Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original paper

Subject Section

Graphlet Laplacians for topology-function and topology-disease relationships

Sam F. L. Windels[1], Noël Malod-Dognin[2] and Nataša Pržulj[1,2,3,*]

[1] Department of Computer Science, University College London, London, WC1E 6BT, United Kingdom
[2] Barcelona Supercomputing Center, 08034 Barcelona, Spain
[3] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXIX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Laplacian matrices capture the global structure of networks and are widely used to study biological networks. However, the local structure of the network around a node can also capture biological information. Local wiring patterns are typically quantified by counting how often a node touches different graphlets (small, connected, induced sub-graphs). Currently available graphlet-based methods do not consider whether nodes are in the same network neighbourhood.
Contribution: To combine graphlet-based topological information and membership of nodes to the same network neighbourhood, we generalize the Laplacian to the Graphlet Laplacian, by considering a pair of nodes to be 'adjacent' if they simultaneously touch a given graphlet.
Results: We utilize Graphlet Laplacians to generalize spectral embedding, spectral clustering and network diffusion. Applying Graphlet Laplacian based spectral embedding, we visually demonstrate that Graphlet Laplacians capture biological functions. This result is quantified by applying Graphlet Laplacian based spectral clustering, which uncovers clusters enriched in biological functions dependent on the underlying graphlet. We explain the complementarity of biological functions captured by different Graphlet Laplacians by showing that they capture different local topologies. Finally, diffusing pan-cancer gene mutation scores based on different Graphlet Laplacians, we find complementary sets of cancer related genes. Hence, we demonstrate that Graphlet Laplacians capture topology-function and topology-disease relationships in biological networks.
Availability: http://www0.cs.ucl.ac.uk/staff/natasa/graphlet-laplacian/index.html
Contact: natasa@cs.ucl.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Systems biology is flooded with large scale "omics" data. Genomic, proteomic, interactomic, metabolomic and other data, are typically modeled as networks (also called graphs). This abundance of networked data started the fields of network biology, allowing us to uncover molecular mechanisms of a broad range of diseases, such as rare Mendelian disorders (Smedley *et al.*, 2014), cancer (Leiserson *et al.*, 2015), and metabolic diseases (Baumgartner *et al.*, 2018). In personalized medicine, network analysis is applied to the tasks of bio-marker discovery (Li *et al.*, 2015), patient stratification (Gligorijević *et al.*, 2016) and drug repurposing (Durán *et al.*, 2017).

Many network analysis methods use the Laplacian matrix as it captures the global wiring patterns of a network (see section 1.1). These methods include spectral clustering, spectral embedding and network diffusion. Each of these families of methods relies on the fact that the eigendecomposition of the Laplacian matrix naturally uncovers network clusters (see section 1.2). Applications of spectral embedding include visualizing genetic ancestry (Lee *et al.*, 2010) and pseudo-temporal ordering of single-cell RNA-seq profiles (Campbell *et al.*, 2015). Applications of spectral clustering include detection of functional sub-network modules in single-cell genomic networks (Bartlett *et al.*, 2017) and

identification of functional modules in co-regulatory networks (Luo *et al.*, 2018). Network diffusion methods are widely used for protein function prediction (Cao *et al.*, 2013) and discovery of disease genes and disease modules, see Cowen *et al.* (2017) for a full review.

1.1 Laplacian matrix definition

The *Laplacian matrix* captures the global structure of a network: for each node it captures the adjacency relationship with other nodes (i.e. who are its neighbours) and its degree centrality (a measure of the node's importance in the network). In a network, $G(V, E)$, two nodes, u and v, are *adjacent* if there exists an edge $(u, v) \in E$ connecting them. The adjacency of all nodes in graph G is represented in an $n \times n$ symmetric *adjacency matrix* A:

$$A(u,v) = \begin{cases} 1 & \text{if } (u,v) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The *neighbourhood* of a node is defined as the set of nodes adjacent to it. A node's *degree* is the size of its neighbourhood, or equivalently, the number of nodes that are adjacent to it. The *degree matrix* of G is defined as the diagonal matrix, D, where $D(u, u)$ is equal to the degree of node u:

$$D(u,v) = \begin{cases} deg(u) & \text{if } u = v \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The Laplacian, $\mathcal{L}$, is defined as $\mathcal{L} = D - A$. The symmetrically normalized Laplacian, L^{sym}, is defined as $L^{sym} = D^{-1/2}\mathcal{L}D^{-1/2}$.

1.2 Laplacian matrix eigendecomposition

j Spectral clustering, spectral embedding and network diffusion analyze networks based on the eigendecomposition of the Laplacian matrix naturally uncovering densely connected sub-networks present in the network. The *eigendecomposition* of $\mathcal{L}$, is defined as $\mathcal{L}U = U\Lambda$, where the i-th column of the $n \times n$ matrix U is known as the i-th *eigenvector* and Λ is the diagonal matrix whose diagonal elements are the corresponding *eigenvalues*. Uncovering densely connected sub-networks present in the network (i.e. network clustering), can be defined as solving the ratio-cut problem: to cut a network into d similar-sized sub-networks whilst minimizing the number of edges being cut. An approximation of this problem is formulated as follows:

$$\underset{U \in \Re^{n \times d}}{\text{minimize}} \quad trace(U^T \mathcal{L} U) \qquad \text{subject to: } U^T U = I, \tag{3}$$

where each column of U is a normalized indicator vector assigning each node to one of the d sub-networks. This problem is solved by d normalized eigenvectors of $\mathcal{L}$ associated with d smallest eigenvalues, illustrating how the Laplacian matrix captures clusters present in the network.

1.3 Matrix alternatives to the Laplacian

Laplacian matrices only capture direct interactions between nodes. To capture the influence of long-range interactions between nodes, Estrada (2012) proposed the *k-path Laplacian* by generalizing the concepts of adjacency and degree. The k-path Laplacian defines a pair of nodes u and v to be *k-adjacent* if the shortest path distance between them is equal to k. Analogously, *k-path degree*, $\deg_k(u)$, generalizes the concept of the degree to the number of length k shortest paths that have node u as an endpoint. The k-path Laplacian, $\mathcal{L}_k^P$, is defined as:

$$\mathcal{L}_k^P(u,v) = \begin{cases} -1 & \text{if } d(u,v) = k \\ \deg_k(u) & \text{if } u = v \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Vicus is an alternative to the Laplacian that captures the intricacies of a network's local structure (Wang *et al.*, 2017) based on network label diffusion. Label diffusion is defined as $P = BQ$, where the $n \times d$ matrix Q assigns the n nodes of network G to one of d possible labels (for labeled nodes), B is an $n \times n$ diffusion matrix, and the reconstructed matrix P is an $n \times d$ matrix used for predicting labels for unlabeled nodes. To give Vicus its 'local' interpretation, the label diffusion process determining B is constrained to diffuse information of each node only to its direct neighbourhood. Under given assumptions and defining Vicus as $\mathcal{L}^V = (I - B^T)(I - B)$, it was shown that Q can be learned as the eigenvectors of $\mathcal{L}^V$. As Q captures the local connectivity between nodes that is implied by the 'localized' diffusion matrix B and can be computed as the eigenvectors of $\mathcal{L}^V$, Vicus is interpreted as a Laplacian matrix. Vicus is applied to protein module discovery and ranking of genes for cancer subtyping (Wang *et al.*, 2017).

1.4 Problem

All Laplacian based applications are based on the same underlying principle of *guilt by association*, inferring information on a given node based on the group of nodes it is most tightly connected with. However, alternative approaches have inferred information on a given node based on the shape of its interaction pattern, typically independent of the identity of the nodes it is interacting with. These alternative approaches are based on graphlets, small connected sub-graphs (see section 2.1 for a formal definition), to capture the local topology around a node in a network. For example, graphlet based methods have been applied to predict protein function (Milenković and Pržulj, 2008; Davis *et al.*, 2015) and to identify new cancer genes (Milenković *et al.*, 2010) directly from the similarities in terms of their interaction patterns in PPI networks.

Alternatives to the Laplacian matrix that take local topology into account have been suggested. The k-path Laplacian captures the influence of long-range interactions between nodes, but ignores short-range interactions. Vicus captures local topology around each node as the strength of its connection to its neighbours after applying a localized label diffusion algorithm. Although Vicus is focused on capturing local topology, it lacks interpretability from a structural perspective.

1.5 Contribution

We introduce the Graphlet Laplacian, allowing us to analyze nodes based on their network neighbourhoods, whilst restricting the pattern of their interactions to that of a prespecified graphlet. Hence, each graphlet (Figure 1-A) has its own corresponding Graphlet Laplacian. We generalize spectral embedding, spectral clustering and network diffusion to utilize Graphlet Laplacians. Through graphlet-generalized spectral clustering of model networks and biological networks, we show that different Graphlet Laplacians capture different local topologies. By applying graphlet-generalized spectral embedding, we visually demonstrate that Graphlet Laplacians capture biological functions as well. We quantify this through graphlet-generalized spectral clustering analysis. We show that Graphlet Laplacians are not only as biologically relevant as alternative Laplacian matrices, but also capture complementary biological functions. Finally, by graphlet-generalized diffusing of pan-cancer gene mutation scores on the human PPI network, we show that Graphlet Laplacians capture complementary disease mechanisms. We compare our results against those based on alternative state the art Laplacian matrices. A similar methodology based on network motifs was presented by Benson *et al.* (2016) for spectral clustering of directed networks. compare our results to those of the standard Vicus.

2 Materials and methods

2.1 Graphlet Laplacian definition

We generalize the Laplacian matrix so that it can capture the local topology of a network around a node in a broader sense than the identity of its direct neighbours. One of the most sensitive methodologies to capture network topology around a node are *graphlets*: small, connected, non-isomorphic, induced sub-graphs of a large network (Pržulj *et al.*, 2004). All graphlets up to four nodes are depicted in Figure 1-A. To illustrate how graphlets can be used to quantify the local topology around a node, consider node a in the dummy network presented in Figure 1-B. Graphlet G_1 (i.e. a three node path) that touches node a can be found in this dummy network twice: via paths a-b-c and a-b-e. Node a is said to touch G_1 twice. By making these counts for a given node over all graphlets, the local network topology for a given node can be quantified by means of a vector, as illustrated for node a in Figure 1-C. Here we see that node a can be found as part of an edge once (G_0), as part of a three node path twice (G_1), never as part of a triangle (G_2) and so on.

Having established that by counting how often a node touches graphlets can be used to quantify its local topology, we go on and generalize the concept of the Laplacian to that of a Graphlet Laplacian by generalizing the definitions of adjacency and degree to ones based on graphlets. First, we define two nodes u and v of G to be *graphlet-adjacent* with respect to a given graphlet, G_k, if they simultaneously touch G_k. Going back to our previous example, we find that nodes a and b are graphlet-adjacent w.r.t. graphlet G_1 twice, as G_1 can be induced on the dummy network twice: via paths a-b-c and a-b-e, each time including both nodes a and b. Similarly, nodes a and c and nodes a and e are graphlet adjacent only once, w.r.t. graphlet G_1.

Given this extended definition of adjacency, we define the graphlet based adjacency matrix as:

$$A_k(u,v) = \begin{cases} a^k_{uv} & \text{if } u \neq v \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where a^k_{uv} is equal to the number of times nodes u and v are graphlet-adjacent w.r.t graphlet G_k. Analogously, the *graphlet degree* generalizes the node degree as the number of times node u touches graphlet G_k. We extend the degree matrix to the *Graphlet Degree matrix* for graphlet G_k,

$$D_k(u,v) = \begin{cases} d^k_u & \text{if } u = v \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where d^k_u is the number of times node u touches graphlet G_k. For an underlying graphlet G_k, we define the *Graphlet Laplacian* $\mathcal{L}^G_k$, as:

$$\mathcal{L}^G_k = D_k - (A_k/\theta). \tag{7}$$

where $\theta = size(G_k) - 1$. As opposed to the Laplacian simply capturing for each node its neighbours, the Graphlet Laplacian $\mathcal{L}^G_k$ captures for each node how strongly (i.e. frequently) each node is connected in the shape of G_k with each of the other nodes. $\mathcal{L}^G_0$ and $\mathcal{L}^G_1$ are illustrated in Figure 1-C. Finally, note that the Graphlet Laplacian for graphlet G_0, $\mathcal{L}^G_0$, is equivalent to the standard Laplacian, $\mathcal{L}$.

2.2 Graphlet Laplacian properties

To allow for an easy interpretation of the Graphlet Laplacian for each graphlet, G_k, we introduce the two-step transformation function, T, which maps graph G to its Graphlet Laplacian representation: $T(G, G_k) = \mathcal{L}^G_k$. First, T converts $G = \{V, E\}$ to a weighted network $G' = \{V, E'\}$, where the weight of each edge (u, v) in G' corresponds to $a^k_{uv}/(size(G_k) - 1)$ measured in G. Next, T converts G' to its standard Laplacian representation. This shows that the Graphlet Laplacian can be interpreted as the Laplacian of an undirected weighted network. Therefore, the Graphlet Laplacian retains the following key properties of the Laplacian:

- The Graphlet Laplacian, $\mathcal{L}^G_k$, is symmetric and positive semi-definite.
- The smallest eigenvalue is 0 and the corresponding eigenvector is the constant vector $\mathbf{1}$. eigenvector
- The Graphlet Laplacian has n non-negative, real-valued eigenvalues: $0 = \lambda^k_1 \leq \lambda^k_2, \ldots, \lambda^k_n$.
- The multiplicity of the eigenvalue 0 equals the number of connected components in G', which we refer to as *graphlet based components*.

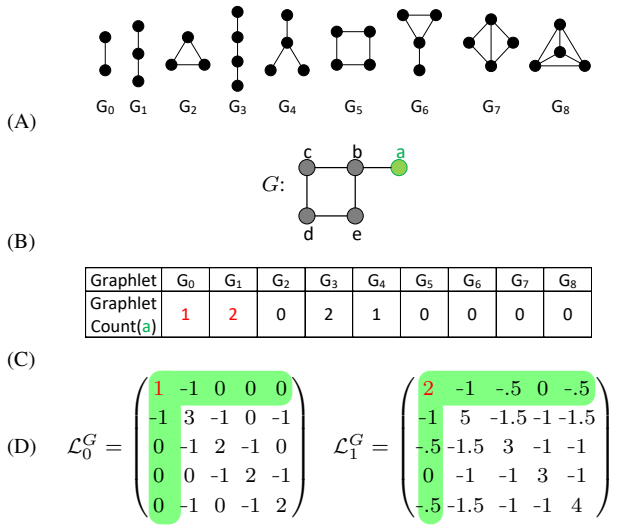

Fig. 1. Illustration graphlets and Graphlet Laplacians. Node a is coloured in green throughout. The graphlet counts of node a for graphlet G_0 and G_1 are coloured in red throughout. **A:** All graphlets with up to 4 nodes, labeled G_0 to G_8. **B:** A dummy network. **C:** A vector of graphlet counts describing the local topology of node a in the example network, G. Node a touches graphlet G_0 via edge (a, b). Node a touches graphlet G_1 twice, via paths a-b-c and a-b-e. **D:** The Graphlet Laplacians for graphlets G_0 and G_1, applied on the network, G, shown in panel B. The diagonal elements correspond to the graphlet counts of each node; e.g. $\mathcal{L}^G_0(1,1)$ is equal to 1, the number of times node a touches graphlet G_0, $\mathcal{L}^G_1(1,1)$ is equal to 2, the number of times node a touches graphlet G_1. The off-diagonal elements correspond to the number of times two nodes touch a given graphlet together, scaled by $size(G_k) - 1$. $\mathcal{L}^G_0(1,2) = -1$, as a and b form G_0 once and $size(G_0) - 1 = 1$. $\mathcal{L}^G_1(1,2) = -1$, as a and b form G_1 twice and $size(G_1) - 1 = 2$.

2.3 Spectral embedding

Spectral Embedding embeds a network in a lower dimensional space, placing nodes close in space if they share many neighbours. Here, we generalize the Laplacian Eigenmap embedding algorithm (Belkin and Niyogi, 2003) so that two nodes are embedded close in space if they frequently simultaneously touch a given graphlet. Given an unweighted network G with n nodes, we find a low dimensional embedding, $Y = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \mathbb{R}^{d \times n}$ such that if nodes u and v are frequently graphlet-adjacent with respect to graphlet G_k, then $\mathbf{y}(u)$ and $\mathbf{y}(v)$ are close in the d-dimensional space by solving:

$$\begin{aligned} \underset{Y}{\text{minimize}} \quad & \sum_{u=1}^{n}\sum_{v=1}^{n} A_k(u,v)\,\|\mathbf{y}_u - \mathbf{y}_v\|^2 \\ \text{subject to :} \quad & YD_k\mathbf{1} = \mathbf{0} \text{ and } YD_kY^T = I, \end{aligned} \tag{8}$$

where A_k is the graphlet-based adjacency matrix of G for graphlet G_k, D_k is the graphlet-based degree matrix of G for graphlet G_k. The columns of Y are found as the generalized eigenvectors associated with the 2^{nd} to $(d+1)^{th}$ smallest generalized eigenvalues solving $Y\mathcal{L}_k^G = \Lambda Y D_k$, where Λ is the diagonal matrix with the generalized eigenvalues along its diagonal.

2.4 Spectral clustering

Spectral clustering uncovers groups of nodes in a network that form densely connected network clusters. By generalizing spectral clustering to Graphlet Laplacian based spectral clustering, we are able to identify network components that are densely connected with respect to a given graphlet. Many different variations of spectral clustering exist (Von Luxburg, 2007). Aiming for a balanced clustering, we generalize normalized spectral clustering as defined by Ng *et al.* (2002) to use different Laplacians including Graphlet Laplacians, all denoted by a generic $\mathcal{L}$ in algorithm 1. We skip the normalization step (i.e. step 1) for Vicus, as Vicus is already normalized.

Algorithm 1 Normalized spectral clustering

Input A network G with n nodes, and a number of clusters d.
Output d clusters of the n nodes of G.

1: Compute the Laplacian matrix, $\mathcal{L}$, and corresponding diagonal matrix, D, for the network G.
2: Compute the normalized Laplacian as: $\mathcal{L}^{sym} = D^{-1/2}\mathcal{L}D^{-1/2}$.
3: Compute the d eigenvectors of $\mathcal{L}^{sym}$ associated with its d smallest eigenvalues: $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_d] \in \mathbb{R}^{n\times d}$.
4: Normalize Y so that each row has unit norm.
5: Cluster the n points $\{\mathbf{y}\}_{u=1}^n$ into d groups using k-means.

For each network, we determine the numbers of clusters, d, by using the rule of thumb: $d \approx \sqrt{n/2}$ (Kodinariya and Makwana, 2013). In the Supplement Section 1, we present the justification for this approach, based on inspection of the spectra of different Laplacian matrices of each network. Because of the heuristic nature of spectral clustering, we perform 20 runs for each clustering and consolidate them into a single clustering applying ensemble clustering (Ghosh and Strehl, 2002).

2.5 Network diffusion

Network diffusion refers to a family of related techniques, which propagate information on nodes through the network. Here, we will focus on generalizing the diffusion kernel to *graphlet based diffusion kernel*. The diffusion kernel is often called the 'heat kernel', as it can be viewed as describing the flow of heat originating from the nodes across the edges of a graph with time. In network biology, nodes typically represent genes and 'heat' on a node represents experimental measurements. For a set of n nodes, these measurements are encoded in vector $P_0 \in \mathbb{R}^n$. Information is diffused as follows: $P = HP_0$, where H is a diffusion kernel. For a given graphlet G_k, we define the graphlet based diffusion kernel, H_α^k, as:

$$H_\alpha^k = e^{-\alpha\mathcal{L}_k^G}, \tag{9}$$

where the parameter $\alpha \in \mathbb{R}$ controls the level of diffusion. This way, diffusion of information on nodes propagates between nodes restrained by how often they form a given graphlet G_k together.

2.6 Topological dissimilarity of networks

2.6.1 Graphlet Correlation Distance

The Graphlet Correlation Distance (GCD-11) is the current state of the art heuristic for measuring the topological distance between networks (Yaveroğlu *et al.*, 2014). First, the global wiring pattern of a network is captured in its Graphlet Correlation Matrix (GCM), an 11×11 symmetric matrix comprising the pairwise Spearman's correlations between 11 different graphlet based counts over all nodes in the network. The Graphlet Correlation Distance between two networks is computed as the Euclidean distance of the upper triangle values of their GCMs.

2.6.2 Non-graphlet based network descriptors

The difference between the following non-graphlet based network descriptors can be used to measure the distance between two networks:

- The *degree distribution* is the distribution of node degrees over all nodes. It is summarized as a vector of counts, i.e. the k^{th} value is the number of nodes that have degree k. To measure the distance between two networks, this vector is first rescaled to reduce the contribution of higher degree nodes. The pairwise distance between two networks is the euclidean distance between their rescaled degree distribution vectors. For more details, see (Yaveroğlu *et al.*, 2014).
- The *diameter* of a connected network is the maximum shortest path distance that is observed among all node pairs. The distance between two networks is the absolute difference of their diameters.
- The *average clustering coefficient* is the total number of three node cliques in the network over the number of possible three node cliques in the network. The distance between two networks is the absolute difference of their average clustering coefficient.

2.7 Cluster enrichment analysis

To assess if a cluster of genes is biologically relevant, we measure if it is statistically significantly enriched in a specific biological annotation term by applying the hyper-geometric test. That is, we consider each cluster as a 'sampling without replacement', in which each time we find a given annotation, we count that as a 'success'. The probability of observing the same or higher enrichment (i.e. successes) of the given annotation purely by chance is equal to:

$$p = 1 - \sum_{i=0}^{X-1} \binom{K}{i}\binom{M-K}{N-i} / \binom{M}{N}, \tag{10}$$

where N is the number of annotated genes in the cluster, X is the number of genes annotated with the given annotation in the cluster, M is the number of annotated genes in the network, and K is the number of genes annotated with the given annotation in the network. An annotation is considered to be statistically significantly enriched if its enrichment p-value is lower than or equal to 5% after application of the Benjamini and Hochberg correction for multiple hypothesis testing.

2.8 Data

2.8.1 Real biological network data collection

We create three types of molecular interaction networks for human and baker's yeast (*S. cerevisiae*) by collecting the following data: experimentally validated protein-protein interactions (PPIs) from IID version 2018-05 (Kotlyar *et al.*, 2016) and BioGRID version 3.4.161 (Stark *et al.*, 2006), genetic interactions from the same version of BioGRID, and gene co-expressions from COXPRESdb version 6.0 (Okamura *et al.*, 2015).

2.8.2 Random model network generation

For each of the following eight random network models we generate ten networks containing 2,000 nodes at edge density of 1.5%: Erdős-Rènyi random graphs (ER) (Erdős Paul and Rényi Alfréd, 1959), generalized random graphs with the degree distribution matching to the input graph

(ER-DD) (Newman, 2010), Barabási-Albert scale-free networks (SF) (Barabási and Albert, 1999), geometric random graphs (GEO) (Penrose, 2003), geometric graphs that model gene duplications and mutations (GEO-GD) (Pržulj *et al.*, 2010), stickiness-index based networks (Sticky) (Pržulj and Higham, 2006), popularity-similarity optimization graphs (PSO) (Papadopoulos *et al.*, 2012) and nonuniform PSO graphs (nPSO) (Muscoloni and Cannistraci, 2018). A summary on the basic properties of these networks and how to generate them can be found in Supplement Section 2.1.

2.8.3 Biological annotations

For each gene in our biological networks, we collect the most specific experimentally validated biological process annotations (BP), cellular component annotations (CC) and molecular function annotations (MF) present in the Gene Ontology (GO) (Ashburner *et al.*, 2000).

2.8.4 Cancer gene annotations

We collect the pan-cancer gene mutation frequency scores computed by Leiserson *et al.* (2015) for the purpose of detecting pan-cancer disease modules. Leiserson *et al.* (2015) collected raw pan-cancer mutation data, such as SNV's, indels and CNA's, from the TCGA database (Kandoth *et al.*, 2013). These data were filtered to exclude statistical outliers and include only the samples (corresponding to a patient) for which SNV and CNA data were available. The resulting data set contains mutations on 11,565 genes across 3,110 patients in cancers across 20 different tissues. Additionally, we collect the sets of known cancer driver genes in all available tissues from IntOGen (Gonzalez-Perez *et al.*, 2013) and Cosmic (Futreal *et al.*, 2004).

3 Results and discussion

We investigate the potential usage of Graphlet Laplacians to analyze network data via embedding, clustering and network diffusion experiments. We consider Graphlet Laplacians for graphlets with up to four nodes. We compare our results to the state of the art Laplacian matrices: the standard Laplacian, the k-path Laplacian and Vicus. We consider path lengths up to three for the k-path Laplacian, corresponding to the maximum size of the considered graphlets underlying the Graphlet Laplacian. We set Vicus' diffusion parameter to 0.9, as this value is recommended in the original paper (Wang *et al.*, 2017) and leads to the largest number of enriched functions (see Section 3.2).

3.1 Graphlet Laplacians capture different local topologies

While the standard Laplacian simply captures the direct neighborhoods of nodes and can be used to cluster densely connected nodes together, the graphlet-based neighborhood captured by our Graphlet Laplacian allows for clustering of nodes that strongly participate in a given graphlet of interest. Because different graphlets capture different local topologies around nodes in a network (e.g., G_3 involve paths while G_8 involves cliques), clusters obtained by using different Graphlet Laplacian are expected to possess different topological features, which we assess as follows.

To assess if two graphlet Laplacians, $\mathcal{L}_i^G$ and $\mathcal{L}_j^G$, capture different topologies, we apply each Laplacian to cluster nodes in a network using Graphlet Laplacian based spectral clustering. The resulting clusters are used to partition the network into two sets of sub-networks, by inducing the sub-networks from each clustering. $\mathcal{L}_i^G$ and $\mathcal{L}_j^G$ capture different topologies if the corresponding sets of sub-networks have significantly different topology, which we measure by the overlap of two distributions: the distribution of GCD-11 distances between the sub-networks produced from $\mathcal{L}_i^G$ with the sub-networks produced from $\mathcal{L}_j^G$ and distribution of GCD-11 distances between the sub-networks produced from $\mathcal{L}_i^G$. The two Graphlet Laplacians capture statistically significantly different topologies if the Wilcoxon-Mann-Whitney U-test (MWU) between the two distributions of distances is lower than or equal to 5% (see Figure 2 for the case of $\mathcal{L}_0^G$ and $\mathcal{L}_4^G$). For each type of model network, we perform this test ten times and report the least significant p-value for each pairwise comparison of Graphlet Laplacian based sub-networks. We also considered the following non-graphlet based network distance measures: degree distribution distance, diameter distance, and average clustering coefficient distance (see section 2.6.2). In general and independent of the network distance measure used, clusters obtained from different Graphlet Laplacians are typically statistically significantly topologically different at the 5% significance level. This is true across all of our biological networks and most of our model networks, with some exceptions in geometric models which are known to have homogeneous structure. Thus, Graphlet Laplacians not only capture network cluster that have different topology, but can also be used to measure the structural homogeneity of a given network.

We illustrate this by investigating how the parameters of the PSO/nPSO model networks influence the topological homogeneity of the networks generated, see Supplementary Figures 7,8 and 9. At a low temperature (i.e. nodes are connected to nearby nodes) and low number of communities (i.e. the angle of each node is sampled from a univariate Gaussian), both types of networks are homogeneous. As temperature increases, newly added nodes are more uniformly connected in space (i.e. are more randomly connected), making the generated PSO and nPSO networks closer to ER networks, thus breaking the homogeneous structure. In nPSO networks, increasing the number of communities increases the homogeneity of the networks. This effect stronger lower temperatures.

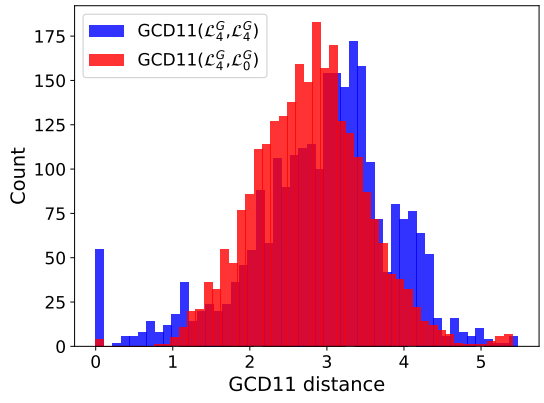

Fig. 2. Comparison of topological distance distributions between sub-networks captured by two different Graphlet Laplacians in the human PPI network. The distribution of GCD-11 distances between the sub-networks from $\mathcal{L}_0^G$ (in blue) is statistically significantly different from the distribution of GCD-11 distances between the sub-networks from $\mathcal{L}_0^G$ and the sub-networks from $\mathcal{L}_4^G$ (in red) with MWU p-values <5%. This means that$\mathcal{L}_0^G$ and $\mathcal{L}_4^G$ capture different topologies in the human PPI network.

3.2 Different Graphlet Laplacians capture different biological functions

In biological networks, genes having similar functions tend to be densely connected to each other (Hartwell *et al.*, 1999), which is why spectral clustering based on the standard Laplacian matrix has been used to uncover functional regions in networks (Bu *et al.*, 2003). Alternatively, graphlets have been used to show that functionally related genes tend to be similarly wired, independent of them being densely connected (Milenković and

Pržulj, 2008). As graphlet Laplacians capture both types of information, they should also capture biological functions.

To informally visualize this, we perform spectral embedding. We focus on the embedding of the yeast GI network, for which we use 14 core biological process annotations defined by Costanzo *et al.* (2016). We illustrate the spectral embedding of the symmetrically normalized $\mathcal{L}_3^G$ Graphlet Laplacian in Figure 3. The embeddings of the other Laplacian matrices of the yeast GI network can be found in the Supplement, Section 3. As seen in Figure 3, the spectral embedding of $\mathcal{L}_3^G$ correctly groups and separates the biological processes of 'nuclear cytoplasmic transport', 'metabolism / mitochondria', 'Golgi / endosome / vacuole sorting' and 'Chrom. seg. / kinetoch. / spindle / micro tub.'. In the supplement, we illustrate that Vicus and the Laplacian fail to find any grouping at all, placing all of the nodes in the same dense cluster. Embeddings based on $\mathcal{L}_2^P$ and $\mathcal{L}_3^P$ succeed in separating different genes into different clusters, but without grouping them in a biologically meaningful way.

Next, we aim to quantify this result by measuring the difference in functions captured by different Graphlet Laplacians. We apply Graphlet Laplacian based spectral clustering for each graphlet on our set of human molecular networks and assess the functional enrichments in terms of the percentage of clusters enriched and the total number of annotations enriched (Figure 4). Additionally, we create a baseline to validate the statistical significance of our enrichment results. We perform the same experiment 100 times with randomized GO-annotations. We do this by swapping the sets of gene annotations in the molecular networks such that no gene has its original set of annotations.

First, we observe that clusterings based on all Graphlet Laplacians but $\mathcal{L}_4^G$ tend to be of similar quality as those based on the standard Laplacian or Vicus, both in terms of percentage of clusters enriched as well as total number of annotations enriched. $\mathcal{L}_2^P$ and $\mathcal{L}_3^P$ capture the lowest amount of functions in PPI networks, both in terms of percentage of clusters enriched and GO-BP annotations enriched. Secondly, in our randomized experiment with randomized GO-annotations, we consistently find 0% of the clusters to be enriched, regardless of the type of Laplacian matrix used. This shows that all Laplacian based enrichments are statistically significant. We find similar results in yeast, see Supplement, Section 5. In the Supplement, we additionally observe that for each network and annotation type, there is always at least one Graphlet Laplacian that shows a larger number of the total number of enriched annotations than Vicus. We conclude that Graphlet Laplacians are at least as biologically relevant as the standard Laplacian, k-path Laplacian and Vicus.

Having established that Graphlet Laplacian based clusters capture biological functions, we quantify the overlap in their enriched functions. In the Supplement, Section 6, we calculate the Jaccard Index between the sets of enriched functions corresponding to each Graphlet Laplacian. For GO-BP enrichments in clusterings on the human PPI and COEX networks, the average Jaccard Index is 0.22 and 0.30 respectively, meaning that different Graphlet Laplacians capture different functions. To further demonstrate this point, we present the number of GO-BP functions that are enriched only in the clustering obtained by a particular Graphlet Laplacian in Figure 5. We observe that each type of Laplacian matrix shows a tendency to capture some distinct biological functions, indicating the link between the biological function and the topology of these diverse molecular networks. The same is observed for GO-MF and GO-CC annotations, both in yeast and human networks (see the Supplement, Section 7). Combining this observation with our previous results, we can conclude that Graphlet Laplacian based spectral clustering allows for distinguishing different sets of similarly wired network components that are not only biologically relevant, but may also capture complementary biological functions.

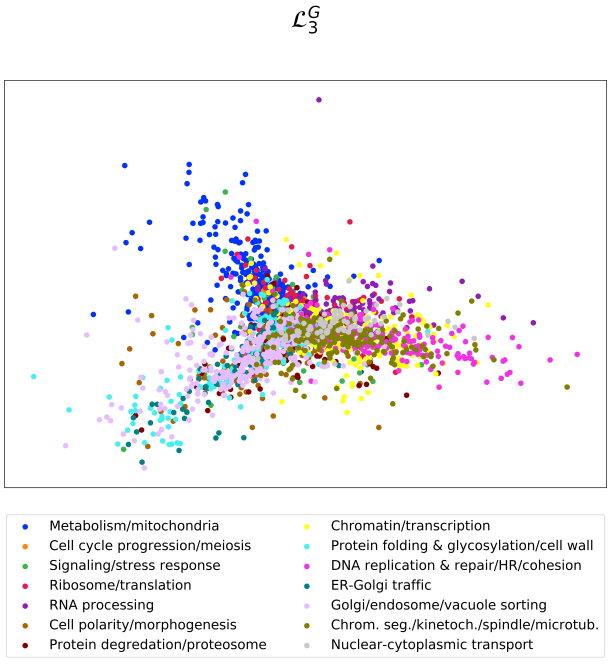

Fig. 3. Capturing biological functions with Graphlet Laplacian $\mathcal{L}_3^G$. 2D spectral embedding of the yeast GI network using the Graphlet Laplacian for G_3. Points represent genes and are color-coded with 14 core biological process annotations defined by Costanzo et al. (2016).

3.3 Different Graphlet Laplacians capture complementary sets of pan-cancer related genes

Laplacian based approaches towards predicting cancer related genes are based on guilt by association: genes which tend to be connected to frequently mutated genes are used as cancer gene predictions. Here we show that by considering the different shapes (i.e. graphlets) by which genes can be connected to frequently somatically mutated genes, complementary cancer mechanisms can be captured.

We do this by diffusing (see section 2.5) the gene mutation frequency scores (see section 2.8.4) on the human PPI network based on different Graphlet Laplacian matrices. Network diffusion is a method underlying many of the different approaches of cancer gene prioritization (Cowen *et al.*, 2017). We prioritize genes as potential cancer related genes according the highest diffused score first. We measure the quality of these scores using the area under the Precision-Recall (PR) curve and the area under the Receiver Operator Characteristic (ROC) curve. We assume a gene is correctly classified as a cancer related gene if it is known to be a cancer driver gene in at least one type of cancer (see section 2.8.4). We observe that accuracy is independent of the Graphlet Laplacian used and on par with the standard Laplacian, with an average area under the PR and ROC curves of 0.21 and 0.78 respectively. In terms of accuracy Graphlet Laplacian based scores consistently outperform those based on k-path or Vicus, which achieve an average area under the PR and ROC curve of 0.17 and 0.74 and 0.14 and 0.73 respectively (see Supplementary Figures 17 and 18).

In Figure 6 we evaluate the overlap between the top hundred highest ranking cancer related genes per Laplacian, measured using the Jaccard Index. We observe five distinct clusters of different Laplacian matrices capturing different sets of cancer related genes. Importantly, diffusion based on three sets of Graphlet Laplacians ($\mathcal{L}_{\{2,5,7\}}^G$, $\mathcal{L}_{\{1,3,4,6\}}^G$ and $\mathcal{L}_{\{8\}}^G$) provide scores dissimilar to those achieved using the standard

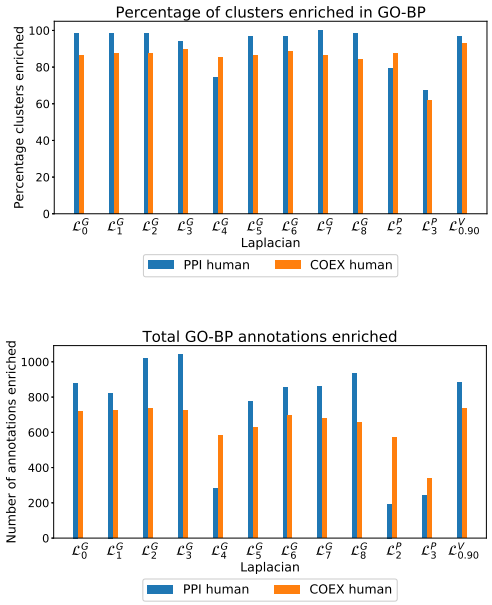

Fig. 4. Cluster Quality. A: For our set of human molecular networks (color-coded), the percentage of clusters enriched in BP annotations, with clusters obtained based on spectral clustering using different Laplacian matrices (x-axis). **B:** For our set of human molecular networks, the total number of enriched GO-BP annotations in clusters obtained based on spectral clustering using different Laplacian matrices (x-axis).

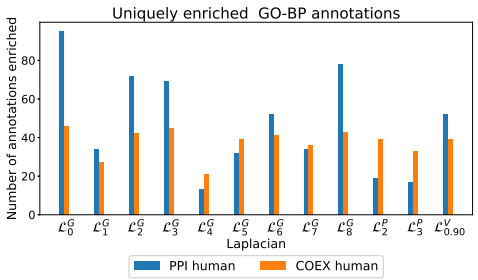

Fig. 5. GO-BP uniquely enriched. The number of annotations that are uniquely enriched in clusterings based on the indicated Laplacian matrix for each biological network (color coded).

Laplacian (the average Jaccard Index of each cluster with the standard Laplacian based scores being 0.79, 0.87 , 0.65 respectively). Conversely, the highest scoring genes based on $\mathcal{L}^P_{\{2,3\}}$ prove to overlap greatly with those based on the standard Laplacian (the average Jaccard Index being 0.91). Vicus based diffusion provides cancer related gene scores dissimilar from all other Laplacian matrices, be it at lower accuracy, as shown above. Similar results are obtained applying graphlet generalized diffusion on the human COEX network, as shown in Supplement, Sections 8 and 9. We conclude that Graphlet Laplacian based diffusion can be used to find complementary sets of cancer related genes.

4 Conclusion

In this paper, we introduce Graphlet Laplacians for simple networks to simultaneously capture graphlet-based topological information and neighborhood membership information. We demonstrate that they can straightforwardly be plugged into current Laplacian based network analysis methods widely used in systems biology, using spectral clustering, spectral embedding and network diffusion as example applications.

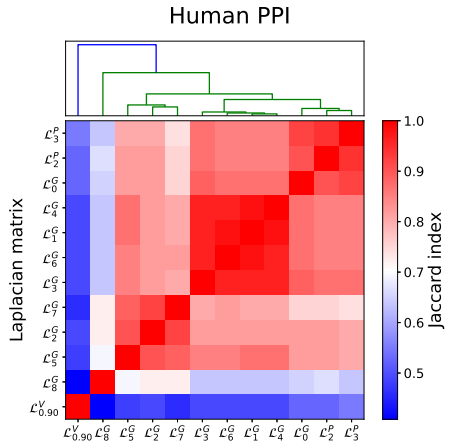

Fig. 6. Overlap of highest scoring cancer related genes. Evaluating the overlap in the sets of the top 100 highest scoring genes based on different Laplacians measured by the Jaccard Index, with scores computed by performing network diffusion of mutation frequency scores on the human PPI network.

Through our generalized spectral embedding and spectral clustering methods on real and model networks, we show that different Graphlet Laplacians capture sub-networks having distinct local topologies and that are enriched in different, but complementary sets of biological annotations. Finally, we show that our generalized network diffusion of pan-cancer gene mutation scores resulted in complementary sets of cancer related genes for gene prioritization dependent on the underlying graphlet. In all the tested applications, our Graphlet Laplacians perform as good as and often better than k-path and Vicus Laplacians, while being directly interpretable.

As indicated, Graphlet Laplacians can directly replace the traditional Laplacian matrix in state-of-the-art network analysis methods, allowing them to consider alternative ways of how nodes are connected. For instance, our Graphlet laplacians could be used to extend embedding methods such as hyper-coalescent embedding (Muscoloni *et al.*, 2017), which may result in more relevant community detections in biological networks and in more accurate analyses of the dynamics of cells' biological processes. Furthermore, Laplacian matrices are used in data-integration frameworks to incorporate prior knowledge (e.g., via so-called graph regularizations in matrix factorization based data integration). Thus, using our Graphlet Laplacians in such data-integration frameworks could result in biologically more accurate patient stratifications, biomarker discoveries, and drug-target interaction predictions (Gligorijević *et al.*, 2016)

Finally, the applications of Graphlet Laplacians are not limited to biology, as the generalized network-analysis tools are applicable in any discipline that uses network representations, including physics, social sciences, and economy.

Funding

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the European Research Council (ERC) Consolidator Grant 770827, the Serbian Ministry of Education and Science Project III44006, the Slovenian Research Agency project J1-8155, the Prostate Project and UCL Computer Science departmental funds.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Bartlett, T. E., Müller, S., and Diaz, A. (2017). Single-cell Co-expression Subnetwork Analysis. *Scientific Reports*, **7**(1), 15066.

Baumgartner, C., Spath-Blass, V., Niederkofler, V., Bergmoser, K., Langthaler, S., Lassnig, A., Rienmüller, T., Baumgartner, D., Asnani, A., and Gerszten, R. E. (2018). A novel network-based approach for discovering dynamic metabolic biomarkers in cardiovascular disease. *PloS one*, **13**(12), e0208953.

Belkin, M. and Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, **15**(6), 1373–1396.

Benson, A. R., Gleich, D. F., and Leskovec, J. (2016). Higher-order organization of complex networks. *Science*, **353**(6295).

Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., and Others (2003). Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic acids research*, **31**(9), 2443–2450.

Campbell, K., Ponting, C. P., and Webber, C. (2015). Laplacian eigenmaps and principal curves for high resolution pseudotemporal ordering of single-cell RNA-seq profiles. *bioRxiv*, page 27219.

Cao, M., Zhang, H., Park, J., Daniels, N. M., Crovella, M. E., Cowen, L. J., and Hescott, B. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PloS one*, **8**(10), e76339.

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., Van Leeuwen, J., Van Dyk, N., Lin, Z. Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., Srikumar, T., Bahr, S., Chen, Y., Deshpande, R., Kurat, C. F., Li, S. C., Li, Z., Usaj, M. M., Okada, H., Pascoe, N., Luis, B. J. S., Sharifpoor, S., Shuteriqi, E., Simpkins, S. W., Snider, J., Suresh, H. G., Tan, Y., Zhu, H., Malod-Dognin, N., Janjic, V., Przulj, N., Troyanskaya, O. G., Stagljar, I., Xia, T., Ohya, Y., Gingras, A. C., Raught, B., Boutros, M., Steinmetz, L. M., Moore, C. L., Rosebrock, A. P., Caudy, A. A., Myers, C. L., Andrews, B., and Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, **353**(6306), aaf1420.

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, **18**(9), 551.

Davis, D., Yaveroğlu, Ö. N., Malod-Dognin, N., Stojmirovic, A., and Pržulj, N. (2015). Topology-function conservation in protein-protein interaction networks. *Bioinformatics*, **31**(10), 1632–1639.

Durán, C., Daminelli, S., Thomas, J. M., Haupt, V. J., Schroeder, M., and Cannistraci, C. V. (2017). Pioneering topological methods for network-based drug–target prediction by exploiting a brain-network self-organization theory. *Briefings in bioinformatics*, **19**(6), 1183–1202.

Erdős Paul and Rényi Alfréd, S. (1959). On random graphs. *Publicationes Mathematicae*, **6**, 290–297.

Estrada, E. (2012). Path Laplacian matrices: Introduction and application to the analysis of consensus in networks. *Linear Algebra and Its Applications*, **436**(9), 3373–3391.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer*, **4**(3), 177.

Ghosh, S. A. and Strehl, A. (2002). Clusters Ensembles-A Knowledge Reuse Framework for Combining Multiple Partition. In *Journal of Mathine Learning Research*, volume 3, pages 583–617.

Gligorijević, V., Malod-Dognin, N., and Pržulj, N. (2016). Patient-Specific Data Fusion for Cancer Stratification and Personalised Treatment. *Pacific Symposium on Biocomputing.*, **21**, 321–32.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, **10**(11), 1081.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, a. W. (1999). From molecular to modular cell biology. *Nature*, **402**(6761 Suppl), C47–C52.

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., and Others (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, **502**(7471), 333.

Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, **1**(6), 2321–7782.

Kotlyar, M., Pastrello, C., Sheahan, N., and Jurisica, I. (2016). Integrated interactions database: Tissue-specific view of the human and model organism interactomes. *Nucleic Acids Research*, **44**(D1), D536–D541.

Lee, A. B., Luca, D., Klei, L., Devlin, B., and Roeder, K. (2010). Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, **34**(1), 51–59.

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., and Others (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, **47**(2), 106.

Li, Z., Qiao, Z., Zheng, W., and Ma, W. (2015). Network cluster analysis of protein–protein interaction network–identified biomarker for type 2 diabetes. *Diabetes technology & therapeutics*, **17**(7), 475–481.

Luo, J., Yin, Y., Pan, C., Xiang, G., and Tu, N. H. (2018). Identifying functional modules in co-regulatory networks through overlapping spectral clustering. *IEEE transactions on nanobioscience*, **17**(2), 134–144.

Milenković, T. and Pržulj, N. (2008). Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, **6**, 257–273.

Milenković, T., Memišević, V., Ganesan, A. K., and Pržulj, N. (2010). Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *Journal of the Royal Society Interface*, **7**(44), 423–437.

Muscoloni, A. and Cannistraci, C. V. (2018). A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. *New Journal of Physics*, **20**(5), 52002.

Muscoloni, A., Thomas, J. M., Ciucci, S., Bianconi, G., and Cannistraci, C. V. (2017). Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nature Communications*, **8**(1), 1615.

Newman, M. E. J. M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856.

Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., and Kinoshita, K. (2015). COXPRESdb in 2015: Coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Research*, **43**(D1), D82–D86.

Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguná, M., and Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature*, **489**(7417), 537.

Penrose, M. D. (2003). *Random Geometric Graphs*. Oxford University Press.

Pržulj, N. and Higham, D. J. (2006). Modelling protein–protein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, **3**(10), 711–716.

Pržulj, N., Wigle, D. A., and Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics*, **20**(3), 340–348.

Pržulj, N., Kuchaiev, O., Stevanović, A., and Hayes, W. (2010). Geometric evolutionary dynamics of protein interaction networks. In *Biocomputing 2010*, pages 178–189. World Scientific.

Smedley, D., Köhler, S., Czeschik, J. C., Amberger, J., Bocchini, C., Hamosh, A., Veldboer, J., Zemojtel, T., and Robinson, P. N. (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, **30**(22), 3215–3222.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, **34**(90001), D535–D539.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, **17**(4), 395–416.

Wang, B., Huang, L., Zhu, Y., Kundaje, A., Batzoglou, S., and Goldenberg, A. (2017). Vicus: Exploiting local structures to improve network-based analysis of biological data. *PLoS Computational Biology*, **13**(10), e1005621.

Yaveroğlu, Ö. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A., and Pržulj, N. (2014). Revealing the hidden language of complex networks. *Scientific Reports*, **4**, 4547.