ASNA 2009
Zürich

Closeness Centrality Extended To Unconnected Graphs : The Harmonic Centrality Index

Yannick Rochat[1]

Institute of Applied Mathematics
University of Lausanne, Switzerland

UNIL | Université de Lausanne
Institut de mathématiques appliquées

[1]yannick.rochat@unil.ch

Abstract

Social network analysis is a rapid expanding interdisciplinary field, growing from work of sociologists, physicists, historians, mathematicians, political scientists, etc. Some methods have been commonly accepted in spite of defects, perhaps because of the rareness of synthetic work like (Freeman, 1978; Faust & Wasserman, 1992). In this article, we propose an alternative index of closeness centrality defined on undirected networks. We show that results from its computation on real cases are identical to those of the closeness centrality index, with same computational complexity and we give some interpretations. An important property is its use in the case of unconnected networks.

1 Introduction

The study of centrality is one of the most popular subject in the analysis of social networks. Determining the role of an individual within a society, its influence or the flows of information on which he can intervene are examples of applications of centrality indices. They are defined at an actor-level and are expected to compare and better understand roles of each individual in the network. In addition, a graph-level index called centralization (i.e. how much the index value of the most central node is bigger than the others) is defined for each existing index .

Each index provides a way to highlight properties of individuals, dependently on its definition. For example, degree centrality attributes high measure to an individual having great influence on its neighbors. Closeness centrality highlights the players who will be able to contact easily all other members of the network. Betweenness centrality gives highest values to individuals through whom information is more likely to pass.

In 1978, Linton C. Freeman wrote a seminal article reviewing three types of centrality (Freeman, 1978). After a short review about centrality, he outlines three indices, interpretable each in a different way, and presented in an elegant form: the index belongs to the interval $[0, 1]$, a value close to 1 signifying high centrality, a value close to 0 signifying low centrality and meaning that the actor plays an accessory role at a graph-level. The index values can be compared between actors of the same graph, or between distinct graphs (which may provide an intuition about the importance of an agent, but often is not exploitable in-between graphs of different types).

Degree centrality is computed by counting the neighbors of each vertex. It is given by

$$c_D(x_i) = \frac{\deg(x_i)}{n-1} \tag{1}$$

with $x_i \in V$, V the set of nodes, $n = |V|$ and $\deg(x_i)$ the degree of node x_i. Any attempt to modify or improve this definition will certainly make it more complex.

Closeness centrality sums distances from a vertex to each other. It is defined as

$$c_C(x_i) = \frac{n-1}{\sum_{j \neq i} \text{dist}(x_i, x_j)} \tag{2}$$

with $x_i \in V$, $n = |V|$ and $\text{dist}(x_i, x_j)$ the distance from node x_i to node x_j. There are algorithms to optimize the calculation of this index implying some approximations (Eppstein & Wang, 2004). Some variants have been proposed (Newman, 2003; Csardi & Nepusz, 2006; Butts, 2009). We will focus on them

later in this document.

Betweenness centrality of a node is proportional to the number of occurrences of itself on all geodesics of the graph. The calculation is made as follows:

$$c_B(i) = \frac{2\sum\sum \frac{g_{jk}(i)}{g_{jk}}}{(n-1)(n-2)} \tag{3}$$

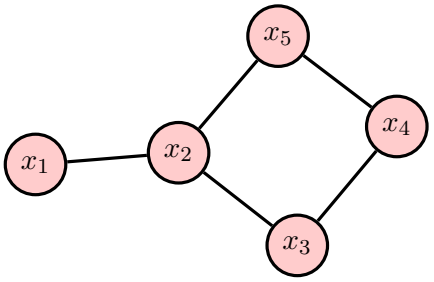
with g_{jk} the number of geodesics from node x_j to node x_k, $g_{jk}(i)$ the number of geodesic from x_j to x_k containing x_i, the double sum being calculated on all pairs (j,k) such that $j \neq i \neq k$ and $j < k$. Ulrik Brandes studied it in depth and provided an algorithm reducing computation time along with some variations (Brandes *et al.* , 2006; Brandes, 2008).

Since then, these three indices have been discussed, others have been proposed and some have become very popular, as Bonacich centrality index, which uses eigenvectors (Bonacich, 1987). More than thirty years later, the three previous indices are still accepted as the norm and in one case at least, some improvements can be done.

Without any increase of computational complexity, this article proposes an alternative to the index of centrality of proximity - the index of centrality harmonic - giving comparable results (ranks are mostly the same) and a possible interpretation on unconnected graphs unlike closeness centrality. In section 2 we define and explain how to compute the harmonic index. In section 3 and 4 we study its behavior in comparison with that of the closeness centrality. Finally in section 5 we give an interpretation of the unconnected case, show some problematic cases and discuss its general interpretation.

2 Classical and other closeness centrality indices

Computation for every node of the closeness centrality index (eq. 2) needs the distances between all pairs of vertices. In the case of graph $\mathcal{G}_1$ (see figure 1a), the geodesic distances between vertices are reported in table 1b.



(a) $\mathcal{G}_1$: undirected 5-nodes graph.

	x_1	x_2	x_3	x_4	x_5
x_1	X	1	2	3	2
x_2	1	X	1	2	1
x_3	2	1	X	1	2
x_4	3	2	1	X	1
x_5	2	1	2	1	X

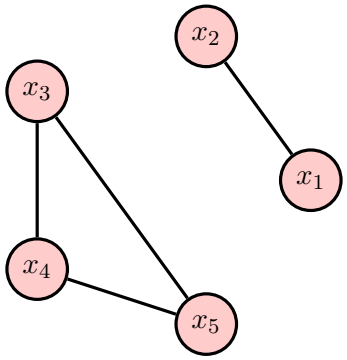
(b) d_{ij} : distance matrix of graph $\mathcal{G}_1$ (fig. 1a).

Figure 1: Example of a connected graph.

We compute the index of node x_1 as an example:

$$c_C(x_1) = \frac{n-1}{\sum_{i \neq 1} d_{i1}} = \frac{5-1}{1+2+3+2} = \frac{1}{2}.$$

Unfortunately, when the graph is unconnected like in fig. (2a), closeness centrality appears to be useless because the distance between two vertices belonging to different components is infinite by convention, which makes the sum in 2 infinite too and therefore its inverse equal to zero. For every vertex of such a graph, there is always another vertex belonging to another component: indices of all vertices of the graph are therefore useless and the calculation of the index is limited to the largest component, omitting the roles played by individuals of other components. But what can we conclude from calculation if the size of the largest connected component is not significantly greater than the second? Degree and betweenness centralities can always be calculated: how can we correct the closeness centrality definition in order to be able to compute it on every kind of network?



(a) $\mathcal{G}_2$: undirected 5-nodes graph.

	x_1	x_2	x_3	x_4	x_5
x_1	X	1	∞	∞	∞
x_2	1	X	∞	∞	∞
x_3	∞	∞	X	1	1
x_4	∞	∞	1	X	1
x_5	∞	∞	1	1	X

(b) d_{ij} : distance matrix of graph $\mathcal{G}_2$ (fig. 2a).

Figure 2: Example of an unconnected graph: sum of the elements of any line or row of matrix (2b) is always infinite.

Some alternatives exist. In (Csardi & Nepusz, 2006), they propose to replace the infinite distance between two vertices belonging to two distinct components by the number of vertices of the graph: the largest geodesic possible in a graph with n vertices is of length $n-1$ (a chain-graph). Hence closeness centrality can be generalized to the formula

$$c_\alpha(x_i) = \frac{n-1}{\sum_{i \neq j} \text{dist}(x_i, x_j) + m\alpha} \tag{4}$$

with vertices $\{x_j\}_j$ chosen in the same connected component as the vertex x_i , $n = |V|$, m the number of vertices unconnected to x_i and $\alpha \in \mathbb{R}_+$ a constant greater than or equal to the diameter of graph.

The innovation proposed by this article is the index of harmonic centrality, briefly described in (Newman, 2003; Butts, 2009) and defined as the sum of the inverted distances instead of the inverted sum of the distances.

$$\sum_{i \neq j} \frac{1}{\text{dist}(x_i, x_j)} \tag{5}$$

The use of the harmonic mean avoid cases where an infinite distance outweighs the others. The index is normalized by noting that on a star graph, the maximum is obtained by the node in the center and is $|V| - 1$. Thus the index of harmonic centrality is defined by

$$c_H(x_i) = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{\text{dist}(x_i, x_j)}. \tag{6}$$

Let's see two examples of calculation, with nodes x_1 and x_3 from $\mathcal{G}_2$:

$$\begin{aligned} c_H(x_1) &= \frac{1}{5-1} \sum_{i \neq 1} \frac{1}{\text{dist}(x_1, x_i)} = \frac{1}{4}\big(\frac{1}{1} + \frac{1}{\infty} + \frac{1}{\infty} + \frac{1}{\infty}\big) = \frac{1}{4} \\ c_H(x_3) &= \frac{1}{5-1} \sum_{i \neq 3} \frac{1}{\text{dist}(x_3, x_i)} = \frac{1}{4}\big(\frac{1}{\infty} + \frac{1}{\infty} + \frac{1}{1} + \frac{1}{1}\big) = \frac{1}{2}. \end{aligned} \tag{7}$$

One notices on this small example that the number of vertices belonging to the same component as the computed vertex increases its harmonic centrality index, because these are all non-zero in the calculation. Thus, the index attaches greater importance to well-connected vertices. Moreover, the index appears to have higher probability of reaching lower values when computed on an unconnected graph, reflecting the inability to communicate between individuals of different components (note that it is far from being systematically the case when comparing two graphs of the same size, one connected and the other not).

The maximum value of the harmonic centralization is defined through the study of the index on a star graph. The index of the node in the center is 1 and of the leaves is $\frac{1}{n-1}(\frac{1}{1} + (n-2)\frac{1}{2}) = \frac{n}{2(n-1)}$. Therefore centralization index of harmonic centrality is

$$C_H = \frac{2(n-1) \sum_i \big(c_H^* - c_H(x_i)\big)}{n} \tag{8}$$

with $c_H^* = \max_j c_H(x_j)$ and $n = |V|$.

3 Methods

In order to study the behavior of closeness and harmonic centralities, we use three types of networks: random graphs generated with Erdös-Rényi model (Erdös & Rényi, 1959; Erdös & Rényi, 1960), scale-free graphs generated with Barabasi-Albert model (Barabasi & Albert, 1999) and some real networks. In each case, we compute the two indices on every node and then compare the ranks inducted by them. We decided to use Spearman's correlation ρ which is appropriate to this decision.

During the simulations, we generate one hundred graphs, determine each ρ and then compute their mean and standard deviation. With the third category of networks, we only calculate ρ. Generating a random graph according to Erdös-Rényi model begins with a set of unconnected vertices, then for each pair of vertices an edge is created with some fixed probability (chosen values: see table 1). Concerning the scale-free networks, the generating method we use begins with a complete graph of order equal to the number of edges it has been decided to create at each time step. Then, a vertex is added at each time step and edges are drawn with preferential attachment (Dorogovtsev & Mendes, 2003): nodes have a probability proportional to their degree of being connected to the new node.

Those choices may be too naive or restricted: in a near future, we hope to make measurements on assortative (i.e. positive correlation of degrees) networks (Newman & Park, 2003; Xulvi-Brunet & Sokolov, 2004), a common property of social networks.

Following networks are studied and appear in table 3:

- Padgett's **florentine** families (Kent, 1978; Wasserman & Faust, 1994). We use its biggest component composed of 15 nodes and 20 edges (see figure 3).

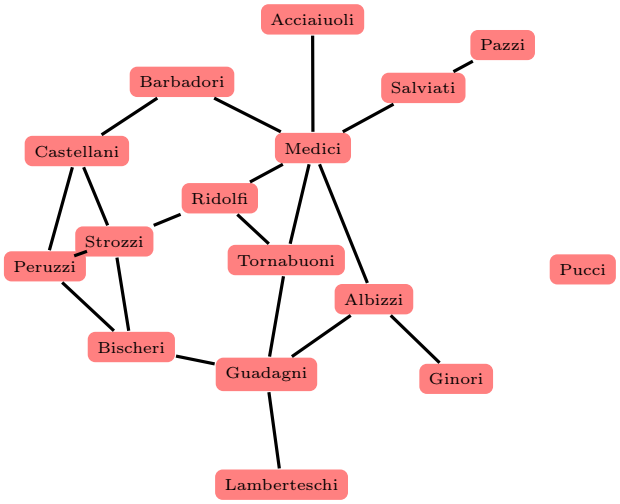


Figure 3: Padgett's florentine families.

- A network of relation between **dolphins** in doubtful Sound, New-Zealand (Lusseau *et al.* , 2003). The graph is connected and owns 62 nodes and 159 edges.

- "*A* ***coauthorship*** *network of scientists working on network theory and experiment*" (Newman, 2006) compiled by Mark Newman, unconnected and made of 1589 nodes and 2742 edges. Like earlier, we only use the biggest component, which owns 379 nodes and 914 edges. Note that this network was compiled from only two sources: two literature reviews of the field (see figure 4).

- A network of friendships in a **karate** club after a scission (Zachary, 1977). 34 nodes and 78 edges.

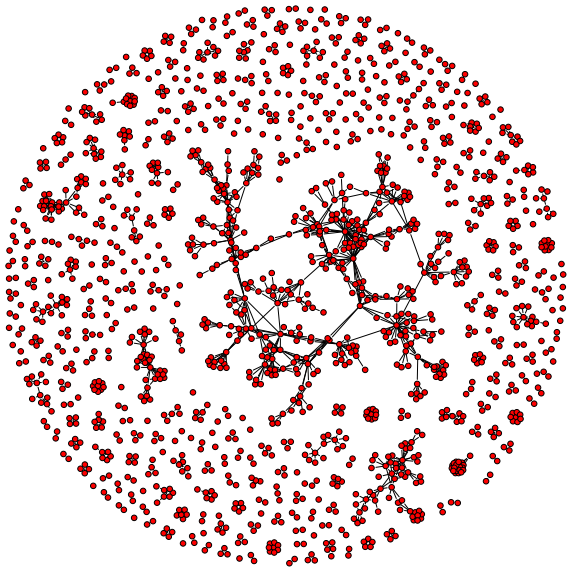


Figure 4: The full network of scientists: $|V| = 1589$, $|E| = 2742$. The giant component is easily discernible.

4 Results

In table 1 are shown the results obtained with random graphs along with some properties of the graphs. In table 2 are shown the results corresponding to scale-free graphs and also some properties. Mean and standard deviation are computed among the ρ's of all the graphs generated.

$\|V\|$	p	$\|E\|$	mean	sd
100	0.1	990	0.9946	0.0013
100	0.2	1980	0.9924	0.0027
100	0.3	2970	0.9999	8.56×10^{-5}
1000	0.01	99900	0.9993	4.68×10^{-5}
1000	0.05	499500	0.9972	0.0002
1000	0.1	999000	0.9999	1.56×10^{-5}

Table 1: Results for simulations of 100 random graphs in each case.

In table 3 are reported ρ's computed on the four graphs presented earlier.

$\|V\|$	$\|E\|$	edges added	mean	sd
100	99	1	0.9812	0.0161
100	198	2	0.9899	0.0024
1000	999	1	0.9933	0.0059
1000	1998	2	0.9988	0.0002
1000	2997	3	0.9989582	9.76×10^{-5}

Table 2: Results for simulations of 100 scale-free graphs in each case.

network	ρ
florentine	0.9514388
dolphins	0.9380984
coauthorship	0.9612777
karate	0.953108

Table 3: Correlations of closeness and harmonic centrality measures.

5 Discussion

Correlation coefficients from tables 1, 2 and 3 are close to 1. In those cases, we conclude that both indices behave the same.

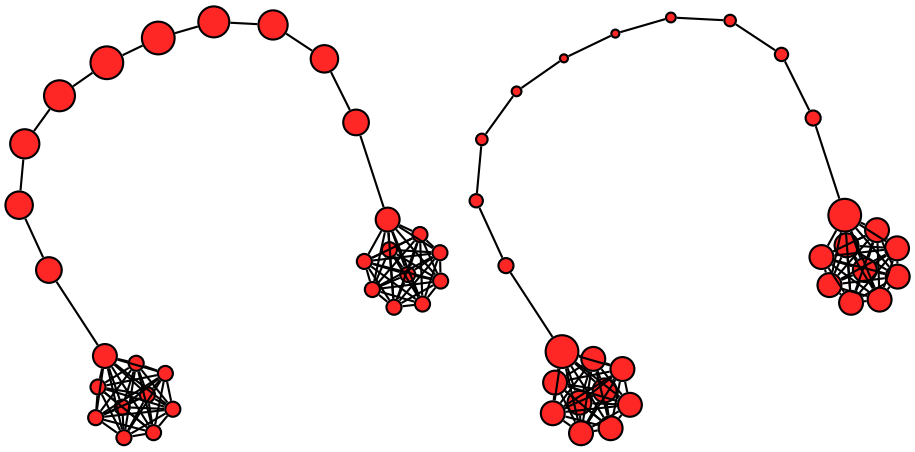
Computational complexity of the harmonic centrality index is $\mathcal{O}(n|E|)$, with n the number of nodes we decided to compute the index for. It is the same as the closeness centrality.

There is a strong assumption when use is made of the closeness centrality: the network has to be connected. This can lead to wrong interpretations of the results: Padgett's sixteen florentine families were especially chosen among more than a hundred! A lot of links are omitted and we can seriously hypothesize that the Pucci family is probably in the same component as the other fifteen families appearing in the graph (see figure 3). Newman's network (see figure 4) leads to the same conclusion because it was compiled from a selected set of collaborations (the two literature reviews). Therefore, with its "connectedness-limitation", closeness centrality won't give any help when trying to understand

the network as a whole. It isn't useful when a study is done on graphs defined from samples.

We have seen that the harmonic centrality behaves similarly as the closeness centrality when the graph is connected, and can also be computed on unconnected graphs. We give some interpretations.

- Nodes close to the one we are interested in will improve its measure (In some degenerated cases, this can also lead to incompatibility with the closeness centrality index, (see figure 5) where $\rho = -0.7954545$! But such a graph is not likely to appear in social network analysis.), which means that being in a dense cluster, even a little one, will assure this individual a higher value of its index. In this case the exploitation of harmonic centrality is no more compatible with closeness centrality.



(a) CLOSENESS (b) HARMONIC

Figure 5: $\mathcal{G}_3$. A degenerate case: $\rho = -0.7954545$! Size of each node vary depending upon the rank of the corresponding index: a big node represents an individual whose rank is small (1 is the best rank), a small node is an individual with a very high rank.

- If the graph is unconnected, a non-zero value doesn't mean the individual can communicate to everyone else, but instead that he can play a certain role in the graph, which can and must be compared to the ones of the

others via the measures. Being in a small component doesn't mean this individual will get the smallest score (see figure 6).

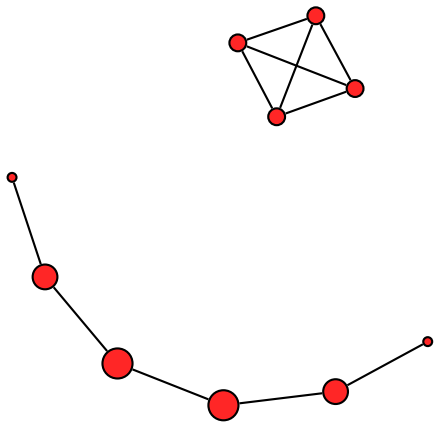


Figure 6: Each node of the biggest component hasn't got necessarily higher harmonic centrality than nodes in other components.

- An isolate will always have an harmonic centrality of 0.

Some computations (not appearing in this article) showed us that the harmonic centrality most often gives higher values on unconnected and sparse graphs than the generalized centrality (See equation 4. Remark that isolates invert this tendency.) Therefore the results from the computation of harmonic centrality can give rise to many more interpretations than an index which gives values very close to 0. For example, the mean of the harmonic centrality computed on the network of scientists is more than twenty times higher than the mean of the generalized centrality (0.01374 against 0.00068).

6 Conclusion

An alternative method of computing centrality has been defined and studied. High correlation of harmonic centrality with closeness centrality makes it a good alternative, in particular because it can be computed and interpreted

on unconnected graphs. Limitations and unexpected behavior of the index find their origin, from what we've seen, only from degenerate and highly improbable cases. Comparisons should be done on other types of graphs, especially those reproducing properties of social networks. How to interpret the results brought by this index may also need a more in depth study: the definition of the harmonic centrality index uses inversions, but not in a similar way as the closeness centrality index. How close those two indices really are is an important question to answer.

Acknowledgements

Thanks to Jean-Philippe Antonietti, Gabor Csárdi, Sarah Dégallier, Jérémie Knüsel and Gilles Steiner for useful discussions. Some data are made available by Mark Newman on his webpage (`www-personal.umich.edu/~mejn/`). Analysis and computation were done using R (`www.r-project.org/`) and igraph (Csardi & Nepusz, 2006). Illustrations were compiled thanks to igraph and TikZ (Tantau, 2009).

References

Barabasi, Albert-Laszlo, & Albert, Réka. 1999. Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

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

Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. 2006 (Aug.). *Maximizing Modularity is hard.*

Brandes, Ulrik. 2008. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, **30**(2), 136–145.

Butts, Carter T. 2009. *sna: tools for social network analysis.* R package version 2.0.

Csardi, Gabor, & Nepusz, Tamas. 2006. The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695, http://igraph.sf.net.

Dorogovtsev, S. N., & Mendes, J. F. F. 2003. *Evolution of networks: from biological nets to the internet and WWW*. Oxford University Press.

Eppstein, David, & Wang, Joseph. 2004. Fast approximation of centrality. *Journal of Graph Algorithms and Applications*, **8**(1), 39–45.

Erdös, P., & Rényi, A. 1959. On random graphs. *Publ. Math. Debrecen*, **6**(290).

Erdös, P., & Rényi, A. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, **5**(17).

Faust, Katherine, & Wasserman, Stanley. 1992. Centrality and prestige: A review and synthesis. *Journal of Quantitative Anthropology*, **4**, 23–78.

Freeman, Linton C. 1978. Centrality in social networks conceptual clarification. *Social Networks*, **1**(3), 215 – 239.

Kent, D. 1978. *The rise of the Medici: Faction in Florence, 1426-1434*. Oxford: Oxford University Press.

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., & Dawson, S. M. 2003. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, **54**, 396–405.

Newman, M. E. J. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, **74**(3), 036104.

Newman, Mark. 2003. The Structure and Function of Complex Networks. *SIAM Review*, **45**(mars), 167–256.

Newman, M.E.J, & Park, Juyong. 2003. Why social networks are different from other types of networks. *Phys. Rev. E*, **68**(3), 036122.

Tantau, Till. 2009 (February). *The TikZ and PGF Packages*. Institut für Theoretische Informatik, Universität zu Lübeck.

Wasserman, Stanley, & Faust, Katherine. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Xulvi-Brunet, R., & Sokolov, I. M. 2004. Reshuffling scale-free networks: From random to assortative. *Phys. Rev. E*, **70**(6), 066102.

Zachary, W. W. 1977. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33**, 452–473.