Topological ranks reveal functional knowledge encoded in biological networks: a comparative analysis

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

[1]Department of Engineering, University of Palermo, Palermo, 90121, Italy, Palermo and [2]Department of Mathematics and Computer Science, University of Palermo, Palermo, 90121, Italy, Palermo
[*]Corresponding author. mariella.bonomo@community.unipa.it, raffaele.giancarlo@unipa.it, simona.rombo@unipa.com

ABSTRACT

Motivation: Biological networks topology yields important insights into biological function, occurrence of diseases and drug design. In the last few years, different types of topological measures have been introduced and applied to infer the biological relevance of network components/interactions, according to their position within the network structure. Although comparisons of such measures have been previously proposed, to what extent the topology *per se* may lead to the extraction of novel biological knowledge has never been critically examined nor formalized in the literature.
Results: We present a comparative analysis of nine outstanding topological measures, based on compact views obtained from the rank they induce on a given input biological network. The goal is to understand their ability in correctly positioning nodes/edges in the rank, according to the functional knowledge implicitly encoded in biological networks. To this aim, both internal and external (gold standard) validation criteria are taken into account, and six networks involving three different organisms (yeast, worm and human) are included in the comparison. The results show that a distinct handful of best performing measures can be identified for each of the considered organisms, independently from the reference gold standard.
Availability: Input files and code for the computation of the considered topological measures and K-haus distance are available at `https://gitlab.com/MaryBonomo/ranking`.
Contact: simona.rombo@unipa.it
Supplementary information: Supplementary data are available at *Briefings in Bioinformatics* online.

1. INTRODUCTION

Biological networks are versatile models, effective in representing and studying the complex relationships occurring among the main players of the cellular life [8, 39, 42, 54]. They may encode biological knowledge related to the physical interactions of cellular components [48], as well as information on genotype-phenotype associations [9]. The density of information implicitly codified in biological networks often results in the complexity of their topology, whose study has proved to yield insights into biological function, occurrence of diseases and drug design [1, 31, 47]. To this regard, many efforts have been payed devoted to the study of specific topological structures, such as network motifs [38], graphlets [44] and communities [43], which may be intended as a sort of "local" descriptors able to bring out relevant biological information hidden inside dense and complex networks.

Biological networks topology has been investigated also for the identification of specific cellular components, or their associations, which may be considered more biologically representative than others, according to their position inside the network. Examples of that are the *essential* protein-protein interactions studied by [30], indispensable for the survival or reproduction of an organism, as well as the *network hubs* by [10] and [5], often associated to regulatory checkpoints, essential to the molecular programs that generate specific cellular phenotypes. Further studies have been proposed on *gene prioritization*, that is, searching for candidate genes underlying biological processes or diseases [2, 29], and on the relevance of biological components inferred by their *centrality* inside the network structure [23, 30, 32, 33, 55].

Despite the fact that an association between the biological relevance of components (interactions) and their position within the network structure has been widely proved in the literature, to what extent this may lead to the extraction of novel biological knowledge has never been critically examined nor formalized.

We present a paradigm for the identification of significant "global" descriptors of a biological network, relying on the characterization of the relevance of nodes and edges across the network structure. To this aim, the main building boxes of our methodological framework are *topological measures* [26], which assign a real weight to nodes or edges based only on the network topology. In particular, it has emerged recently that systematic explorations of different topological measures, with a comprehensive understanding of their dependence on the specific network type and application context, may contribute to accelerate the solution of difficult problems (e.g., drug repositioning) and their clinical translation [7]. Here we propose to build compact hierarchical views from the *topological ranks* of nodes and edges induced by such measures, in two different assets, *static* and *dynamic*. While in the former case the network is considered in its entirety, in the latter one weights and rankings are assigned dynamically during the relevance discovery process.

We propose a methodology for the evaluation of topological ranks obtained from different measures, that relies on two different criteria: (1) *statistical significance*, via Montecarlo Hypothesis Test, and (2) *biological relevance*, quantified by comparing topological ranks against those obtained from external knowledge (e.g., gold standards). The former is a sort of *internal* criteria, which allows to discriminate the most significant ranks independently from the specific application context. The latter criteria aims to measure to what extent hidden information may be retrieved from a biological network taken as a whole, and intuitively this depends also on the specific type of information one is looking for. It is worth pointing out that the proposed methodology is general enough to apply also in contexts different from biological systems.

An extensive set of outstanding topological measures, mostly coming from the literature together with some novel ones proposed here, is considered for the comparison. Also, six biological networks involving three organisms (*C. elegans*, *S. cerevisiae*, *H. sapiens*) and seven different gold standards have been included in the experimental analysis.

The proposed evaluation of topological ranks leads to the following main insights.

- *Hierarchical views based on edges are more powerful than those relying on nodes* in capturing hidden knowledge from biological networks. This confirms and provides still more rigorous and extensive evidence of what observed in [5], where the synergistic interaction between specific cellular components has revealed to be decisive in the occurrence and progress of some diseases.
- *The dynamic asset is more successful than the static one*, confirming and generalizing the research presented by [26], where this is shown for a topology measure only, in the specific context of community detection.
- *It is possible to identify a handful of topological measures performing well according to both the considered criteria*, depending on the type of network and organism under consideration.

As for the third point, it seems that measures based on clustering coefficient [45, 50] perform well almost unanimously on the considered networks. However, for each organism a specific handful of measures performing better than others can be further identified. Interestingly, to this respect, it seems that if the type of knowledge one is trying to infer changes (i.e., with reference to different gold standards), which are the most significant and best performing topological ranks remains unaffected.

As a final remark, we notice that the proposed topological views are **succinct global representations** that induce a lossy compression of the original network, based on "relevance" rather than syntactic patterns. Such a novel and emerging approach to lossy compress networks may be considered worth of investigation in its own right.

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 1 (Biological Network) A biological network $\mathcal{N}$ is an undirected graph $\langle V, E\rangle$ such that:

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

The biological networks considered in this study may be roughly distinguished in two main categories: *genotype-phenotype associations* networks and *physical-interaction* networks. For each of the two categories, the specific networks considered here are described in detail in the following paragraphs, and their main structural features are summarized in Table 1 of the Supplementary Material.

2.1.1. Genotype-Phenotype Association Networks

We consider three types of networks in this category.

Gene Disease Network. The Gene Disease Network (***GDN***) is a one-mode projection of the Diseasome bipartite network [27], where two genes are connected by an edge if mutations in them are associated with the occurrence/progress of at least one common disease.

Human Disease Network. Another possible one-mode projection of the Diseasome is the Human Disease Network (***HDN***), where nodes are associated to diseases and edges to the presence of gene mutations involved in both the two corresponding diseases.

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 [28], 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.

2.1.2. Physical-Interaction Networks

We consider different networks in this category, all of the same type.

Protein-Protein Interaction Networks Three different yeast PPI datasets are accounted for. The first two PPI networks, namely ***D1*** and ***D2***, have been built by [53] by filtering two networks, one used by [21] and another containing yeast protein interactions generated by six individual experiments, to delete unreliable interactions. The third PPI network, ***Y2H***, is built upon interactions obtained by high-throughput yeast two-hybrid screening [52], where self-edges have been eliminated according to [3].

2.2. Topological Measures

Two biological 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 $(\phi(u), \phi(v)) \in E'$. Similarly to the definition in [13], we provide the following.

Definition 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 $\mathcal{N}$.

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

In the following, the topological measures considered as primitives for the proposed framework are shortly summarized. All such measures are formally defined in Section *Topological Measures* of the Supplementary Material.

2.2.1. Edge Topological Measures

Let $i, j \in V$ be two nodes and $(i, j) \in E$ be an edge of $\mathcal{N}$.

Incremental edge topological measures. Three of these measures considered here quantify how much the i's and j's neighborhoods overlap. Larger the overlap, higher the weight assigned to (i, j). In particular, the *Topological Overlap Measure* (***TOM***) considers only the immediate neighbors, whereas its Generalized version ***GTOMm*** [46, 51] includes all the neighbors at distance $\leq m$. While TOM normalizes the size of common neighborhood over the smallest between i and j neighborhoods, *Edge Clustering Value* (***ECV***) by [49] 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 [34]. Also *Dispersion* [6] extends the latter, taking into account both the size and the *connectivity* of i,j's common neighborhood. Intuitively, it quantifies 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, we call them ***KB1***, ***KB2*** and ***KB3***, respectively (see details in the Supplementary Material).

Decremental edge topological measures. *Edge Betweenness* (***EB***) by [26] is the fraction of shortest paths in the network $\mathcal{N}$ containing the edge (i, j). *Edge Clustering Coefficient* (***ECC3***) by [45] 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***) by [35] 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 $\mathcal{N}$.

Incremental node topological measures. *Node Clustering Coefficient* (***NCC***) by [50] measures how much densely connected is the i's neighborhood. According to *Eigenvector Centrality* (***EGC***) by [11], i can acquire high centrality either by having a high degree itself, or by being connected to other highly-important nodes.

Decremental node topological measures. *Betweenness Centrality* (***BC***) [20] quantifies the extent to which node i lies on *geodesic* (shortest) paths between other pairs of vertices. *Subgraph Centrality* (***SGC***) [15] quantifies the centrality of i based on the number of subgraphs it belongs to. *κ-Path Centrality* (***KPC***) by [4] 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. Topological Rank and Views

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. We consider first the case of edges.

Definition 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 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 $S_i = E \setminus (\bigcup_{j=1}^{i} E_j)$.

Definition 5 ($i\%$ percentage incremental (decremental) view) Let p be the largest 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 4 and 5, is that the former is focused on the partitions built on the rank (e.g., they may be associated to fixed value intervals of the considered measure), whereas 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.

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 *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 corresponding to $\mathcal{V}$. Formal details are omitted for brevity.

2.3.1. Computing Topological Ranks

The definitions provided above assume that a rank, reflecting "relevance" of a node or an edge, is given. Here we provide algorithms that compute a rank based on a generic topological measure, belonging to one of the classes introduced in Section *Topological Measures*. The function being either incremental or decremental implies the same characterization for the views obtained via the corresponding rank. 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 *Algorithms* 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 of w on each of the 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. 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 [26] which, in this formulation, can be considered as a special case of Algorithm 2. Indeed, [26] 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

Suppose that a rank view induced by a specific topological measure is given. 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 [12, 14, 18, 19, 25, 40, 41, 43], 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 a ranking associated to some gold standard, obtained via information not dependent on the topology of the input network. The rank induced by a topological function can thus be compared against the gold standard rank. 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 [24] for analogous applications of this test in the biological domain), 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 ranks and (c) the specification of H_0 for the statistical significance test. Those points are presented next, focusing only on the edge rank and incremental case, since the equivalent edge rank and decremental cases are analogous.

2.4.1. Gold Standards

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

Gold Standards G1 and G2. In exploring GDN, it seems natural to expect that one would like first to see edges corresponding to the 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 (G1) [27], and (2) GO terms (G2) [22] between two genes (with references to the biological process vocabulary only).

Gold Standard G3. For the HDN gold standard ranking, how many SNPs are common to a pair of diseases is considered, according to [27].

Gold Standard G4. For WGN, in analogy with GDN, a gold standard is considered such that the weight of each edge is the number of defects of phenotype that two genes have in common.

Gold Standards G5, G6 and G7. For PPI networks, gold standard rankings have been associated to the number of biological complexes two proteins participate together. To this aim, three reference sets of yeast complexes have been considered here, each specifically selected for the networks in analysis [53]: G5 for D1, G6 D2 and G7 for Y2H, respectively. G5 includes 81 complexes of sizes at least 5, created from MIPS [36]. G6 is made of 162 hand-curated complexes (size no less than 4 proteins) from MIPS [37]. Finally, G7 includes 975 known and curated complexes from `ftp://ftpmips.gsf.de/yeast/catalogues/complexcat`.

2.4.2. Performance of a Rank View via a Measure of Agreement Between Ranks

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 ranks 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 as 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 the sequence of those ranks were a permutation of the edge numbers, then we could easily compare them via standard methods such as Kendall rank index [17]. Unfortunately, since there may be ties, i.e., more than one edge may be associated to the same integer representing its ranking, we cannot use the mentioned index directly, as well as many others (see discussion in [16]). A rank with ties is referred to usually as partial. Among the many possibilities, we have chosen to use K_{haus}, a distance functions on partial ranks defined by [16] and that belongs to a class of distances specifically designed for partial ranks. Given two partial ranks, K_{haus} counts the number of inversions in the ranks, 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.

2.4.3. Statistical Significance of a Rank Comparison

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}. 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, without replacement, from the set of edges of the network. We perform a MonteCarlo simulation consisting of 100 iterations for both models. In each iteration and for each model, we compute K_{haus} between $\mathcal{E}_g$ and the randomly generated permutation. Then we set the significance level at 1% as a measure of relevance.

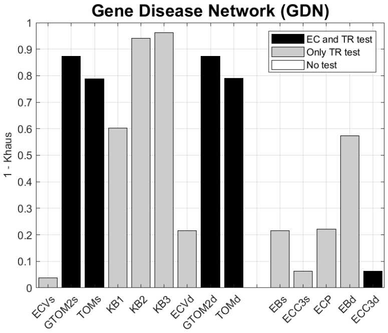

Fig. 1. Performance and statistical significance for the rankings returned by topological measures for GDN w.r.t. the gold standard G1.

3. RESULTS

The full set of results is reported in Section *Results* of the Supplementary Material, where a table is reported for each network, associated gold standard and type of rank (induced by edges or by nodes). Tables show both performance and statistical significance of the rank obtained by the considered topological measures, in both static and dynamic settings (when the latter has been evaluated). Results indicate that most measures deliver a rank 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

3.1.1. Gene Disease Network: Results w.r.t. gold standard G1

Edge rank. The histogram in Figure 1 graphically illustrates the most representative example of the results obtained for this network. In particular, the value of $1 - K_{haus}$ is shown on the vertical axis for the considered topological measures, when the gold standard G1 is considered. The histogram shows that the best compromise between biological relevance and statistical significance is represented by *GTOM2* and *TOM*, in both the static and the dynamic settings (with a slight improvement in the latter case). Also *KB3*, closely followed by *KB2*, has good performance, although both the associated rankings do not pass the EC test.

Edge equivalent rank. From Table 3 of the Supplementary Material, it is evident that only *NCC* (static/dynamic) passes both EC and TR tests, however its performance is not high. *NB* dynamic reaches the best performance, although it passes only the TR test.

3.1.2. Gene Disease Network: Results w.r.t. gold standard G2

Edge rank. Table 4 of the Supplementary Material shows that, also in this case, *GTOM2* and *TOM* (static/dynamic) pass both EC and TR test, and *ECV* as well. However, the performance of all considered measures is worse, on average, than in the case of G1. Therefore, it seems that the involvement of gene pairs in common biological processes is more difficult to be inferred from GDN, than their influence on common diseases. This is possibly due to the complexity of cell processes, and to the fact that genes whose mutations are involved in the occurrence or progress of the same diseases, may act on different (e.g., complementary) biological processes.

Equivalent edge rank. Table 5 of the Supplementary Material shows that *NCC* (static/dynamic) is the only measure passing both significance tests, also for G2. However, in contrast with the case of the gold standard G2, this time the best performance is reached by *NCC* dynamic.

3.1.3. Human Disease Network: Results w.r.t. gold standard G3

Edge rank. Results (Tables 6 and 7 of Supplementary Material) are similar to those obtained for GDN on G1, although here *GTOM2* dynamic reaches more markedly the best performance, among those measures passing the significance tests for edge rank.

Edge equivalent rank. *NCC* dynamic is the only one passing both EC and TR tests, and it also outperforms all other measures, as in the case of GDN with G2. However, the performance of measures is on average slightly worse for HDN than for GDN, possibly due to the fact that the former is sparser than the latter (link density equal to 0.009 and 0.017, respectively).

3.1.4. Worm Gene Network: Results w.r.t. gold standard G4

Edge rank. Figure 2 shows that decremental measures remarkably outperform incremental ones, in terms of both performance and statistical significance. The best performing measure is *EB*, immediately followed by *ECC3* in the static case. This can be in part explained by the fact that the WGN is a very dense graph (link density equal to 0.897), as opposed to the GDN and HDN variants that are very sparse (0.017 and $0,009$, respectively).

Equivalent edge rank. Table 9 of the Supplementary Material shows that only *NCC* (static) passes both EC and TR tests, although the performance of all measures is worse than in edge rank, in analogy with results obtained for GDN and HDN.

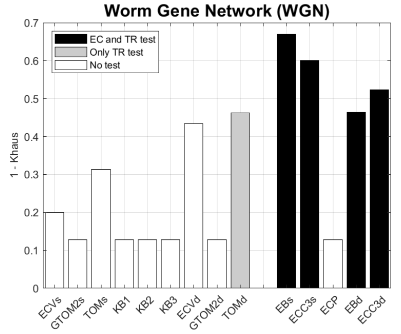

Fig. 2. Performance and statistical significance for the rankings returned by topological measures for WGN.

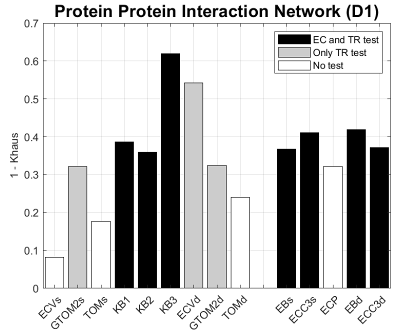

Fig. 3. Performance and statistical significance for the rankings returned by topological measures for the PPI network D1.

3.1.5. Protein-Protein Interaction Networks: Results w.r.t. gold standards G5, G6, G7

Edge rank. As already obtained for the previously discussed networks, also for the PPI networks the best performance of topological measures is reached on the less sparse network, that is, $D1$ (see Figure 3), having link density equal to 0.01 against the 0.007 and 0.001 of $D2$ and $Y2H$, respectively. However, results on the three considered PPI networks are comparable (Tables 10-15 of the Supplementary Material), in particular those of $D1$ and $D2$, where both incremental and decremental measures pass the statistical significance tests and the best performing measure is *KB3* for edge rank. As for $Y2H$, the best performance is reached by ECV, although only the decremental measures passes both EC and TR tests.

Edge equivalent rank. In analogy with all other analyzed networks, edge equivalent rank performs worse than edge rank. EGC is the only incremental measure returning statistical significant results, together with NB and other decremental measures. However, NCC dynamic reaches very good performance on $Y2H$, although it passes only the TR test.

3.2. Topological Views Applications

Capturing external knowledge. While in the previous paragraphs the closeness between ranks associated to the topological measures and a gold standard has been analyzed, here the attention is turned on a further exploration of the ability of topological measures in capturing information not directly encoded in the network, through the compact views induced by their respective ranks. In particular, the measures best performing in the case of D1 w.r.t. G5 have been considered, and the intersection between the complexes intercepted by edges involved in the topological and G5 ranks is computed at different view percentages (see Table 18 of the Supplementary Material). It is evident from these results that, even if the agreement between edges involved at the same percentage view is not always large, the agreement in terms of captured external knowledge (i.e., complexes) is in some cases highly pronounced. To this respect, KB3 confirms its best performance, being able to capture the 71% of complexes involved in the gold standard already at the 15% view, and although only the 9% of edges are in common between the two ranks at that view. At the same view also KB2 reaches good performance (the 50% of common complexes with the 5% of common edges), and all three measures based on dispersion perform very well at the 30% and 60% views.

Network clustering. Table 19 of the Supplementary Material shows a further application of the proposed paradigm, based on the consideration that topological views intrinsically induce a clustering of the considered network. In particular, as a byproduct of the research presented here, we have also investigated if the connected components intercepted at a certain percentage view may be associated to biologically significant groups. To this aim, PPI networks have been chosen as a case study and, for each topological measures and different percentage views (between 10% and 50%), the connected subgraphs induced by the edges involved in the views are compared against known complexes (the same discussed in Section *Protein-Protein Interaction Networks*). Results corresponding to the only measures for which both EC and TR tests are passed are shown in Table 16 of the Supplementary Material. In particular, the ability in discovering known biologically significant protein complexes through topological views is measured in terms of Precison, Recall and Fmeasure, as in [43]. Interestingly, for each of the three PPI networks, it is possible to identify at least one topological measure whose performance is superior to those of clustering methods proposed in the literature specifically for this task [43]. Moreover, EB in its dynamic setting is the best performing one on all three networks, and it is able to obtain very good accuracy even at only the 10% percentage view. These further results on one side are in agreement and generalize previous literature, such as the approach by [26] that shows how

EB in a dynamic setting can be successfully applied for community detection. On the other side, they open the way for the investigation of novel approaches for biological network clustering, where topological ranks may act as a boost.

4. DISCUSSION

Interesting insights follow from the above results.

First of all, ***edge rank** has shown to be more successful than edge equivalent rank in inferring the importance of network components*, when compared against external knowledge. Indeed, for all analyzed networks the edge ranks are closer to gold standard ranks than equivalent edge ranks. This leads to the consideration that edge centrality may be considered more representative of the information content intrinsically encoded into the biological network model, than node centrality, confirming and formalizing the intuitions by [30] and [5].

Another important finding is that *ranks computed in the **dynamic setting** often outperform those obtained in the static one*, in some cases even remarkably (e.g., for human). This confirms and generalizes the study presented by [26], where the dynamic asset, intended exactly as in our approach, is shown to outperform the static one in the context of a community detection algorithm based on the edge betweenness.

Interestingly, for both edge and equivalent edge rank, *it is possible to identify a measure*, *ECC3* (dynamic) and *NCC*, respectively, that is common to the three considered organisms in returning both statistically and biologically relevant ranks. Both *ECC3* and *NCC* are based on ***clustering coefficient***.

*There is a different **handful of best performing measures** for each of the three families of networks considered here*, and corresponding to a different organism (see Table 1 for the case of edge rank). That is, which measures perform better and return significant ranks strictly depends on the information intrinsically encoded inside the network, which also influences its topology. More specifically, for the human genotype-phenotype networks, previously shown to be difficult to cluster [27], measures based on neighborhoods outperform the others. On the genotype-phenotype network of worm, that is the most dense of those analyzed here, the best performing measure is that optimizing the network modularity [20]. Dispersion based measures [6] are the best compromise between statistical significance and biological relevance for the yeast protein-protein interaction networks considered in our analysis.

In summary, the previous result highlights that a carefully chosen set of topological measures provide ranks that closely follow the ones obtained from the considered gold standard, which encode biological knowledge. Operationally, this allows to consider only a percentage of a network detaining the most important features of the gold standard.

The results we have obtained also bring to light the subtle "multi-agent" relationship one should be aware of when using topological measures. The first two "agents" involved are rather straighforward, i.e., the information encoded in the network topology, and the ability of a measure to capture this latter. The third agent is more subtle since it involves the correlation between the information encoded in the network topology and that one is trying to discover, here represented by the gold standard. With regards to this third agent, the experiments on GDN seem to suggest that which measures perform better remains approximately unaffected for the same network, and with reference to different gold standards.

We have shown that, *even if the topological ranks are not syntactically identical to the gold standard ones, they can be considered semantically equivalent.* Indeed, the associated knowledge (e.g., set of intercepted biological complexes at different views) is not too much different in the two cases.

Also, it seems that the proposed paradigm works better on denser networks, possibly due to the fact that the encoded information is larger than for sparse networks.

5. CONCLUSION

We have proposed a comparative analysis of a set of outstanding topological measures, finalized to show which are the best performing ones in ranking nodes/edges of biological networks, according to their corresponding functional relevance. Although only some of the existing biological network types have been accounted for, the methodology presented here for the comparison of topological measures applies also to other types of biological networks which have not been included in this analysis (e.g., molecular regulatory networks). The provided overview confirms and sistematically summarizes previous results of the literature, still leading to novel conclusions. Moreover, it opens the avenue to further investigations, such as the study of lossy compression in biological networks, based on the succinct global representations induced by the choice of the most relevant topological views, rather than the entire network. Also, the introduced paradigm seems to be successful in boosting important tasks in the context of network analysis, such as network clustering. This could be further explored also for other applications. Another interesting open issue is to study if there are specific network classes for which static and dynamic ranks induce always the same partitions, and other ones for which partitions are always different in the two cases. Moreover, it has been shown that ranks based on edge topological measures outperform those based on node ones, in the proposed comparative analysis. This could be further investigated to understand if there are other problems for which this behaviour changes, e.g., studying

Table 1. Best performing measures for edge rank and the three considered organisms (human, worm and yeast), distinguished by those based on clustering coefficient (CC), neighborhoods (N), modularity (M) and dispersion (D) in the static (s) and dynamic (d) assets.

	Best Performing Measures			
Organism	**CC**	**N**	**M**	**D**
H. sapiens	ECC3 (d)	GTOM2(s, d)	–	–
		TOM (s, d)	–	–
C. elegans	ECC3 (d)	–	EB (s, d)	–
S. cerevisiae	ECC3 (d)	–	EB (s, d)	KB1,KB2,KB3

which proteins are more relevant in the occurrence and progress of human diseases. Finally, we plan to study in the future adaptive multicriteria approaches for the generation of new and best performing topological ranks, starting from the ones discussed here.

Funding

Part of the research presented here has been funded by the MIUR-PRIN research project “Multicriteria Data Structures and Algorithms: from compressed to learned indexes, and beyond”, grant n. 2017WR7SHH, by the “Algorithms, methods and tools for knowledge discovery in the context of Precision Medicine” project, funded by Indam (GNCS 2020), and by the MISE-PON AMABILE (B76G20000880005).

Biographical Notes

Mariella Bonomo is PhD student in Information and Communication Technologies at University of Palermo. Her research interests are focused on the analysis of complex networks in different domains, such as bioinformatics, social advertising and precision medicine.

Raffaele Giancarlo is Full Professor of Computer Science at Univesity of Palermo. He works on the design and analysis of algorithms and data structures for the solution of problems in the big data domain, with a main focus on biological data analysis and data compression.

Daniele Greco received a joint MS DEgree in Computer Science from University of Palermo and Universitè Paris-Est Marne-la-Vallèe, UPEMLV, in 2018. His main expertise is on data analysis in the biological domain.

Simona E. Rombo is Associate Professor of Computer Science at University of Palermo. Her research focuses on bioinformatics, including biological networks analysis. She is also CEO of Kazaam Lab, an innovative startup that provides software services based on big data and artificial intelligence technologies for precision medicine.

REFERENCES

1. M. L. Acencio and N. Lemke. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinform.*, 10:290, 2009.
2. S. Aerts, D. Lambrechts, S. Maity, et al. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544, 2006.
3. Y. Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in Networks. *Nature*, 466:761–764, 2010.
4. T. Alahakoon, R. Tripathi, N. Kourtellis, et al. K-path Centrality: A New Centrality Measure in Social Networks. In *Proc. of the 4th Workshop on Soc. Net. Syst.*, SNS ’11, pages 1:1–1:6, New York, NY, USA, 2011. ACM.
5. A. Aytes, A. Mitrofanova, C. Lefebvre, et al. Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer Cell*, 25(5):638 – 651, 2014.
6. L. Backstrom and J. Kleinberg. 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 &#38; Social Computing*, CSCW ’14, pages 831–841, New York, NY, USA, 2014. ACM.
7. A. Badkas, S. De Landtsheer, and T. Sauter. Topological Network measures for drug repositioning. *Briefings in Bioinformatics*, 22(4), 12 2020.
8. A. L. Barabasi. Scale-free Networks: A Decade and Beyond. *Science*, 325(5939):412–413, 2009.
9. A. L. Barabasi, N. Gulbahce, and J. Loscalzo. Network Medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68, 2011.
10. K. Basso, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37(4):382–390, 2005.
11. P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2:113–120, 1972.
12. M. Bonomo, A. La Placa, and S. E. Rombo. Prediction of lncRNA-disease associations from tripartite graphs. In *VLDB Workshops, DMAH 2020*, volume 12633 of *LNCS*, pages 205–210. Springer, 2020.
13. U. Brandes and T. Erlebach. *Network Analysis: Methodological Foundations (LNCS)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.
14. X. Chen and G. Yan. Novel human lncrna-disease association inference based on lncrna expression profiles. *Bioinform.*, 29(20):2617–2624, 2013.

15. E. Estrada and J. A. Rodríguez-Velázquez. Subgraph centrality in complex networks. *Phys. Rev. E*, 71:056103, May 2005.
16. R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17:134–160, 2003.
17. R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, 20:628–648, 2006.
18. V. Fionda, L. Palopoli, S. Panni, and S. E. Rombo. Protein-protein interaction network querying by a "focus and zoom" approach. In Mourad Elloumi, Josef Küng, Michal Linial, Robert F. Murphy, Kristan Schneider, and Cristian Toma, editors, *Proc. of Bioinformatics Res. and Develop.(BIRD) 2008, Vienna, Austria, July 7-9*, volume 13 of *Communications in Computer and Information Science*, pages 331–346. Springer, 2008.
19. V. Fionda, L. Palopoli, S. Panni, and S. E. Rombo. A technique to search for functional similarities in protein-protein interaction networks. *Int. J. Data Min. Bioinform.*, 3(4):431–453, 2009.
20. L. C. Freeman. Centrality in Social Networks conceptual clarification. *Social Networks*, 1(3):1978–1979, 2012.
21. A. C. Gavin et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, 2006.
22. Gene-Ontology-Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 11 2014.
23. R. Giancarlo, D. Greco, F. Landolina, and S. E. Rombo. Network Centralities and Node Ranking. *Encyclopedia of Bioinf. and Comp. Biol.*, 1:950–957, 2019.
24. R. Giancarlo, S. E. Rombo, and F. Utro. Epigenomic *k*-mer dictionaries: shedding light on how sequence composition influences *in vivo* nucleosome positioning. *Bioinformatics*, 31(18):2939–2946, 2015.
25. R. Giancarlo and F. Utro. Algorithmic paradigms for stability-based cluster validity and model selection statistical methods, with applications to microarray data analysis. *Theoretical Computer Science*, 428:58–79, 2012.
26. M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. of the National Academy of Sciences*, 99:7821–7826, 2002.
27. K.-I Goh, M. E. Cusick, D. Valle, et al. The human disease network. *Proc. of the National Academy of Sciences*, 104(21):8685–8690, 2007.
28. R. A. Green, H.L. Kai, A. Audhya, et al. A High-Resolution C. elegans Essential Gene Network Based on Phenotypic Profiling of a Complex Tissue. *Cell*, 145:470–482, 2011.
29. D. Guala and E. L. L. Sonnhammer. A large-scale benchmark of gene prioritization methods. *Scientific Reports*, 7(1), 2017.
30. X. He and J. Zhang. Why Do Hubs Tend to Be Essential in Protein Networks? *PLOS Genetics*, 2(6):1–9, 06 2006.
31. V. Janjic and N. Przulj. Biological function through network topology: a survey of the human diseasome. *Briefings in Functional Genomics*, 11(6):522–532, 09 2012.
32. B. H. Junker, D. Koschützki, and Falk Schreiber. Exploration of Biological Network centralities with CentiBiN. *BMC Bioinform.*, 7:219, 2006.
33. D. Koschützki and F. Schreiber. Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. *Gene Regulation and Systems Biology*, 2:GRSB.S702, 2008.
34. P. V. Marsden and K. E. Campbell. Measuring Tie Stength. *Social Forces*, 63:482–501, 1984.
35. P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Mixing Local and Global Information for Community Detection in Large Networks. *J. Comput. Syst. Sci.*, 80(1):72–87, February 2014.
36. H. W. Mewes et al. MIPS: a database for genomes and protein sequences. *Nuc. Ac. Res.*, 28(1):37–40, 2000.
37. H. W. Mewes et al. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Research*, 34(suppl1):D169–D172, 2006.
38. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, 2002.
39. S. Panni and S. E. Rombo. Searching for repetitions in biological networks: methods, resources and tools. *Briefings in Bioinformatics*, 16(1):118–136, 2015.
40. L. Parida, C. Pizzi, and S. E. Rombo. Irredundant tandem motifs. *Theoretical Computer Science*, 525:89–102, 2014.
41. C. Pizzuti and S. E. Rombo. *PINCoC*: A co-clustering based approach to analyze protein-protein interaction networks. In *Intelligent Data Engineering and Automated Learning - IDEAL 2007, 8th Int. Conf., Birmingham, UK, December 16-19, 2007, Proceedings*, volume 4881 of *LNCS*, pages 821–830. Springer, 2007.
42. C. Pizzuti and S. E. Rombo. Multi-functional Protein Clustering in PPI Networks. In Mourad Elloumi, Josef Küng, Michal Linial, Robert F. Murphy, Kristan Schneider, and Cristian Toma, editors, *Proc. of Bioinformatics Res. and Develop. (BIRD) 2008, Vienna, Austria, July 7-9*, volume 13 of *Communications in Computer and Information Science*, pages 318–330. Springer, 2008.
43. C. Pizzuti and S. E. Rombo. Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352, 2014.
44. N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 01 2007.
45. F. Radicchi, C. Castellano, F. Cecconi, et al. Defining and identifying communities in networks. *Proc. of the National Academy of Sci.*, 101:2658–2663, 2004.
46. E. Ravasz, A. L. Somera, D. A. Mongru, et al. Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297:1551–1555, 2002.
47. M. Santolini and A. L. Barabási. Predicting perturbation patterns from the topology of biological networks. *Proc Natl Acad Sci USA*, 115(27):E6375–E6383, 2018.
48. R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.
49. J. Wang, M. Li, J. Chen, and Y. Pan. 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, May 2011.
50. D. J. Watts. *Small worlds.* Princeton University Press, Princeton, 1999.
51. A. Yip and S. Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, 8:22, 2007.
52. H. Yu et al. High-Quality Binary Protein Interaction Map of the Yeast Interactome network. *Science*, 322(5898):104–110, 2008.
53. N. Zaki, J. Berengueres, and D. Efimov. Prorank: A Method for Detecting Protein Complexes. In *Proc. of the 14th Annual Conference on Genetic and Evolutionary Computation*, GECCO'12, pages 209–216, 2012.
54. Y. Zhang, Z. Wang, and Y. Wang. Multi-hierarchical profiling: an emerging and quantitative approach to characterizing diverse Biological Networks. *Briefings in Bioinformatics*, 2016.
55. E. Zotenko, J. Mestre, D. P. O'Leary, and T. M. Przytycka. 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, 2008.