Subject Section

Inferring Biological Function from Network Topology via Topological Ranking- Supplementary Material

Raffaele Giancarlo 1,*, Daniele Greco 2,* and Simona E. Rombo 1,*

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

1 Biological Networks

Network	Type of nodes	No. Nodes	No. Edges	Link density
\overline{GDN}	Gene	903	6,760	0.01659
HDN	Disease	516	1,118	0.009
GDN_{GO}	Gene	801	5,277	0.01647
WGN	Gene	554	137,918	0.897
YD1	Protein	990	4,687	0.01
YD2	Protein	1,443	6,993	0.007
Y2H	Protein	1,966	2,705	0.001

Table 1. Basic structural features of the considered biological networks

2 Topological measures

2.1 Edges

2.1.1 Incremental measures

Topological Overlap Measure. Given an edge (i,j), the Topological Overlap Measure (TOM) and its Generalized version (GTOMm) (Ravasz $et\ al.$, 2002; Yip and Horvath, 2007) score the degree to which i's and j's neighbors overlap. The larger the overlap, the higher the weight assigned to (i,j). TOM considers only the immediate neighbors, whereas GTOMm includes all the neighbors at distance $\leq m$:

$$w_{ij} = \begin{cases} \frac{|N_m(i) \cap N_m(j)| + a_{ij}}{\min\{|N_m(i)|, |N_m(j)|\} + 1 - a_{ij}} & \text{if } i \neq j \\ 1 & \text{otherwise} \end{cases}$$
 (1)

In Equation 1, $N_m(i)$ denotes the set of i's neighbors reachable with a shortest path of length at most m from i, and $a_{ij}=1$ if there exists an edge connecting vertices i and j, $a_{ij}=0$ otherwise.

Edge Clustering Value. Similarly to TOM, Edge Clustering Value (ECV) (Wang *et al.*, 2011) quantifies how much the *i*'s and *j*'s neighborhood overlan.

$$w_{ij} = \frac{|N_1(i) \cap N_1(j)|^2}{|N_1(i)| \cdot |N_1(j)|}.$$
 (2)

Unlike TOM which normalizes the size of common neighborhood over the smallest between i and j neighborhoods (see Equation 1), ECV is equal to 1 if and only if i and j have the same *exact* neighbors. It is worth noting that both TOM and ECV can be interpreted as a biological, neighborhoodnormalized versions of Granovetter's *embeddeness* measure, historically used to characterize tie-strength in social networks (Marsden and Campbell, 1984).

Dispersion. For a given edge (i,j), the Dispersion measure (Backstrom L, 2014) extends Granovetter's tie-strength measure (Marsden and Campbell, 1984), taking into account both the size and the *connectivity* of i,j's common neighborhood. Intuitively, it quantifies of how "not well"-connected is the i,j's common neighborhood within G_i , the subgraph induced by i and its neighbors. More formally, let G_i be the subgraph induced in G by $\{i\} \cup N_1(i)$. In G_i , let j be a neighbor of i (i.e. $j \in N_1(i)$), and denote with $C_{ij}^{(i)} = N_1(i) \cap N_1(j)$ the set of common neighbors of i and j within the induced subgraph G_i . The absolute dispersion (Backstrom L, 2014) is defined as follows:

$$disp(i,j) = \sum_{s,t \in C_{ij}^{(i)}} d_v(s,t)$$
(3)

where d_v is a boolean function such that $d_v = 1$ if and only if s and t are not connected by a path of length ≤ 2 in G_i . The authors define two

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

¹ Department of Mathematics and Computer Science, University of Palermo, Palermo, Italy

²Department, University of Pisa, Pisa, Italy

^{*}To whom correspondence should be addressed.

2 Giancarlo et al.

enhanced versions of dispersion: *parametric* dispersion (Equation 4) and *recursive* dispersion (Equation 5):

$$param(i, j, \alpha, \beta, \gamma) = \frac{(disp(i, j) + \beta)^{\alpha}}{emb(i, j) + \gamma}$$
 (4)

$$rec(i,j) \leftarrow \frac{\sum_{w \in C_{ij}^{(i)}} x_w^2 + 2\sum_{s,t \in C_{ij}} d_v(s,t) x_s x_t}{emb(i,j)}$$
 (5)

where emb(i,j) is equal $|C_{ij}^{(i)}|$ (see details in (Backstrom L, 2014)). It is easy to show that both these measures are not symmetric. Since in our context we are considering undirected graphs, we unambiguously assign a weight to the edge (i,j) by defining and applying the following three variants:

1.**rec_max**. Assigns a dispersion weight to edge (i,j) as $w_{ij} = rec_max(i,j) = MAX\{rec(i,j), rec(j,i)\}.$

2.**rec_min**. Assigns a dispersion weight to edge (i,j) as $w_{ij} = rec_min(i,j) = MIN\{rec(i,j), rec(j,i)\}$.

3.param_sym. Computes a *symmetric* parametric variant of $w_{ij} = param(i, j, \alpha, \beta, \gamma)$ with parameters $\alpha = 0.61$, $\beta = 0$, $\gamma = 5.0$ (i.e., parameters used by authors, see details in (Backstrom L, 2014)), considering the common neighborhood C_{ij} in G (and not in the induced subgraphs G_i , G_j , as in the original – non-symmetric– definition).

2.1.2 Decremental measures

Edge Betweenness. Given an edge e_{ij} , the Edge Betweenness (EB) (Girvan and Newman, 2002) is the fraction of shortest paths in the network \mathcal{N} containing the edge e_{ij} :

$$w_{ij} = \sum_{s,t \in V} \frac{\sigma_{st}(e_{ij})}{\sigma_{st}} \tag{6}$$

In Equation 6, σ_{st} represents the total number of shortest paths connecting nodes s, t ($s \neq t$) in the network, whereas $\sigma_{st}(e_{ij})$ counts only the shortest paths between the same nodes containing the edge e_{ij} . The higher w_{ij} , the more likely the edge e_{ij} acts as a bridge connecting separate communities, thus representing an inter-community edge.

Edge Clustering Coefficient. Given the edge e_{ij} , the Edge Clustering Coefficient (ECC3) (Radicchi *et al.*, 2004) is the number of *triangles* the edge e_{ij} belongs to, divided by the number of triangles that might potentially include it:

$$w_{ij} = -\frac{z_{i,j}^3 + 1}{s_{i,j}^{(3)}} = -\frac{|N_1(i) \cap N_1(j)| + 1}{\min[(|N_1(i)| - 1)(|N_1(j)| - 1)]}$$
(7)

In Equation 7, $z_{i,j}^3$ is the number of triangles built on the edge e_{ij} and $s_{i,j}^{(3)}$ is the maximal possible number of them. The minus sign is explained as follows. Many triangles exists within dense communities. Therefore, the higher $|w_{ij}|$ the more likely e_{ij} lies within a dense community, being an intra-community edge. Nevertheless, in order to identify communities (Radicchi et al., 2004) use $|w_{ij}|$ in a Girvan-Newman fashion (Girvan and Newman, 2002): at each step of the divisive algorithm, edges with the lowest $|w_{ij}|$ are removed, eventually splitting the original network into separate connected components. Here, we directly take the opposite of the original measure to rank first edges e_{ij} having the lowest $|w_{ij}|$ (that is: the inter-community ones). These edges represent "less important" elements according to our framework.

Edge Centrality Proximity Distance. Edge Centrality Proximity Distance (ECPd) (De Meo *et al.*, 2014) is based on the notion of Edge Centrality (Equation 9) as introduced in (DeMeo *et al.*, 2012):

$$w_{ij} = 1 - \sqrt{\sum_{k=1}^{n} \frac{(L^{\kappa}(e_{ik}) - L^{\kappa}(e_{kj}))^2}{d(k)}}$$
 (8)

$$L^{\kappa}(e) = \sum_{s \in V} \frac{\sigma_s^{\kappa}(e)}{\sigma_s^{\kappa}} \tag{9}$$

In Equation 8, $L^{\kappa}(e_{ik})$ is the Edge Centrality, the fraction of times a random walker traverses the edge e_{ik} running through a random simple path of length at most κ . Thus, $(L^{\kappa}(e_{ik})-L^{\kappa}(e_{kj}))^2$ expresses a proximity between nodes i and j: the probability that a message propagated to node i reaches also node j, with node k being a common neighbor of the two. Therefore, Equation 8 represents a distance between nodes i and j: the higher w_{ij} the more likely the nodes i and j belong to different communities and e_{ij} being an inter-community edge.

2.2 Nodes

2.2.1 Incremental measures

Node Clustering Coefficient. Given a node i, Node Clustering Coefficient (NCC) (Watts, 1999) expresses how much densely connected is the i's neighborhood. Let k_i be the number of neighbors of i, and let n_i be the number of edges connecting such neighbors:

$$x_i = \frac{2n_i}{k_i(k_i - 1)}. (10)$$

In Equation 10, the denominator is equal to the maximum value for k_i (recall $\mathcal N$ is an undirected graph). As a result, the greater the value of x_i , the closer the i's neighborhood to be a clique.

Eigenvector Centrality. Given a node i, the Eigenvector centrality (EGC) (Bonacich, 1972) is a measure of topological "importance" (Brandes and Erlebach, 2005) for the node i in the network \mathcal{N} . Specifically, node i can acquire high centrality either by having a high degree or by being connected to other highly-important.

$$x_i = \frac{1}{\lambda} \sum_{j \in G} A_{ij} \ x_j \tag{11}$$

In Equation 11, λ is the largest positive eigenvalue of the adjacency matrix A, satisfying the equation $A \mathbf{x} = \lambda \mathbf{x}$, with $\mathbf{x} = (x_1, x_2, ..., x_n)$ being the vector of node centralities. Notably, Google's PageRank (Page *et al.*, 1999) is a variant of Eigenvector Centrality.

2.2.2 Decremental measures

Betweenness Centrality. Given a node i, Node Betweenness (BC) (Freeman, 2012) quantifies the extent to which node i lies on *geodesic* (shortest) paths between other pairs of vertices:

$$x_i = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{12}$$

In Equation 12, $\sigma_{st}(v)$ represents the number of shortest paths from node s to node t that pass through i, whereas σ_{st} is the total number of shortest paths between the same nodes. Intuitively, the higher x_i the more likely i lies on a path between nodes in different communities.

Subgraph Centrality. Given a node i, Subgraph Centrality (SGC) (Estrada and Rodríguez-Velázquez, 2005) quantifies the centrality of node i based on the number of subgraphs it belongs to:

$$x_i = \sum_{k=0}^{\infty} \frac{(A^k)_{ii}}{k!} \tag{13}$$

Since each closed path is associated with a connected subgraph, EGC counts the number of closed walks in the network. In Equation 13, $(A^k)_{ii}$ is the number of closed paths of length k, starting and ending on node i. Closed walks are weighted such that smaller walks are given higher weights (see (Estrada and Rodríguez-Velázquez, 2005)). Therefore, smaller subgraphs are given higher weights than larger ones, which makes SGC able to quantify the extent to which node i participates in network motifs within real-world networks. (Milo $et\ al.$, 2002; Yeger-Lotem $et\ al.$, 2004)

 κ -Path Centrality. Given a node i, the κ -Path Centrality (KPC) (Alahakoon et al., 2011) is defined as the sum, over all possible source nodes s, of the probability that a message originating in s goes through i, assuming the message runs along random simple paths of length at most κ -

$$x_i = \sum_{s \neq i} \frac{\sigma_s^{\kappa}(i)}{\sigma_s^{\kappa}} \tag{14}$$

In Equation 14, $\sigma_s^{\kappa}(i)$ is the number of messages originating at node s passing through node i, whereas σ_s^{κ} is the total number of messages originated from node s. Despite a similar formulation, KPC differs substantially from NB. In particular, KPC does not assume information flows necessarily across shortest paths, as NB does.

3 Algorithms

Algorithm 1 Returns a *static* edge rank \mathcal{E} of network \mathcal{N} , based on the topological measure w.

```
1: function rank-static(\mathcal{N}, w)
2: for each (i, j) \in E
3: compute the weight function w(i, j)
4: \mathcal{E} \leftarrow sort E in non-increasing order of weight
5: \mathcal{E} \leftarrow group edges of E with equal rank
6: return \mathcal{E} = (E_1, E_2, ..., E_k)
7: end function
```

Algorithm 2 Returns a *dynamic* edge rank \mathcal{E} of network \mathcal{N} , based on the topological measure w.

```
1: function rank-dynamic(\mathcal{N}, w)
     \mathcal{E} \leftarrow \emptyset
2:
3:
      i \leftarrow 1
       \mathbf{while}\ (|E|>0)\ \mathbf{do}
5:
          for each (i, j) \in E
6:
               compute the weight function w(i,j)
          Let E_i be the set of edges in E having maximum weight
7:
8:
          Append E_i to \mathcal E and delete edges in E_i from E
9:
          i \leftarrow i + 1
10:
        end while
        return \mathcal{E} = (E_1, E_2, ..., E_k)
12: end function
```

4 Benchmark Datasets

4.1 Protein-Protein Interaction Networks

Yeast PPIN. [(Daniele): TODO: briefly describe the networks Yeast-D1, Yeast-D2, Yeast Y2H and cite sources].

3

Benchmarks for the Yeast PPIN. We consider three reference sets of yeast complexes: Cmplx1 for YEAST-D1 (D1), Cmplx2 for YEAST-D2 (D2) and Cmplx3 for Y2H (Y2H), respectively. Cmplx1 includes 81 complexes of sizes at least 5, created from MIPS (Mewes *et al.*, 2000). Cmplx2 is made of 162 hand-curated complexes (size no less than 4 proteins) from MIPS (?). Finally, Cmplx3 includes 975 known and curated complexes from [(Daniele): TODO: source?].

5 Statistical Significance of a Topological Ranking

- 5.1 Precision, Recall and F-Measure Indices
- 5.2 Kendall index in Case of Ties

6 Results

6.1 Human and Gene Disease Networks

Definire Global Comparison

View Type	F	Rank Type	K_haus	EC test	TR test
I	I ECV		0.921352		√
I	GTOM2	static	0.263199	\checkmark	\checkmark
I	TOM	static	0.394417	\checkmark	\checkmark
I	kb_rec_max	static	0.512545		\checkmark
I	kb_rec_min	static	0.157832		\checkmark
I	kb_param_sym	static	0.109155		\checkmark
I	ECV	dynamic	0.587569		\checkmark
I	GTOM2	dynamic	0.258508	\checkmark	\checkmark
I	TOM	dynamic	0.387743	\checkmark	\checkmark
D	EB	static	0.904365		✓
D	ECC3	static	0.879348		\checkmark
D	ECP	static	0.902387		\checkmark
D	EB	dynamic	0.733648		\checkmark
D	ECC3	dynamic	0.88796		\checkmark

Table 2. Global Comparison for the Human Disease Network HDNG using Edge Ranks. The column $View\ Type$ specifies the view, either (I)ncremental or (D)ecremental. The columns F and Rank give the topological measure together with the ranking type, respectively.. Acronyms for the measures are in Section 2. K_haus reports the corresponding normalized K_{haus} distances obtained: the smaller the value, the more similar the topological ranking to the functional ranking. The final two columns indicate whether the EC and TR tests were passed.

6.2 Worm Desease Network

qui le tabelle che ha Greco di tutti tests su questo network

References

Alahakoon, T., Tripathi, R., Kourtellis, N., Simha, R., and Iamnitchi, A. (2011). K-path centrality: A new centrality measure in social networks. In *Proceedings of* 4 Giancarlo et al.

View Type	F	Rank Type	K_haus	EC test	TR test
$:GC_HDN_Edges_results$ heightI	NCC	static	0.558474	✓	✓
I	EGC	static	0.573012		✓
I	NCC	dynamic	0.432467	✓	✓
I	EGC	dynamic	0.673747		✓
D	NB	static	0.553613		✓
D	SGC	static	0.742343		
D	KPC	static	0.774174		
D	NB	dynamic	0.557036		✓
D	SGC	dynamic	0.664536		✓
D	KPC	dynamic	0.838758		

Table 3. Global Comparison for the Human Disease Network HDNG using Edge Equivalent Ranks. The legend is as in Table 2.

View Type	F	Rank Type	K_haus	EC test	TR test
I	I ECV		0.96157		√
I	GTOM2	static	0.126684	\checkmark	\checkmark
I	TOM	static	0.211708	\checkmark	\checkmark
I	kb_rec_max	static	0.398346		\checkmark
I	kb_rec_min	static	0.058925		\checkmark
I kb_param_sym		static	0.037909		\checkmark
I	ECV	dynamic	0.783565		\checkmark
I	GTOM2	dynamic	0.126097	\checkmark	\checkmark
I	TOM	dynamic	0.210773	\checkmark	\checkmark
D	EB	static	0.784118		<u>√</u>
D	ECC3	static	0.937585		\checkmark
D	ECP	static	0.778692		\checkmark
D	EB	dynamic	0.426541		\checkmark
D	ECC3	dynamic	0.937758		\checkmark

Table 4. Global Comparison for the Gene Disease Network $GDNG_DSDS$ using Edge Ranks. The legend is as in Table 2.

the 4th Workshop on Social Network Systems, SNS '11, pages 1:1–1:6, New York, NY, USA. ACM.

Backstrom L, K. J. (2014). Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on facebook. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 831–841, New York, NY, USA. ACM.

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

Brandes, U. and Erlebach, T. (2005). Network Analysis: Methodological Foundations (Lecture Notes in Computer Science). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

De Meo, P., Ferrara, E., Fiumara, G., and Provetti, A. (2014). Mixing local and global information for community detection in large networks. *J. Comput. Syst. Sci.*, **80**(1), 72–87.

DeMeo, P., Ferrara, E., Fiumara, G., and Ricciardello, A. (2012). A novel measure of edge centrality in social networks. *Knowledge-Based Systems*, 30, 136 – 150.

Estrada, E. and Rodríguez-Velázquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E*, **71**, 056103.

Freeman, L. C. (2012). Centrality in social networks conceptual clarification. *Social Networks*, **1**(3), 1978–1979.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99, 7821–7826.

Marsden, P. V. and Campbell, K. E. (1984). Measuring tie stength. *Social Forces*, **63**, 482âL"501.

Mewes, H. W. et al. (2000). MIPS: a database for genomes and protein sequences. Nuc. Ac. Res., 28(1), 37–40.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, 298(5594), 824–827.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101, 2658–2663.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and BarabÃisi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297, 1551–1555.

Wang, J., Li, M., Chen, J., and Pan, Y. (2011). A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3), 607–620.

Watts, D. (1999). Small worlds. Princeton University Press, Princeton.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U., and Margalit, H. (2004). Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proceedings of the National Academy of Sciences*, **101**(16), 5934–5939.

Yip, A. and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, **8**, 22.

short Title 5

View Type	F	Rank Type	K_haus	EC test	TR test
I	NCC	static	0.7739	√	✓
I	EGC	static	0.457653		\checkmark
I	NCC	dynamic	0.746356	\checkmark	\checkmark
I	EGC	dynamic	0.932883		\checkmark
D	NB	static	0.243135		√
D	SGC	static	0.954153		\checkmark
D	KPC	static	0.984258		
D	NB	dynamic	0.223707		\checkmark
D	SGC	dynamic	0.929725		\checkmark
D	KPC	dynamic	0.98898		

Table 5. Global Comparison for the Gene Disease Network $GDNG_DSDS$ using Edge Equivalent Ranks. The legend is as in Table 2

View Type F Rank Type K_haus EC test TR test ECV static 0.960733 GTOM2 0.145038 Ι static 0.240163 TOM Ι static kb_rec_max static 0.4478460.074278 kb_rec_min static 0.043585 kb_param_sym static I ECV dynamic 0.764334 GTOM2 dynamic 0.144126 TOM 0.23884 I dynamic D EB static 0.811777D ECC3 static 0.934783 D ECP 0.80787 static D EB dynamic 0.452198 D ECC3 dynamic 0.938017

View Type	F	Rank Type	K_haus	EC test	TR test
I	NCC	static	0.801589	√	√
I	EGC	static	0.673495		\checkmark
I	NCC	dynamic	0.767608	\checkmark	\checkmark
I	EGC	dynamic	0.936067		\checkmark
D	NB	static	0.274195		√
D	SGC	static	0.945941		\checkmark
D	KPC	static	0.98083		
D	NB	dynamic	0.240112		\checkmark
D	SGC	dynamic	0.919067		\checkmark
D	KPC	dynamic	0.985727		

Table 7. Global Comparison for the Gene Disease Network $GDNG_GODS$ using Edge Equivalent Ranks. The legend is as in Table 2

"output" — 2020/7/28 — page 5 — #5

.

Giancarlo et al.

View Type	View Type F		K_haus	EC test	TR test
I	I ECV		0.633149	√	√
I	GTOM2	static	0.628768	\checkmark	\checkmark
I	TOM	static	0.592387	\checkmark	\checkmark
I	kb_rec_max	static	0.548323		\checkmark
I	kb_rec_min	static	0.667305		
I	kb_param_sym	static	0.682106		
I	ECV	dynamic	0.505902		\checkmark
I	GTOM2	dynamic	0.629108	\checkmark	\checkmark
I	TOM	dynamic	0.592576	\checkmark	\checkmark
D	EB	static	0.556917		√
D	ECC3	static	0.637379		\checkmark
D	ECP	static	0.543236		\checkmark
D	EB	dynamic	0.538452		\checkmark
D	ECC3	dynamic	0.650556		

View Type	F	Rank Type	K_haus	EC test	TR test
I	NCC	static	0.529577	✓	√
I	EGC	static	0.458644		\checkmark
I	NCC	dynamic	0.508407	\checkmark	\checkmark
I	EGC	dynamic	0.635614		\checkmark
D	NB	static	0.599601		√
D	SGC	static	0.643931		\checkmark
D	KPC	static	0.663018		
D	NB	dynamic	0.611015		\checkmark
D	SGC	dynamic	0.623985		\checkmark
D	KPC	dvnamic	0.667948		

Table 9. Global Comparison for the Gene Disease Network $GDNG_GOGO$ using Edge Equivalent Ranks. The legend is as in Table 2

.