An Introduction to Network Centrality

with Applications in R

David Schoch

2018-12-10

Contents

Preface **5**

1 Introduction **7**

2 Centrality Indices **9**

2.1 Degree 9
2.2 Betweenness (and variants) 9
2.3 Closeness (and variants) 11
2.4 Eigenvector (and variants) 13
2.5 Matrix exponential 14
2.6 Others 14
2.7 References 14

Preface

Chapter 1

Introduction

Chapter 2

Centrality Indices

The purpose of network centrality is to identify important actors or entities in a network. Structural importance is determined by so called *measures of centrality*, commonly defined in terms of indices $c : V \to \mathbb{R}$ interpreted as

$$c(u) > c(v) \iff u \text{ is more central than } v.$$

Since the meaning of structural importance is by no means unambiguous, a vast amount of different indices have been introduced over time. In addition, any mapping $c : V \to \mathbb{R}$ induces a ranking of nodes, but not every such ranking might represent a plausible concept of structural importance.

2.1 Degree

Degree centrality is the most simple form of a centrality index. It is defined as

$$c_d(u) = \{v : \{u, v\} \in E\} = |N(u)|$$

Degree centrality is a purely *local* measure since it only depends on the direct neighborhood of a node. A simple application example is popularity in friendship networks, i.e. "who has the most friends?". The definition of degree centrality can easily be adapted for directed and weighted networks. For directed networks,

$$c_d^+(u) = \{v : (u, v) \in E\} = |N^+(u)| \quad \text{and} \quad c_d^-(u) = \{v : (v, u) \in E\} = |N^-(u)|$$

are the *out-degree* and *in-degree*, respectively. *Weighted degree*, sometimes also refered to as *strength*, is defined as

$$c_{wd}(u) = \sum_{v \in N(u)} w_{uv}.$$

2.2 Betweenness (and variants)

Betweenness was independently developed by Anthonisse (1971) and Freeman (1977) and is an extension of *stress centrality*, introduced by Shimbel (1953). Shimbel assumes that the number of shortest paths containing a node u is an estimate for the amount of "stress" the node has to sustain in a network. The more shortest paths run through a node the more central it is. Formally, stress centrality is defined as

$$c_{stress}(u) = \sum_{s,t \in V} \sigma(s, t|u),$$

where $\sigma(s,t|u)$ is the number of shortest paths from s to t passing through u. Instead of the absolute number of shortest paths, betweenness centrality quantifies their relative number. The relative count is given by

$$\delta(s,t|u) = \frac{\sigma(s,t|u)}{\sigma(s,t)},$$

where $\sigma(s,t)$ is the total number of shortest paths connecting s and t. The expression $\delta(s,t|u)$ can be interpreted as the extent to which u controls the communication between s and t. Betweenness is then defined as

$$c_b(u) = \sum_{s,t\in V} \delta(s,t|u)$$

The interpretation of betweenness is not only restricted to communication. More generally, betweenness quantifies the influence of vertices on the trans- fer of items or information through the network with the assumption that it follows shortest paths. Many different variants of shortest path betweenness have been proposed to incorporate additional assumptions, e.g. the specific lo- cation of a vertex u on a shortest s,t-path or its length. Some of these variants are given in the following table.

proximal source	$c_{bs}(u) = \sum_{s,t\in V} \delta(s,t|u) \cdot A_{su}$
proximal target	$c_{bt}(u) = \sum_{s,t\in V} \delta(s,t|u) \cdot A_{ut}$
k-bounded distance	$c_{bk}(u) = \sum_{dist(s,t)\leq k} \delta(s,t|u)$
length-scaled	$c_{bd}(u) = \sum_{s,t\in V} \frac{\delta(s,t|u)}{dist(s,t)}$
linearly-scaled	$c_{bl}(u) = \sum_{s,t\in V} \delta(s,t|u) \frac{dist(s,u)}{dist(s,t)}$

Details on these variants can be found in Brandes (2008). Other variants of the betweenness concept rely on different assumptions of transfer in networks besides shortest paths.

A measure based on network flow was defined by Freeman et al. (1991). The authors assume information as a flow process and assign to each edge a non-negative value representing the maximum amount of information that can be passed between the endpoints. The aim is then to measure the extent to which the maximum flow between two vertices s and t depends on a vertex u. Denote by $f(s,t)$ the maximum (s,t)-flow w.r.t. constraints imposed by edge capacities and the amount of flow which must go through u by $f(s,t|u)$. Similar to shortest path betweenness, *flow betweenness* is then defined as

$$c_f(u) = \sum_{s,t\in V} \frac{f(s,t|u)}{f(s,t)}.$$

The index was introduced as a betweenness variant for weighted networks but it can easily applied to unweighted ones. In the case of simple undirected and unweighted networks, the maximum (s,t)-flow is equivalent to the number of edge disjoint (s,t)-paths and $f(s,t|u)$ is the number of paths u lies on.

A variant based on walks instead of paths was proposed by Newman (2005). His *random walk betweenness* calculates the expected number of times a random (s,t)-walk passes through a vertex u, averaged over all s and t. Newman shows, that his variant of betweenness can also be calculated with a current-flow analogy by viewing a graph as an electrical network. Random walk betweenness is then equiva- lent to the amount of current that flows through u. The index is thus also known as *current flow betweenness*. Details and formal definitions of his versions can be found in the literature (Newman, 2005; Brandes and Fleischer, 2005).

Two variants based on the *randomzed shortest path* (RSP) framework (Yen et al., 2008; Saerens et al., 2009; Kivimäki et al., 2014) were proposed by Kivimäki et al. (2016). The variants are refered to as *simple RSP betweenness* and *RSP net betweenness*. The derivation of both indices is rather involved and goes beyond the scope of this tutorial. The interested reader should consult the original work for details. One aspect worth mentioning, though, is that both variants include a tuning parameter β. Both variants converge to the traditional version of betweennes for $\beta \to \infty$. For $\beta \to 0$, simple RSP betweenness converges to the expected number of visits to a node over all absorbing walks with respect to the unbiased random walk probabilities. This means for undirected networks, that the index converges to degree. RSP net betweenness, on the other hand, converges to Newmann's random walk betweenness.

All variants of betweenness can be described in a more general form considering a flow of information analogy. Depending on the assumption of how information is 'flowing' between s and t, the set $P(s,t)$ contains all possible information channels to transmit the piece of information. This set might contain all shortest (s,t)-paths if the information has to be trans- mitted as fast as possible or all random (s,t)-walks when the delivery time does not matter. In principle, any kind of trajectory on a graph can be thought of as an information channel. The set $P(s,t|u)$ contains all information channels where the vertex u is in a position to control the information flow. For shortest path betweenness, u is in a controlling position if it is part of an information channel and for proximal target betweenness if it presents the information to the target t. In the former case $P(s,t|u)$ comprises all elements of $P(s,t)$ that contain u as an intermediary and in the latter all elements that contain the edge (u,t) . Again, the position of control could be defined as any location on a trajectory. A measure of relative betweenness is then defined with aggregation rules over the two specified sets, commonly the fraction of their cardinalities. This fraction can also be weighted according to specified rules, e.g. as in length scaled betweenness. Aggregating over all possible sources and targets, we can thus define a *generic betweenness* index as

$$c_{bg}(u) = \sum_{s,t \in V} \frac{|P(s,t|u)|}{|P(s,t)|} \cdot w(s,t).$$

Thus, many other variants are possible, for instance also k-betweenness mentioned by Borgatti and Everett (2006), where $P(s,t)$ is the set of all (s,t)-paths of length at most k.

There exist many more betweenness-like indices that where not covered here, but are listed below in no particular order:

- *communicability betweenness* based on the matrix exponential (Estrada et al., 2009)
- α and β betweenness, which are closely related to current flow betweenness (Avrachenkov et al., 2013, 2015),
- *ranking betweenness*, which combines betweenness with the idea of PageRank (Agryzkov et al., 2014)
- range-limited forms of betweenness (Ercsey-Ravasz et al., 2012)
- *bridgeness* (Jensen et al., 2016)
- *super mediator* (Saito et al., 2016)
- *BridgeRank* (Salavati et al., 2018)

2.3 Closeness (and variants)

Closeness centrality was first mentioned by Bavelas (1950) and later formally defined by Sabidussi (1966). Closeness is defined as the reciprocal of the sum of distances from a node to all other nodes in the network, that is

$$c_c(u) = \frac{1}{\sum\limits_{t \in V} dist(u,t)}.$$

Vertices in a network are considered to be central if they have a small total distance to all other vertices in the network. By definition of graph-theoretic distances, closeness is ill-defined on unconnected graphs. A close variant ap- plicable to both connected and unconnected graphs is given by

$$c_{hc}(u) = \sum_{t \in V} \frac{1}{dist(u,t)}.$$

This variant was proposed by various researcher. Among the first are Gil-Mendieta and Schmidt (1996) who refer to it as *power index*. Rochat (2009) later introduced it as *harmonic closeness*.

As in the case of betweenness, many different variations of closeness have been proposed, mostly to correct for the fact that the "classical" closeness is not properly defined on unconnected networks. Moxley and Moxley (1974) introduced an index which penalizes the number of not reachable nodes. Their *adjusted index of centrality* is defined as

$$c_{aic}(u) = \frac{\sum_{s \in V} \sum_{t \in V} dist(s,t) + \rho n_s}{\sum_{t \in V} dist(u,t) + \rho n_u},$$

where n_u are the number of non-reachable nodes by u and ρ is a penalizing factor.

vf-irmeicrn-98 introduce *integration* as an index which evaluates how well a vertex is integrated in a network. It is defined as

$$c_{int}(u) = \frac{\sum_{t \in V} (diam(G) + 1 - dist(u,t))}{n-1},$$

where $diam(G)$ is the diameter of the network.

Dangalchev (2006) suggests

$$c_{rc} = \sum_{t \in V} \frac{1}{2^{dist(u,t)}}$$

as another variant.

There also exist at least two parametrized versions of closeness. Jackson (2010) introduced *decay centrality* as

$$c_{dc} = \sum_{t \in V} \alpha^{dist(u,t)}$$

where $\alpha \in (0,1)$. *Generalized closeness*, proposed by Agneessens et al. (2017), is parametrized in a slightly different way. Formally,

$$c_{gc}(u) = \sum_{t \in V} dist(u,t)^{-\alpha},$$

where $\alpha \geq 0$. The index apporaches classic closeness for $\alpha \to 1$ and converges to degree for $\alpha \to \infty$.

Hage and Harary (1995) introduced *eccentricity*, which does not rely on summing up all distances, but simply taking the inverse of the maximum, that is

$$c_{ec} = \frac{1}{\max\{dist(u,t) : t \in V\}}.$$

While distances in networks are commonly defined via shortest paths, other concepts, such as *random walks* (Noh and Rieger, 2004), have also been used to design closeness-like indices. An index of particular interest is *information centrality*, proposed by Stephenson and Zelen (1989). The index is based on counting all paths between two vertices and the edge overlap among these paths. Afterwards, a matrix is formed that contains the lengths of all paths on the diagonal and the overlap on the off diagonal entries. This matrix is inverted and a harmonic mean of each row is formed. The authors interpret this procedure from an information-theoretic point view. They argue that the information content of a path is inversely propor- tional to the length of a path and the edge overlap represents a covariance among paths. Note that these calculations do not have to be performed explicitly but can be derived by inverting a matrix

$$C = (L + J)^1,$$

where L is the Laplacian matrix and J the matrix of all ones. With the matrix C, information centrality equates to

$$c_{inf}(u) = \left(C_{uu} + \frac{T - 2R}{n}\right)^{-1},$$

where

$$T = \sum_{v \in V} C_{vv} \quad \text{and} \quad R = \sum_{v \in V} C_{uv}.$$

Note that T and R are the same for all vertices, such that the induced ranking only depends on $C_u u$.

2.4 Eigenvector (and variants)

Eigenvector centrality was introduced by Bonacich (1972), yet earlier versions can already be found in Wei (1952) and Berge (1958). The index is part of the class of *feedback centralities.* Measures in this class assume that the centrality of a node is conditional on the centrality of its neighbors. Nodes are highly central if they are connected to other highly central nodes. If we define the centrality of a vertex as the sum of the centrality scores of its adjacent vertices, we obtain

$$c_e(u) = \sum_{t \in V} A_{ut} c_e(u).$$

The resulting system of equation by considering the eigenvalue problem $\lambda_1 x = Ax$. eigenvector centrality is then given by the eigenvector x_1 associated with the principal eigenvalue λ_1. Note that the principal eigenvector can also be computed with the power iteration by repeatedly multiplying A to an arbitrary vector b_0 until convergence, i.e.

$$\lim_{k \to \infty} \frac{A^k b_0}{\|A^k b_0\|} = x_1$$

Since an entry A^k_{uv} is the number of (u, v)-walks of length k, eigenvector centrality of a node u can also be seen as the limit proportion of walks starting at u.

Google's *PageRank* is undoubtedly one of the most famous adaptations of eigenvector centrality for directed graphs (Page et al., 1999). It is defined as

$$c_{pr}(u) = \sum_{v \in N^-(u)} \alpha \frac{c_{pr}(v)}{c_d^+(v)} + (1 - \alpha)$$

where α is damping factor, commonly set to 0.85. Note that an equivalent index was already introduced earlier by Friedkin and Johnsen (1990) (see Friedkin and Johnsen (2014) for a discussion).

A feedback centrality dating back to 1953 was introduced by Katz (1953). Similarly to eigenvector centrality, all walks emanating from a node u are summed up but longer walks are penalized by an attenuation factor α. For- mally, *Katz status* is defined as

$$c_{katz}(u) = \sum_{k=0}^{\infty} \sum_{t \in V} \alpha^k A^k_{ut}.$$

In order for the series to converge, α has to be chosen such that it is smaller than the reciprocal of the largest eigenvalue of A. In this case, Katz status can be calculated with the closed form

$$c_{katz}(u) = \left((I - \alpha A)^{-1} \cdot \mathbf{1}_n\right)_u,$$

where $\mathbf{1}_n$ is the all one vector of length n. A close variant is Bonacich's β-centrality, whose definition also allows for a negative attenuation factor β (Bonacich, 1972). It is given by

$$c_{\alpha,\beta}(u) = \alpha \sum_{k=1}^{\infty} \sum_{t \in V} \beta^k A_{ut}^{k-1},$$

where α is merely a scaling factor, such that it can be omitted without altering the induced ranking. With $|\beta| \leq \frac{1}{\lambda_1}$, a closed form is given by

$$c_{\alpha,\beta}(u) = \left((I - \beta A)^{-1} A \cdot \mathbf{1}_n\right)_u.$$

Katz status and eigenvector centrality can be considered as positive feedback centralities, since the centrality of a vertex is higher if it is connected to other vertices with a high centrality score. In contrast, Bonacich's -centrality with a negative is a negative feedback centrality, since vertices are considered central, if they are connected to vertices with low centrality score. This kind of centrality is particularly of interest in bargaining situations since bargaining power comes from being in a better position than negotiating partners.

2.5 Matrix exponential

There exist a variety of indices which are based on the matrix exponential $\exp(A)$ of the adjacency matrix. The first of this kind, *subgraph centrality*, was introduced by Estrada and Rodríguez-Velázquez (2005). It is closely related to eigenvector centrality and Katz status, since it also involves counting walks. The difference is that only closed walks are considered and longer walks are inversely weighted by the factorial of their length, i.e.

$$c_s(u) = \sum_{k=0}^{\infty} \frac{A_{uu}^k}{k!}.$$

The weighting by factorials is a convenient choice since it guarantees conver- gence of the series. Its closed form is given by the matrix exponential, such that

$$c_s(u) = \exp(A)_{uu}.$$

The authors also consider variants, where only walks of even or odd length are considered, giving rise to *even* and *odd subgraph centrality* defined as

$$c_{se}(u) = \sum_{k=0}^{\infty} \frac{A_{uu}^{2k}}{(2k)!} \quad \text{and} \quad c_{so}(u) = \sum_{k=0}^{\infty} \frac{A_{uu}^{2k+1}}{(2k+1)!}.$$

All three indices can also be expressed with the spectral decomposition of A:

$$c_s(u) = \sum_{k=0}^{n} \exp(\lambda_j) x_j^2(u)$$

$$c_{se}(u) = \sum_{k=0}^{n} \cosh(\lambda_j) x_j^2(u)$$

$$c_{so}(u) = \sum_{k=0}^{n} \sinh(\lambda_j) x_j^2(u),$$

where $x_j(u)$ is the uth entry of the eigenvector x_j associated with the eigen- value λ_j.

2.6 Others

2.7 References

Bibliography

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

Agryzkov, T., Oliver, J. L., Tortosa, L., and Vicent, J. (2014). A new betweenness centrality measure based on an algorithm for ranking the nodes of a network. *Applied Mathematics and Computation*, 244(Supplement C):467–478.

Anthonisse, J. (1971). The rush in a directed graph. *Stichting Mathematisch Centrum. Mathematische Besliskunde*, BN 9/71.

Avrachenkov, K., Litvak, N., Medyanikov, V., and Sokol, M. (2013). Alpha current flow betweenness centrality. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 106–117. Springer.

Avrachenkov, K. E., Mazalov, V. V., and Tsynguev, B. T. (2015). Beta Current Flow Centrality for Weighted Networks. In Thai, M. T., Nguyen, N. P., and Shen, H., editors, *Computational Social Networks*, Lecture Notes in Computer Science, pages 216–227. Springer International Publishing.

Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Berge, C. (1958). *Théorie Des Graphes et Ses Applications*. Dunod, Paris.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1):113–120.

Borgatti, S. P. and 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. and Fleischer, D. (2005). Centrality Measures Based on Current Flow. In *STACS*, volume 3404, pages 533–544. Springer.

Dangalchev, C. (2006). Residual closeness in networks. *Physica A: Statistical Mechanics and its Applications*, 365(2):556–564.

Ercsey-Ravasz, M., Lichtenwalter, R. N., Chawla, N. V., and Toroczkai, Z. (2012). Range-limited centrality measures in complex networks. *Physical Review E*, 85(6):066103.

Estrada, E., Higham, D. J., and Hatano, N. (2009). Communicability betweenness in complex networks. *Physica A: Statistical Mechanics and its Applications*, 388(5):764–774.

Estrada, E. and Rodríguez-Velázquez, J. A. (2005). Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41.

Freeman, L. C., Borgatti, S. P., and 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. and Johnsen, E. C. (1990). Social influence and opinions. *The Journal of Mathematical Sociology*, 15(3-4):193–206.

Friedkin, N. E. and Johnsen, E. C. (2014). Two steps to obfuscation. *Social Networks*, 39:12–13.

Gil-Mendieta, J. and Schmidt, S. (1996). The political network in Mexico. *Social Networks*, 18(4):355–381.

Hage, P. and Harary, F. (1995). Eccentricity and centrality in networks. *Social Networks*, 17(1):57–63.

Jackson, M. O. (2010). *Social and Economic Networks*. Princeton university press.

Jensen, P., Morini, M., Karsai, M., Venturini, T., Vespignani, A., Jacomy, M., Cointet, J.-P., Mercklé, P., and Fleury, E. (2016). Detecting global bridges in networks. *Journal of Complex Networks*, 4(3):319–329.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kivimäki, I., Lebichot, B., Saramäki, J., and Saerens, M. (2016). Two betweenness centrality measures based on Randomized Shortest Paths. *Scientific Reports*, 6.

Kivimäki, I., Shimbo, M., and Saerens, M. (2014). Developments in the theory of randomized shortest paths with a comparison of graph node distances. *Physica A: Statistical Mechanics and its Applications*, 393(Supplement C):600–616.

Moxley, R. L. and Moxley, N. F. (1974). Determining Point-Centrality in Uncontrived Social Networks. *Sociometry*, 37(1):122–130.

Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54.

Noh, J. D. and Rieger, H. (2004). Random Walks on Complex Networks. *Physical Review Letters*, 92(11).

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Techreport.

Rochat, Y. (2009). Closeness centrality extended to unconnected graphs: The harmonic centrality index. Technical report.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.

Saerens, M., Achbany, Y., Fouss, F., and Yen, L. (2009). Randomized Shortest-Path Problems: Two Related Models. *Neural Computation*, 21(8):2363–2404.

Saito, K., Kimura, M., Ohara, K., and Motoda, H. (2016). Super mediator – A new centrality measure of node importance for information diffusion over social network. *Information Sciences*, 329(Supplement C):985–1000.

Salavati, C., Abdollahpouri, A., and Manbari, Z. (2018). BridgeRank: A novel fast centrality measure based on local structure of the network. *Physica A: Statistical Mechanics and its Applications*, 496:635–653.

Shimbel, A. (1953). Structural parameters of communication networks. *The bulletin of mathematical biophysics*, 15(4):501–507.

Stephenson, K. and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37.

Wei, T.-H. (1952). *The Algebraic Foundations of Ranking Theory*. PhD thesis, University of Cambridge.

Yen, L., Saerens, M., Mantrach, A., and Shimbo, M. (2008). A Family of Dissimilarity Measures Between Nodes Generalizing Both the Shortest-path and the Commute-time Distances. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 785–793, New York, NY, USA. ACM.