ties are independent of each other—just a big bucket of connected dyads, like a sample of mother-daughter pairs. But what makes networks interesting is that sets of dyads interlink by sharing nodes, forming paths through the network. Because of paths, almost any node can affect any other, creating what otherwise might be seen as "spooky action at a distance," as Einstein described quantum mechanics. The crux of the walk structure perspective is that some nodes will, because of the structure of paths, be more or less influential in the network as a whole.

But let us back up and discuss paths more carefully. A path is a kind of traversal through a network that begins at a node, ends at a node, and moves only between adjacent (i.e., tied) nodes. By convention, the length of a traversal is defined as the number of links traversed. For simplicity, we assume that ties have no direction so that the A—B tie is the same object as the B—A tie. But the term *path* has a very specific meaning in graph theory, namely a traversal that never repeats a node. Hence the sequence A—B—C—D—E is a path, but the sequence A—B—C—D—B—E is not. A traversal that allows revisiting of nodes (but not links) is a trail. The sequence A—B—C—D—B—E is a trail, but A—B—C—D—B—C—E is not, because it repeats the B-to-C move. An unrestricted traversal, which can repeat nodes and links, is a walk. The sequence A—B—C—D—B—C—E is a walk, as is the sequence A—B—A—B—A—B—C—B, etc. Obviously walks can be infinite in length. Also obviously, trails are a kind of walk, and paths are a kind of trail. Among paths between a given pair of nodes, the shortest path is called a geodesic. Note there can be multiple geodesics linking the same two nodes.

The walk structure perspective says that what a centrality measure measures is the involvement of a node in the traversal structure of a graph (which is to say the walk structure, since walks are the most general kind of traversals). The more involvement, the more central.

0004829569.INDD 336 Dictionary: NOAD 6/15/2020 1:13:28 PM

OUP • UNCORRECTED PROOF

How measures differ is in what kinds of traversals they pay attention to—unrestricted walks, trails, paths, geodesics—and what property of the traversals they focus on (e.g., their number or their length). The perspective also pays attention to the role of the node in these traversals. Are they starting/ending points? Or interior nodes in the sequence? We shall refer to measures focusing on the former as radial measures and measures focusing on the latter as medial measures.

To illustrate, let us look at various measures of centrality in light of these different traversals, starting with degree. Degree is just the number of ties a node has. For our purposes, it is useful to think of a tie as a path (indeed, a geodesic) of length 1. So degree centrality counts the number of paths of length 1 that emanate (or terminate) from a node. Note that degree is a radial measure: we are counting paths that emanate from (or terminate at) a node.

Another well-known measure is closeness. Sabidussi (1966) and Freeman (1978/1979) both define closeness as the sum of graph-theoretic distances from a node to all others, where distance is defined as the length of a geodesic connecting two nodes. While the sum of distances is simple, it is more intuitive to think in terms of the average distance. More central nodes are closer, on average, to all others than peripheral nodes. Since closeness is an inverse measure of centrality (larger numbers indicate less centrality), a popular alternative is reciprocal closeness. Whereas classical average closeness for node i can be defined as $\left(n-1\right)^{-1}\sum_j d_{ij}$, where d_{ij} is the distance from i to j, reciprocal closeness averages the reciprocals of distances: $\left(n-1\right)^{-1}\sum_j d_{ij}^{-1}$. Both measures of closeness are radial measures like degree but differ from degree in that, instead of counting qualified paths, they assess their typical length. This distinction is important, but we must be careful to understand what it is not. Both degree and closeness make use of length, but in different ways. Degree uses length to bound the set of paths to be considered (specifically, paths of length 1), but then counts them rather than assesses their length. Consider the many obvious variants of degree. As early as 1966, Donald Sade proposed the k-reach measure, which counts the number of paths (emanating from a node) of length less than or equal to k, a parameter to be determined by the researcher. Or we could count all of the geodesics, regardless of length, that emanate from a node. In both cases, length is used to identify suitable paths, but the property measured is the number or frequency of qualifying paths.[3]

Agneessens, Borgatti, and Everett (2017) offer a variant of reciprocal closeness that yields a weighted average of distances. This is accomplished by adding a negative exponent to the distances: when the exponent is large, the measure weights short distances more than long ones (see Equation 1). When the exponent is 1, the measure reduces to reciprocal closeness. Of interest in this context is that, while the measure is clearly about assessing typical length rather than counting paths, Agneessens et al. note that when the exponent is set to infinity, the formula yields degree centrality. The point is that our distinction between length and count measures can break down in special cases (see also Brandes, Borgatti and Freeman, 2016).

$$c_i\left(\delta\right)=\frac{\sum_j d_{ij}^{-\delta}}{n-1} \qquad \text{Equation 1}$$

Betweenness centrality is defined by the formula in Equation 2 (Freeman, 1978/1979). In the formula, g_{ikj} is the number of geodesic paths from i to j that pass through k. The summation is

OUP • UNCORRECTED PROOF

across all possible pairs *i* and *j*. Ignoring the denominator inside the sum, the measure counts the number of shortest paths from anywhere to anywhere that pass through *k*. In short, the measure defines geodesics as the traversals of interest, and then looks at how often *k* is an interior node in these traversals. It is a medial measure of centrality, and the property measured is frequency. Of course, the actual formula divides g_{ikj} by g_{ij}, which is the total number of geodesics from *i* to *j*. So a more accurate description of the measure is that it sums the share of best paths that *k* sits on. It counts all of the geodesics that include *k* as an interior node, but it discounts those for which there is an alternative geodesic between the same pair. Other betweenness measures have been devised that don't rely on geodesics. For example, Newman (2005) proposes random walk betweenness, which counts all walks, rather than just geodesics.

$$b_k = \sum_{i,j} \frac{g_{ikj}}{g_{ij}} \qquad \text{Equation 2}$$

An interesting betweenness measure is flow betweenness (Freeman, Borgatti, & White, 1991). For nonvalued graphs, the measure attempts to count the number of edge-independent paths between all pairs of nodes that pass through a given node. This measure introduces a new descriptive element in characterizing nodes' involvement in the traversals of a graph. The individual paths being counted are true paths in that nodes are not repeated. However, the unit of analysis is the set of paths from a source node to a target, which are required to be edge independent.[4] Thus, the subset of paths used for flow betweenness have the property that they do not share edges. In addition, there is no requirement that the paths be geodesics. One reason for considering whole sets of paths from one node to another is to model traversals in which something flowing through a network may be in multiple places at once. This happens when what is flowing is copied rather than transferred across ties. Thus, multiple paths are followed at the same time. This issue is discussed further in the section on the third perspective.

Beta centrality, also known as Bonacich power (Bonacich, 1987; Bonacich & Lloyd, 2001), is defined in Equation 3, where *R* is the adjacency matrix of the graph, and **1** is a column vector of ones so that *R*1 is just the row sums of *R*. The measure draws on the earlier work of Katz (1953) and Hubbell (1965). Beta centrality is just Katz's measure divided by β, and Hubbell's measure is Katz's plus 1.

$$b = \left(I - \beta R\right)^{-1} R1 \qquad \text{Equation 3}$$

Note that $\left(I - \beta R\right)^{-1}$ is just a transformation of the data matrix *R*. Under the right conditions, $\left(I - \beta R\right)^{-1}$ can be written as the infinite sum shown in Equation 4. Following Equation 3, we can then multiply the infinite sum by *R* and get a matrix we will call *W*, as shown in Equation 5.

$$\beta^0 R^0 + \beta^1 R^1 + \beta^2 R^2 + \cdots \qquad \text{Equation 4}$$

$$W = \left(I - \beta R\right)^{-1} R = \beta^0 R^1 + \beta^1 R^2 + \beta^2 R^3 + \cdots \qquad \text{Equation 5}$$

The matrix *W* is revealing. Ignoring the betas, it is a sum of powers of the adjacency matrix. These have a meaning. The (*i*,*j*)th cell of the R^1 matrix (which is just the adjacency matrix)

0004829569.INDD 338 Dictionary: NOAD 6/15/2020 1:13:33 PM

OUP • UNCORRECTED PROOF

gives the number of walks of exactly length 1 from *i* to *j*. The R^2 matrix gives the number of walks of exactly length 2. The R^3 matrix gives the number of walks of exactly length 3, and so on. This means that, summed up, *W* gives the total number of walks of all lengths from every node to every other, and if we take the row sums, beta centrality is then a frequency measure of the radial type.[5] Of course, without the betas, every cell of *W* would be infinity, since walks can be of infinite length. So let's consider the betas. For clarity of exposition, let's temporarily assume that beta is less than 1—say, 0.5. Then walks of length 1 will be weighted $0.5^0 = 1$, which means they are given full weight. Walks of length 2 will be weighted $0.5^1 = 0.5$, which means they will be weighted only half as much as walks of length 1. The R^3 matrix will be weighted 0.5^2 or 0.25, and so on. Walks of even the modest length of 10 will be weighted only 0.000977, and walks of infinite length will be weighted 0. Thus, beta centrality counts the number of walks of all lengths emanating from a node, weighted inversely by their length.

Now, must beta be less than 1 for this to work? No. For the infinite sum to converge, the absolute value of beta must be less than $1/\lambda$, where λ is the largest eigenvalue of *R*. But λ can always be rescaled to be 1 by simply dividing every value in *R* by λ. For example, if the largest eigenvalue of a matrix is 45, we can divide the cells of the matrix by 45, and the λ of the new matrix will be 1. As a result, if we prefer to think of beta as a fraction like 0.5, we can do so by rescaling the matrix so that λ is 1.

It is useful to observe that when beta is ~~0~~, beta centrality is just degree. We can see in Equation 3 that if beta is ~~0~~, the equation reduces to $R\mathbf{1}$, which is simply degree. As beta increases, longer walks are included, but weighted by (shortness of) length. Therefore, we can think of this as a radial frequency measure that counts walks weighted (rather than selected) by length. When the absolute value of beta approaches $1/\lambda$ closely (from below), the resulting centrality scores become indistinguishable from eigenvector centrality. The formula for eigenvector is shown in Equation 6, where *v* is the eigenvector. Rewriting it as Equation 7 makes it clear that, with eigenvector centrality, a node's score is proportional to the sum of the scores of its neighbors, and $1/\lambda$ is the proportionality constant.

$$Rv = \lambda v \qquad \text{Equation 6}$$

$$v_i = \frac{1}{\lambda} \sum_j r_{ij} v_j \qquad \text{Equation 7}$$

More measures could be discussed, but it seems reasonable to stop here and present a three-dimensional typology of centrality measures (Borgatti & Everett, 2006). The dimensions are clear. The first dimension is what kind of traversals the measure pays attention to. The choices discussed here have been geodesics, paths, trails, and unrestricted walks. The second dimension is the node's position along these traversals: is it an endpoint[6] or an interior node? We have referred to measures relying on the former as radial and measures relying on the latter as medial. The third dimension is what property of traversals we are assessing: their frequency or their length? This is perhaps the fuzziest of the dimensions, since many measures appear to take account of both. However, there is a subtle distinction between assessing the typical length of a path, as in closeness or information centrality (Stephenson & Zelen, 1989), and using length to bound the set of paths to be counted, as in degree and betweenness. We see the latter as frequency measures. The beta centrality family

0004829569.INDD 339 Dictionary: NOAD 6/15/2020 1:13:33 PM

OUP • UNCORRECTED PROOF

Table 17.1 Three-Dimensional Typology of Centrality Measures

Radial measures				
	Geodesics	Paths	Trails	Walks
Length	Closeness,[a] ARD,[b] generalized closeness[c]	Information[d]		IEC,[k] Markov FPT[e]
Length × Freq				Katz, Hubbell, beta,[f] PN,[g] eigenvector, TEC[k]
Frequency	Degree, PII[h]	k-Reach	GPI[i]	
Medial measures				
	Geodesics	Paths	Trails	Walks
Length				
Length × Freq				
Frequency	Betweenness, generalized betweennesses[j]	Flow betweenness		RWB,[l] MEC[k]

[a] Freeman (1978/1979). [b] Average reciprocal distance. [c]Agneessens et al. (2017). [d]Stephenson & Zelen (1989). [e] First passage time. [f] Bonacich (1987). [g] Everett and Borgatti (2014). [h] Smith et al. (2014). [i] Markovsky, Willer, and Patton (1988). [j] Brandes (2008). [k] Friedkin (1991). [l] Newman (2005).

of measures are harder to classify. Effectively, the scores are the sum of products of the number of walks and (an inverse function of) the lengths of the walks. This is an ambiguous number. If normalized by the number of walks, we would call it the average (inverse) length of walks emanating from a node, weighted by frequency. If normalized by the sum of inverse lengths, we would call it the average number of walks, weighted by inverse length. Since we don't normalize by either, the nature of the quantity is indeterminate. Table 17.1 presents the typology, with the beta centrality measures occupying a mixed category of length and frequency. The table includes some measures that have not been discussed here.

The Contribution/Induced Centrality Perspective

The contribution perspective follows directly from the walk structure perspective presented in the last section but is in a certain sense more disciplined. The idea is to quantify the extent to which a node (or set of nodes) contributes to a notable characteristic of the graph as a whole. To do this, we identify a structural characteristic of the network as a whole (known as a graph invariant), then remove a given node, recalculate the graph invariant, and take the difference. This difference captures the node's contribution to the graph invariant and is taken to be its centrality score. We do the same thing for each node in turn to calculate every node's centrality. Centralities calculated in this way are known as induced centralities (Everett & Borgatti, 2010) or vitalities (Koschützki et al., 2005).

0004829569.INDD 340 Dictionary: NOAD 6/15/2020 1:13:33 PM

OUP • UNCORRECTED PROOF

A well-known example of an induced centrality is flow centrality (Freeman et al., 1991). The idea behind flow centrality is that a network is a system of pipe-like channels through which things flow. The valued adjacency matrix gives the capacities of all the pipes. The classical problem, posed by Ford and Fulkerson (1956), is: given the selection of a source *S* and a target *T*, what is the maximum amount of stuff that can flow from *s* to *t*? For graphs with unit capacities on all pipes, the problem can be shown to be equivalent to the number of edge-independent paths from *s* to *t*, where a set of edge-independent paths is a set in which no path shares a link. The idea of flow centrality is to measure how much of the flow from *s* to *t* (and, indeed, from any node to any node) passes through a given node *k*. Because flow solutions are not necessarily unique, the method Freeman et al. devised to calculate this flow was an induced centrality. First they calculated maximum flows from every node to every other, summing these up. Then they removed a node from the network and recalculated the sum of maximum flows. The difference between these sums was interpreted as the amount of flow passing through that node. In hindsight, it is not clear that the method exactly captures the conceptualization, but it is in any case a clear example of an induced centrality.

It should be obvious that degree centrality is an induced centrality. If we take as the graph invariant the number of edges in the graph and now remove a node, the change in the number of edges in the graph is exactly the number of edges incident on the node, which is to say, its degree. However, it can be shown that not all accepted measures of centrality are induced centralities. Everett and Borgatti (2010) present the case of the graph in Figure 17.1 (left). Let us assume we will be measuring the centrality of nodes *a* and *b*, in which case we will be removing them, one at a time, from the graph. The graphs without *a* and without *b* are shown in the middle and right-hand sections of Figure 17.1. Note that these two graphs are isomorphic. Whatever graph invariant we calculate on them, the result will be exactly the same, which means that *a* and *b* will have the same induced centrality score for any invariant we can imagine. Interestingly, when we calculate off-the-shelf measures of centrality on this graph, we find that all measures give identical scores to *a* and *b* except for closeness measures. For example, the standard Freeman closeness centrality gives *b* a

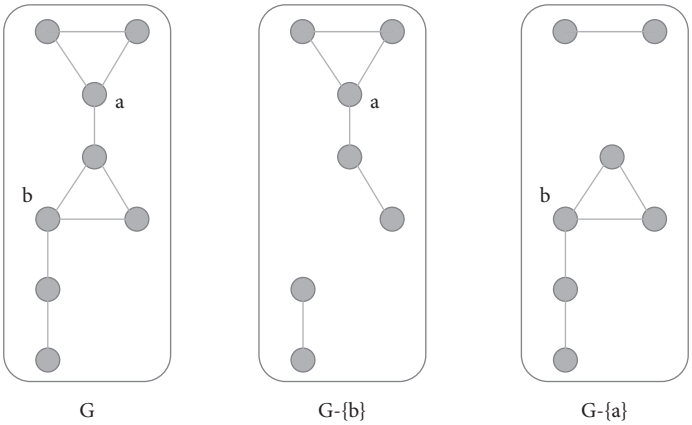

FIGURE 17.1 Removing node *a* or node *b* yields identical residual graphs, but nodes *a* and *b* have different closeness scores.

0004829569.INDD 341 Dictionary: NOAD 6/15/2020 1:13:34 PM

OUP • UNCORRECTED PROOF

slightly better score. So does information centrality (Stephenson & Zelen, 1989), as well as three-step reach. As a result, none of the closeness measures can be conceptualized as induced centralities.

This result has implications for the question raised earlier about what properties a centrality measure must conform to. Given the principle that isomorphic nodes must have the same score, we might have been tempted to make a stronger statement, namely that any centrality measure must give the same score to two nodes whose removal yields isomorphic residual graphs. Clearly, this fails for closeness. The reverse fails as well: if two nodes have the same centrality score, it is not the case that their residual graphs G-{*a*} and G-{*b*} are isomorphic.

An interesting set of induced centralities can be built on conventional centralities. For example, suppose we take betweenness centrality and calculate it for every node.[7] Now we sum the betweennesses to obtain a single number for the entire graph—a graph invariant. Suppose we use this to define an induced centrality. It might seem like the results would be trivial—each node contributed their betweenness centrality to the sum, and so removing the node, recalculating every remaining node's betweenness, and recalculating the sum would perhaps subtract from this sum precisely that node's betweenness, just like degree. But it's not so. Consider the network of marriage ties (Padgett & Ansel, 1993) shown in Figure 17.2. Notice that part of the Medici's betweenness comes from enabling Pazzi and Salviati to connect to the rest of the network. But more than that, the Medici's presence gives Salviati betweenness, because the Salviati family connect Pazzi to the Medici, and

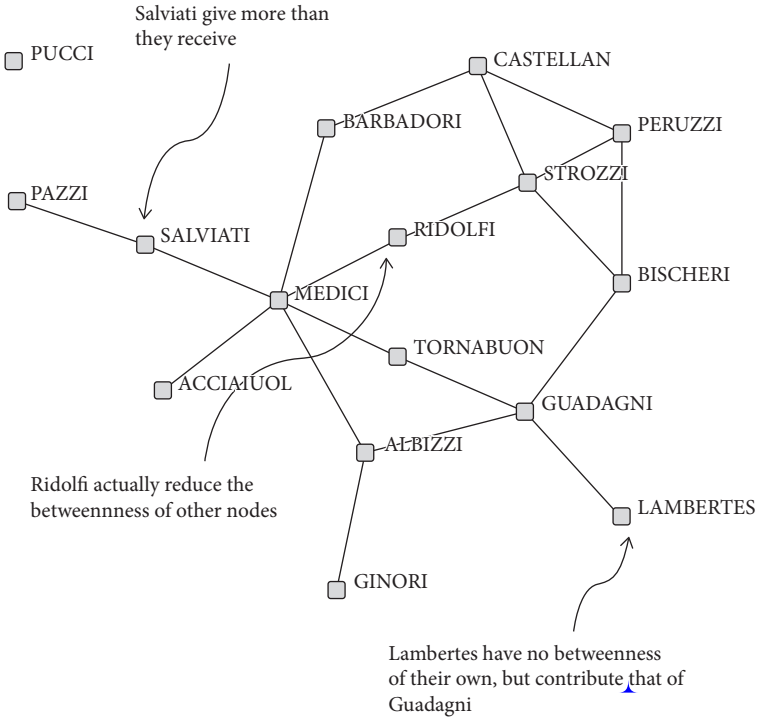

FIGURE 17.2 Marriage ties among Florentine families during the Renaissance.

0004829569.INDD 342 Dictionary: NOAD 6/15/2020 1:13:34 PM

OUP • UNCORRECTED PROOF

Table 17.2 Direct and Indirect Contributions to the Total Betweenness of the Network

Family	Total	Direct	Indirect
Medici	73	48	25
Guadagni	38	23	15
Albizzi	43	19	24
Salviati	57	13	44
Ridolfi	9	10	−1
Bischeri	11	10	1
Strozzi	14	9	5
Barbadori	14	9	5
Tornabuoni	13	8	5
Castellani	18	5	13
Peruzzi	24	2	22
Pazzi	35	0	35
Ginori	28	0	28
Acciaiuoli	24	0	24
Lamberteschi	29	0	29
Pucci	0	0	0

from there, the world. So, removing the Medici doesn't just remove their direct contribution to the sum of betweennesses, but also Salviati's contribution.

Thus, when the graph invariant is defined as the sum of betweennesses, the induced centrality can be decomposed into two parts: the direct contribution of a node's own betweenness and the indirect contribution of increasing others' betweenness. Table 17.2 shows, for each node, the induced or total contribution, their direct betweenness, and the induced minus the direct, which we label "indirect." It is interesting to compare node profiles on their direct and indirect contributions. The Medici make a strong indirect contribution, but it is small relative to other nodes. Several nodes, from Pazzi down to Lamberteschi, make no direct contribution of their own, but make strong indirect contributions. That is, they have no betweenness themselves, but they enable other nodes to have betweenness. Pendants (nodes of degree 1) nearly always have this quality. Perhaps the most interesting node in the network is the Ridolfi family. They have a direct betweenness score of 10, but when removed from the network we find that their contribution to the sum of all betweenness is 9, meaning that their indirect contribution is actually negative. In other words, their presence reduces the betweenness of the nodes around them, unlike every other node in the graph.

There are a number of appealing things about induced centralities. One is that we needn't limit ourselves to measuring the centrality of individual nodes. We can do groups of nodes. For example, to measure the centrality of the marketing department in an organization, we can simply delete the whole department and see how much the graph invariant changes. Moreover, following Borgatti (2003, 2006), we can combine the induced centrality with a

0004829569.INDD 343 Dictionary: NOAD 6/15/2020 1:13:34 PM

OUP • UNCORRECTED PROOF

combinatorial optimization algorithm to identify optimal sets of nodes who make the most contribution—as an ensemble—to the graph invariant. Another advantage of induced centralities is that we can work with two-mode networks just as comfortably as one-mode networks. It simply doesn't matter what kind of network we calculate the graph invariant on. In the two-mode Davis, Gardner, and Gardner (1941) data, we can use this approach to measure the centrality of each person and each event, although we might still be interested in separate normalizations for each mode, à la Borgatti and Everett (1997). Yet another advantage is that induced centralities can be calculated on edges instead of nodes: we simply calculate the invariant, remove an edge, recalculate the invariant, and call the difference the centrality of that edge. It is its contribution to the network characteristic as a whole. Finally, we can generalize the induced centrality approach to make use of multiple graph invariants. For instance, suppose we have different invariants that we think capture different aspects of the cohesion of a graph. We want to measure each node's contribution to this multivariate construct of cohesion. An obvious approach is to take the sum of squared differences between the vector of invariants for the graph as a whole and the vector derived from the graph with the node removed.

Above all, though, the great benefit of induced centralities is that they can be generated on the fly, enabling researchers to use a measure of centrality that is precisely appropriate for a given research context. We needn't limit ourselves to off-the-shelf measures that may have been designed with assumptions and purposes very different from the current situation. At the same time, because induced centralities are all constructed in the same way, we needn't treat a measure based on a different invariant as an entirely new measure with unknown properties. In fact, we can think of induced centrality as a single but parameterized measure. In this it is like beta centrality, in which the user chooses the value of beta. In the case of induced centrality, the user-selected parameter is the graph invariant.

Following this thread of customizing measures, it is worth considering the "key player" line of thought presented in Borgatti (2003, 2006). These papers point out that when we consider specific reasons for detecting key players, we find that off-the-shelf measures are not optimal. Two classes of reasons or aims are considered, labeled KPP-Pos and KPP-Neg. KPP-Pos refers to the general aim of aiding a network by increasing cohesion and facilitating flows. KPP-Neg refers to the aim of fragmenting a network and preventing the nodes from coordinating or sharing. Consider the network in Figure 17.3. The most central node by all standard measures is node 1. Yet removing node 1 from the network would not fragment it. If that were the goal, node 8 would be the best choice, even though it is not the most central.

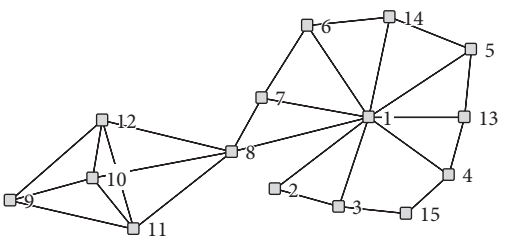

FIGURE 17.3 Node 1 is most central, but removing node 8 is more damaging.

0004829569.INDD 344 Dictionary: NOAD 6/15/2020 1:13:34 PM

OUP • UNCORRECTED PROOF

Table 17.3 Reasons for Identifying Key Players

DISRUPT	We want to remove them—to maximally disrupt the network.
ENHANCE	We want to help them—to make the network as a whole function better (diffuse info; coordinate well).
INFLUENCE	We want to identify key opinion leaders—to influence the network.
LEARN	We want to know who is in the know—so we can question or surveil them.
REDIRECT	We want to remove/prune them—to redirect flows in the network toward our preferred players.

We can extend this line of thinking by considering additional or more specific goals. Table 17.3 lists several possibilities. The "disrupt" and "enhance" aims correspond to KPP-Neg and KPP-Pos. The "influence" goal refers to identifying nodes who could be seeded with an attitude, belief, behavior, or virus that we want to spread maximally through the network. The "learn" goal is, in a sense, the inverse of the "influence" goal. We want to identify nodes that are likely to have received the most (non-redundant) flow from the network as a whole. Finally, the "redirect" goal is about pruning the network to redirect flows in the network to different nodes. In a police/defense setting this might be to redirect flows to nodes that we can more conveniently surveil. Alternatively, we might want to prune the network to make it more compact. For the sake of brevity we omit examples, but suffice it to say that, for each goal, we can construct networks where the standard measures of centrality are not as optimal as bespoke measures.

The Flow Outcomes Perspective

The first perspective we presented—involvement in the walk structure of the graph—was about the different kinds of traversals that could be made on a graph. At the time, we didn't ask what these traversals represented—what was traversing and why? Our third perspective on centrality takes this as its starting point.

Consider a coin moving through an economy. A person gives it to a store clerk to pay for gum. Later, the clerk gives it to another customer in change. The customer gives it to a friend who will use it in a vending machine. The coin keeps moving through the economy until eventually it is lost or taken out of service by the government. We might think of the coin as traversing a very large network. In reality, this is a bit of a stretch—in all likelihood, no tie existed between the person and the store clerk before the coin passed between them. It is more like the movement of the coin created the illusion of a tie on the spot. On the other hand, the coin can't just go anywhere from anywhere. While it is possible to mail a coin to a far-off land, in general the coin will pass from hand to hand, and when it is in the hands of a given individual, there is a limited set of others that have any probability of receiving it. Within that set, it may be exceedingly difficult to predict who it will go to, and so we might consider that it could go to any of them with uniform probability. So we might model the coin's movement as a random walk through a network and observe that at any given point in time there are millions of coins taking random walks through the same network.

0004829569.INDD 345 Dictionary: NOAD 6/15/2020 1:13:34 PM

OUP • UNCORRECTED PROOF

Viewed this way, it seems natural to pose a variety of questions about the traversals, which we refer to as flow outcomes. For example, when a coin reaches a given node, we might ask what the probability distribution is of the number of hops it took to get there. We might be especially interested in the mean of that distribution, that is, the expected number of hops from the mint to the first time it reaches a given person (known as mean first passage time or FPT in the Markov chain literature). Or we might ask how often we expect the coin to visit a given node. Similarly, we might ask the probability of reaching a node before reaching another (such as a terminal node indicating exit from the system, as when the coin is retired). Underlying these questions is an assumption that the answers will be different for different nodes, depending on their position in the network structure. This suggests conceiving of centrality measures as related to flow outcomes, and indeed perhaps the measures could be defined in terms of estimating flow outcomes. This is the essence of the flow perspective on centrality (Borgatti, 2005).

To start, let's consider a trivial example. Suppose we define the following flow regime, which we will call the package regime. In this regime, we require that things can only be in one place at one time, and that they only flow along geodesics (i.e., shortest paths). On average, how long does it take for something to go from a given node to any other, randomly chosen, node in the network? If we assume that it takes one unit of time to traverse each link, the answer is Freeman's closeness centrality divided by $N - 1$, which is to say, the average geodesic distance from the node to all others.

Let us continue with this particular flow regime and add one item. If there are multiple equally short paths from one node to another, the package chooses uniformly at random among them. Now suppose we run a simulation with every pair of nodes serving as a source and a target, and we count the number of times the package passes through each node. The expectation of this quantity is exactly betweenness centrality. This means that, across thousands of simulations, the average number of times something passes through a given node converges on its betweenness centrality.

The point here is that both the closeness and betweenness formulas give the expected values of two different flow outcomes in a network, given a particular flow regime. Closeness gives the expected time until first arrival, and betweenness gives the frequency of arrival.[8] What is immediately obvious when we take this perspective is how unrealistic the flow regimes underlying betweenness and closeness are. The concept of shortest paths involves both having a target and complete knowledge of the network. Neither of these is likely for most flows that we encounter. Consider, for example, the flow of information—such as gossip—through an organization. In transmitting the information, a given node might have a distal target in mind but has little control over whether or how the information gets there. A look at empirical gossip flows suggests a regime very different from the package regime that betweenness is consistent with. Gossip does not travel along shortest paths, nor is it limited to actual paths at all. Paths do not repeat nodes, but a bit of gossip can easily reach the same node multiple times. It is less likely, however, that node A tells node B the same piece of gossip more than once, suggesting graph-theoretic trails. Moreover, unlike the package regime, there's no reason to expect that gossip can only be in one place at one time. If I tell you something interesting, setting off a chain of transmission, I still have the gossip and can send it to someone else as well. Similar considerations apply to contagious diseases.

Table 17.4 presents a number of flow regimes that readily come to mind. One question that arises is what happens if we apply a measure consistent with one flow regime to a

0004829569.INDD 346 Dictionary: NOAD 6/15/2020 1:13:35 PM

OUP • UNCORRECTED PROOF

Table 17.4 A Sample of Flow Regimes

Name	Traversal	Contagion	Description
Used book	Trail	Move	I read a paperback. When finished, I pass it along to a friend, who does the same, until someone keeps it. A person might receive it twice but (normally) doesn't pass it to same person as before.
News/ gossip	Trail	Copy	I hear some gossip and pass it on. Then I tell someone else. Each node does the same but normally doesn't tell the same person multiple times.
Itinerant	Path	Move	I am an itinerant mathematician like Erdös. I live and work with you until your spouse kicks me out. Then I go to someone else's house. Each time, I burn my bridges and can't return.
Virus	Path	Copy	A virus infects a host, who then infects others. Infected people develop antibodies so they can't get it again.
Coin	Walk	Move	A coin moves through the economy until it is lost or retired. It can only be in one place at a time.
Attitude	Walk	Copy	My views affect yours, which in turn affects others. We are all continually affecting each other.
Travel	Geodesic	Move	People traveling by air to some destination prefer to make the fewest number of stops.

research context that is more consistent with a different flow regime. For example, if we are looking at the flow of information and calculate Freeman's betweenness centrality, how badly would its estimates be of the number of times a given bit of information flowed over a certain node? Borgatti (2005) provides an example where it does quite poorly. The data are the marriage ties among Florentine families shown in Figure 17.2. A simulation is programmed with the gossip regime and run thousands of times. The average number of times something flows over a node is calculated and compared with betweenness centrality. Both betweenness and the simulation give the Medici the highest score. But whereas betweenness puts the Strozzi family right in the middle of the pack, the gossip simulation puts the Strozzi in second place. This result seems more in keeping with historical reality in the sense that the Strozzi were the previously dominant family and still a major power at the time of the Medici rise.

Discussion

The three perspectives we have presented are clearly related. All are about traversals through a graph, and which ones tend to involve a given node. The walk structure perspective is perhaps the most agnostic. Yes, it claims that a node is more central to the extent it participates in the walk structure of the graph, but beyond that it is just a typology of measures based on the kinds of traversals the measures pay attention to, and what aspect of the traversals—number or length or both—that they record. The main benefit of this perspec-

0004829569.INDD 347 Dictionary: NOAD 6/15/2020 1:13:35 PM

OUP • UNCORRECTED PROOF

tive is that it enables us to put all centrality measures into a single common framework. It is the one characterization that leaves no measure out. It is also the only one that is means based rather than outcome based, meaning that we define measures in terms of how they are calculated and what kinds of math are used.

The second perspective takes from the first perspective the notion that centralities measure involvement in the network, but the focus on traversals is incidental. Rather, the focus is on a kind of goal—a graph invariant—and measuring the extent that a given node contributes to the goal. Most of the goals we can readily imagine have something to do with the cohesion of a network, such as the average distance among all pairs of nodes. (And most cohesion measures have something to do with the number and shortness of paths linking pairs.) It seems likely that the first perspective includes the second as a special case. That is, all induced centralities can be located in the typology of the first perspective, but not all measures discussed in the context of walk perspective can be derived as induced centralities. As a general principle, any measure that gives different values to points whose removal yields isomorphic residual graphs cannot be an induced centrality, since any induced centrality must give the same value to nodes with isomorphic neighborhoods.

The walk perspective is useful for imposing some order on the set of proposed centrality metrics but excludes none. As a result, we may define centrality measures that defy interpretation. For example, suppose we count the number of paths emanating from a node in which adjacent nodes alternate in degree, for example, a path like A—b—C—d—E, where lowercase letters indicate low-degree nodes and capital letters indicate high-degree nodes. It's a measure, and we can classify it, but we don't know what to do with it.

In contrast, induced measures carry at least one interpretation built into the measure: it is the extent to which each node contributes to some graph property, such as fragmentation. However, the meaningfulness of a measure depends on the meaningfulness of the graph invariant. If we choose a silly graph invariant, we will get a highly interpretable, but still silly, measure. For example, if the graph invariant is the number of nodes with degree 5, nodes with high centrality will be those that contribute to that property, and nodes with low centrality will be those that take away from that property.

The flow perspective takes a step beyond the induced perspective in ensuring meaningfulness. The flow perspective restricts measures to just those that represent a flow outcome, such as time of first arrival, frequency of arrival, probability of arrival as a function of time, and so on. An area for future research is to ask, as we did with induced centralities, whether every measure of centrality can be seen as measuring the outcome of a flow process whose rules merely need to be uncovered. Depending on how richly we define our flow regime, we can define more complicated outcomes such as the quality of whatever is flowing when it arrives, given that distortions are introduced along the way. If the observed signal equals true signal plus noise, how much of the true signal can be inferred by averaging the signal observed from many sources?

A key advantage of the flow approach is that we are modeling an actual process. To the extent the process matches the reality of some situation, we obtain a measure that not only ranks nodes but also estimates a known quantity, such as the number of unique bits of information each node can be expected to receive by a certain time. As discussed earlier in the context of betweenness, estimating such quantities from measures that are inconsistent with the actual social process at hand can give very poor results.[9] Moreover, through simulation, we can easily examine unique flow regimes that no one has studied before (and for

0004829569.INDD 348 Dictionary: NOAD 6/15/2020 1:13:35 PM

OUP • UNCORRECTED PROOF

which no measure has been constructed). Ideally, we would use simulation results to guide the development of a closed-form equation that would capture the concept exactly. This is unlikely in practice, as only measures based on shortest paths and unrestricted walks are computationally easy. But it really doesn't matter: the stable simulation results averaged over thousands of runs give us our measures directly. This may be costly with big data, but not nearly as costly as graph-theoretic measures that involve tracing all paths.

Final Note

A leitmotif in this chapter, particularly evident in the discussions of the induced and flow perspectives, is the construction of centrality measures that are appropriate to a given theoretical task and/or research design.[10] In the section on key player approaches, it was noted that off-the-shelf measures, such as betweenness or beta centrality, are almost always suboptimal for any given purpose, simply because they were not designed specifically for that purpose. The inadequacy is compounded exponentially when the job requires identifying a set of nodes working together. Thus, this chapter is very much a vote for bespoke measures constructed specifically for the study at hand. On the one hand, this may seem obvious. If you are studying a new construct such as cyber-bullying, you wouldn't use a survey from the 1950s about bullying in the playground. You might start with that, but you would customize the instrument for the job. On the other hand, we know (from working with countless workshop participants) that the main reason people choose a measure of centrality is that somebody famous in their field used that measure, indemnifying the present research against criticism. The second major reason is that they tried all the measures in their software package and used the one that happened to yield a significant result. This is obviously a bad approach because, with dozens of named measures on the market, the risk of spurious correlation is extremely high. Yet in an age in which researchers refer to their publications as "hits," both of these practices are extremely hard to argue against, especially in the case of young assistant professors looking for tenure. All we can say is that for a paper to have legs, it needs to have integrity. If you want your paper to be cited for years to come, use the right measure for the job.[11]

Acknowledgments

This paper began as a presentation to the 12th Applied Statistics International Conference (2015) in Ribno, Slovenia. We are grateful to the organizers for having given us the opportunity to elaborate these ideas.

Notes

1. As discussed in the section on network flows, we could devise a measure that contains a random element, perhaps calculated via simulation. In that case it is the expected value that must be the same for isomorphic nodes.
2. Although not strictly necessary, we typically use binary relations, meaning that the ties connect *pairs* of nodes, as opposed to triples, quads, etc. Graphs with nonbinary ties are

0004829569.INDD 349 Dictionary: NOAD 6/15/2020 1:13:35 PM

OUP • UNCORRECTED PROOF

called hypergraphs. Note that we use tie, line, link, and edge interchangeably, just as we treat nodes, vertices, points, and actors as synonyms.

3. Of course, there are also count measures that don't make use of length in any way. For example, we could just count the number of all paths or trails emanating from a node, although this would be computationally prohibitive.
4. Two paths are edge independent if they don't contain any of the same links. For example, the path A—B—C—D is not edge independent of A—E—B—C—F—D because the B—C link is used in both.
5. We can also take column sums. When $R(i,j)$ can be viewed as the amount of influence that i has on j, then the row sums are a measure of the total direct and indirect influence that i has. When $R(i,j)$ can be viewed as the amount of respect that i accords j, then the column sums can be viewed as a measure of the status of j.
6. Again, for simplicity, we consider only undirected graphs, so the term *endpoint* suffices and refers to either end of a traversal. But for directed graphs we would distinguish origins from termini.
7. Instead of the sum of betweenness scores, we can also think of this quantity as the sum of all entries in the Freeman dependency matrix (Freeman, 1980), which is also related to the Wiener index (Brandes, Borgatti, and Freeman, 2015).
8. More accurately, it is the frequency with which something passes through a node on its way somewhere else. If we add in the number of times that the node is the target of the traversal, then we must add N to betweenness.
9. However, there are some cases where extreme flow regimes, such as those calling for shortest paths, can be useful. For example, in an epidemiological setting, closeness centrality gives a kind of worst-case scenario for how quickly a sexually transmitted disease could reach a given node. This is certainly useful for planning.
10. Here we echo the call by Friedkin (1991), who argues for deriving centrality measures from a theoretical model of a given social process.
11. Note that in supporting customized measures, we are not necessarily calling for new standard measures to be added to the shelf. For example, if someone tweaks betweenness centrality for a specific application (e.g., by ignoring paths longer than k), there is no need to name it so-and-so centrality and publish it as a new measure. Part of the point of the walk structure perspective is to make family resemblances obvious and to attempt to discourage the need to name each of an infinite number of variations on the same concept. Similarly, if someone selects a novel graph invariant and bases an induced centrality on it, there is generally no need to name it and publish it as a standalone measure. The whole point of induced centrality is that it is a parameterized measure, the parameter being the choice of invariant.

References

Agneessens, F., Borgatti, S. P., & Everett, M. G. (2017). Geodesic based centrality: Unifying the local and the global. *Social Networks*, *49*, 12–26.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, *92*(5), 1170–1182.

Bonacich, P., & Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, *23*(3), 191–201.

Borgatti, S. P. (2003). The key player problem. In R. Breiger, K. Carley, & P. Pattison (Eds.), *Dynamic social network modeling and analysis: Workshop summary and papers* (pp. 241–252). Washington, DC: National Academy of Sciences Press.

0004829569.INDD 350 Dictionary: NOAD 6/15/2020 1:13:35 PM

OUP • UNCORRECTED PROOF

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71.
Borgatti, S. P. (2006). Identifying sets of key players in a network. *Computational, Mathematical and Organizational Theory*, *12*(1), 21–34.
Borgatti, S. P., & Everett, M. G. (1997). Network analysis of 2-mode data. *Social Networks*, *19*(3), 243–269.
Borgatti, S. P., & Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social Networks*, *28*(4), 466–484.
Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, *30*(2), 136–145.
Brandes, U., Borgatti, S. P., & Freeman, L. C. (2016). Maintaining the duality of closeness and betweenness centrality. *Social Networks*, *44*, 153–159.
Davis, A., Gardner, B. B., & Gardner, M. R. (1941). *Deep South: A social anthropological study of class and caste.* Chicago, IL: University of Chicago Press.
Everett, M. G., & Borgatti, S. P. (2010). Induced, endogenous and exogenous centrality. *Social Networks*, *32*(4), 339–344.
Everett, M. G., & Borgatti, S. P. (2014). Networks containing negative ties. *Social Networks, 38*, 111–120.
Ford, L. R., & Fulkerson, D. R. (1956). Maximal flow through a network. *Canadian Journal of Mathematics*, *8*(3), 399–404.
Freeman, L. C. (1978/1979). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.
Freeman, L. C. (1980). The gatekeeper, pair-dependency and structural centrality. *Quality and Quantity, 14*(4), 585–592.
Freeman, L. C., Borgatti, S. P., & White, D. R. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, *13*(2), 141–154.
Friedkin, N. E. (1991). Theoretical foundations for centrality measures. *American Journal of Sociology*, *96*(6), 1478–1504.
Hubbell, C. H. (1965). An input-output approach to clique identification. *Sociometry*, *28*(4), 377–399.
Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43.
Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). Centrality indices. In U. Brandes, *&* T. Erlebach, (Eds.), *Network Analysis: Methodological Foundations is* (pp. 16–61). Berlin: Springer Science & Business Media.
Markovsky, B., Willer, D., & Patton, T. (1988). Power relations in exchange networks. *American Sociological Review*, *53*, 220–236.
Newman, M. E. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, *27*(1), 39–54.
Padgett, J. F., & Ansell, C. K. (1993). Robust action and the rise of the Medici, 1400-1434. *American Journal of Sociology*, *98*(6), 1259–1319.
Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.
Schoch, D., & Brandes, U. (2016). Re-conceptualizing centrality in social networks. *European Journal of Applied Mathematics*, *27*(6), 971–985.
Smith, J., Halgin, D., Kidwell, V., Labianca, G., Brass, D., & Borgatti, SP. (2014). Power in politically charged networks. *Social Networks* 36: 162–176
Stephenson, K., & Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, *11*(1), 1–37.

0004829569.INDD 351 Dictionary: NOAD 6/15/2020 1:13:35 PM