Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Compact Views Based on Topological Measures for Biological Networks Exploration

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

[1] Department of Mathematics and Computer Science, University of Palermo, Palermo, Italy
[2] Department, University of Pisa, Pisa, Italy

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.
Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Contact: name@bio.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biological networks are powerful models widely used in the literature to represent various biological phenomena (Barabasi, 2009; Pizzuti and Rombo, 2014; Zhang *et al.*, 2016). They encode biological knowledge related, for example, to physical interactions among cellular components (e.g., protein-protein interaction networks, Singh *et al.* (2008)), as well as information on genotype-phenotype associations (e.g., the diseasome, Barabasi *et al.* (2011)). Interestingly, although important biological insights have been obtained from the analysis of biological networks (e.g., ...), the extraction of new useful knowledge is not yet fully satisfying, due to different factors.

A first important aspect is related to the quality of data biological networks rely on. As an example, many protein-protein interactions annotated in public databases have been identified with high-throughput techniques, and not always the accuracy of corresponding annotations is sufficiently high. Moreover, it is well known that, while some organisms are well characterized (e.g., *S. cerevisiae*), for others is not the same (e.g., *D. melanogaster*). This has important implications on the different *structure* of same type networks associated to different organisms, causing difficulty in proposing approaches that perform well on both model organisms and less characterized species. In the last few years, a further level of complexity has emerged in the analysis of biological networks, due to the always larger amounts of annotations which have been made available, making the networks much larger in size than in the past.

In this scenario, it is worth recalling that the original purpose of biological networks is to encode biological information on the reciprocal relationships among cellular components, in order to infer novel knowledge which may lead to a more complete understanding of the cellular life. Therefore, if one would be able to generate ”compact descriptors” of the biological networks, such that only the more relevant information would be carried on, then possible analysis techniques exploited in cascade could benefit of that. Indeed, similar network descriptors would allow to obtain: (1) robustness to the not always high quality of data and organisms difference in characterization; (2) efficiency, due to the chance of discarding less significant portions of the network from the analysis; (3) effectiveness, by identifying those elements in the networks which are ”more important” than others for drawing conclusions on the complex processes governing the cellular life.

The research presented here provides contribution in the direction highlighted above. In particular, a novel framework is proposed for biological networks exploration, based on the network topology. The goal is to provide hierarchical views of the most important nodes/edges in the network, such that the importance is indicative of the functional information encoded by those network elements. In order to have a clear picture of the approach, consider the following example, built on a completely different application context. Assume that a network represents logistics connections among the monuments of a city. If some tourists

have planned a very short visit to the city, e.g., a weekend, they will probably be interested in visiting only the most important monuments, then a small fraction of the network will be necessary, i.e., the view on the most important monuments. If they will spend at least a week in the city, then a larger fraction of the network needs to be explored, e.g., the view on the monuments including the most important ones and some others, and so on. If a person has decided definitely to move to this city, it is reasonable to think that, at a certain moment, the entire network will have been explored.

In order to build such compact views of edges/nodes on biological networks, it is needed to characterize nodes and edges according to their "relevance". To this aim, the notion of *topological measure* (Girvan and Newman, 2002) is of great interest here. Intuitively, how much important a network element is in order to infer significant biological knowledge has necessarily to be reflected by the specific network topology and how that element is connected inside the network. On the other hand, it is natural to ask whether such an importance inferred by the network topology has to be intended as representative of the only biological information directly encoded by the specific network under consideration, or instead it is effective in order to bring to light novel knowledge that was hidden in the network modelled data.

An extensive set of topological measures, mostly coming from the literature (e.g., ...), is considered here and exploited as a pool of basic primitives for the construction of hierarchical views on biological networks. In particular, such views may be obtained from *topological rankings* of nodes and edges, based on the considered measures. One of the most important contributions of this research is the study of how much the rankings based on the topology of a specific network, reflect the functional importance of the corresponding nodes/edges, according to some external information (e.g., gold standards). AGGIUNGERE ALTRO SE NECESSARIO.

2 Materials and Methods

2.1 Biological Networks

According to the specific problem under consideration, biological networks may rely on undirect/direct graphs, hyper-graphs, bipartite graphs, etc.. Here we define a *biological network* as follows.

Definition 2.1 (Biological Network)**.** A biological network $\mathcal{N}$ is an undirect graph $\langle V, E \rangle$ such that:

- V is the set of nodes, usually associated to cellular components (e.g., proteins, genes, etc.).
- E is the set of edges, representative of biological relationships between components associated to the corresponding linked nodes.

The biological networks considered in this study are described in detail in the following paragraph, and their main structural features are summarized in Table 1 of the Supplementary Material.

2.1.1 Gene Disease Network

The Gene Disease Network (GDN) is a one-mode projection of the Diseasome bipartite network (Goh *et al.*, 2007), where two genes are connected by an edge if mutations in the respective genes are associated with the occurrence/progress of at least one common disease. Another possible one-mode projection of the Diseasome is the Human Disease Network (HDN), for which the results on topological measures comparison confirm the findings obtained on the dual GDN, therefore they are provided in the Supplementary Material and not discussed here.

2.1.2 Worm Gene Network

In analogy with GDN, the Worm Gene Network (WGN) is obtained as a one-mode projection of a bipartite graph obtained for *C. elegans* in **?**, by placing in one class 554 essential genes and on the other 94 phenotypic defects. A gene is connected to a defect if its inhibition (via breakdown experiments) is involved in the development of the defective phenotype. As for WGN, its nodes are the genes, and there is an edge connecting two genes if they have at least one defect of phenotype in common.

2.1.3 Protein-Protein Interaction Networks

COMPLETARE QUI, NO SUPPLEMENTARY.

2.2 Topological Measures

Intuitively, a topological measure assigns a real weight to nodes or edges of a biological network $\mathcal{N}$, based only on the topological structure of $\mathcal{N}$. Topological measures have been used in a variety of domains as tools to identify, e.g., relevant elements (Zotenko *et al.*, 2008; Giancarlo *et al.*, 2019) or dense substructures (Fortunato, 2010; Girvan and Newman, 2002), or to characterize the global organization of complex networks (Watts, 1999; Jeong *et al.*, 2001).

Formally, recall that two networks $\mathcal{N} = \langle V, E \rangle$ and $\mathcal{N}' = \langle V', E' \rangle$ are isomorphic ($\mathcal{N} \simeq \mathcal{N}'$) if there exists a bijection $\phi : V \to V'$ such that $(u, v) \in E$ if and only if $(\phi(u), \phi(v)) \in E'$.

In the remaining part of this section, we refer to a biological network $\mathcal{N} = \langle V, E \rangle$ for all definitions. Similarly to the definition in (Brandes and Erlebach, 2005), we provide the following.

Definition 2.2 (Topological measure)**.** Let X be V or E. A real-valued function $w : X \to \mathbb{R}$ is a topological measure if and only if: $\forall x \in X,\ \mathcal{N} \simeq \mathcal{N}' \implies w_{\mathcal{N}}(x) = w_{\mathcal{N}'}(\phi(x))$, where $w_{\mathcal{N}}(x)$ denotes the value $w(x)$ in $\mathcal{N}$. **questa definizione secondo me non si capisce troppo bene**

For the purpose of this study, among the many measures one can define, two classes are of interest: *incremental* and *decremental*. Measures in the first class are meant to capture the intuition that the higher the weight assigned by w to an element of X, the more "important" that element is in the network. The dual holds for measures in the second class. **quindi le decrementali che fanno, misurano la "non importanza"???**

In the following the topological measures considered as primitives for the proposed framework are shortly summarized. It is worth pointing out that a novel measure based on work by (**?**) is also proposed here. QUAL'E' DI QUELLE CHE MENZIONIAMO??? All such measures are formally defined in Section 2 of the Supplementary Material.

2.2.1 Edge Topological Measures

Let $i, j \in V$ be two nodes and $(i, j) \in E$ be and edge of $\mathcal{N}$. Three *incremental* topological measures are considered for (i, j), all quantifying how much the i's and j's neighborhood overlap, although with some differences. The larger the overlap, the higher the weight assigned to (i, j). In particular, the *Topological Overlap Measure (TOM)* considers only the immediate neighbors, whereas its Generalized version *GTOMm* (Ravasz *et al.*, 2002; Yip and Horvath, 2007) includes all the neighbors at distance $\leq m$. While TOM normalizes the size of common neighborhood over the smallest between i and j neighborhoods, *E TOMm pure? Edge Clustering Value (ECV)* (Wang *et al.*, 2011) 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, neighborhood-normalized versions of Granovetter's *embeddeness* measure, historically used to characterize tie-strength in social networks (Marsden and Campbell, 1984). Also *Dispersion* (Backstrom L, 2014) extends the latter, 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, i.e., the subgraph induced by i and its neighbors. Three main variants of Dispersion are considered here (see Section 2 in the Supplementary Material). As for the *decremental* topological measures, the three following are considered. *Edge Betweenness (EB)* (Girvan and Newman, 2002) is the fraction of shortest paths in the network $\mathcal{N}$ containing the edge (i, j). *Edge Clustering Coefficient (ECC3)* (Radicchi *et al.*, 2004) is the number of *triangles* the edge (i, j) belongs to, divided by the number of triangles that might potentially include it. *Edge Centrality Proximity Distance (ECPd)* (De Meo *et al.*, 2014) is based on computing the fraction of times a random walker traverses an edge, running through a random simple path of length at most κ.

2.2.2 Node Topological Measures

Let $i \in V$ be a node of $\mathcal{N}$. The incremental topological measures for i are: *Node Clustering Coefficient (NCC)* (Watts, 1999), measuring how much densely connected is the i's neighborhood; *Eigenvector Centrality (EGC)* (Bonacich, 1972), such that i can acquire high centrality either by having a high degree itself, or by being connected to other highly-important nodes. The decremental topological measures are the following. *Betweenness Centrality (BC)* (Freeman, 2012) quantifies the extent to which node i lies on *geodesic* (shortest) paths between other pairs of vertices. *Subgraph Centrality (SGC)* (Estrada and Rodríguez-Velázquez, 2005) quantifies the centrality of i based on the number of subgraphs it belongs to. Finally, *κ-Path Centrality (KPC)* (Alahakoon *et al.*, 2011) is 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.

2.3 Edge/Node Rank Views via The Topology of a Network

In this section we provide explicit definitions for edge (node, resp.) ranks and their associated views, together with procedures for generating them with the use of topological measures.

2.3.1 Generic Rank Views

We consider first the case of edges.

Definition 2.3 (Edge Rank)**.** An *edge rank* of $\mathcal{N}$ is an ordered list $\mathcal{E} = (E_1, E_2, \cdots, E_k)$ of subsets of E such that they are a partition of E.

Intuitively, by displaying the subgraphs induced by incrementally considering, in the order given, the sets in $\mathcal{E}$, one can get incremental views of $\mathcal{N}$, according to the priority, i.e, "relevance" , of the edges given by the ranking. A decremental view can be obtained analogously by removing edges according to $\mathcal{E}$. In this latter case, the priority of the ranking indicates irrelevance. Formally, one can define a sequence of *views* of $\mathcal{N}$, based on $\mathcal{E}$, as follows.

Definition 2.4 (i-th incremental (decremental) view)**.** Given an integer $1 \leq i \leq k$, the *i-th incremental view* of $\mathcal{N}$ *w.r.t.* $\mathcal{E}$ is the subgraph $\mathcal{N}_i$ of $\mathcal{N}$ induced by the set $S_i = \bigcup_{j=1}^{i} E_j$. The *i-th decremental view* is defined analogously, except that $S_i = E \setminus (\bigcup_{j=1}^{i} E_j)$.

A useful alternative to the above views is the following.

Definition 2.5 ($i\%$ percentage incremental (decremental) view)**.** Let p be the maximal integer such that the cardinality of $S_{i\%} = \cup_{j=1}^{p} E_j$ is at most $i\%$ of the edges in E. The $i\%$ incremental view of $\mathcal{N}$ is defined as the subgraph $\mathcal{N}_{i\%}$ of $\mathcal{N}$ induced by the set $S_{i\%}$. The $i\%$ *percentage decremental view* is defined analogously, except that $S_{i\%} = E \setminus (\bigcup_{j=1}^{p} E)$.

The difference between views in Definitions 2.4 and 2.5, is that the latter uses *at most* a specified number of edges (nodes) in $\mathcal{N}$, therefore it is focused on the "size" of the view one wants to generate. On the other hand, incremental views according to Definition 2.4 account for the ..., and not for the number of detected edges or nodes. **Quindi, ricapitolando: la vista percentuale ci fa scegliere l'occupazione di rete che vogliamo includere nella vista, invece l'altra il numero di partizioni. Ma a che serve? E invece, una vista che ci consenta di scegliere il grado di rilevanza non ce l'abbiamo? Quindi posso dire "100 archi", oppure "20 partizioni", ma non posso scegliere "rilevanza pari almeno ad x", oppure "rilevanza pari ad almeno il 20 per cento dal valore massimo?" e simili?**

As for nodes, the definitions of rank and views are analogous to the ones given above for edges and therefore omitted. It is worth noting that, in terms of views, the one corresponding to a node rank $\mathcal{V}$ reduces to an edge rank $\mathcal{E}^*$, that we refer to as *node equivalent edge rank*. Indeed, informally, given $\mathcal{V}$, one can construct $\mathcal{E}^*$ by progressively growing the sets of edges in $\mathcal{E}^*$ as they are inserted/removed in the view coresponding to $\mathcal{V}$. Formal details are omitted for brevity.

2.3.2 Rank Views induced by Topological Measures

The definitions of views given in Section 2.3.1 assume that a ranking, reflecting "relevance" of a node or edge, is given. Here we provide algorithms that compute a ranking based on a generic topological measure, belonging to one of the classes introduced in Section 2.2. The function being either incremental or decremental implies the same characterization for the views obtained via the corresponding ranking. Only the case of edges is discussed, since the corresponding case for nodes is analogous.

A topological measure can be used to generate edge ranks of $\mathcal{N}$ of two different types: *static* and *dynamic*. The corresponding procedures are described next and sketched by Algorithms 1 and 2 in Section XXX of the Supplementary Material, respectively. Both algorithms take as a parameter a topological measure w. In the static case, one pass is made to compute the value $w(i, j)$ returned by the considered topological measure, on all edges of the input $\mathcal{N}$. Then, the edges are sorted in non-increasing order of weight and partitioned into groups in non-decreasing order of rank. In the dynamic case, $|E|$ steps are executed. At each step, the edges with the highest score are grouped together and the result is appended to an initially empty list. Those edges are deleted from the network **Ma quanti se ne rimuovono? Con che criterio?**. Then, the process is repeated on the resulting network. Finally, from the groups of edges in the list, the corresponding ranking is produced.

The difference between static and dynamic ranking is that, while the former refers to the "absolute" values produced by the topological measures for each edge/node, the latter is able to capture the "relative" importance that an edge (a node) has with respect to its neighbors. This can be useful to intercept edges/nodes with an important role in their topological context, but whose identity is someway hidden by other edges/nodes scoring much higher values. Once that highest score edges/nodes are discarded, the role of these hidden important ones may come out. An interesting example of that is provided by the Girvan-Newman community detection algorithm (Girvan and Newman, 2002) which, in this acception, can be considered as a special case of Algorithm 2. Indeed, Girvan and Newman (2002) stated that a dynamic version of Edge Betweeness is more effective than a static one in their case, since it is able to progressively identify those edges to be deleted from the input network, in order to better separate the clusters to be produced in output.

2.4 Experimental Methodology

Given a rank view induced by a specific topological measure, the study presented here aims to understand how much it is representative not only of the biological knowledge directly encoded by the network topology, but also of novel, hidden, functional information. As usual in both supervised and unsupervised classification contexts (see, e.g., (Furfaro *et al.*, 2017;

?; **?**; Pizzuti and Rombo, 2014)), the "performance" of the rank view in discovering hidden functional knowledge may be evaluated by using *external* criteria. Here, an external criterion relies on the existence of some gold standard ranking, obtained via information not dependent on the topology of the input network. The ranking induced by a topological function can be thus compared against the gold standard ranking. Once that quantification is available, it is also important to assess how much statistically significant it is. To this end, one can resort to a MonteCarlo Hypothesis Test (see **?**), where the null hypothesis H_0 is that the mentioned quantification is due to chance. That is, its value is no better than the one obtained by a random ranking.

In summary, the above approach requires three ingredients: (a) gold standards (b) a measure of agreement between rankings and (c) the specification of H_0 for the statistical significance test. Those points are presented next, focusing only on edges and incremental case, since node ranking and decremental case are analogous.

2.4.1 Gold Standards

For each of the considered biological networks, at least a gold standard has been defined, as follows.

Gene Disease Network In exploring GDN, it seems natural to expect that one would like to see first edges corresponding to most strongly correlated gene pairs. Among the many possible weight assignments, we use very simple and intuitive ones which are meant to encode a biological tie-strength proportional of the number of common (1) diseases implied by SNPs (DS) (Goh *et al.*, 2007), and (2) GO terms (GO) (Consortium, 2014) between two genes (with references to the only biological process vocabulary). It is worth pointing out that, in the latter case, if an edge of GDN does not correspond to at least 4 common GO terms between the node proteins, then it has been eliminated from the network. The resulting network features are summarized in Table 1 of the Supplementary Material (see GDN_{GO}). **Credo sia stato fatto per evitare archi con peso zero, ma non sono piÃ¹ certa che fosse necessario. Tuttavia, se la scelta ci convince, dovremmo forse motivare perchÃ¨ archi con peso nullo non ci piacciono?**

Worm Gene Network The weight of each edge is the number of defects of phenotype that two genes have in common. (CHECK)

Protein-Protein Interaction Networks QUI CHE AVETE FATTO???

2.4.2 Performance of a Rank View via a Measure of Agreement Between Rankings

Let w be an edge topological measure and let g be the function that encodes a certain gold standard. That is, g assigns weights to the edges of the network based on knowledge not depending on its topology. Intuitively, the closer are the edge rankings produced by w and g, the better the performance of w with respect to the gold standard encoded by g. We formalize such an intuition in two different ways, outlined informally next. The first evaluates w globally with respect to any arbitrary gold standard function g, a very demanding test. The second assumes that g is much simpler, i.e., a zero-one function, meant to label with a one all and only edges of $\mathcal{N}$ satisfying the same biological property or requirement. Now, w is evaluated in the progressive setting, measuring its ability to discover very soon the edges of $\mathcal{N}$ labeled with a one. It is to be remarked that the demanding task of evaluating w globally with respect to an arbitrary function g can be also phrased in the progressive setting as well. However, that would be a further increase in difficulty for w to perform well, since in that setting the ranking of w must closely track that of g in order to perform well. As a matter of fact, at the early stages of this research, we have experiments with this latter approach (global-progressive) to soon realize its unfairness. Indeed, preliminary experiments in that setting have indicated that all functions do poorly.

A detailed description of the global and progressive performance evaluation modes follows.

Global Comparison Assume that each edge is numbered with an integer in $[1, |E|]$. Let $\mathcal{E}_w = (E_{1,w}, E_{2,w} \cdots, E_{k,w})$ and $\mathcal{E}_g = (E_{1,g}, E_{2,g} \cdots, E_{p,g})$ be the rankings coming out of w and g, respectively. If those ranking were a permutation of the edges numbers, then we could easily compare them via standard methods such as Kendall rank index **? RG to add**. Unfortunately, since there may be ties, e.g., more than one edge can be in a given class of each ranking, we cannot use the mentioned index directly. A ranking with ties is referred to usually as partial. Among the many possibilities, we have chosen to use K_{haus}, a distance functions on partial rankings defined by Fagin *et al.* (2003) and that belongs to a class of distances specifically designed for partial rankings. Given two partial rankings, K_{haus} counts the number of inversions in the rankings, excluding ties. It is normalized so that it has value in $[0, 1]$, where zero indicated identity. In order to assess how close is $\mathcal{E}_w$ to $\mathcal{E}_g$, we use K_{haus}. The lower its value, the better the performance of w with respect to the gold standard g.

Progressive Comparison Let E_g be the set of edges of the network that are labeled with a one by the gold standard function g. For a given topological function w, consider a progressive generation of percentage views starting, say, at 10% and increments by 10%. Each of those views is evaluated in terms of the percentage of the edges in E_g that are present in the view. The lower the percentage of the view for which 100% is reached, the better the performance of w.

2.4.3 Statistical Significance of a Topological Ranking

Consider $\mathcal{E}_w = (E_{1,w}, E_{2,w} \cdots, E_{k,w}), \mathcal{E}_g = (E_{1,g}, E_{2,g} \cdots, E_{p,g})$ and K_{haus}. That is, the case of global comparison. We use two null models. The first referred to as *total random* and denoted by TR, in which a random permutation of the edges of the network is generated. The second, referred to as *equal classes* and denoted by EC, in which each class of $\mathcal{E}_w$ is assigned the same number of edges it has, but this time chosen randomly **NON CHIARO, TRA CHI SONO SCELTI RANDOM?**. We perform a MonteCarlo simulation consisting of 100 iterations for both models and set the significance level at 1%. In each iteration and for each model, we compute K_{haus} between $\mathcal{E}_g$ and the randomly generated permutation. For the progressive comparison, the procedure is analogous and the same null models are generated. However, now the score is different since for each percentage view obtained from the random permutation, we have to consider the percentage of edges of E_g that are in that view.

3 Results

The full set of results is reported in Section XXX of the Supplementary Material. The results reported in those tables indicate that most measures deliver results that are significant with respect to TR. This is not entirely surprising, since the measures selected in this study are among the most natural and prominent for community detection. However, when one imposes a more stringent null model for the hypothesis test, i.e., EC, only a handful of them is still significant.

3.1 Global Comparison

They are reported in Tables **??**-2 Surprisingly, the set of "successful measures" is nearly the same for each of the benchmark networks used in this part of this study. No decremental measure is among them and there seems to be very little difference between static and dynamic setting. This is

interesting since decrementality and dynamicity were perceived as two key components of a high performing community detection algorithm Girvan and Newman (2002). In terms of distance from the functional rankings, both GTOM and TOM perform very well in all tests, with the exception of a reasonable performance in one. **[(RG): ora qualche conclusione]**.

Network	Case	View Type	F Rank Type	Performance
$GDNG\_DSDS$	Edges	I	GTOM2 static	$K_{haus} = 0.126684$
$GDNG\_DSDS$	Edges	I	GTOM2 dynamic	$K_{haus} = 0.126097$
$GDNG\_DSDS$	Edges	I	TOM dynamic	$K_{haus} = 0.210773$
$GDNG\_DSDS$	Edges	I	TOM static	$K_{haus} = 0.211708$
$GDNG\_DSDS$	Nodes*	I	NCC dynamic	$K_{haus} = 0.746356$
$GDNG\_DSDS$	Nodes*	I	NCC static	$K_{haus} = 0.7739$

Table 1. **Most statistically robust topological measures for Global Comparison in** GDN**.** The column *Case* indicates whether the result is obtained by means of an Edge ("Edges")or Equivalent Edge Rank (this latter referred as "Nodes*"). *View Type* specifies the view, either (I)ncremental or (D)ecremental. *F Rank* gives the topological measure together with the ranking type. Acronyms for the measures are in Supplementary Section **??**. *Performance* reports the corresponding normalized K_{haus} distances obtained: the smaller the value, the more similar the topological ranking to the functional ranking.

3.2 Progressive Comparison

3.3 Human and Gene Disease Networks

Network	Case	View Type	F Rank Type	Performance
GDN_GOGO	Nodes*	I	NCC dynamic	$K_{haus} = 0.508407$
GDN_GOGO	Nodes*	I	NCC static	$K_{haus} = 0.529577$
GDN_GOGO	Edges	I	TOM static	$K_{haus} = 0.592387$
GDN_GOGO	Edges	I	TOM dynamic	$K_{haus} = 0.592576$
GDN_GOGO	Edges	I	GTOM2 static	$K_{haus} = 0.628768$
$0,009\ GDN\_GOGO$	Edges	I	GTOM2 dynamic	$K_{haus} = 0.629108$
GDN_GOGO	Edges	I	ECV static	$K_{haus} = 0.633149$

Table 2. **Most statistically robust topological measures for Global Comparison in** $GDNG\_GOGO$. The legend is as in Table **??**.

The full set of results is reported in Supplementary Section **??**. Quite surprisingly, the results reported in those tables indicate that most of the incrememntal measures that was possible to test do not pass even the TR test. This can be in part explained by the fact that the WGN is a very dense graph (link density of 0.897), as opposed to the HDN and GDN variants that are very sparse (link density of 0.009 and **[(RG): to be computed]**, respectively. Even more interesting is the fact that now the decremental measures are statistically significant in regard to both EC and TR. Those results are reported in Table 3.

Network	Case	View Type	F Rank Type	Performance
WGN	Edges	D	EB dynamic	$K_{haus} = 0.33$
WGN	Edges	D	ECC3 dynamic	$K_{haus} = 0.40$
WGN	Nodes*	I	NCC static	$K_{haus} = 0.44$

Table 3. **Most statistically robust topological measures for Global Comparison in WGN**. The legend is as in Table **??**. [(RG): La tabella non compl,eta, mancano EB ed ECC3 static, inioltre i valori riportati non corrispondono a quelli della tesi]

3.4 Discussion

Secondo me, punti salienti sono i seguenti.

-Tranne casi eccezionali, tutte le misure son buone nel caso random. Quindi, per stabilire su casi reali se il risultato e' affidabile bisogna ricorrere ad una procedura di validazione. Per il momento si puo attingere a quelle per clustering networks e communities, ma sarebbero piu adatte quelle per ranking (di cui si sa poco). Questo e' un problema interessante sul quale lavorare. Ovviamente, un ranking puo essere visto come unsupervised clustering.

-giustificare perche c'e una netta separazione tra le misure buone per PPI e quelle buone per gli altri networks. Probabilmente abilita di clustering.

-Discutere se c'e un vantaggio di dynamic vs static. Newman e girvan dicono che quella e' una componente essenziale del loro metodo, che altrimenti non sarebbe buono (discutono un esempio). Probabilmente cio e' legato alla presewnza di clusters molto evidenti e pronunciati.

- dire qualcosa sulla complessitá di tempo, in relazione a community detection ?In oerticolare, il tempo del sorting degli archi puó risultare dominante ia nel caso statico che dinamico per funzioni "graziose". QualÃ¨ il trade-off tra tempo e predizione ?

-Come mai incremental e decremental si separano ? Di nuovo, densita clusters.

NCC static emerge come una buona misura quando la "funzione" puo essere riassunta da un intorno di nodi e vertici. Poco adatta al clustering.

EB si prende rivincita sulle PP

discutere qualcosa di bio che si ewcvince dagli esperimenti

4 Conclusion

To be written To be written. To be written . To be written . To be written . To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written written

Acknowledgements

To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written To be written.

Funding

To be written To be written To be written To be written To be written To be written To be written To be written

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 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.

Barabasi, A.-L. (2009). Scale-free networks: A decade and beyond. *Science*, **325**(5939), 412–413.

Barabasi, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat Rev Genet*, **12**(1), 56–68.

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.

Consortium, T. G. O. (2014). Gene Ontology Consortium: going forward. *Nucleic Acids Research*, **43**(D1), D1049–D1056.

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.

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

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, **17**, 134–160.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, **486**, 75–174.

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

Furfaro, A., Groccia, M. C., and Rombo, S. E. (2017). 2d motif basis applied to the classification of digital images. *Comput. J.*, **60**(7), 1096–1109.

Giancarlo, R., Greco, D., Landolina, F., and Rombo, S. (2019). Network centralities and node ranking. *Encyclopedia of Bioinformatics and Computational Biology*, **1**, 950âŁ“957.

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.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and BarabÃ¡si, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, **104**(21), 8685–8690.

Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**(6833), 41–42.

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

Pizzuti, C. and Rombo, S. E. (2014). Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, **30**(10), 1343–1352.

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 BarabÃ¡si, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Singh, R., Xu, J., and Berger, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, **105**(35), 12763–12768.

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.

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

Zhang, Y., Wang, Z., and Wang, Y. (2016). Multi-hierarchical profiling: an emerging and quantitative approach to characterizing diverse biological networks. *Briefings in Bioinformatics*.

Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*, **4**(8), 1–16.