Comparison of Centralities for Biological Networks[*]

Dirk Koschützki and Falk Schreiber
Bioinformatics Center Gatersleben-Halle
Institute of Plant Genetics and Crop Plant Research
Corrensstraße 3
06466 Gatersleben, Germany.
koschuet,schreibe@ipk-gatersleben.de

Abstract: The analysis of biological networks involves the evaluation of the vertices within the connection structure of the network. To support this analysis we discuss five centrality measures and demonstrate their applicability on two example networks, a protein-protein-interaction network and a transcriptional regulation network. We show that all five centrality measures result in different valuations of the vertices and that for the analysis of biological networks all five measures are of interest.

1 Introduction

Centrality analysis is particularly useful in analyzing biological networks and hence in helping to understand the underlying biological processes. It has been shown that central vertices in protein-protein interaction networks are often functionally important and the deletion of such vertices is related to lethality [JMBO01]. In [WS03] three different types of centralities are defined and applied to metabolic, protein-protein interaction and domain sequence networks. Fell and Wagner discuss the possibility that metabolites with highest degree may belong to the oldest part of the metabolism [FW00].

However, it has also been shown that the degree of a vertex alone, as a specific centrality measure, is not sufficient to distinguish lethal proteins clearly from viable ones [Wu02], that in protein networks there is no relation between network connectivity and robustness against amino-acid substitutions [HCW02], and that for biological network analysis several centrality measures have to be considered [WS03]. To assist scientists in the exploration of biological networks, we discuss and compare five different centrality measures. Some of these are already known in biological sciences, others are transferred from different fields of sciences such as social network analysis. The application of these measures shows that some correlate strongly in one network and weakly in another. As a result, we conclude that for the analysis of biological networks several measures should be considered.

[*] This work was supported by the German Ministry of Education and Research (BMBF) under grant 0312706A.

This paper is organized as follows: in Sect. 2 we define the graph model on which we operate. Section 3 introduces the five centralities in networks, all are explained using one example graph. These measures are applied to typical biological networks in Sect. 4.

2 Definitions

A undirected *graph* $G = (V, E)$ consists of a finite set V of *vertices* ($n = |V|$) and a finite set $E \subseteq V \times V$ of *edges* ($m = |E|$). An edge $e = (u, v) \in E$ connects two vertices u and v. The vertices u and v are said to be *incident* with the edge e and *adjacent* to each other. The set of all vertices which are adjacent to u is called the neighborhood $N(u)$ of u ($N(u) = \{v : (u, v) \in E\}$). A graph is called *loop-free* if no edge connects a vertex to itself. An *adjacency matrix* A of a graph $G = (V, E)$ is a $(n \times n)$ matrix, where $a_{ij} = 1$ if and only if $(i, j) \in E$ and $a_{ij} = 0$ otherwise. The adjacency matrix of any undirected graph is symmetric.

The *degree* $d(v)$ of a vertex v is the number of its incident edges. Let $(e_1, \ldots, e_k)$ be a sequence of edges in a graph $G = (V, E)$. This sequence is called a *walk* if there are vertices $v_0, \ldots, v_k$ such that $e_i = (v_{i-1}, v_i)$ for $i = 1, \ldots, k$. If the edges e_i are pairwise distinct and the vertices v_i are pairwise distinct the walk is called a *path*. The *length* of a walk or path is given by its number of edges, $k = |(e_1, \ldots, e_k)|$. A *shortest path* between two vertices u, v is a path with minimal length, all shortest paths between u, v are called *geodesics*. The *distance* ($\mathrm{dist}(u, v)$) between two vertices u, v is the length of a shortest path between them. Two vertices u, v of a graph $G = (V, E)$ are called *connected* if there exists a walk from vertex u to vertex v. If any pair of different vertices of the graph is connected, the graph $G = (V, E)$ is called *connected*. If a walk starts at vertex u, chooses uniformly at random one of the incident edges of the current vertex until it finally reaches the target v then we call this walk a *random walk* between u and v.

In the remainder of this paper we consider only non-trivial[1] undirected loop-free connected graphs. This restriction is required for a common definition of all centrality measures covered in this paper. Some of the centralities can easily be expanded to cover directed or unconnected graphs, even an extension towards weighted edges is possible.

3 Centralities in Networks

Formally a centrality is a function $\mathcal{C}$ which assigns every vertex $v \in V$ of a given graph G a value $\mathcal{C}(v) \in \mathbb{R}$. As we are interested in the ranking of the vertices of the given graph G we choose the convention that a vertex u is more important than another vertex v iff $\mathcal{C}(u) > \mathcal{C}(v)$. In the following sections we explain five different centrality measures and show an example graph and the corresponding centrality values.

[1] Graphs of at least two vertices and one edge.

3.1 Degree

An obvious order of the vertices of a graph can be established by sorting them according to their degree. The corresponding centrality measure *degree-centrality* ($\mathcal{C}_d$) is defined as $\mathcal{C}_d(v) := d(v)$. See the work of Freeman [Fr79] for a long list of references to the usage of degree-centrality in social network analysis. For biological network analysis degree-centrality for example is used in [JMBO01] to correlate the degree of a protein in the network with the lethality of its removal. See Fig. 1 for an example graph and Table 1 for the corresponding centrality values.

3.2 Eccentricity

The next three definitions of centralities all operate on the concepts of paths within the given graph. The simplest definition uses the distance between vertices. The *eccentricity* ecc of a vertex u is defined as $\mathrm{ecc}(u) := \max_{v \in V} \mathrm{dist}(u, v)$ and the *eccentricity-centrality* ($\mathcal{C}_e$) as $\mathcal{C}_e(u) := \frac{1}{\mathrm{ecc}(u)}$. The reciprocal of $\mathrm{ecc}(u)$ is used to assure that more central vertices have a higher value of $\mathcal{C}_e$, because such central vertices are the ones with the smallest eccentricity value. An application of eccentricity within the biological context is shown in [WS03]. Again, as for degree-centrality, Fig. 1 and Table 1 show an example.

3.3 Closeness

In contrast to eccentricity, closeness-centrality uses not only the maximum distance between the vertex of interest and all other vertices but uses the sum of the distances of this vertex and all other vertices. The *closeness-centrality* is defined as $\mathcal{C}_c(u) := \frac{1}{\mathrm{sumdist}(u)}$ with $\mathrm{sumdist}(u) = \sum_{v \in V} \mathrm{dist}(u, v)$. Closeness-based centrality measures are cited extensively in the work of Freeman [Fr79]. Wuchty *et al.* [WS03] apply this centrality to different biological networks and show the correspondence with the service facility location problem.

3.4 Random Walk Betweenness

Within networks a communication between two vertices u, v may be visible to a third vertex w if this vertex lies in the path of the communication between these vertices. To measure the centrality of a vertex the ability to observe communication is a feasible approach. Different methods to model communication are conceivable. There are for example communications over shortest paths, paths with maximum flow and random walks. All of these are potential models for *betweenness* and are covered in the literature [Fr77, FBW91, Ne03].

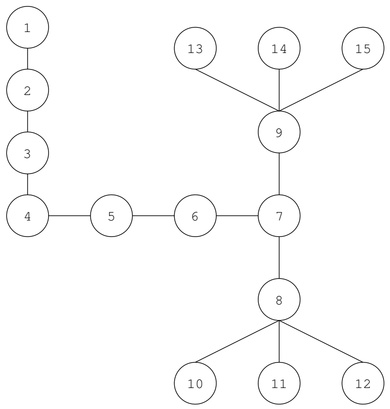

Figure 1: A graph to show the five different centrality measures

Newman's random walk approach models information transmission and therefore matches problems often modelled in biological networks. For the *random-walk betweenness centrality* ($\mathcal{C}_r$) the centrality of a vertex w is equal to the number of times that a random walk from u to v goes through w, averaged over all u and v [Ne03]. A detailed coverage of the required calculations for the random-walk betweenness is beyond the scope of this paper and the reader is directed toward [Ne03].

3.5 Bonacich's Eigenvector Centrality

A different approach to order the vertices of a graph was suggested by Bonacich [Bo72]. His idea is based on the assumption that the value of a single vertex is determined by the values of the neighboring vertices. In contrast to the previous measures not only the position of a vertex within the graph is considered but also the centrality values of its neighbors.

Bonacich suggested the following definition: $\mathcal{C}_\lambda(u) := \sum_{v \in N(u)} \mathcal{C}_\lambda(v)$. If we consider the adjacency matrix representation of the graph, this is equivalent to $\mathcal{C}_\lambda(v_i) := \sum_{j=1}^{n} a_{ij} \mathcal{C}_\lambda(v_j)$. This leads directly to the well known problem of eigenvector computation $\lambda S = AS$ and the eigenvector of the largest eigenvalue is the *eigenvector-centrality* ($\mathcal{C}_\lambda := S$) [Bo72].

3.6 Centralities shown on an example graph

To demonstrate that all of the explained centrality measures usually give different results we use the example graph shown in Fig. 1. For this graph Table 1 shows all centrality values for the five centrality measures.

Vertex	$\mathcal{C}_d$	Vertex	$\mathcal{C}_e$	Vertex	$\mathcal{C}_c$	Vertex	$\mathcal{C}_r$	Vertex	$\mathcal{C}_\lambda$
8	4	5	0.2500	7	0.0286	7	0.7429	7	0.5021
9	4	4	0.2000	6	0.0263	6	0.5619	8	0.4563
7	3	6	0.2000	8	0.0238	5	0.5143	9	0.4563
2	2	3	0.1667	9	0.0238	8	0.4762	6	0.2761
3	2	7	0.1667	5	0.0233	9	0.4762	10	0.1927
4	2	2	0.1429	4	0.0200	4	0.4476	11	0.1927
5	2	8	0.1429	10	0.0182	3	0.3619	12	0.1927
6	2	9	0.1429	11	0.0182	2	0.2571	13	0.1927
1	1	1	0.1250	12	0.0182	1	0.1333	14	0.1927
10	1	10	0.1250	13	0.0182	10	0.1333	15	0.1927
11	1	11	0.1250	14	0.0182	11	0.1333	5	0.1517
12	1	12	0.1250	15	0.0182	12	0.1333	4	0.0830
13	1	13	0.1250	3	0.0169	13	0.1333	3	0.0448
14	1	14	0.1250	2	0.0143	14	0.1333	2	0.0230
15	1	15	0.1250	1	0.0120	15	0.1333	1	0.0097

Table 1: The centrality values for the example graph. The vertices are ordered by descending centrality value

4 Application and Discussion

We applied the presented centrality measures to two biological networks, a protein-protein-interaction (PPI) network and a transcriptional regulation (TR) network. The PPI network is based on the April 2004 release of the *Homo sapiens* network from the DIP-Database [SMS$^+$04]. It models proteins as vertices and interactions as edges. The TR network of *Escherichia coli* models operons as vertices and regulation between transcription factors and operons as edges, see [SOMMA02] for details and the link to the data.

For the PPI-network we removed 51 self interactions as the graph has to be loop-free for the calculation of the eigenvector-centrality $\mathcal{C}_\lambda$. Furthermore, we removed vertices and edges not connected to the giant component, and the resulting graph consisted of 563 vertices and 870 edges. For the TR network we used the data from [SOMMA02] and applied the same strategy: we removed self regulation and used the giant component for the analysis. As the regulation data is directed we used the underlying undirected graph. The activation and repression information at the edges was not used as our framework is based on unweighted graphs. Finally, the graph for the TR network consisted of 325 vertices and 453 edges.

The calculation of the five centrality measures for both graphs was done on a modern desktop PC (3 GHz, 2 GB Ram, Linux 2.6.x, Java 1.4.2) and took between less than a second ($\mathcal{C}_d$) and several minutes ($\mathcal{C}_r$). This was expected as the complexity of the different algorithms ranges from $O(n)$ to $O((n+m)n^2)$.

To analyse a network we calculated the presented centralities for all vertices. Then for each centrality all vertices were ordered by descending centrality value and, for vertices with the same centrality value, by ascending vertex-label. The vertices were enumerated from 1 to n, this number is the position of the vertex according to the centrality. Finally, we build a new table which for every vertex and every centrality contains the position of the vertex according to the centrality.

With the calculated positions we made several analyses of the correlation. It shows that some of the measures, e.g. eccentricity $\mathcal{C}_e$ and eigenvector $\mathcal{C}_\lambda$, are highly correlated in the PPI-network (see Fig. 2) while others are only weakly correlated. Within the TR network a strong correlation between eigenvector $\mathcal{C}_\lambda$ and closeness $\mathcal{C}_c$ was observed (see Fig. 3), but in contrast to the PPI network there is only a weak correlation between $\mathcal{C}_e$ and $\mathcal{C}_\lambda$. Tables 2 and 3 show all correlation coefficients based on Pearson's method. Note that the strong correlation of degree $\mathcal{C}_d$ and random-walk betweenness $\mathcal{C}_r$ is discussed in [Ne03].

In conclusion, the analysis of biological networks clearly benefits from the application of several centrality measures. Our next step lies in the comparison of different centralities measures with biological information.

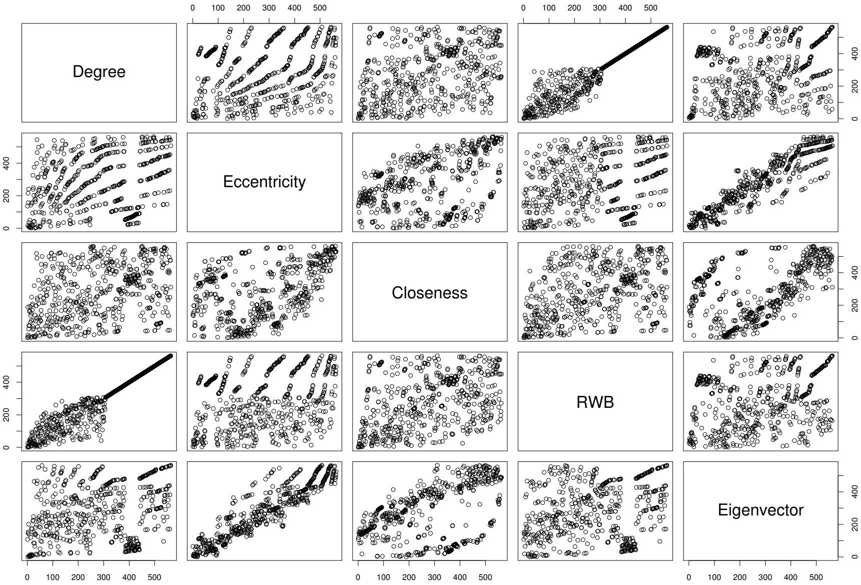

Figure 2: Scatter plot matrix of the centrality positions for the PPI network

	$\mathcal{C}_d$ Degree	$\mathcal{C}_e$ Eccentricity	$\mathcal{C}_c$ Closeness	$\mathcal{C}_r$ RWB	$\mathcal{C}_\lambda$ Eigenvector
$\mathcal{C}_d$	1.0000	0.2794	0.3396	0.9534	0.2703
$\mathcal{C}_e$	0.2794	1.0000	0.4231	0.2776	0.9248
$\mathcal{C}_c$	0.3396	0.4231	1.0000	0.3843	0.4726
$\mathcal{C}_r$	0.9534	0.2776	0.3843	1.0000	0.2627
$\mathcal{C}_\lambda$	0.2703	0.9248	0.4726	0.2627	1.0000

Table 2: Correlation coefficients for the centrality positions for the PPI-network

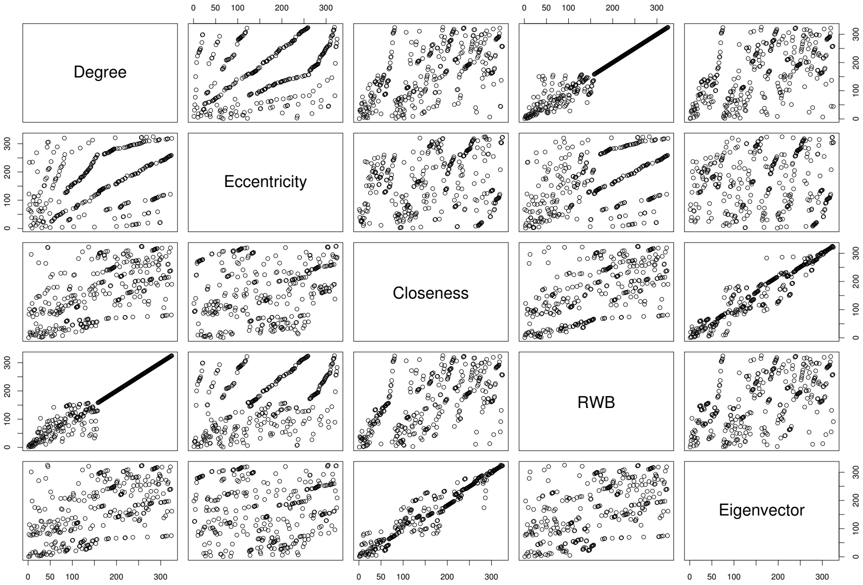

Figure 3: Scatter plot matrix of the centrality positions for the TR network

	$\mathcal{C}_d$ Degree	$\mathcal{C}_e$ Eccentricity	$\mathcal{C}_c$ Closeness	$\mathcal{C}_r$ RWB	$\mathcal{C}_\lambda$ Eigenvector
$\mathcal{C}_d$	1.0000	0.3974	0.5861	0.9700	0.5499
$\mathcal{C}_e$	0.3974	1.0000	0.2208	0.4172	0.0514
$\mathcal{C}_c$	0.5861	0.2208	1.0000	0.5856	0.9552
$\mathcal{C}_r$	0.9700	0.4172	0.5856	1.0000	0.5164
$\mathcal{C}_\lambda$	0.5499	0.0514	0.9552	0.5164	1.0000

Table 3: Correlation coefficients for the centrality positions for the TR network

References

[Bo72] Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*. 2:113–120. 1972.

[FBW91] Freeman, L. S., Borgatti, S. P., and White, D. R.: Centrality in valued graphs: a measure of betweenness based on network flow. *Social Networks*. 13:141–154. 1991.

[Fr77] Freeman, L. C.: A Set of Measures of Centrality Based on Betweenness. *Sociometry*. 40(6):35–41. 1977.

[Fr79] Freeman, L. S.: Centrality in social networks: Conceptual clarification. *Social Networks*. 1:215–239. 1979.

[FW00] Fell, D. A. and Wagner, A.: The small world of metabolism. *Nature Biotechnology*. 18:1121–1122. 2000.

[HCW02] Hahn, M. W., Conant, G., and Wagner, A.: Molecular evolution in large genetic networks: connectivity does not equal importance. Technical report. Santa Fe Institute. 2002. 02-08-039.

[JMBO01] Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N.: Lethality and centrality in protein networks. *Nature*. 411:44. 2001.

[Ne03] Newman, M. E. J.: A measure of betweenness centrality based on random walks. arXiv cond-mat/0309045. 2003.

[SMS$^+$04] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D.: The database of interacting proteins: 2004 update. *Nucleic Acids Res*. 32(1):449–451. 2004.

[SOMMA02] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U.: Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*. 31(1):64–68. 2002.

[WS03] Wuchty, S. and Stadler, P. F.: Centers of complex networks. *J Theor Biol*. 223(1):45–53. 2003.

[Wu02] Wuchty, S.: Interaction and domain networks of yeast. *Proteomics*. 2:1715–1723. 2002.