GigaScience, 7, 2018, 1–31

doi: 10.1093/gigascience/giy014
Advance Access Publication Date: 19 February 2018
Review

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

REVIEW

Bipartite graphs in systems biology and medicine: a survey of methods and applications

Georgios A. Pavlopoulos [1,*,†], Panagiota I. Kontou[2,†], Athanasia Pavlopoulou[3], Costas Bouyioukos[4], Evripides Markou[2] and Pantelis G. Bagos [2,*]

[1]Lawrence Berkeley Labs, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA, [2]University of Thessaly, Department of Computer Science and Biomedical Informatics, Papasiopoulou 2–4, Lamia, 35100, Greece, [3]Izmir International Biomedicine and Genome Institute (iBG-Izmir), Dokuz Eylül University, 35340, Turkey and [4]Université Paris Diderot, Sorbonne Paris Cité, Epigenetics and Cell Fate, UMR7216, CNRS, France

*Correspondence address. Georgios A. Pavlopoulos, Lawrence Berkeley Labs, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. Tel: +302231066936; E-mail: g.pavlopoulos@lbl.gov http://orcid.org/0000-0002-4577-8276; and Pantelis G. Bagos, University of Thessaly, Department of Computer Science and Biomedical Informatics, Papasiopoulou 2–4, Lamia, 35100, Greece. Tel: +302231066914; E-mail: pbagos@compgen.org http://orcid.org/0000-0003-4935-2325
†These authors contributed equally to this work.

Abstract

The latest advances in high-throughput techniques during the past decade allowed the systems biology field to expand significantly. Today, the focus of biologists has shifted from the study of individual biological components to the study of complex biological systems and their dynamics at a larger scale. Through the discovery of novel bioentity relationships, researchers reveal new information about biological functions and processes. Graphs are widely used to represent bioentities such as proteins, genes, small molecules, ligands, and others such as nodes and their connections as edges within a network. In this review, special focus is given to the usability of bipartite graphs and their impact on the field of network biology and medicine. Furthermore, their topological properties and how these can be applied to certain biological case studies are discussed. Finally, available methodologies and software are presented, and useful insights on how bipartite graphs can shape the path toward the solution of challenging biological problems are provided.

Keywords: biological networks; graph theory; systems biology; bipartite graphs; ecological networks; network medicine

Background

Today, in the big-data and OMICS era, established high-throughput technological advances, integrative biology, and bioinformatics have significantly changed our view on how to tackle difficult biological problems toward the understanding of more complex biological systems. For example, yeast-two-hybrid [1] and protein chips [2] have enabled biologists to experimentally detect the complete protein interactome or protein–protein interactions (PPIs) for certain organisms [3–7]. Microarrays and RNA-seq [8] have accelerated the discovery of differentially expressed genes across different conditions (i.e.,

Received: 9 July 2017; **Revised:** 15 January 2018; **Accepted:** 13 February 2018

© The Author(s) 2018. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

disease vs control) and the study of developmental processes, as well as pharmacogenomic responses and the evolution of gene regulation in different species. In this way, through the generation of gene regulation networks, new knowledge about gene behavior and unknown functions can be unraveled. Furthermore, the latest drug screening and mass spectrometry techniques allow for a massively parallel protein–compound interaction identification and exploration, whereas genome sequencing technologies [9] have exponentially increased the number of newly sequenced genomes. Therefore, exploration and discovery of new genes, new lineages of life, identification of single nucleotide polymorphisms (SNPs) or variations causative for genetic disorders [10], population genetics, characterization of the genetic material recovered from environmental metagenomic samples [11], and direct interspecies genome comparisons have opened new research fields while, simultaneously, have changed the landscape of bioentity associations known until today.

PPIs, gene expression, gene regulation, literature co-occurrences, evolutionary relationships, signal transduction, metabolic pathways, and others are often captured in network representations, where a node represents a bioentity and an edge the relationship between them. PPIs, for example, are represented as simple undirected graphs, whereas gene signal transduction and regulation networks as directed graphs (digraphs). Additionally, gene expression networks can be found as weighted graphs, pathways as petri-nets, and gene regulation together with literature co-occurrences as semantic graphs. Finally, multi-edged networks can hold information about nodes that are connected in multiple ways. For example, 2 proteins might co-occur in biomedical literature, share common domains, have a certain degree of sequence similarity, be evolutionary related, and interact physically. For a better understanding of the definitions of the aforementioned networks, as defined by graph theory, more detailed descriptions are available elsewhere [12,13].

In this review, we thoroughly discuss the potential and the usability of bipartite graphs for analyzing biological networks. To our knowledge, this is the first extensive investigation into bipartite graphs, given that other studies have focused on generic graph analysis. A bipartite graph, also referred to as a "bigraph," comprises a set of graph vertices decomposed into 2 disjoint sets such that no 2 graph vertices within the same set are adjacent. As discussed by Burgos et al. [14] and Kontou et al. [15], applications of such bipartite graphs can range from the representation of enzyme-reaction links in metabolic pathways to gene–disease associations or an ecological network. While network analyses have focused mainly on unipartite (1-mode) networks, considerably less attention has been paid to the deeper study of bipartite networks and their potential in biological sciences.

Many nonbiological real-world networks may be naturally viewed and modeled by a bipartite graph structure. Perhaps the oldest example of such bipartite network originates from the analysis of Deep South data, also known as the "Southern Women" data, collected in 1941, representing a set of women attending social events over a period of 9 months [16]. Other notable examples studied extensively in the literature include, for instance, the actors–movies network, where each actor was linked to the movies he/she appeared in [17,18]; the scientists–papers network, where the scientists were linked to the papers they authored [18–20]; the board–directors network, where the members of the board of directors are linked to the companies they lead [21,22]; the peer-to peer exchange networks in which peers are linked to the data they provide [23]; the world cities hosting branches of multinational firms [24]; the supreme court justices joining majority opinions [25]; and the legislators sponsoring bills [26]. Moreover, during recent years, the bipartite graph has been used extensively in internet technology and applications since it has been used to model the relationship between queries and URLs in query logs [27], between video shots and tags [28], for entities and co-lists in web pages [29], for users and items in recommendation [30], for behavior analysis of internet traffic [31], and for detecting network traffic anomalies [32].

The current review is structured as follows: we provide a mathematical definition of a bipartite graph; we comment on its topological properties; we summarize several projection strategies to generate 2 unipartite networks from a bipartite network; we discuss the theoretical properties and the importance of the projections, as well as the potential biological applications of them; we describe several real-life network types and how these can be analyzed using the graph theory related to bipartite graphs; we describe models and algorithms for bipartite graphs; and, finally, we comment on the advantages of available software dedicated to analyze bipartite networks.

Bipartite Graphs

Definition

A graph $G = (U, V, E)$ is *bipartite* (or *bigraph* or *2-mode network*) if its vertices can be divided into 2 disjoint sets, U and V, such that every edge (E) connects a vertex in U to 1 in V (Figure 1A, B). Vertex sets U and V are usually termed as the parts of the graph. Equivalently, a graph that does not contain any odd-length cycles is by definition a bipartite graph, whereas bipartite graphs are also equivalent to 2 colorable graphs. Among the various types of graphs, trees, acyclic graphs, and circular graphs with an even number of vertices, are by definition bipartite. A bipartite graph represents a special case of a k-partite graph with $k = 2$. If a bipartite graph is not connected, it may have more than 1 bipartition; in this case, the (U, V, E) notation is helpful in specifying 1 particular bipartition that may be of importance in an application. If $|U| = |V|$, that is, if the 2 subsets have equal cardinality, then G is called a balanced bipartite graph. If all vertices on the same side of the bipartition have the same degree, then G is called biregular.

Bipartite graphs can be efficiently represented by biadjacency matrices (Figure 1C, D). The biadjacency matrix B that describes a bipartite graph $G = (U, V, E)$ is a (0,1)-matrix of size $|U| \times |V|$, where $B_{ik} = 1$ provided there is an edge between i and k, or $B_{ik} = 0$, otherwise. Biadjacency matrices can be used to describe equivalences between bipartite graphs, hypergraphs, and directed graphs. In most cases, biadjacency matrices are (0,1)-matrices and the networks are, therefore, unweighted. However, in some applications, as in the case of ecological networks, matrices with $B_{ik} > 1$ are also used to represent a weighted bipartite network.

Properties of bipartite graphs

Bipartite graphs, as opposed to generic networks that have their own topological characteristics, comprise a distinct category with their very own unique properties. Given that network metrics for unipartite networks have been studied extensively, herein attention is given to network metrics used specifically for bipartite graphs. A short commentary on such topological features is provided below [12,33,34]. Of note, there are dozens of specialized metrics for bipartite ecological networks, some of which are discussed in-depth by Dormann et al. [35]. Notably, known tools dedicated to automated topological

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

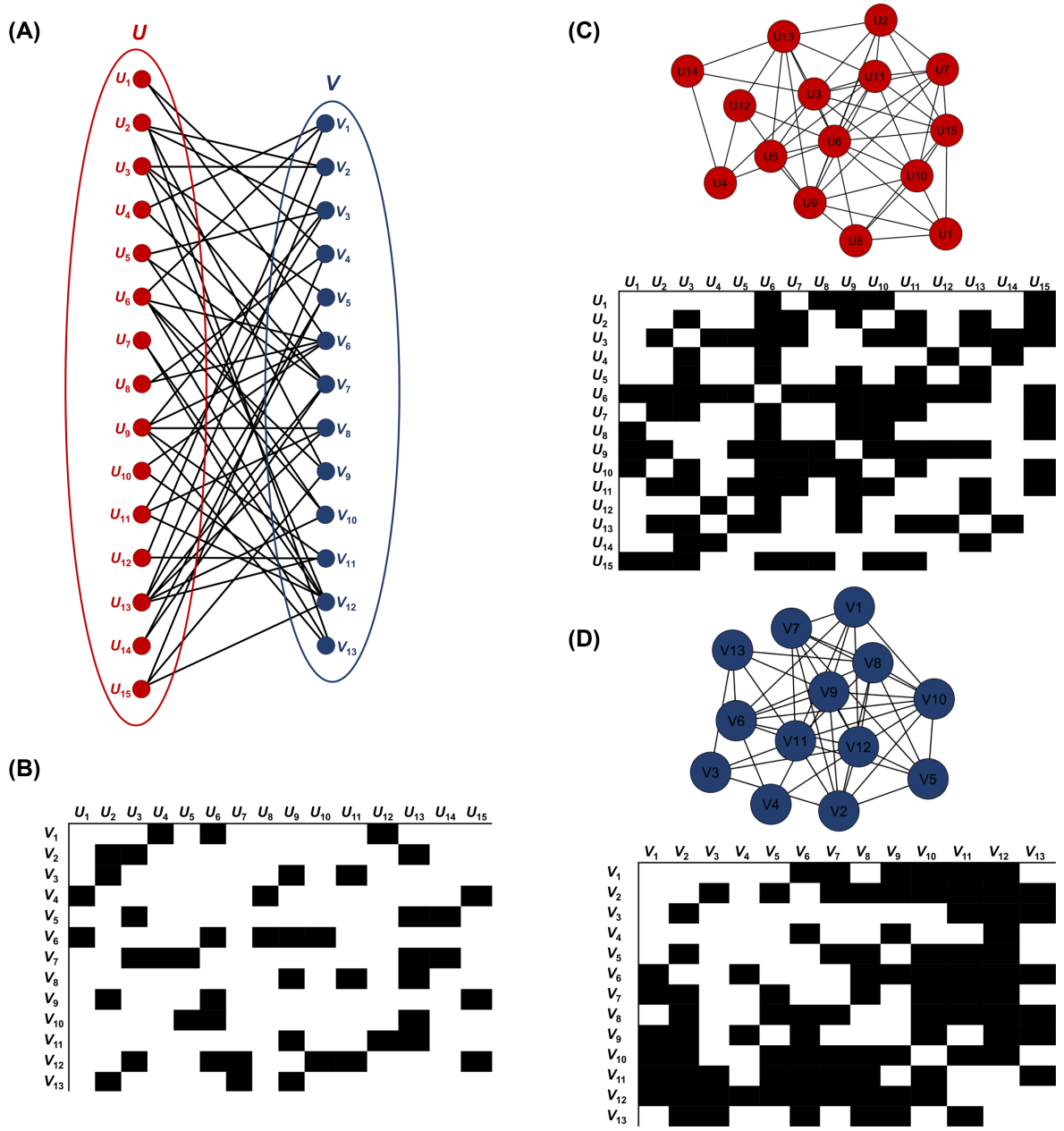


Figure 1: Construction of unipartite networks from a bipartite network. (A) The bipartite network. (B) The biadjacency matrix of the bipartite network. (C) The first unipartite network with its adjacency matrix. (D) The second unipartite network with its adjacency matrix. The adjacency matrices are symmetrical across the diagonal line.

analysis for generic networks are the Network Analysis Profiler (NAP) [36], Cytoscape's Network Analyzer [37], the Stanford Network Analysis Platform (SNAP) [38], and the igraph library ([39]). Although these are not bipartite graph-specific, they do offer a wide spectrum of functions and modules related to topological analyses.

Degree

In a simple undirected graph, the degree or degree centrality is defined as the number of edges incident upon a node. Nodes with the highest degree (i.e., connected to more nodes) are considered as "hubs." In a directed graph, the degree can be calculated as the sum of the in-degree (number of incoming edges) and the out-degree (number of outcoming edges). As opposed to a fully connected graph $G = (V, E)$, which can have a maximum of $|V|(|V| - 1)/2$ connections, in a bipartite graph, the maximum degree of a node can be equal to the number of nodes from the opposing set ($max[deg(u)] = |V|$ or $max[deg(V)] = |u|$). Furthermore, the sum of the degrees of the first part is equal to the sum of the degrees of the opposing part, and both are equal to the cardinality of the edge set:

$$\sum_{v \in V} \deg(v) = \sum_{u \in U} \deg(u) = |E|$$

Closeness centrality

Closeness centrality is a measure to determine whether a node can communicate with other nodes within the network readily and through short paths. Hence, the more central a node is, the

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

closer it is to all other nodes. Closeness centrality is inversely proportional to the shortest path length between 2 nodes. In a bipartite graph, a node can have a minimum distance "1" from vertices of the opposing set and "2" from vertices of the same set. Moreover, due to the bipartite structure, all paths between nodes of the same set are of even length, a property that rather complicates the calculation of several measures.

Betweenness centrality

The nodes with high betweenness centrality are the ones that serve as bridges between 2 highly connected communities. An all-against-all shortest path calculation is often required in order to estimate betweenness centrality reliably. Hence, each node increases its centrality score every time it is involved in a shortest path. The nodes with very high betweenness centrality scores are the ones that serve as mediators between 2 or more neighborhoods. In a bipartite graph, paths can originate and terminate at a node of each vertex set.

Eigenvector centrality

Eigenvector centrality is a measure to identify the nodes that are connected to "important nodes," such as hubs, within a network. The eigenvector centrality of a node is proportional to the sum of centralities of the nodes it is adjacent to. Bipartite eigenvector centrality is further reviewed by Daugulis [40].

Clustering coefficient. The global clustering coefficient indicates the tendency of a network to form tight clusters. Similarly, the local clustering coefficient shows the tendency of a node to belong to a cluster. While this is a useful measure for a generic network, applying the 2 clustering coefficients directly to a bipartite network is meaningless. Handling a 2-mode network as a 1-mode network is not recommended, as projected 2-mode networks tend to have more and larger fully connected cliques [41]. Moreover, the conventional clustering coefficient cannot be used in bipartite networks, where cycles of size 3 are absent. Instead, other coefficients based on the fraction of cycles with size 4 have been proposed, with similar clustering properties [42]. To overcome such problems, a number of clustering coefficients for 2-mode networks have been proposed elsewhere [33,42–45].

Nestedness Nestedness is an important property of ecological networks. It is usually defined as a pattern of interactions in which "specialists" (e.g., pollinators that visit few plants) interact with subsets of the species with which "generalists" (e.g., pollinators that visit many plants) interact. Nestedness is not a metric in itself but a concept that, at least to date, has not been formally defined through mathematical relationships. This probably explains the fact that there are several distinct metrics by which it can be measured. In mathematical terms, nestedness can be defined as a property of the previously mentioned biadjacency matrix B. If B is a perfectly nested binary matrix, then there exists a permutation of rows and columns such that the set of edges in each row i contains the edges in row $i+1$, while the set of edges in each column j contains those in column $j+1$. In particular, the rows and columns of B can be sorted (with $B_{1,j} > 0\ \forall j$ and $B_{i,1} > 0\ \forall i$) such that $B_{i,j} \leq \min(B_{i,j-1}, B_{i-1,j})$, a property that can be extended to quantitative matrices as well [46]. Thus, in general, a nested structure corresponds to a systematic arrangement of non-zero entries in the binary matrix often used to represent a network. However, measuring the nestedness of a given network is not always straightforward, and there are several detection methods for identifying nested patterns among other possible matrix arrangements [47]. A schematic representation is shown in Figure 2.

The most widely used metric of nestedness is the nestedness temperature, $T = 1 - N$, which quantifies whether the observed arrangement of 1's and 0's deviates from the arrangement given by an isocline that describes a perfect nestedness benchmark. Contributions of unexpected absences and presences in the upper-left and bottom-right sides, respectively, are weighted by their squared Euclidian distances from the isocline [48]. Similar metrics have been presented by Araujo and coworkers [49], whereas fast algorithms and software for calculating T were presented by Guimarães and Guimaraes [50].

An additional metric, C, is based on the concept of "species richness" and, unlike T, quantifies nestedness exclusively between rows [51].

NODF (nestedness metric based on overlap and decreasing fill) was developed later in order to overcome 2 major disadvantages of previous methods: marginal totals may differ among columns and/or rows and the presences (1's) in less-filled columns and rows may coincide with those found in the more-filled columns and rows, respectively. Therefore, *NODF* has some important features that distinguish it from the preceding metrics, i.e., it calculates nestedness independently among rows and columns, which allows the evaluation of nestedness only among sites (i.e., species composition) or among species (i.e., species occupancy), whereas it is able to evaluate how nested 1 or more columns (or rows) is in relation to other ones [52].

A modified version of *NODF*, termed *WNODF* (where "W" stands for "weighted"), was also developed later to handle quantitative matrices [53] Other approaches have also been developed for the same task, including methods that rely on the eigenvalues and the spectral radius of the matrix [46,54].

Modularity Modularity is another feature usually found in ecological networks. Modularity occurs when certain groups of nodes (usually species) within a network are much more highly connected to each other than they are to other nodes of the network, with weak interactions among different modules (Figure 3). Modularity measures the tendency of a network to divide into modules (also called groups, clusters, or communities). In networks with high modularity, the nodes within modules are densely connected but sparsely connected in different modules. The most widely used measure of modularity is calculated from the (symmetric) adjacency matrix A by:

$$Q(\mathbf{A}) = \frac{1}{W} \sum_{C \in P} \sum_{i,j \in C} \left[A_{ij} - \frac{k_i k_j}{W} \right]$$

Where $W = \sum_{i,j} Aij$ and $k_i = \sum_j A_{ij}$ is the degree of node i. The indices i, j run over the nodes of the graph, whereas C runs over the communities (modules) of the partition. Since the adjacency matrix is not symmetric, $A = (O, B/B^T, O)$ can be applied. Hence, modularity enables the detection of modules (or clusters) in the first place, and it can be further optimized by particular optimization algorithms to detect community structure in networks (see below) [55].

Of particular note, networks can be both highly nested and highly modular [56] (Figure 4).

Internal links and pairs. A usual approach for the analysis of bipartite graphs consists of deriving unipartite graphs (projections) from the underlying bipartite structure (Figure 1C and D). However, this is associated with important loss of information and

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

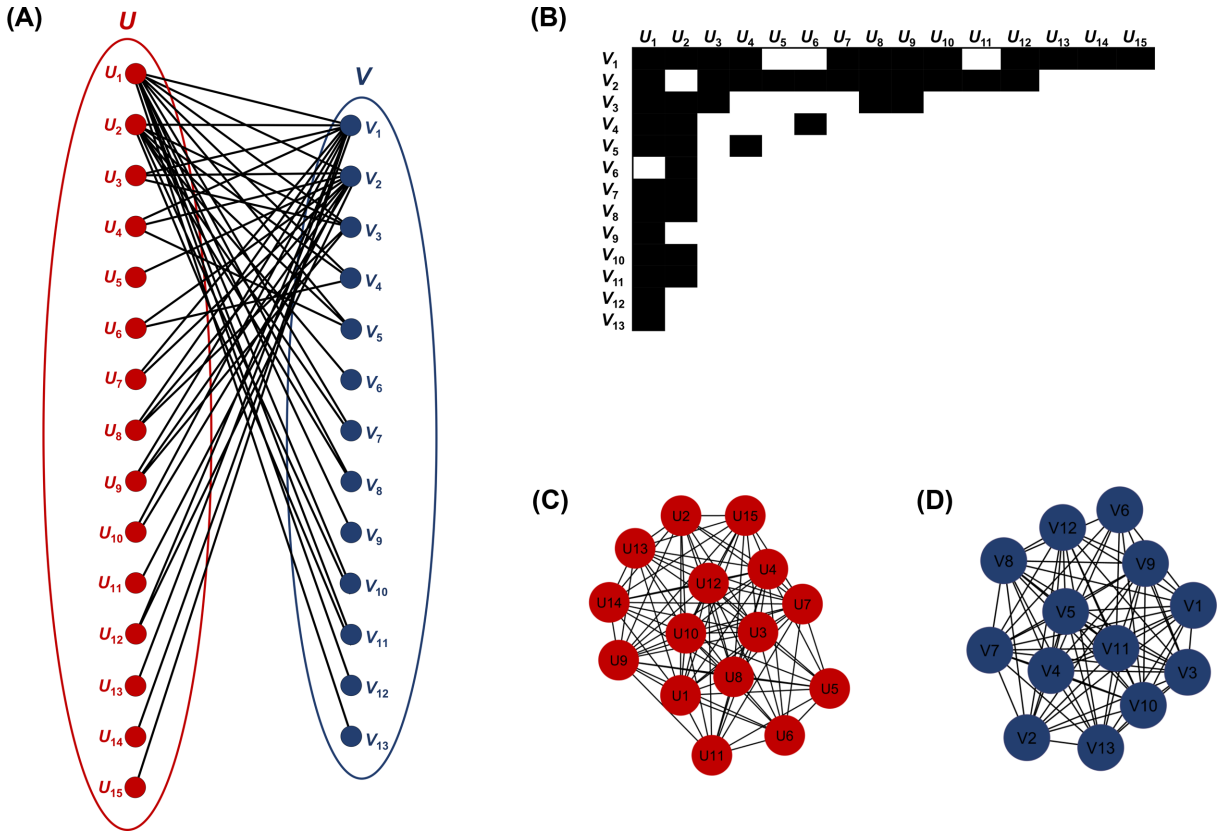


Figure 2: Network nestedness. Example of (A) a bipartite network, (B) the biadjacency matrix of the bipartite network, and (C,D) the projected unipartite networks.

data storage issues (a detailed description of projection is provided below).

Allali and coworkers [57] introduced the *internal links* and *pairs* as metrics useful for analyzing a bipartite graph, thereby providing an understanding of the projection of the bipartite graph. Specifically, in a bipartite graph $G = (U, V, E)$, (u, v) is a U-internal ($\perp$-) pair of G only if by adding the new link (u, v) to G does not change its U-projection; it is a U-internal link provided that the removal of the link (u, v) from G does not change its U-projection. The number of U-internal links of a node is called "U-internal degree." The authors illustrated the relevance of these concepts in several real-world bipartite networks, highlighting their discriminative ability when benchmarked against random graphs. Internal links and pairs can be useful metrics for both modeling complex networks and storing them in a compact format [57].

Bipartivity Many biological systems are naturally modeled as bipartite networks. However, there are also networks that although they are not naturally bipartite, they appear to be closer to bipartite compared to what can be expected by a completely random network; for instance, networks formed by 2 types of nodes that have a preference for interactions with nodes of the other type, such as networks of sexual relationships. It is possible to test whether a graph is bipartite and to return either a 2-color graph (if it is bipartite) or an odd cycle graph (if it is not) in linear time by using a depth-first search algorithm. The main idea is to assign to each vertex a color that is different from the color of its parent in a depth-first search tree in a preorder traversal of the tree. In this way, a 2-color spanning tree consisting of edges connecting vertices to their parents is generated, although some of the non-tree edges may not be properly colored.

Bipartivity is a measure that quantifies how close a given network is to being bipartite. Two such measures were provided first by Holme and coworkers [58]. The first measure is based on the optimal 2-coloring of the network [58]. The exact value of this quantity is Nondeterministic Polynomial time (NP)-complete; therefore, an optimal calculation is not possible. They proposed instead an approximate solution by a simulated annealing approach. The latter is based on the count of odd circuits that, in most cases, can be calculated in polynomial time. Later, Estrada and Velásquez provided a different measure, $\beta(G)$, based on the spectral decomposition of the biadjacency matrix [59]. This measure is easy to compute and allows the calculation of individual node contributions to global bipartivity, which is based on the concept of closed walks. Pisanski and Randić have taken into consideration the so-called Szeged index (Sz) and the revised Szeged index ($Sz*$), both of which can be considered generalizations of the Wiener number to cyclic structures. They found that the quotient of the 2 indices, termed $\sigma(G)$, can be used as a novel measure for characterizing the degree of bipartivity of networks because the 2 indices assume the same values for bipartite graphs but different values for nonbipartite graphs. Thus, they proposed $\sigma(G) = Sz/Sz*$ as a measure of bipartivity and they also provided empirical evidence that it is in good agreement with $\beta(G)$ [60].

Ecological indices. In this section, some metrics that are routinely being used in ecological bipartite network analysis are mentioned. The symbol L indicates the number of realized links, whereas $|U|$ *and* $|V|$ denote the number of species of each

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

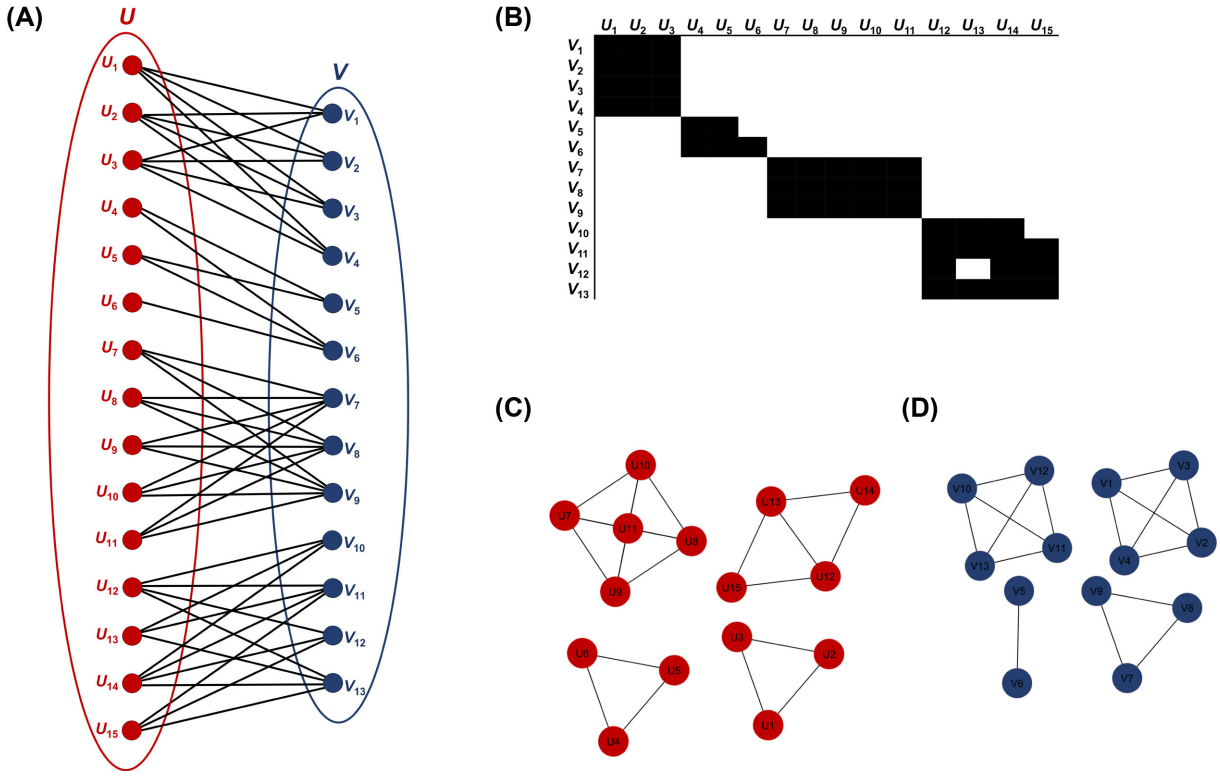


Figure 3: Network modularity. Example of (A) a bipartite network, (B) the biadjacency matrix of the bipartite network, and (C, D) the corresponding unipartite networks.

party in bipartite networks (e.g., hosts [U] vs. parasites [V]). *Connectance* (C) is the fraction of all possible links that are realized, $C = L/(|U| * |V|)$, which represents a standard measure of food web complexity. The related *linkage density* is defined as $D = L/(|U| + |V|)$.

In a food web of $|U|$ consumers and $|V|$ prey species, the mean number of prey species (links) per consumer is termed *generality*, given by $G = L/|U|$, and the mean links per prey *vulnerability*, given by $V = L/|V|$. The *web-asymmetry* defines the balance between numbers in the 2 levels and it is given by $W = (|V| - |U|)/(|U| + |V|)$, where positive numbers indicate more low-trophic level species and negative more high-trophic level species. Most of these metrics also have a weighted counterpart, whereas there are also several other metrics designed for quantitative interactions, such as *Shannon's evenness* (for measuring interactions), H_2 (a network-level measure of specialization based on the deviation of a species' realized number of interactions and that expected from each species' total number of interactions), and *niche overlap* (the mean similarity in the interaction patterns between species of the same trophic level). The reader can refer to key publications for more information on the topic of ecological indices [35,61].

Projection

In a bipartite network, the nodes are divided into 2 disjoint sets (U, V), and the edges (E) connect nodes that belong to different sets. From a bipartite network, it is possible to derive 2 projected networks, where each one is composed of only 1 set of nodes. This approach for analyzing bipartite networks is termed "projection," i.e., deducing relationships between nodes of the same type. In other words, in order to study the relationships among a particular set of nodes, the bipartite network has to be compressed by 1-mode projection.

The U 1-mode projection ("U-projection" for short) is composed of a network containing only U-nodes, where 2 U-nodes are connected when they have at least 1 common neighboring V-node. Conversely, the V-projection is a network of V-nodes in which 2 V-nodes are connected when they have at least 1 common neighboring U-node. Some authors argue that bipartite projections are easier to analyze compared to their original bipartite network because they are 1-mode networks and hence there is no need to develop new techniques to analyze the bipartite networks. However, because bipartite projections are usually weighted networks, the analysis of these projections is not so straightforward. Projecting a bipartite network into a 1-mode network merely transforms the problem of the analysis of a bipartite structure into the problem of analyzing a weighted one, not an easy task. Indeed, the projection transformation is associated with loss of information, including the specific identity of the V-nodes responsible for the linkages between U-nodes. Nonetheless, bipartite projection constitutes an important methodological tool in network science, and its use is recommended in cases where processing a natively 1-mode network is impossible or impractical.

In particular, a $|U| \times |V|$ biadjacency matrix B, defining a bipartite network $G = (U, V, E)$, can be projected onto an $|U| \times |U|$ unipartite (U-projected) or 1-mode network, denoted by P_U, as BB^T (the projection on V, denoted as P_V, is similarly obtained by B^TB). The ability to construct unipartite networks from bipartite ones in this way also leads to the question whether the mathematical properties of the projected networks can be inferred only by knowing the bipartite structure. Several authors have studied the mathematical properties of such projected networks in relation to the properties of the bipartite network. One

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

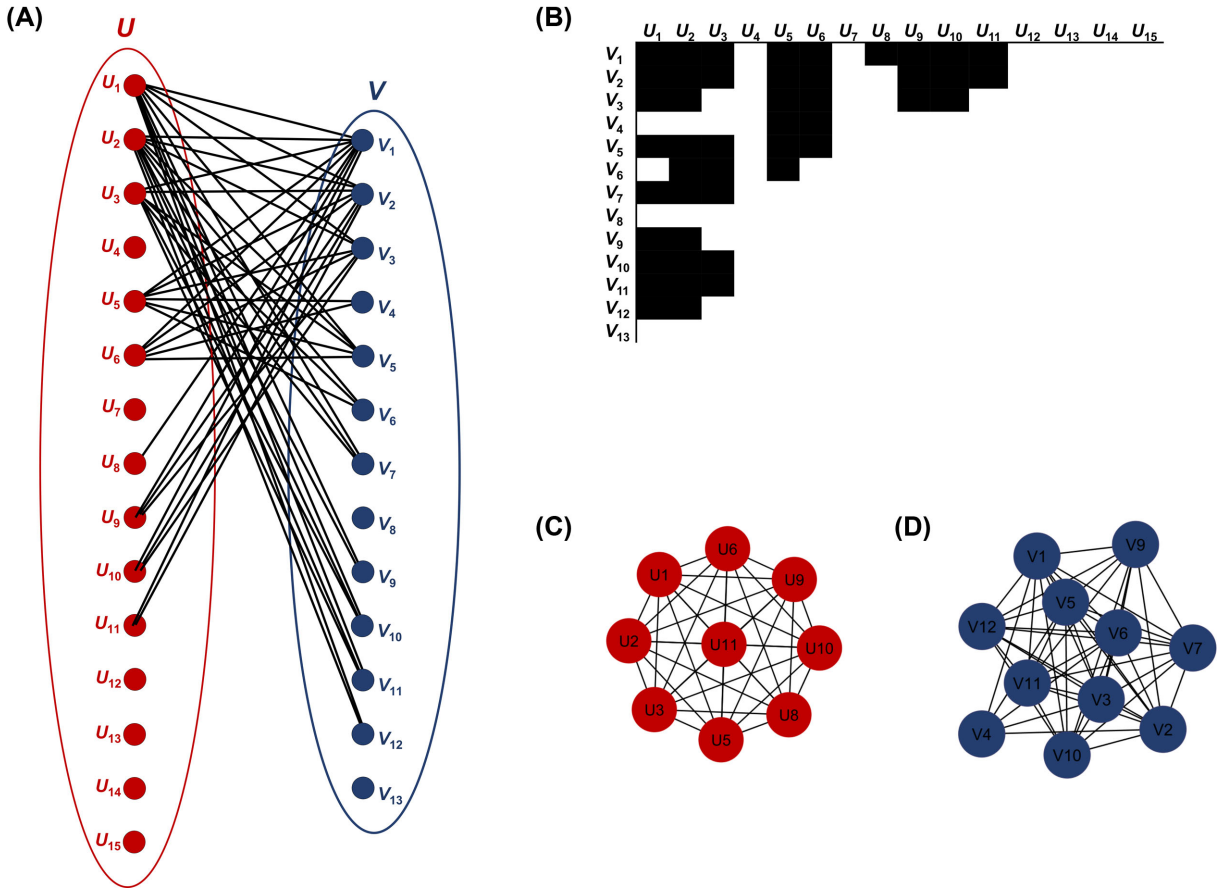


Figure 4: Mixed network (nested + modular). Example of (A) a bipartite network, (B) the biadjacency matrix of the bipartite network, and (C, D) the projected unipartite networks.

important feature of the edge weights constructed this way in a projected network is their constrained range of possible values. The range of weight values of an edge between nodes i and j in a bipartite projection (W_{ij}) can be expressed as a function of these nodes' degrees (i.e., k_i and k_j) and the total number of nodes of the other partition ($|U|$):

$$\min(k_i, k_j) - (|U| - \max(k_i, k_j)) \leq W_{ij} \leq \min(k_i, k_j)$$

However, in general, higher-degree nodes tend to have stronger edges compared to lower-degree nodes. Additionally, it is widely known that the degree distribution of the nodes in a partition of a bipartite network influences the degree distribution of its 1-mode projection on that partition. Moreover, Mukherjee and coworkers have shown that in a projected network, the degree distribution of the other partition (V) also has a very strong influence on the degree distribution of the 1-mode projection on U [62]. They also showed that if partition U corresponds to a peaked distribution, then it is possible to derive closed-form expressions for the 1-mode degree distribution. Other authors went a few steps further in order to calculate the degree distribution analytically [63,64]. The most complete treatment was given by Nacher and Akutsu [64] who studied the case of scale-free distributions for both sets of nodes (denoted by S-S) and that of scale-free and exponential degree distribution (denoted by S-E) for the 2 sets of nodes. They presented a mathematical analysis demonstrating that it is possible to infer the degree distributions of projected networks given the information contained in the original bipartite network, thereby deriving some simple relationships. For instance, a bipartite network with 2 sets of nodes with degree distributions $P_U(k) \propto k^{-\gamma 1}$ and $P_V(k) \propto k^{-\gamma 2}$ exhibits a V-projection that follows a power-law $k^{\max(-\gamma 1+1,\ -\gamma 2)}$ for node degree, where $\gamma 1$ and $\gamma 2$ indicate the power law exponents of the distribution of U and V nodes, respectively, in the bipartite network. On the other hand, a bipartite network with 2 sets of nodes with degree distributions $P_U(k) \propto k^{-\gamma 1}$ and $P_V(k) \propto \exp(-\lambda k)$ leads to a V-projection, defined by a power-law $k^{-\gamma 1+1}$ node degree distribution. The analytical results were confirmed by computer simulations performed using artificially constructed networks [64].

Various methods of bipartite network projection have been proposed in the literature [17,33,65–70], and they all involve the use of a threshold, and, in most cases, they yield weighted unipartite networks. Usually, edges, the weights of which exceed the threshold value, are retained, while those with weights that are below the threshold value are omitted. The methods greatly vary, however, on the way threshold values are identified. The simplest and most widespread approach for extracting the backbone of bipartite projections is through the application of an unconditional (or global) threshold. In particular, a single weight threshold is selected and applied to all edges in the bipartite projection, and edges are retained in the backbone network only if their weight in the bipartite projection exceeds this predefined threshold. The most commonly used

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

weight threshold of zero preserves all edges with a non-zero weight, whereas others have used different thresholds, including these sets at the percentage of the maximum observed edge weight or at the mean observed edge weight. The unconditional threshold approach, although widely used, suffers from several shortcomings. In general, if the presence of any shared connections to V-nodes is considered adequate for inferring that an edge exists between 2 U-nodes, then an unconditional threshold should be used for backbone network extraction. If, however, an instance of shared V-nodes is not sufficient to infer that an edge exists between 2 U-nodes, then unconditional threshold backbones may be problematic. The structure of a backbone extracted by using an unconditional threshold depends heavily on the selected threshold value; moreover, certain structural features of unconditional threshold backbones of bipartite networks are systematically biased. Thus, this approach in which a universal threshold is applied indiscriminately to all edge weights can yield a 1-mode projection with several undesirable properties [66].

Several methods with thresholds conditioned on the U-nodes' degree are available and include in the backbone edges the weights of which exceed weight values expected in a null model. All methods begin with a standard projection and then use a statistical model to assess the significance of the weights [71,72]. Some methods involve normalization of the edge weights in the bipartite projection in a way that adjusts U-nodes' varying numbers of interactions with V-nodes and transforms the edge weights into measures that assess the tendencies or revealed preferences to co-occur. To this end, Bonacich suggested normalization [73], and Borgatti and Halgin used the Pearson correlation coefficient [74], whereas other methods relied on the hypergeometric distribution to perform a test for the statistical significance of edge weights, conditioned on each U-nodes' number of interacting V-nodes (i.e., row marginals in the bipartite network) [66,75,76].

Although the aforementioned methods are used for the improvement of unconditional thresholds, they have also been criticized because they implicitly treat V-nodes interchangeably. In such cases, those methods are not suitable for inferring U-nodes' relationships because they fail to consider V-nodes' differing degrees. To overcome the limitations of unconditional and U-nodes' degree conditioned threshold approaches, a null model is required to identify the distribution of expected edge weights that would be observed if U-nodes were linked to V-nodes randomly. This linking process is conditioned on (or constrained by) both the U-nodes' and V-nodes' degrees. The most widely used model is the fixed degree sequence model (FDSM), which compares the observed projection edge weights to the distribution of possible edge weights that might be observed if all U-nodes' and all V-nodes' degrees were fixed at their values in the empirical data. For instance, Zweig and Kauffman presented a systematic approach that evaluates the significance of the *co-occurrence* for each pair of nodes [70]. In principle, the FDSM yields a distribution of expected edge weights that is conditioned on both U-nodes' and V-nodes' degrees. However, in practice, FDSM risks overconditioning or imposing too many assumptions on the null model. To address this problem, Neal proposed the stochastic degree sequence model (SDSM) method that uses a Monte Carlo approach to assess the statistical significance of edge weights against a null model that is conditioned on each U-nodes' number of interacting V-nodes and each V-nodes' number of interacting U-nodes (i.e., both row and column marginals in the bipartite network) [67].

Bipartite Biological Networks

In this section, a brief description of the most important classes of biological networks that possess a native bipartite structure and data and the methods pertinent to bipartite biological networks used are provided. The objectives of the analysis in each case and the specific outcomes obtained from such analyses are outlined. The bipartite networks described below were arbitrarily classified by the authors into 4 broad categories, namely, ecological networks, molecular networks, biomedical networks, and epidemiological networks.

Ecological networks

Ecological networks (Figure 5A) are representations of the biotic interactions in ecosystems in which species are indicated by nodes that are connected by pairwise interactions that can be either trophic or symbiotic. Ecological networks are used to describe and compare the structures of real-world ecosystems. These network models are used to investigate the effects of network structure on properties such as ecosystem stability. A fundamental goal of ecological research is to unravel the mechanisms that influence the stability of fragile ecosystems. Thus, the relationship between ecosystem complexity and stability is a major topic of interest in ecology. The use of ecological networks makes it possible to analyze the effects of the network properties described above on the stability of an ecosystem. Ecological networks can be further subdivided into 3broad types: food webs (FWs), mutualistic webs (MWs), and host–parasitoid webs (HPWs). Although all 3 types contain trophic interactions, studies of FWs, according to the most strict definition, typically focus on predator–prey interactions where consumers that are usually bigger than their resources are involved [77].

Traditional FWs originate from the population biology school of thought and they focus on trophic links among organisms, particularly predator–prey and primary consumer-basal resource feeding relationships. Historically, research in ecological networks began from descriptions of trophic relationships in aquatic FWs. However, recent work has explored FWs, as well as webs of mutualists, and, as a result, several important properties of ecological networks have been identified. The energy flux through the web and the relationships between mass and numerical abundance of each species are common themes investigated in FWs. In general, FWs have high complexity, measured as connectance (i.e., the proportion of all possible links that are realized in a network), and smaller size compared to other biological networks [78]. Of note, FWs can have a native bipartite structure only when 2 layers are involved (i.e., plants and herbivores). However, quite often, they consist of ambiguously defined trophic levels connected by a number of links of intraguild predation and thus cannot be viewed as a single bipartite graph. In such cases, in order to perform analyses that rely on the bipartite structure (such as for nestedness), one needs to extract and analyze the bipartite subwebs embedded in them [79].

HPWs also originate from the population biology school of thought but they concentrate on a special type of predator–prey relationship, namely, between parasitoids and their hosts [80]. The term "parasitoid" is used to describe insects (usually parasitic wasps) that develop as larvae on the tissues of other arthropods (usually terrestrial insects), which they eventually kill. These networks are particularly well suited for a quantitative analysis because the number of hosts killed and the number of parasitoid individuals produced can be observed directly. Another advantage of these networks is that they are usually

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

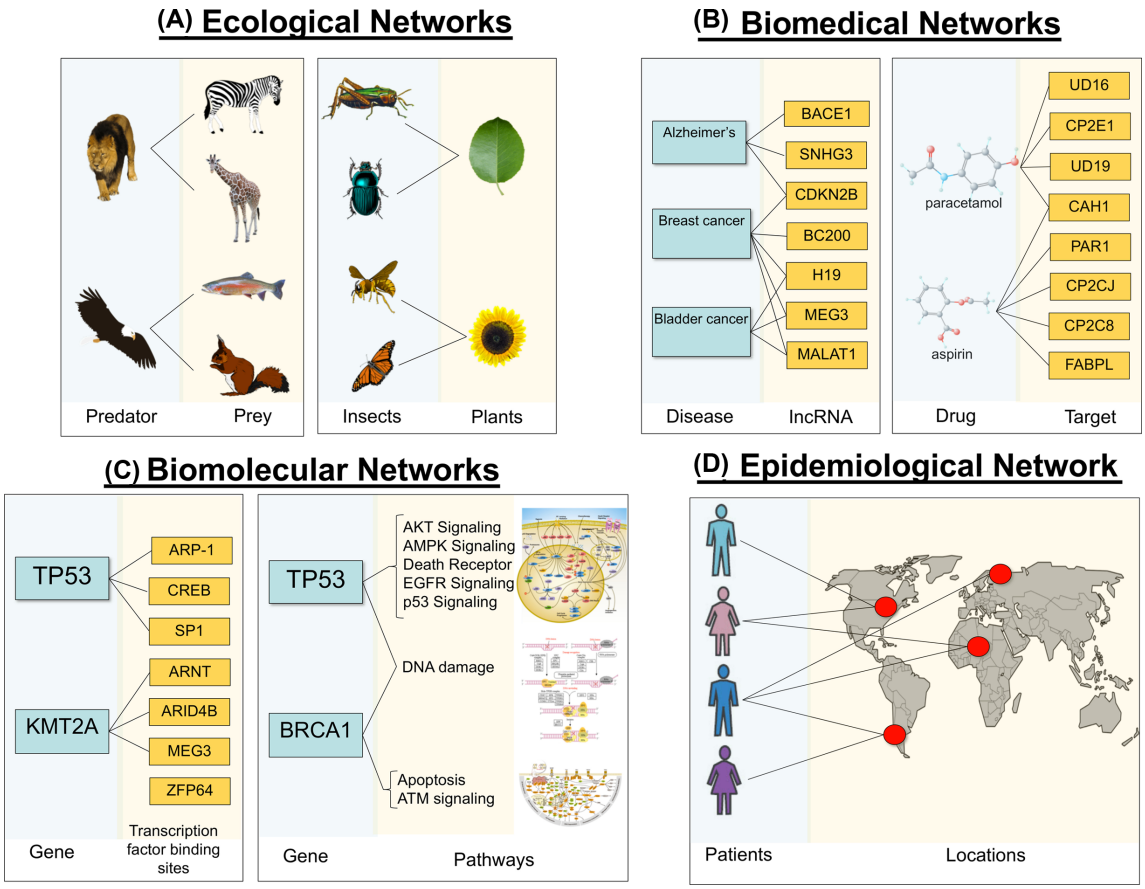


Figure 5: Overview and examples of various types of networks. (A) Ecological networks. An example of a predator–prey (left) and host–parasite example (right) network. (B) Biomedical networks. An example of a disease–gene (left) and a drug–target (right) network. (C) Biomolecular networks. An example of a gene–transcription factor binding site (left) and a gene–pathway (right) network. (D) Epidemiological network. An example of a patient–location network.

resolved at the level of species, avoiding potential problems with the use of "trophic species" of FWs, in which species that share predators and prey are clustered together. An obvious disadvantage of these networks is that they, by definition, focus on a small subset of the ecological community and are therefore not well suited for studying energy fluxes through the ecosystem [81].

MWs are used to study ecosystem properties relevant to pollination and seed dispersal, rather than population dynamics or energy fluxes. Among the various MWs studied in the literature, a significant portion is devoted to pollination networks, which depict the interactions between plants and their animal pollinators [82]; frugivore networks, which contain the interactions between plants and their animal seed dispersers; and ant–plant networks, which examine the interactions between plants that provide food and/or shelter for ants, which in turn provide protection for the plants [83]. Specialism tends to be a common feature in most MWs, at least compared to FWs, and this is probably even more the case for endosymbiotic systems. Most networks of plant–animal mutualism involve a small number of species. An analysis of 52 mutualistic networks showed that their nestedness is high. This pattern suggests a scale-free network. Thus, in mutualistic networks, the edges are not placed randomly. Furthermore, species communities with higher complexity (greater number of interactions) nestedness increase with the complexity (number of interactions) of the network, since for a given number of species, communities with more interactions are markedly more nested [84].

Nestedness, as we have already discussed, is considered to be an important topic in the study of ecological networks. A bipartite network, such as the one between plants and their mutualistic animals, is nested if specialists interact with species that form well-defined subsets of the species that generalists also interact with. A nested structure usually implies that there is a core of generalist species interacting among themselves and a tail of specialists interacting with most of the generalist species [84]. Within FWs, especially in aquatic systems, nestedness appears to be related to body size, because the diets of smaller predators tend to be nested subsets of those of larger predators. There seems to be 2 extremes, these are freshwater FWs that tend to have many generalists and HPWs that tend to have many specialized parasitoids. The nested structure of mutualistic networks is suggested to play a role in network stability. Additionally, recent analyses have shown that ecological networks are also modular and that the modularity co-occurred with nestedness [85]. Moreover, the correlation between nestedness and modularity depends on network connectance [56]. Although mathematical and computational analyses have suggested that nestedness increases species richness as well, an empirical analysis of 59 datasets representing mutualistic plant–pollinator networks showed that this statement may be incorrect. A simpler metric, the number of mutualistic partners of a species, has been found to be a much better predictor of species survival and, hence, community persistence. These results suggest that nestedness is, at best, a secondary

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

factor rather than a causative one for biodiversity in mutualistic communities [86].

The degree distribution of ecological networks is also debatable among scientists. The type of degree distribution (exponential or power-law form) is typically considered indicative of the overall architecture of the network. Early works suggested that the distribution of connections, $P(k)$, is skewed with long tails indicative of power-law scaling. Such features suggest that communities might be self-organized in a nonrandom fashion that might have important consequences in their resistance to perturbations (such as species removal) [87]. Others have indicated deviations from the "scale-free" topologies, a fact that is thought to result from nonmatching biological attributes of species that prevent the occurrence of certain interactions (the so-called forbidden links). Large-scale analysis for topological patterns in 29 plant–pollinator and 24 plant–frugivore networks showed that most of the plant–animal mutualistic networks show species connectivity distributions following a truncated power-law (broad-scale networks) and only a few show scale-free properties. It is suggested that plant–animal mutualistic networks follow a build-up process based on the preferential attachment of species [88]. The skewed degree distributions of bipartite mutualistic and antagonistic networks are usually assumed to show that ecological or co-evolutionary processes constrain the relative numbers of specialists and generalists in the network. Such constraints in adding links, including morphological mismatching between mutualistic partners, restrict the number of interactions established, thereby resulting in deviations from scale invariance. Other simpler models that do not require the existence of nonmatching species traits have also been proposed [89]. Finally, differences have been found between the degree distributions of mutualistic and antagonistic networks, suggesting that different processes are restricting these 2 classes of networks, especially the largest mutualistic networks. Probably, spatial and temporal heterogeneity largely affect the structure of the larger networks [90].

Of particular note, early ecological networks (e.g., FWs) were binary networks that merely depicted the presence or absence of feeding interactions and not the quantity of those interactions. This lack of quantification has been long recognized as a weakness in ecological network research, since not all species and interactions are equally important because not all are equally abundant. The availability of quantified webs highlighted the importance of link strength, establishing the notion that the strength of the interaction plays an important role in stability, with many weak and few strong links leading to stable but potentially complex webs. More recently, the focus has shifted again from exploring the magnitude of complexity and the strength of interactions to approaches for understanding the specific configuration of complexity (e.g., clustering, the importance of loops or motifs, and so on) [77]. All of the above are also verified by the analysis of the trends in establishing new metrics and algorithms suitable for quantitative networks, as well as the development of new methods for community detection or evaluation of the dynamic properties of the system (see below) [91].

Biomedical Networks

Contrary to ecological networks, *biomedical bipartite networks* (Figure 5B) are more abstract since the one partition of the network is usually composed of molecular components found in cells and the other of various indicators of human diseases. In particular, the one partition is usually composed of genes (or their protein products), drugs, or environmental exposures; the opposing partition is usually comprised of diseases, symptoms, or adverse drug effects. Thus, biomedical networks have introduced the network analysis techniques into the classical biomedical literature, since they use methods of network analysis in order to model factors that influence human diseases, traditionally analyzed with standard statistical methods. Therefore, this network-based approach in medicine offers a platform to explore not only the molecular complexity of a single disease but also to explore the molecular relationships among distinct pathophenotypes, identify new disease susceptibility genes, uncover the biological significance of disease-associated mutations, and identify drug targets and biomarkers for complex diseases [92].

The *gene–disease network* (*diseasome*), the archetype of this type of network, is a bipartite graph in which the first set of nodes consists of diseases and the opposing one of disease-associated genes [93]. A disease and a gene are connected by a link only if the gene is implicated in the particular disease. Given a bipartite network, one can construct by projection the human disease network, i.e., the network of human diseases where diseases with common genetic components are connected, or the human disease gene network, i.e., the network of human genes where genes participating in common human disorders are connected. The first diseasome was created based on the list of human disorders, disease genes, and associations between them obtained from the OMIM (Online Mendelian Inheritance in Man) database by Goh and coworkers [94]. However, although OMIM is one of the major repositories holding genetic association data for Mendelian diseases, it mainly archives rare disorders of high penetrance [95]. This parameter is of importance since multigenic diseases of low penetrance may have different properties that have to be taken into account. Other subsequent studies, such as the ones conducted by Barrenas et al. [96] and Liu et al. [97], partially overcome this issue by integrating gene-disease association data from multiple resources. Of importance, in the study conducted by Goh et al., [94] neither the disease concepts nor the gene terms were standardized, whereas in the studies conducted by Barrenas et al. [96] and Liu et al. [97], an effort was made to homogenize the disease concepts but not the gene terms. When data from genome-wide association studies (GWAS) are used, one is also able to construct a *gene–phenotype network* linking genetic polymorphisms to intermediate phenotypes, such as cholesterol levels or blood pressure [98]. Such approaches can be useful, especially in the context of identifying the causal pathways linking genetic variation, intermediate phenotypes, and diseases (Mendelian randomization), but the data on phenotypes are rather sparse. Recently, Kontou et al. [15], performed a similar analysis by combining data from OMIM and 2 other primary resources containing information of gene-disease associations: the National Institutes of Health's Genetic Association Database (GAD) [99], which contains data from genetic association studies that mostly target multigenic diseases of low penetrance, and the National Human Genome Research Institute (NHGRI) catalog of published GWAS [100], which includes a manually curated collection of published GWAS, with more than 100 000 assayed SNPs and SNP-disease associations. GAD and GWAS are, therefore, complementary to OMIM. Moreover, since disease name heterogeneity and ambiguity in all 3 repositories would not allow for a direct data comparison, the naming conventions described in the *International Classification of Diseases* (ICD-10) were used. Finally, in order to maintain a uniform nomenclature, all genes from the 3 databases were converted to the official Human Genome

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

Organisation (HUGO) Gene Nomenclature Committee [101] gene symbols.

Such network-based approaches for the discovery of gene-disease associations have enabled biomedical researchers to not only investigate the genetic complexity of a particular disease but also the relatedness among apparently discrete disease phenotypes [92,102]. Diseases have been found to be highly connected genetically, displaying many connections between both individual disorders and disorder classes. In other words, there seems to be a widespread genetic relatedness across many diverse domains of human disorders, transcending traditional disease categorization. Moreover, disease networks can provide the foundation for predicting causative genes, thereby unraveling a disease's underlying molecular mechanisms and enabling the design of new therapeutic strategies [92,102]. Genes associated with similar disease phenotypes have a higher propensity to interact physically with each other, forming distinct disease-specific functional modules [103,104]. Connections between disorders are also not completely random. Rather, disorders tend to form clusters on the basis of similar pathophysiology. Conversely, diseases with similar phenotypes have an increased tendency to share genes [94]. To achieve this global connectivity, complex diseases, such as diabetes and obesity, play the role of "connectors," bridging in this way disorders from different classes. Finally, of particular note, some genes are associated with only a few diseases, whereas others are implicated in numerous diseases, and likewise, some diseases are influenced by only 1 to 2 genes, and others are caused by dozens of genes [93].

Gene–disease networks have also been constructed for various classes of related diseases, including autoimmune diseases [105], neurological diseases [106], cardiovascular diseases [107], and others. Tissue specificity is also considered in gene–disease networks, since clinical manifestations of diseases are usually restricted to specific tissues. Although some disease-associated genes are expressed only in certain tissues, the expression patterns of disease genes alone cannot explain the observed tissue specificity of diseases. By extending the diseasome, a network-based approach was used by Hayasaka and colleagues [108] to investigate how different brain areas are associated with genetic disorders and genes. In particular, the authors constructed a tripartite network with genes, diseases, and the affected brain areas. In the resulting network, a disproportionately large number of gene-disease and disease-brain associations were attributed to a small subset of genes, diseases, and brain areas. Furthermore, a small number of brain areas were found to be associated with a large number of the same genes and diseases. These core brain regions encompassed the areas identified by previous genome-wide association studies and suggest potential areas of focus for the future imaging genetics research. These ideas were implemented in the so-called disease–tissue network, which is an obvious extension of the diseasome, in order to include information regarding tissue specificity. The primary hypothesis here is that for a disease to manifest itself in a particular tissue, a whole functional subnetwork of genes (disease module) needs to be expressed in that tissue. The expression patterns of disease genes were combined with the human interactome, and the results indicated that genes expressed in a specific tissue tend to be localized in the same neighborhood of the interactome. On the contrary, genes expressed in different tissues are segregated in distinct network neighborhoods. Most importantly, Kitsak et al. [109] showed that it is the integrity and the completeness of the expression of the disease module that determines disease manifestation in selected tissues. This approach led to the construction of a disease–tissue network that offers a predictive map of the statistically significant disease-tissue associations. This approach allowed the researchers to examine known disease-tissue relationships and predict newly definable disease-tissue associations.

Further extending diseasome, a large-scale biomedical literature database (including PubMed and National Center for Biotechnology Information's [NCBI's] MeSH terms) was used to construct a *symptoms–disease network* (Human Symptoms Disease Network) and investigate the connection between clinical manifestations of diseases and their underlying molecular interactions [110]. In the projected network, the link weight of 2 diseases quantifies the degree of similarity of their respective symptoms. The authors integrated disease–gene association and PPI data and found that the symptom-based similarity of 2 diseases correlates strongly with the number of shared genetic associations and the extent to which their associated proteins interact. Moreover, the diversity of the clinical manifestations of a given disease can be related to the connectivity patterns of the underlying PPI network. Such approaches could be useful in the identification of unexpected associations between diseases, in disease etiology research, and in drug design.

Another important extension of the diseasome is based on the identification of environmental factors that influence diseases. The majority of diseases (especially the polygenic ones) are, in part, caused or influenced by human interaction with harmful environmental substances. Traditionally, epidemiological studies have investigated such exposures, whereas the identification of gene–environment interactions represents an important area of genetic epidemiology [111]. The *exposure–disease network* was compiled using a global repository of the Centers for Disease Control and Prevention, which contains literature surveys on matching environmental chemical substances exposure with human disorders. The bipartite network contained links from 60 substances to more than 150 disease phenotypes. The analysis of the bipartite network and the projected networks identified mercury, lead, and cadmium as being associated with the largest number of disorders. On the other hand, breast cancer, fetal abnormalities, and non-Hodgkin's lymphoma were found to be associated with most of the environmental chemicals. Moreover, tobacco smoke compounds, parabens, and heavy metals tend to be connected, implying common disease-causing factors; however, this is not the case for fungicides and phyto estrogens [112].

Furthermore, the diseasome was extended to include drugs. The *drugs–target network* (*drugome*) consists of a bipartite graph that links approved drugs with the their target proteins (the gene products) [113]. The network produced in this way connects most drugs into a highly interlinked giant component, with strong local clustering of similar drugs. Topological analyses of this network quantitatively showed an overabundance of drugs that target already targeted proteins, confirming the prevalence of the so-called me-too drugs on the market. To analyze the relationships between drug targets and disease gene products, the shortest distance between both sets of proteins was measured in models of the human interactome network. Although an enrichment for etiological drugs, which directly target the disease-causing component, was clearly observed, still a majority of existing drugs target components as far away from the disease-causing genes as a random target would do, suggesting a predominance of palliative-acting drugs. Finally, a significant shift toward the closer-to-target drugs approved after 1996 from those approved before 1996 was observed, supporting a recent trend toward rational drug design [113]. The diseasome can be further supplemented by a drugome.

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

Traditionally, new targets for drugs have been predicted on the basis of molecular or cellular features, by exploiting, e.g., similarity in drug chemical structure or activity across cell lines. An inference method based on the similarity of the drug–target bipartite network topology similarity, managed, however, the prediction of new targets for existing drugs. Thus, outperformed both drug-based similarity and target-based similarity inference methods. By using this method, 5 old drugs (montelukast, diclofenac, simvastatin, ketoconazole, and itraconazole) were found to have polypharmacological effects on human estrogen receptors or dipeptidyl peptidase-IV, whereas simvastatin and ketoconazole showed potent antiproliferative activities on breast cancer cell lines [114].

In a fashion similar to the drug–disease network, the *vaccine–disease* and the *vaccine–gene* networks were constructed by Zhang and coworkers [115]. From these networks, those genes that interact with many vaccines and, conversely, those vaccines associated with many genes were identified as hubs. These findings correlated with existing knowledge and generated new hypotheses on the fundamental interaction mechanisms involving vaccines, diseases, and genes. Similar approaches were based on phenotypic side effect similarities (the *drug–side effects network*) in order to infer whether 2 drugs share a target. Campillos and colleagues [116] tested several such unexpected drug–drug relationships on 746 marketed drugs, validated the implied drug–target relations by *in vitro* binding assays, and found 11 drugs that exhibited significant activity. Nine of those were tested and confirmed in cell assays, documenting the feasibility of using phenotypic information to infer molecular interactions and hinting at new uses of marketed drugs [116]. Going a step further, a multilevel network (the *process–drug–side effect network*) was built by merging the *drug–biological process network* and the *drug–side effect network*. By analyzing the process–drug–side effect network, meaningful relationships between biological processes and side effects were inferred in an efficient manner [117].

Biomolecular Networks

Bipartite graphs provide an appropriate abstraction to represent relationships and associations between different classes of biological molecules and therefore have been extensively used for studying and modeling interactions between biomolecules (Figure 5C). Unlike biomedical networks, which represent relationships between abstract terms such as “diseases” or “phenotypes,” molecular networks illustrate interactions that occur physically between biomolecules and take place inside all various cell compartments. These interactions are reconstructed by using computational and mathematical methods of analysis applied on multiomics data generated from high-throughput experiments.

Data from high-throughput proteomics experiments (i.e., yeast-two-hybrid [Y2H], Immunoprecipitation-Mass Spectrometry [IP-MS], and tandem affinity-purification/mass spectrometry [TAP-MS]) are extensively modeled using bipartite graphs [118–120]. Bipartite graph models are utilized in different levels of analysis of PPI data, including assignment of individual peptides to proteins, as well as analysis and detection of protein complexes.

Peptide-to-protein assignment modeling by bipartite graphs

Any type of high-throughput proteomics experiment that uses MS reports a list of all the detected peptides and a measure of their abundance. The subsequent analysis requires the assignment of each identified peptide to the corresponding protein and an estimate of its abundance. Therefore, as many peptides are assigned to a single protein and many proteins share the same peptides, a bipartite graph between peptides and proteins is constructed to carry out the analysis. This network between peptides and proteins is then processed using a series of algorithms that operate on this bipartite structure to find the most appropriate protein assignment for each peptide. Inferring the correct proteins from these complex bipartite graphs is a difficult problem; therefore, methods based on empirical Bayesian analysis, reverse database search, and calculation of expectation values have been developed. Even though protein identification is the most widely used method for the analyses of MS-based proteomics data, the available software tools for identifying proteins are still not perfect [121,122]. A detailed review of available protein identification methods is provided by Nesvizhskii [123].

Protein complexes in protein-protein interaction

The modeling of protein complexes as networks plays the most important role in advancing our understanding of protein functions and elucidating the dynamics of cellular supermolecular organization. However, protein interaction data generated by high-throughput experiments such as Y2H and TAP-MS are challenged by the presence of high numbers of false positives and high false discovery rates [124]. Similar to the peptide–protein bipartite network, co-complex relations of proteins participating in different complexes are modeled as bipartite graphs in TAP-MS experiments. Here, each individual protein is included in 1 set of nodes, and the set of complexes in which it participates in comprises the second set of nodes. In recent years, there has been a growing number of efforts to incorporate interdomain knowledge to support large-scale analysis of PPI networks. A representative study incorporates Gene Ontology (GO) semantic terms and topological features of the baits and prey proteins to calculate pairwise similarities of baits and generate “seeds” of clusters. Then, each seed cluster is extended to recruit prey proteins that are significantly associated with the same GO terms. Next, network clique and motifs algorithms are applied to identify the protein complexes [118]. An additional typical technique uses network community structure detection algorithms together with 2 well-established machine learning algorithms to predict the protein-complex bipartite network in *Saccharomyces cerevisiae*. Communities were detected by a modularity detection algorithm, and the community-assisted method has outperformed a neighboring assisting method [125]. A more recent approach [126] involves a method inspired from spectral analysis, where the network power graph analysis is applied for the identification of complete biclique motifs. These motifs corresponded well to protein complexes, and a revisit of a characteristic study led to the prediction of the catalytic and regulatory subunits of the casein kinase II complex, as well as the untangling and identification of new protein interactions in the nucleosome.

Gene regulatory networks and *gene co-expression networks*, realized by the physical interaction (binding) of Transcription Factors (TFs) to the regulatory regions of target genes, can readily be modeled as bipartite graphs, where 1 layer of nodes represents the regulatory genes and a second layer of nodes represents target genes. Consequently, every edge in the graph represents a regulatory relation in the form of binding of each regulatory gene product (encoded by the regulatory gene) to the regulatory region of the target gene. Moreover, the respective

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

weights associated with each regulatory edge may represent the influence or the interaction strength between a TF and the regulated gene. An important property of network connectivity, i.e., versatility, emerges when bipartite graphs are used to describe data derived from transcriptomics experiments. In the process of discovering the simplest (sparser) bipartite network able to describe the data [127], a biologically meaningful distinction between versatile and nonversatile networks was made. Versatile networks can describe any type of data and thus are indistinguishable from one another, whereas nonversatile networks require a limited set of data due to constrains imposed by the data. This limitation, however, can be utilized for the reconstruction of network topologies and regulatory signals and to get a glimpse into the biological meaning of the regulatory interactions.

Modeling gene transcription regulation with bipartite graphs facilitated the development of network reconstruction methods involving the decomposition of the gene expression matrix. Typically, a matrix of dimensionality *N*x*M* (*N* genes and *M* samples) is broken down to regulatory signals and regulatory strengths. Established matrix decomposition methods, such as principal component analysis (PCA), independent component analysis (ICA), and singular value decomposition (SVD), have been applied to reduce the dimensionality of the gene expression matrix and, therefore, reconstruct regulatory interactions. However, all PCA-, ICA-, and SVD-based methods use statistical assumptions such as orthogonality and statistical independence and perform decompositions that are difficult to be interpreted in biomolecular systems. Nevertheless, the bipartite network representation of Gene Regulatory Networks (GRNs) permitted the development of a family of methods termed network component analysis (NCA), first introduced by Liao et al. [128]. NCA-based methods are able, under certain constraints, to find scaled reconstructions of the gene expression matrix in 2 matrices $|A|$ and $|P|$, where $|A|$ (an *N*x*L* matrix) contains the regulatory strengths of *L* regulatory genes on the *N* regulated genes and $|P|$ (an *L*x*M* matrix) contains the regulatory signals of *L* regulators in *M* conditions. The criteria that have to be met in order for NCA to be able to perform matrix decomposition include full rank of the matrix $|A|$ (full column rank of matrix $|A|$ must also be maintained even after the removal of a regulatory node, which implies that each column of $|A|$ must have at least *L-1* zeros) and full row rank of matrix $|P|$ Modeling of GRNs as bipartite graphs and their decomposition with NCA have been extensively applied, as NCA's criteria are easily fulfilled by a broad spectrum of biological systems. For a comprehensive review of the different algorithmic approaches and the different biological applications of the NCA-based methods on biological systems, see Wang et al.[129].

In an effort to extend NCA, Ye and coworkers incorporated genetic variation data in the form of SNPs, together with gene expression and ChIP-Chip data, for the concentrations and binding site affinities of TFs in a framework that predicted accurately trans- and cis-acting SNPs. Here, the trans-acting SNPs correspond to the products of regulatory genes and cis-acting SNPs correspond to the binding sites of these products in the target genes of the bipartite network. [130].

Furthermore, a recent method named, CONDOR [131], uses the modular structure of the bipartite graph to associate SNPs with genes' functions. In brief, this method utilizes the community structure of the bipartite graph (i.e., hubs and local clusters) in order to associate expression quantitative traits loci (eQTLs) with the genomic context. The context here is defined not only in terms of genes in the immediate proximity of significant genetic variants but also in terms of the functionally implicated genes through the bipartite network structure analysis. The method exploits genome-wide eQTL analysis in a way that is not restricted to the immediate neighbors of the eQTL-SNP gene.

Furthermore, bipartite network analysis of transcription regulation has been applied to comparative studies of GRNs [132]. In this study, the projected networks of transcription factors and regulated genes (RGs) from *Escherichia coli* and *S. cerevisiae* have been compared to find common characteristics and differences. The connectivity patterns of these 2 networks were found to be very similar. To better understand the differences, randomized versions of the original networks have been constructed. The difference of the TF to RG ratios among species has been found to be the most significant, highlighting a major organizational difference in transcription regulation between prokaryotes and eukaryotes.

Other gene expression regulation networks

The increasing availability and decreasing price of the high-throughput experiments resulted in the generation of an growing number of datasets that involve different types of biological entities including TFs, mRNAs, proteins, and regulatory sequences (e.g., enhancers, repressors). Bipartite graphs provide a suitable structure to model and analyze this wealth of data. In a network topology analysis work, the SICORE algorithm has been proposed [133] for the identification of regulation as well as coregulation effects using protein arrays, miRNAs, and gene expression. In a similar multidata integrative method, data from the human protein interaction network were combined with those from the transcription regulatory network to characterize coregulatory modules [134]. The method entailed a probabilistic statistical model that evaluated whether a cluster of coregulated proteins is likely to form a transcriptional regulatory module in an integrated network.

A novel class of noncoding RNAs has been discovered recently, the long noncoding RNAs (lncRNAs), more than 200 nucleotides in length, a feature that sets them apart from the other small regulatory RNAs. Evidently, bipartite graphs provide a suitable model to study the structural roles of lncRNAs, as 1 layer of nodes can represent the lncRNAs and the second layer the proteins they interact with. A method termed "lncRNA–protein bipartite network inference" has been developed recently [135] that is proposed to be the first to allow the construction of such networks. The method relies on the extraction of characterized lncRNAs–protein interactions from online databases and the usage of a propagation technique to assign each protein a score that is specific for each lncRNA, thereby providing a full set of ranked lists of interacting proteins for every lncRNA.

The bipartite metabolite-reaction representation of the metabolism is a reliable model to represent *metabolic networks* and metabolomics data, where data can be assigned separately with 1 layer of nodes representing the metabolites and the second layer representing the reactions. This representation avoids most of the erroneous assignments of isozymes as well as multifunctional enzymes. A tool introducing active module analysis of metabolic bipartite networks (AMBIENT) has been proposed as an effective means to analyze high-throughput data in a metabolic context [136]. Moreover, in a biomedical study [137], bipartite KEGG *pathways–gene networks* have been investigated together with the detection of differentially expressed genes (DEGs) from microarray experiments. The approach comprised a machine learning method that combines classification from both DEG-derived networks and bipartite KEGG pathways. The generated model was then applied to a series of cancer datasets

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

and was able to robustly reduce the frequently high number of false positives occurring in single DEG experiments.

To summarize, bipartite networks are invaluable in modeling and studying biomolecular networks for 2 major reasons. First, they provide a straightforward abstraction as the 2 different layers of nodes correspond directly to 2 different sets of biomolecular entities with distinct properties. Second, there are several powerful analytical methods from graph theory and linear algebra that, by taking into account particular types of bipartite network connectivity, provide solutions to the data representation and complexity reconstruction problems of the multiomics, high-dimensional, high-throughput biological data.

This section reviewed the most commonly used methods and examples, but it remains important to highlight a unifying method applicable to all bipartite biomolecular graphs, i.e., the power graph analysis. Power graphs are topological transformations of biomolecular networks into less redundant representations. This is achieved by exploiting the abundance of bicliques as topological motifs that are elementary, essential, and embedded in the structure of biological networks. Power graph analysis is an analytical tool that can easily be generalized and applied to directed, undirected, and bipartite networks [138]. However, it always returns a bipartite graph that describes a complex, "hairy ball"–like network by bipartite structures. Power graph analysis for the identification of protein complexes has notable application in the analysis of bipartite GRNs. Moreover, power graph analysis allows the decomposition of a bipartite network into a union of significant motifs, such as the star motif, the clique motifs, and bicliques [126]. This decomposition was used to discover a hierarchy of clusters of transcription factors linked to a hierarchy of clusters of target genes, thereby permitting reproduction of the results of a laborious combined experimental and computational previous study [139] where only the bipartite network structure of the transcription regulatory network in yeast was used as input.

Epidemiological Networks

Another distinct type of bipartite networks, as far as the type of data analyzed and the goals of the analysis are concerned, are those that are directly related to epidemiology (Figure 5D). These networks share some features with biomedical networks, with the focus on human diseases being the most important. However, the main difference lies in the fact that the data are collected and analyzed on an individual patient basis. In general, network analysis in public health and epidemiology resembles the classic approach of social networks analysis and has been used mainly to study disease transmission, especially for HIV infections/AIDS and other sexually transmitted diseases (STDs) [140]. Bipartite structures can be built based on individuals who are classified by gender, location, infectious agent, or comorbidities.

In one case, the *sexual contact network* can be represented as a bipartite graph, in which males form one part of the graph and females the other [141]. Such approaches can be valuable in the understanding of sexual behavior and the evolution of intimate relationships over time [142], as well as the modeling and simulation of STDs, especially HIV infections/AIDS [143–145]. Other theoretical studies have shown that, apart from the dependence between the epidemic threshold and the average and variance of the degree distribution of the network, there is a cutoff value for the infectivity of each population, below which no epidemic outbreak can occur, regardless of the value of the infectivity of the other population [146].

Vectorborne diseases, for which transmission occurs exclusively between vectors and hosts, can also be modeled as bipartite networks. In such models, theoretical work suggests that the spreading of the disease strongly depends on the degree distribution of the 2 sets of nodes and it is sufficient for 1 set to have a scale-free degree distribution with a slow enough decay for the network to have an asymptotically vanishing epidemic threshold [147].

Another case in which the bipartite network can model the spreading of a disease is when 1 set of nodes consists of geographic locations (clusters) in which the epidemic occurred, and the second set consists of the infected cases within a given time period. In this network, which is analyzed by projection, 2 locations are associated if they are both connected to common infected cases in the same period, and the number of infected cases is considered as the weight of the links [148].

Finally, a *comorbidities network* is a prominent example of an epidemiological network. Comorbidities, i.e., the co-occurrence of diseases, can provide valuable information regarding the underlying biological mechanisms of multifactorial diseases and can help to elucidate the effects of environmental exposures, such as diet, lifestyle, and medication, on diseases. By linking network dynamics to real-life data, patient data could provide a valuable basis for generating hypotheses concerning the mechanisms of disease and prove useful in drug repurposing and the development of targeted therapeutic strategies [149]. However, this type of information is conceptually different from the one encountered previously in biomedical networks, since it needs individual patient data in order to be compiled. In particular, detailed information from each patient is needed, and the adjacency matrix of the generated bipartite network closely resembles the traditional epidemiological datasets (the rows represent the patients and the columns the diseases). Projection of this bipartite network can also result in a unipartite network with the correlations of various comorbidities, the so-called phenotypic disease network (PDN) [150]. Large datasets of this type, which could be useful for network analysis, are difficult to be found, in general. However, since worldwide health transaction data are now often collected electronically, disease co-occurrences are currently analyzed quantitatively [151], and in some cases, these data cover entire nations [152]. In the most notable example of PDN, more than 30 million patients' electronic health records compiled from Medicare claims were used. By analyzing the co-occurrence of diseases and mortality, researchers found that disease progression can be studied using network methods, offering the opportunity to enhance our understanding of the origin and evolution of human diseases. Additionally, the dataset that was made publicly available represents the largest relational phenotypic resource that is publicly available to the research community [150]. Such data [153] can be used in other analytical techniques that are in use in traditional epidemiology (e.g., in meta-analysis of summary data). Other network analyses, such as analyses of co-morbidities of hip-fractures elderly patients, provided unexpected results that would be difficult to obtain otherwise, since patients with more serious comorbidities seem to have better follow-up that reduces the risk of readmission, whereas those with relatively less-serious specific comorbidities may have less stringent follow-up, leading to unanticipated incidents that precipitate readmission [154].

Models and Algorithms for Bipartite Graphs

In this section, some general related problems in bipartite graphs and the problem–solution algorithms are first described.

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

Then, some important properties of bipartite graphs that arise from viewing them as dynamic systems, such as percolation and controllability, are discussed.

Odd cycle transversal

A graph $G = (V, E)$ and a number k are given. Does there exist a set of at most k vertices, the removal of which from G would cause the resulting graph to be bipartite? The problem is NP-complete [155], i.e., there is no algorithm that can solve it within a polynomial time with respect to the size of the input, unless $P = NP$. The problem is fixed-parameter tractable, i.e., there is an algorithm, the running time of which can be bounded by a polynomial function of the size of the graph multiplied by an exponential function of k [156]. More specifically, the time for this algorithm is $O(3^k |E||V|)$ [157]. The name *odd cycle transversal* is attributed to the fact that a graph is considered as bipartite if and only if it has no odd cycles. Hence, deleting vertices from a graph in order to obtain a bipartite graph, one needs to "hit all odd cycle" or find a so-called odd cycle transversal set.

Edge bipartization

In a given graph $G = (V, E)$, it is possible to delete at most k edges so that the graph remains bipartite. This problem is also NP-complete and fixed-parameter tractable, and it can be solved in time $O(2^k |E|^2)$ [158].

Matching

A *matching* in a graph is a subset of its edges, where no 2 edges share an endpoint. In many cases, it is simpler to find a specific matching in bipartite graphs than in arbitrary graphs. A matching in a bipartite graph is called *perfect* if for every node of the graph there is an edge in the matching. Given a matching M, if $M+e$ is not a match for any edge e, then M is called *maximal matching*. A matching consisting of a maximum number of edges is called *maximum matching*. While a maximum matching is maximal, a maximal matching is not necessarily maximum. A maximal matching can be easily found by a greedy algorithm in any graph, while a maximum matching in a bipartite graph can be found in $O(\sqrt{(|V|)}|E|)$ time using the Hopcroft-Karp algorithm [159]. In weighted bipartite graphs, a *maximum weight matching* can be found within $O(|V|^2|E|)$ time using the *Hungarian* algorithm [160]. While the largest cardinality maximal matching (i.e., a maximum matching) can be found within polynomial time, a *minimum maximal matching* cannot be found in polynomial time unless $P = NP$. However, the number of edges in any maximal matching is at most twice the number of edges of the minimum maximal matching, and therefore the minimum maximal matching can be approximated within a factor of 2 in polynomial time. Although a perfect matching can be easily found in bipartite graphs by finding a maximum matching, counting the number of different perfect matchings in a bipartite graph appears to be very difficult. In fact, this problem is #P-complete, i.e., if there is a polynomial algorithm that solves it, then $P = NP$ [161].

Stable marriage. The *stable marriage problem* refers to an interesting problem related to bipartite graphs, which may have applications in biology. Let M and W be 2 sets of men and women, respectively, with $|M| = |W| = n$. Each man m in M has a preference $p_m(m, w)$ for each woman w in W and, conversely, each woman w in W has a preference $p_w(w, m)$ for each man m in M, so that:

- for all m, w: $1 \le p_m(m, w) \le n$
- for all m and any $w_1 \ne w_2$: $p_m(m, w_1) \ne p_m(m, w_2)$
- for all w, m: $1 \le p_w(w, m) \le n$
- for all w and any $m_1 \ne m_2$: $p_w(w, m_1) \ne p_w(w, m_2)$

In other words, each man (or woman) has a list of distinct preferences for each woman (or man). If $p_m(m, w) = 1$, then m first prefers w, while if $p_m(m, w) = 2$, then w is the second choice of m, and so on. The goal of the problem is to find n marriages between men and women so that every marriage is stable. A marriage (m, w) is not stable if and only if there is another married couple (m', w') so that $p_m(m, w) > p_m(m, w')$ and $p_w(w', m') > p_w(w', m)$. In other words, a marriage (m, w) is not stable if there is another married couple (m', w'), where m prefers w' than his wife and w' prefers m than her husband.

More formally, in a stable marriage problem, given a complete bipartite graph $G(V, U, E)$, where $|V| = |U| = n$ and each edge (v, u), where v (resp. u) belongs to V (resp. U), have been assigned 2 values: a value $p_m(v, u)$ and a value $p_w(u, v)$, defined as above. The question is whether there is a stable perfect matching in G that represents n stable marriages (defined as above) between the sets V, U. Of note, there are $n!$ different perfect matchings in G. The answer is that there is a fast algorithm [162] that always returns a stable perfect matching. The algorithm is as follows:

– In the case there is an unmarried woman:
  - Each unmarried woman proposes to the man that prefers that most among those that have not already rejected her,
  - Each man selects the woman who prefers the best among the women that proposed to him and rejects the rest proposals.

It has been proven that the above algorithm always returns stable marriages for all men and women within a $O(n^2)$ number of proposals, where n is the number of men (or women). A variant of the problem where the order of preferences is not strict, i.e., there are men (or women) that equally prefer other women (or men), has also been studied [163]. For more information regarding the problem and its variants, refer to a survey conducted by Iwama and Miyazaki [164].

Other general problems

Finding *the longest path* (i.e., finding a simple path of a maximum length) is NP-complete in bipartite graphs, in contrast to the shortest path that can be solved in polynomial time on any arbitrary graph. Moreover, the *girth of a graph* is defined as the length of the shortest cycle contained in the graph. Since bipartite graphs may contain only even cycles, the girth of a bipartite graph is an even number (or 0). Given a bipartite graph $G(V, U, E)$ of girth g, there is an algorithm for counting the number of cycles of length g, $g+2$, $g+4$, within $O(gn^3)$ time, where $n = max(|U|, |V|)$ [165]. In the *k-path partition* problem, the task is to partition a given graph G into the minimum number of paths, each of which has a length at most k. In bipartite graphs, the *k-path partition* problem is usually NP-complete while polynomial-time algorithms are known for specific families of bipartite graphs [166]. Furthermore, given a bipartite graph G, a *biclique* of G is a subgraph of G that is also a complete bipartite graph. Finding a *biclique* of a maximum number of vertices can be done in polynomial time [167], while finding a *biclique* of a maximum number of edges is NP-complete [168]. There is also a large body of literature on the methods for the optimal drawing of bipartite graphs [169–175]. An important algorithmic problem that arises

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

in this respect is drawing a bipartite graph in a way that minimizes crossing edges [176–178].

Percolation

Recent work on network theory has addressed the problem of resilience of networks by the random or targeted deletion of nodes or edges. From the perspective of statistical physics, "*percolation* is the simplest process showing a continuous phase transition." Percolation models on random bipartite graphs offer a simple illustration of this process. Percolation has been examined on graphs with a general degree distribution and has given accurate solutions to various cases, including bond percolation, site percolation, and models in which occupation probabilities depend on the degrees of the vertices [179]. From this point of view, the failure of a biomedical network could be considered as a percolation process, and the determination of the cutoff number of failed nodes/edges required to break down the whole network could be a particularly useful criterion for network failure. [180]. Percolation has been studied mainly in unipartite graphs, but recently the process has been described also in bipartite graphs [181]. In the particular model, throughout the percolation process, the links between nodes with degrees k and q are preserved with a probability proportional to $(kq)^{-\alpha}$, where α is positive so edges between hubs have greater probability to fail. The entire node/edge removal process was studied by using a theory of generating functions, and equations for the macroscopic description of the system were deduced.

Link prediction

The problem of *link prediction* refers to seeking a function of 2 vertices that denotes the similarity or proximity of the vertices. Link prediction enhances our understanding of the associations between nodes in bipartite networks. In general, there are several algorithms that can be used to extract missing information, identify spurious interactions, evaluate network-evolving mechanisms, and so on [182]. However, common link prediction functions for general (e.g., unipartite) graphs are defined using paths of length 2 between 2 nodes. Since in a bipartite graph adjacency vertices can only be connected by paths of odd lengths, these functions are not applicable. Instead, a certain class of graph kernels (spectral transformation kernels) can be generalized to bipartite graphs, where the positive semidefinite kernel constraint is relaxed by using the odd component of the underlying spectral transformation [183]. Other methods have also been developed, including those based on machine learning [184] and those that make use of the concept of internal links [185].

Graph ranking

Consider an edge-weighted graph $G = (V, E)$ where each edge of G has been assigned a positive real number. A set of preferences or order relationships among nodes of G is also given. This set is usually represented as a possibly directed weighted graph $G' = (V, E')$, where $E' \subseteq E$. Each edge of E' has been assigned a positive real number with the following interpretation: if $(u, v) \in E'$, then u must be ranked higher than v, and the penalty for misordering such a pair is given by the weight of edge (u, v) of G'. The goal is to rank the nodes of G so as to minimize the ranking error [186]. Ranking is a general problem in graphs of arbitrary structure, but the special structure imposed by the bipartite nature triggered the development of specialized algorithms [187]. Moreover, regularization-based algorithms have appeared, which find ranking functions that minimize regularized versions of the ranking error [188].

k-partite graphs

As mentioned in the beginning of the section Bipartite Graphs, a bipartite graph is a special case of a k-partite (or multipartite) graph for $k = 2$. More formally, a k-partite graph consists of k nonempty and disjoint sets of nodes $U_1, \ldots, U_k$ where a node $u \in U_i$ can share an edge with a node $v \in U_j$ only if $i \neq j$. In other words, any edge of the graph can only connect nodes in different sets (i.e., node sets U_i are independent). Bipartite and tripartite (i.e., for $k = 3$) graphs are probably the most studied families of k-partite graphs. Let us now list some interesting problems related to k-partite graphs. The problem of recognizing whether an arbitrary graph is k-partite is equivalent to the problem of deciding whether the nodes of the graph can be colored using at most k colors so that each node has been assigned 1 color and any 2 adjacent nodes have been assigned different colors. While recognizing that a bipartite graph can be easily done in polynomial time, recognizing a k-partite graph for any $k > 2$ is *NP*-complete. However, recognizing a complete k-partite graph (i.e., a k-partite graph where any 2 nodes in different node sets share an edge) can be done within polynomial time for any k. In fact, many problems that are *NP*-complete in arbitrary graphs and sometimes even in k-partite graphs can be solved within polynomial time in complete k-partite graphs, e.g., the maximum clique problem, the maximum independent set problem, the graph isomorphism problem, and the Hamiltonian cycle problem. A nice recent work on multipartite graphs with devoted sections to the applications of such graphs in biology can be found in Phillips [189] and Phillips et al. [190].

Community detection

A network is considered to have *community structure*, or clustering, if the nodes of the network can be grouped into (potentially overlapping) sets of nodes in a way that each set of nodes is densely connected internally, i.e., having many edges joining nodes of the same cluster and comparatively few edges joining nodes of different clusters. In the case of *nonoverlapping* community detection, the network is divided naturally into groups of nodes with dense connections internally and sparser connections between groups. There are plenty of methods available for detecting communities, ranging from traditional clustering methods (e.g., hierarchical, spectral) to divisive algorithms and methods that maximize the criterion of modularity [191]. In the case of bipartite networks, community detection has also received considerable attention [192]. It is well understood that community detection is related to the modularity of a network, which quantifies the extent to which vertices cluster into community groups, relatively to a null model network [193]. Moreover, research on the community structure in bipartite graphs has yielded new metrics for the clustering coefficient [45,194], and several specialized methods have been proposed for community detection [195–197] including algorithms for overlapping communities [198] as well as for quantitative biadjacency matrices [199]. A conceptually similar condition encountered mainly in gene expression studies is biclustering (also referred to as coclustering in the literature). Biclustering consists of simultaneous partitioning of the set of samples and the set of their attributes (usually gene expression) into classes. Samples and genes classified together are supposed to have a high relevance to each other. The goal is to find submatrices where the genes

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

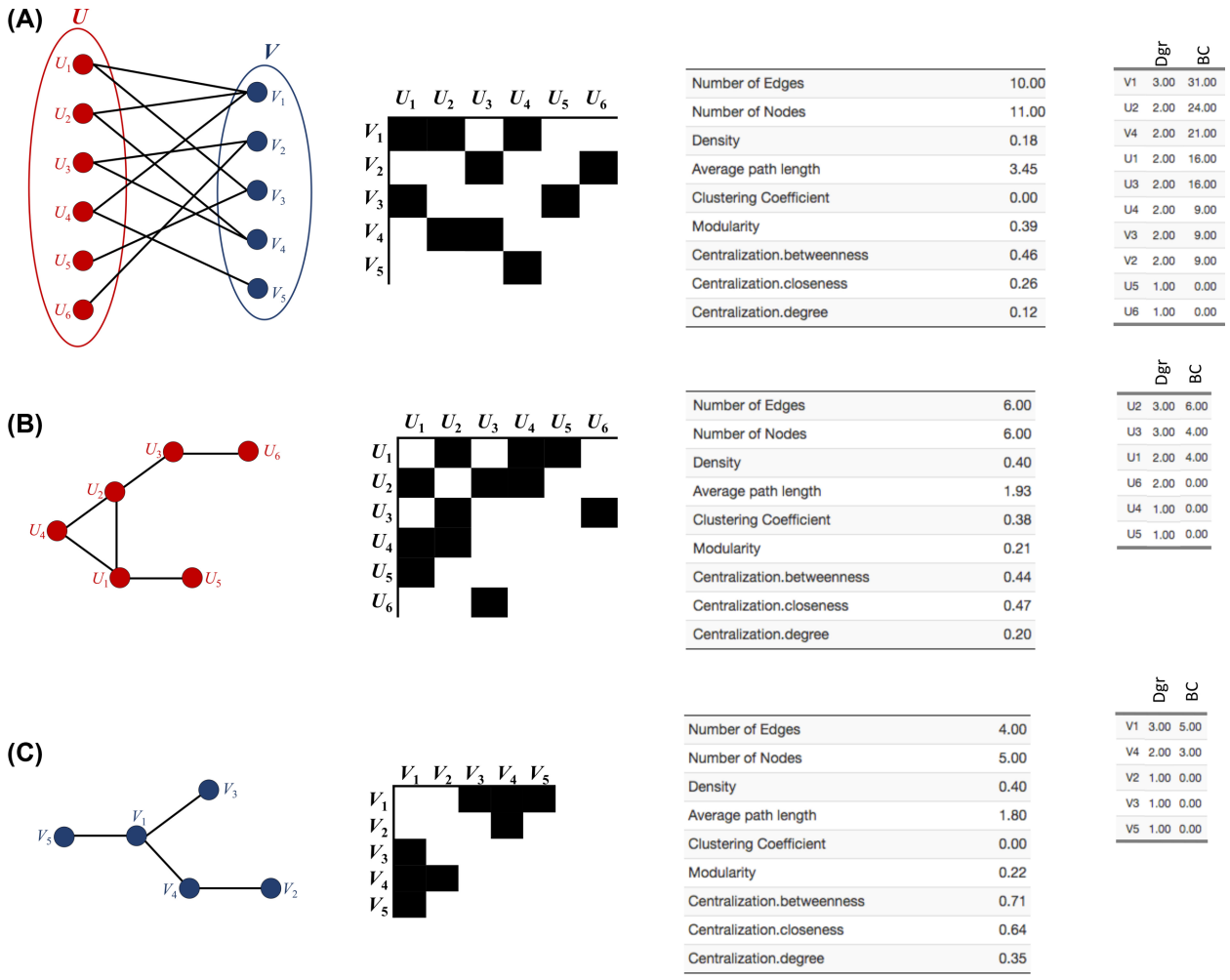


Figure 6: Numerical examples. (A) A small bipartite network, its adjacency matrix, several calculated topological features for the whole graph, and node ranking according to degree and betweenness centrality. Information relevant to projected unipartite networks (B and C).

exhibit highly correlated activities for every condition. The various biclustering methods for gene expression data are reviewed by Prelić et al. and Busygin et al. [200,201]. In general, biclustering methods are thought/presumed to have several advantages over conventional hierarchical clustering approaches; there are also considerable performance differences between the 2 methods. Thus, it would be interesting to test the application of biclustering methods in the task of community detection in bipartite graphs.

Controllability

Controllability describes the ability to drive a dynamic system (e.g., a network) from an initial state to a desired final state in finite time, with a suitable choice of inputs. The controllability of general directed and weighted complex networks has recently been the subject of intense study by several research groups. Investigation of the controllability of complex networks has led to the identification of the set of driver nodes with time-dependent control that can guide/drive the system's entire dynamics. Applications in several real networks revealed that the number of driver nodes is determined mainly by the network's degree distribution. Sparse heterogeneous networks are the most difficult to control, but dense and homogeneous networks can be controlled using only a few driver nodes. Counterintuitively, the driver nodes tend to avoid the high-degree nodes [202]. Furthermore, an analytical framework to address the controllability of bipartite networks is based on the dominating se –based approach, which identifies the topologies that are relatively easy to control with the minimum number of driver nodes. Such approaches offer a promising framework to control bipartite networks and study their undesired behavior [203].

Examples of Bipartite Network Analysis

In previous sections, we discussed some of the topological features of bipartite graphs, and we presented the main categories of biological bipartite networks. Here, we present some examples of bipartite network analysis, using both artificial data and real data.

Detection of patterns using topological features

We used the NAP application [36] in order to give numeric calculations of several of the topological features and metrics described earlier. In Figure 6 we visualize a small bipartite graph with its 2 projected networks and show numeric calculations about their density, the average path length, the clustering

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

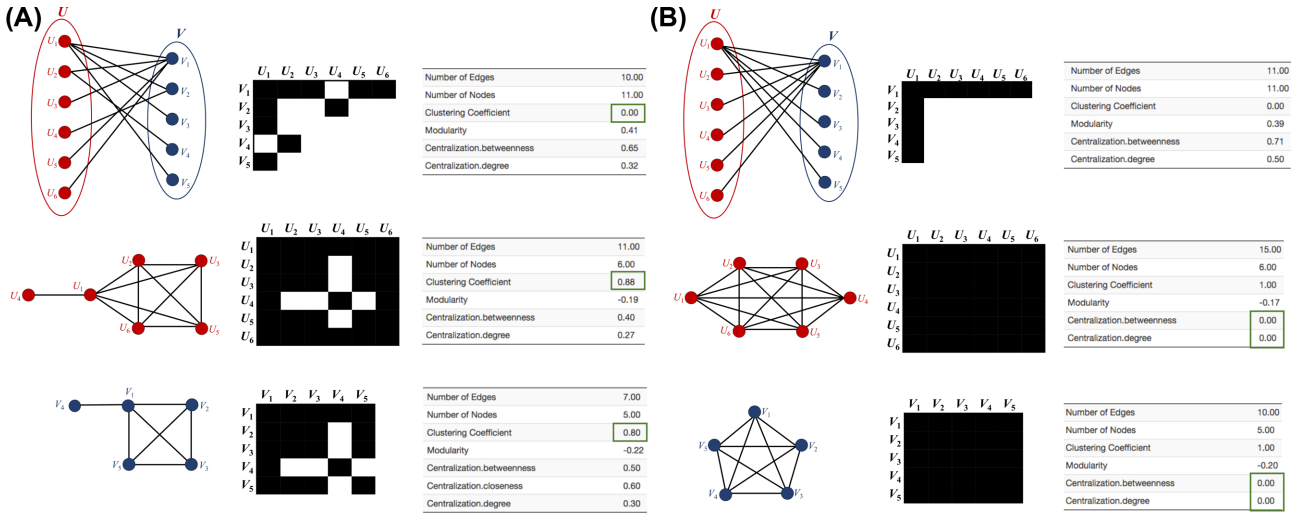


Figure 7: Two examples of the way some topological features of the projected unipartite networks are affected by the bipartite graph's nestedness. (A) Nested bipartite graph. (B) Fully nested bipartite graph. The higher the nestedness of the bipartite graph, the more connected the projected networks. Maximum nestedness leads to fully connected unipartite networks (cliques).

coefficient, the modularity betweenness centrality, the closeness centrality, and the average connectivity degree. Furthermore, we show an automatically calculated ranking of the nodes related to their connectivity and betweenness centrality; nodes with higher rank appear first. In Figure 7, we show how some of these features of the 2 projected networks change in relation to the bipartite graph's topology. For example, in Figure 7A, it is shown that the more nested a bipartite graph is, the lower the clustering coefficient is, as it does not tend to form clusters. Similarly, the higher the nestedness of the bipartite graph, the higher the betweenness centrality of its projected networks. A fully nested bipartite graph that generates 2 fully connected networks (cliques) is shown in Figure 7B. In addition, we observe how nestedness affects the betweenness centrality as well as the centralization degree (hubs). Both are zero since there are no hubs and no nodes bridging communities. In Figure 8 the extent to which the modularity of a bipartite graph can affect the topology of the 2 projections is shown, which also allows us to infer similarity conclusion.

Analysis of the gene–disease network

As mentioned earlier, network-based approaches for the discovery of gene–disease associations have enabled biomedical researchers to investigate the genetic complexity of a particular disease and the relatedness among apparently discrete disease phenotypes. To illustrate the steps that need to be followed in the analysis of such a network, we used as a test case the bipartite networks that contained the associations between the human diseases and the genes that confer susceptibility to these disease from the study conducted by Kontou et al. [15]. The particular analysis was performed by combining data from OMIM and 2 other primary resources containing information of gene–disease associations, the NIH's GAD and the NHGRI catalog of published GWAS. In the original publication, the datasets were combined, but for purposes of illustration, here we used only the GAD dataset in order to avoid confusion.

The original data can be found in Kontou et al. [204], and the reader can directly upload the biadjacency matrix to a network analysis tool in order to calculate topological features of bipartite networks and visualize the bipartite structure. For the analysis, the visual representation, and the projection, we used NAP [36], igraph [205], and bigraph [206], as well as BiLayout [207] and PowerClust [208]. For a detailed presentation of the tools, see the corresponding section. Figure 9A shows numeric calculations regarding the density, diameter, clustering coefficient, modularity, betweenness and closeness centrality, connectance, generality, and vulnerability of the GAD bipartite network. The mean number of genes per disease is 14.82, whereas the mean number of diseases per gene is 1.21. The statistical properties of the network are captured by the proposed metrics, which reveal a moderately dense, asymmetric network (few genes, many diseases), with modular architecture and having a moderate degree of betweenness centralization and low closeness centralization. These properties dictate further the properties of the projected unipartite networks. The disease–disease network is denser, with smaller diameter and a clustering coefficient equal to 0.44, whereas the gene–gene network is wider, with smaller density but larger tendency to form clusters (coefficient equal to 0.75). Betweenness and closeness centralities are comparable for the 2 projected networks. In Figure 9C–G we also present various visualizations of the bipartite structure as well as of the projected networks.

Bipartite Graphs, the Biomedical Big-Data Era, and Some Hints for the Future

Today, we live in the era of big data, where the exponential growth of information in the biosphere is evident. The protein and genome landscapes change continuously as new and hypothetical proteins and genome fragments appear every day. Integrated Microbial Genomes (IMG) [11] today includes approximately 6000 bacterial, 1500 archaeal, approximately 300 eukaryotic, approximately 8000 viral isolate genomes, approximately 1200 genome fragments, 6500 metagenomes, and approximately 2000 metatranscriptomes. Based on a very approximate estimate, this corresponds to approximately 70 million proteins coming from the isolate genomes and approximately 4–10 billion proteins coming from the metagenomes and metatranscriptomes. In addition, the UniProtKB/TrEMBL release of 15 February 2017 [209] contains approximately 7 750 000 sequence entries. Moreover, Uniparc contains approximately

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

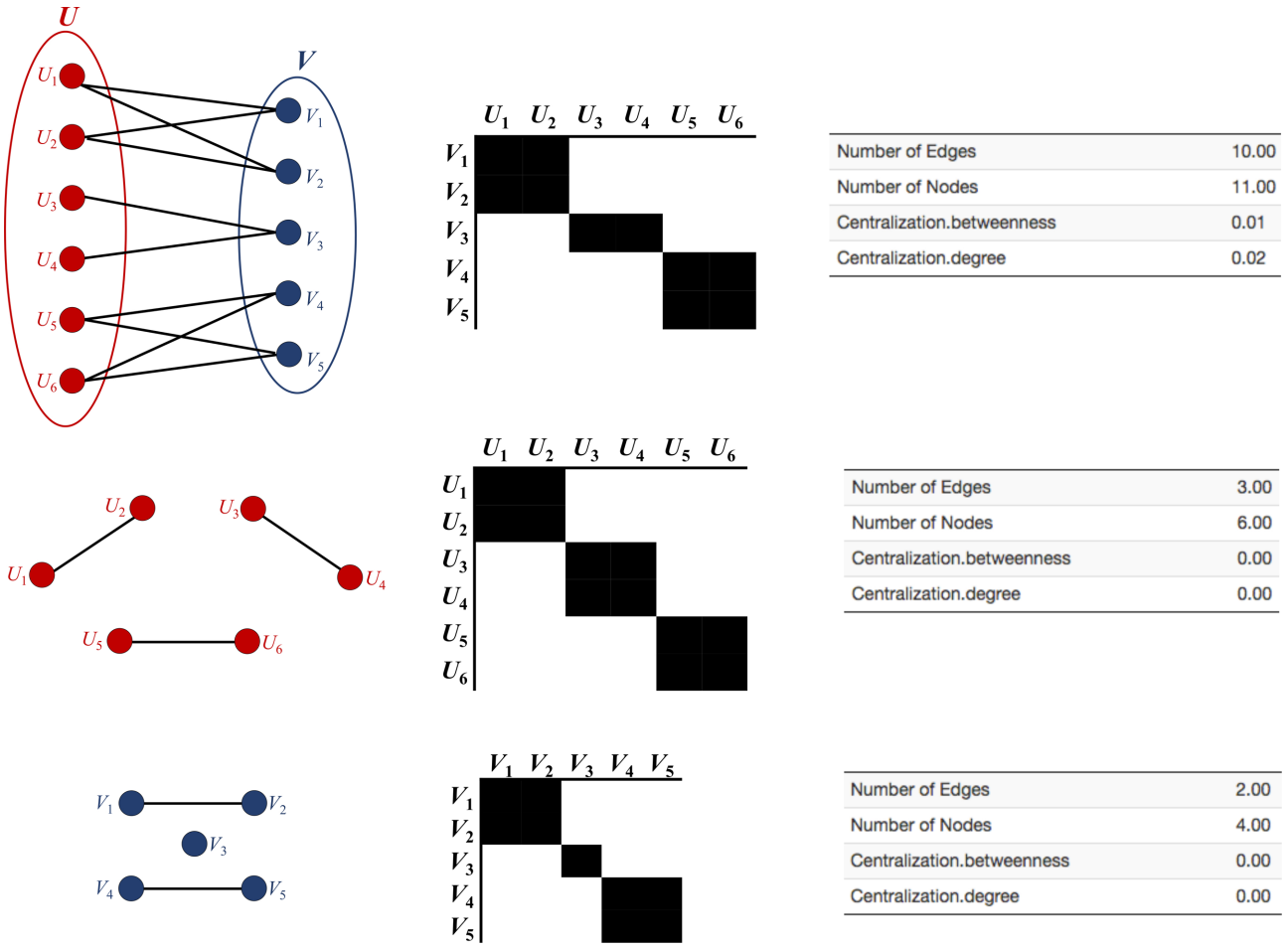


Figure 8: Example of the extent to which a bipartite graph's modularity affects the unipartite projected networks.

150 000 000 protein entries. Protein Family Database (PFAM) [210], version 31.0, a database of a large collection of protein families that organizes proteins into families by similar domains, consists of approximately 17 000 entries. Today, NCBI hosts 1 billion sequences corresponding to 2.2 trillion bases, and RefSeq alone hosts more than 100 million complete accessions. Moreover, PubMed hosts more than 27 million articles today. Also, other databases that host results from high-throughput experiments increase in size every day. Thus, comparative genomics and integrative biology are areas that already have and are expected to experience a boom in the coming years and to dominate other areas within the broader big-data spectrum (e.g., internet of things).

Another important attribute of the biomedical bipartite networks, which reflect the abstract nature of the entities that they contain, is that they make extensive use of data integration techniques and rely on incorporating data from multiple sources (e.g., diseases, SNPs, gene expression, PPIs, clinical symptoms, pharmaceutical drugs), contrary to the ecological and molecular networks. This highlights the need for the creation of publicly available biological databases containing high-quality data. Biological databases, in general, play a central role in bioinformatics, since they offer scientists the opportunity to access a wide variety of biologically relevant data [211]. Furthermore, they are indispensable in the context of network medicine and systems biology and medicine, since the primary data from several databases need to be integrated in order to achieve the desired result [212,213]. Biological databases continue to grow, and the need for data integration techniques, as well as the potential applications in network medicine and systems biology, also increases [214,215]. Initiatives for standardization and construction of ontologies is also of paramount importance in this respect. Currently, available and up-to-date databases exist for a large variety of data, including SNPs [216], RNAs [217], PPIs [218,219], biomolecular pathways [220], drugs [221–223], and diseases [224]. However, the gene–disease relationships, which form the basis of biomedical networks, are considered especially problematic since genetic association studies are characterized by nonreplicability [225,226] and most approaches to collecting data for gene–disease analysis are based on the clearest gene–disease associations derived from the literature. In this respect, OMIM and the GWAS catalog are indispensable resources, but the recent discontinuation of GAD signifies the need for a more sophisticated resource that will contain replicated and unbiased genetic association data.

Today's high-performance computing capabilities allow for analysis of massive networks, but scalability, analysis, and visualization remain a bottleneck [227]. For example, in terms of visualization, layouting a bi- or *n*-partite network remains a challenge. While efficient layout algorithms such as the OpenOrd [228] and Yifan-Hu [229] can be applied on generic networks, limited efforts have been made to lay out large-scale *n*-partite networks, thus rendering the visualization of such networks with current methods unattractive.

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

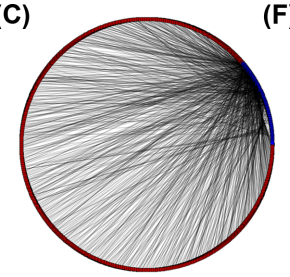
(A)

Bipartite Network		
	Nodes	888 (791G, 98D)
	Edges	1023
	Radius	1.00000
	Density	0.00000
	Diameter	14.00000
	Modularity	0.86233
	Clustering Coefficient	0.00000
	Centralization Betweeness	0.25123
	Centralization Closeness	0.00000
	Linkage Density	1.21352
	Connectance	0.01915
	Generality	14.82609
	Vulnerability	1.32170

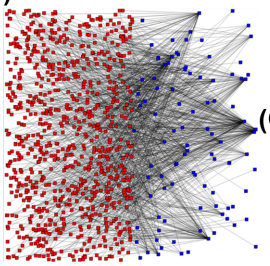
(B)

GENE	DISEASE
ABCD1	SCHIZOPHRENIA
ADAM30	NON-INSULIN DEPENDENT_DIABETES_MELLITUS
ADAMTS9	NON-INSULIN-DEPENDENT_DIABETES_MELLITUS
BIN1	ALZHEIMER'S_DISEASE
BMP4	MALIGNANT_NEOPLASM_OF_COLON
BOD1	CROHN'S_DISEASE
BPI	BIPOLAR_AFFECTIVE_DISORDER
BPI	SCHIZOPHRENIA
CCDC62	PARKINSON'S_DISEASE
FIGN	HYPERTENSION
RPL17P45	OBESITY
ST13P4	MULTIPLE_SCLEROSIS
ZNF646	PARKINSON'S_DISEASE
ZMIZ1	CROHN'S_DISEASE
ZMIZ1	MULTIPLE_SCLEROSIS
ST13P1	NON-INSULIN - DEPENDENT_DIABETES_MELLITUS
ST13P1	SCHIZOPHRENIA

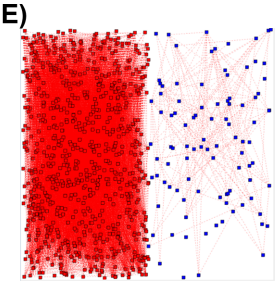
(C)



(D)



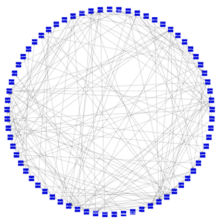
(E)

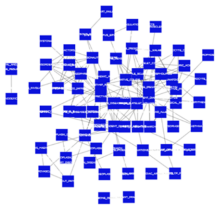


(F)

Disease-Disease network		
	Nodes	69
	Edges	168
	Radius	1.00000
	Density	0.07214
	Diameter	6.00000
	Clustering Coefficient	0.44321
	Centralization Betweeness	0.21000
	Centralization Closeness	0.00414

CROHN'S_DISEASE	MULTIPLE_SCLEROSIS
BIPOLAR_AFFECTIVE_DISORDER	SCHIZOPHRENIA
NON-INSULIN-DEPENDENT_DIABETES_MELLITUS	SCHIZOPHRENIA





(G)

Gene-Gene network		
	Nodes	774
	Edges	14903
	Radius	1.00000
	Density	0.03843
	Diameter	10.00000
	Clustering Coefficient	0.73574
	Centralization Betweeness	0.16384
	Centralization Closeness	0.00000



ADAM30	ADAMTS9
BOD1	ZMIZ1
ABCD1	BPI
CCDC62	ZNF646
BPI	ST13P1
ABCD1	ST13P1
ADAM30	ST13P1

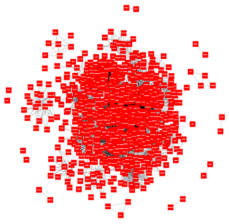


Figure 9: A test case of a bipartite gene–disease network from the genetic association database. (A) Topological features of the whole bipartite network. (B) Data example of bipartite network (gene–disease). (C) Circular visualization of the bipartite network (genes, red; diseases, blue) using PowerClust. (D) Random visualization of the bipartite network showing the directed connections between the 2 disjoint sets of nodes using PowerClust. (E) Random visualization of the bipartite network showing the indirect connections between the 2 disjoint sets of nodes using PowerClust. (F) Topological features of the projected disease–disease network and an example of the monopartite network and different types of visualization. (G) Topological features of the projected gene–gene network and an example of the monopartite network and different types of visualization.

Therefore, the need for efficient visualization and layouting emerges.

In terms of network analysis, clustering is one of the most active research fields. While a plethora of generic clustering algorithms exist, great efforts have been made in the biomedical area to incorporate such algorithms within established network visualization tools. For example, Cytoscape's ClusterMaker plugin [230] includes attribute cluster algorithms such as AutoSOME clustering [231], Eisen's hierarchical and k-means clustering [232], as well as topology-based clustering algorithms, such as affinity propagation [233], community clustering (GLay) [234], MCODE [235], MCL [236], Spectral Clustering of Protein Sequences [237], and transitivity clustering [238]. While these efforts have proven to be very fruitful, often users misuse these algorithms without taking into consideration the topological characteristics of the network. As bipartite graphs come with their own properties, the implementation of scalable clustering algorithms that take advantage of their topology would be very powerful.

Overall, analysis, layout, and visualization adjusted to bipartite and further extended to n-partite graphs are still in their infancy and constitute a big gap in the biomedical field. Therefore, we believe that efficient and scalable tools covering these needs would become protagonists in the field in the future.

Software and tools for bipartite graphs

In this section, we discuss software applications and libraries that are available for the analysis and visualization of bipartite and n-partite networks. While tools for analysis and visualization of unipartite biological networks of general use are presented and analyzed elsewhere [239–242], Table 1 summarizes their functionalities, and Figure 10 shows how they can be used for visualizing bi- and n-partite graphs. However, in most cases, specialized software is needed either in the form of a plugin for an existing tool or as a completely different package.

Cytoscape [243] is an open-source, bioinformatics-oriented software platform mainly implemented to analyze and visualize generic interaction networks. Although it does not specialize in bipartite graphs, some functionality for visualizing and processing such graphs is available through several plugins [244]. Nevertheless, it comes with a plethora of simple and more sophisticated layout algorithms. Therefore, given a bipartite graph $G = (U, V, E)$, vertices of the disjoint sets U and V can be selected, placed separately, and organized by using local simple grid, hierarchical, or circular layouts. In addition, in order to easily follow the nodes of each layer, vertices of different groups can be colored accordingly.

DisGeNET [245] is a Cytoscape plugin designed to analyze human gene–disease association networks. DisGeNET allows users

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

Table 1 A summary of the tools dedicated to bipartite graph analysis and their properties

Tool	Software	Library	Usage	URL
Cytoscape	X		Generic network analysis tool	http://www.cytoscape.org/
DisGeNET	X		Cytoscape's plugin to analyze disease–gene interactions	http://www.disgenet.org/web/DisGeNET
BiLayout	X		Bipartite layout	http://bilayout.bioinf.mpi-inf.mpg.de
Pajek	X		Generic analysis and visualization tool	http://vlado.fmf.uni-lj.si/pub/networks/pajek/
NetworkX	X		Analysis of several types of graphs including bipartite graphs	https://networkx.github.io/
UCINET	X		Social networks; NetDraw is specialized for bipartite graphs	https://sites.google.com/site/ucinetsoftware/home
Gephi	X		Generic network analysis tool	https://gephi.org/
FALCON	X		Analysis of ecological networks	https://github.com/sjbeckett/FALCON
Arena3D	X		Visualization of multilayered graphs	http://arena3d.org/
BicAT	X		Analysis of networks based on biclustering techniques	http://www.tik.ee.ethz.ch/sop/bicat/
GeneWeaver	X		Integration of functional genomics experiments	https://geneweaver.org/
ONEMODE	X		Stata module for producing 1-mode projections of a bipartite network	http://fmwww.bc.edu/repec/bocode/o/onemode.ado
Circos	X		Data visualization using a circular layout	http://circos.ca/
Hiveplots	X		Data visualization using radially distributed linear axes	http://www.hiveplot.com/
Networksis		X	Tool to simulate bipartite networks	https://cran.r-project.org/web/packages/networksis/index.html
enaR		X	Provides algorithms for the analysis of ecological networks	https://cran.r-project.org/web/packages/enaR/
Netpredictor		X	Prediction of missing links in any given bipartite network	https://github.com/abhik1368/Shiny_NetPredictor
biGRAPH		X	Extension of the igraph library for bipartite graphs	https://cran.r-project.org/src/contrib/Archive/biGraph/
BiRewire		X	Bipartite network rewiring through N consecutive switching steps	https://bioconductor.org/packages/release/bioc/html/BiRewire.html
DEsubs		X	Visualization of disease-perturbed subpathways	http://bioconductor.org/packages/release/bioc/html/DEsubs.html

to access a gene–disease database containing integrated data from diverse public resources. DisGeNET presents the gene–disease networks (diseasome) as bipartite graphs and provides the option to view gene–gene and disease–disease networks derived from the diseasome. Advanced search options permit the generation of subnetworks and the analysis of sets of diseases associated through common genes.

BiLayout is a Java plugin that is used to compute a bipartite network layout for 2 groups of nodes. BiLayout allows some simple actions, such as selecting 1 of the groups, showing and hiding unconnected nodes, exporting groups of nodes, and resetting the network. The mouse-over effect allows the user-friendly and customized visualization of all neighbors of a certain node.

Pajek [246] is a free, noncommercial Windows (32-bit) program package for analysis and visualization of large networks (networks containing up to 1 billion vertices and an unlimited number of edges). Pajek implements several methods for the visualization of bipartite graphs and for the analysis of the unipartite projections of the bipartite graph.

NetworkX [247] is a software package for the generation, processing, and analysis of several types of graphs, including bipartite graphs. A node attribute named "bipartite" with values 0 or 1 enables the identification of the corresponding set of each node. The user has to make sure that there are no links between nodes that belong to the same set. Although NetworkX requires user intervention for creating bipartite networks, it provides several options for bipartite network drawing, projection, and data analysis.

UCINET [248] is a commercial software package for Windows designed primarily for the analysis of social network data. It is accompanied by the NetDraw network tool that can handle visualization of bipartite networks. However, the tool contains several options to calculate network metrics that are optimized for the analysis of unipartite graphs. Nevertheless, UCINET also contains modules for projecting the bipartite networks.

Gephi [249] is one of the best open-source visualization and exploration software programs for all kinds of graphs and networks. It can easily render networks that consist of up to 100 000 nodes and 1 000 000 edges.

FALCON [250] is a software package devoted to the analysis of ecological networks and allows user-friendly and efficient calculations of network metrics, such as nestedness scores, using state-of-the-art measures and models. The FALCON code is available in 3 programming languages (R, MATLAB, Octave) and allows users to install further measures and null models easily.

Arena3D [251,252] is an interactive and freely available 3D generic tool, mainly intended to visualize multilayered graphs. It uses a layered display to separate different levels of

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

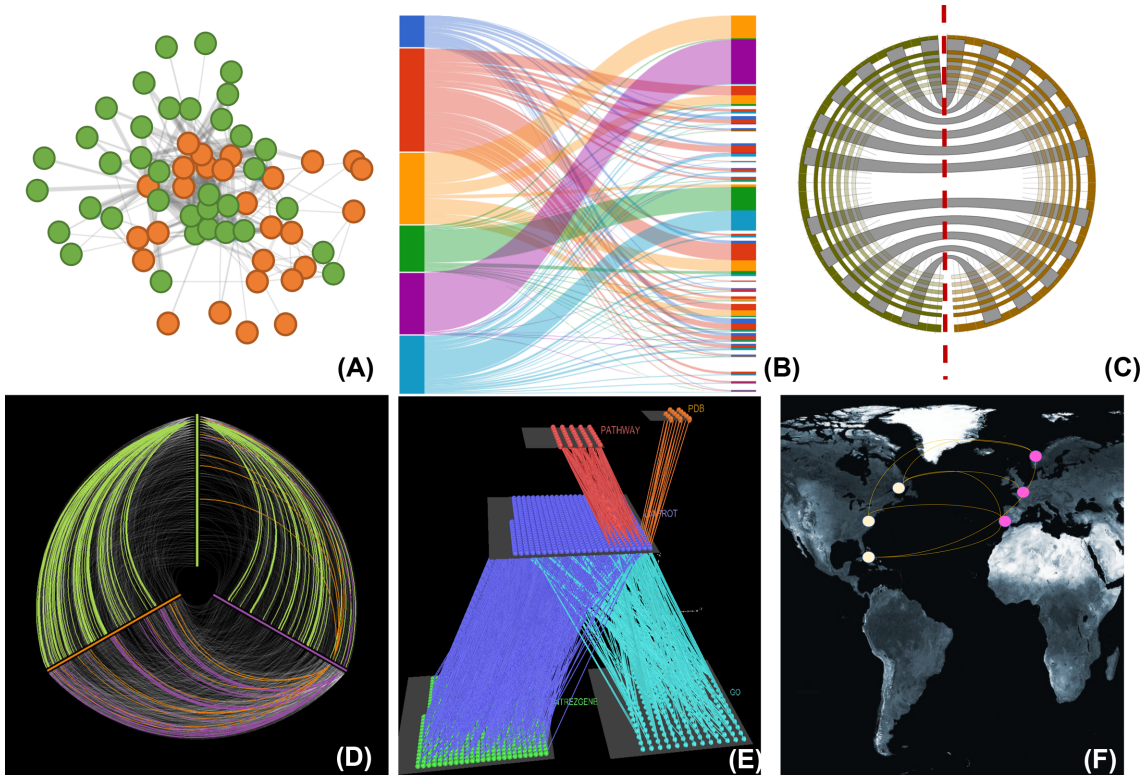


Figure 10: Various types of visualizations of *n*-partite networks. (A) Visualization using a generic network tool such as Cytoscape. Nodes from each group are colored accordingly. (B) Vertical bipartite visualization. (C) Circular visualization using a Circos-like view often used in genomics. (D) A hive plot view visualizing a tripartite graph. (E) Visualization of a multilayered network using Arena3D. (F) Visualization of a bipartite network over a world map.

information while emphasizing the connections between them. Among other functionalities (i.e., great variety of clustering algorithms), Arena3D can be utilized to visualize intra- and internetwork connections, show gene expressions levels, and handle time course data in a phenotypic context. Arena3D's concept can be easily adjusted to visualize bipartite graphs as vertices of the disjoint sets U and V of a bipartite graph $G = (U, V, E)$ that can be separated onto different layers and colored accordingly. Connections across the different layers can easily be loaded and visualized simultaneously. While nodes can be placed anywhere manually, clustering across layers can place the vertices of each layer in a way that crossovers between lines can be minimized. Although Arena3D might be too advanced for the visualization of simple bipartite graphs, it is highly recommended for *n*-partite graphs, where *n* layers can be placed anywhere and in various orientations in 3D space, thereby offering very sophisticated visualizations. An example is shown in Figure 10E.

The *Biclustering Analysis Toolbox* (BicAT) [253] is a software platform for the analysis of gene interconnection networks, as well other types of data (e.g., proteomics data), based on biclustering techniques in a single graphical interface. Furthermore, BicAT offers a variety of facilities (e.g., filtering of biclusters) for data preparation, review, processing, and post analysis. The user is able to choose the optimal/their preferred biclustering algorithm among different algorithms. The program allows the users to install further extensions or algorithms.

GeneWeaver [254] is an online software package for the integration of functional genomics experiments. It contains a set of interactive tools for analysis and visualization of gene sets, gene set descriptions, and gene set association scores from multiple species. It differs from conventional gene set overrepresentation analysis tools in that it allows users to evaluate intersections among all combinations of a collection of gene sets, including, but not limited to, annotations to controlled vocabularies. Gene sets can come from many different sources (e.g., microarray experiments, gene ontology annotations, text mining tools, list of specific genes).

ONEMODE [255] is a Stata module capable of producing 1-mode projections of a bipartite network. This package offers the most complete collection of algorithms for projection, such as methods for unconditional (global) threshold, methods with thresholds conditioned on the U-nodes' degree, methods for controlling U-nodes' differing numbers of interacting V-nodes, the FDSM, and the SDSM.

Circos [256] is a tool widely used in comparative genomics to visualize structural variations and direct comparisons between genomes. It uses a circular ideogram layout to facilitate the display of relationships between pairs of positions by the use of ribbons, which encode the position, size, and orientation of related genomic elements. A potential use of Circos in terms of bipartite network visualization is shown in Figure 10C.

Hiveplots [257] is a rational method for drawing and visualizing networks. Nodes are mapped to and positioned on radially distributed linear axes. While the purpose of the hive plot is to establish a new baseline for visualization of large networks, we believe that it is a very suitable tool for visualizing large-scale *n*-partite, especially tripartite graphs. An example of the application of Hiveplots is shown in Figure 10D.

Other generic visualization tools that could potentially be adjusted to efficiently visualize bipartite graphs are the 2D standalone applications such as graphVizdb [258], Ondex [259], Proviz [260], VizANT [261], GUESS [262], UCINET [263], MAPMAN

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

[264], PATIKA [265], Medusa [266], and Osprey [267], as well as 3D visualization tools such as BioLayout Express [268].

R packages

R is a software environment and a programming language for statistical analysis supported by the R Foundation for Statistical Computing. The R language is widely used among researchers for developing statistical software and data analysis. R is freely available under the GNU General Public License. R contains several packages that can handle bipartite networks. Some of them are oriented toward the analysis of ecological networks (e.g., Networksis, enaR, Bipartite), whereas other tools were designed for more general network analyses.

Networksis [269] is a package for R built for the analysis of ecological networks, as well as the generation of seed graphs for Markov chain Monte Carlo simulations. The tool provides several methods and many options to visualize and analyze bipartite networks. It offers the option to calculate a series of indices summarizing the bipartite network topology. Finally, given that the ability to simulate graphs with given properties is important for the analysis of networks, the package can be used to compare results to null models. Networksis uses sequential importance sampling that has been shown to be particularly effective in estimating the number of graphs adhering to fixed marginals and in estimating the null distribution of graph statistics.

enaR [270] is an R package for Ecosystem Network Analysis (ENA). It is a suite of analytical tools for studying the structure and dynamics of energy and matter fluxes through distinct ecological compartments.

BipartiteR [271] is an R package containing utilities to visualize bipartite networks and compute a set of indices that are often used to describe different aspects of FWs, e.g., pollination webs or predator–prey webs.

Netpredictor [272] is an R package (available also as an R Shiny web application) designed for the prediction of missing links in bipartite networks. The package provides a set of tools for calculating missing links in both bipartite and unipartite networks. Also, Netpredictor allows computation of several bipartite network properties, calculation of significant interactions between 2 sets of nodes using permutation-based testing, and visualization of communities for 2 different sets of nodes.

biGRAPH [206] is an R package extension to the well-known igraph package (which is the method of choice for handling unipartite graphs) that provides a set of methods specifically designed for the analysis of bipartite graphs, including the projection of bipartite graphs handling the problem of information loss. In addition, clustering and community detection among vertex subsets is supported by providing metric distance calculations based on flexible (weighted) neighborhoods. The latest version of the software package contains some of the metrics for bipartite graphs proposed by Borgatti and Everett [34], including measures for density, vertex centrality, and centralization with respect to each vertex subset.

tnet [273] is an R package that, among others things, can handle the analysis of bipartite networks. Although this tool contains several projection methods, it is optimally designed to handle bipartite weighted networks.

BiRewire [274] is an R package in Bioconductor that implements the switching algorithm for the randomization of bipartite graphs retaining their node degrees (i.e., network rewiring). BiRewire can be also used for the randomization of general presence (1)-absence (0) matrices, where the presence distributions must be preserved. Specifically, BiRewire enables users to generate bipartite graphs from any "0–1" matrix, as well as rewired versions of these graphs.

DEsubs [275] is an R package designed to extract differentially expressed, disease-associated subpathways from a pathway network generated from RNA-seq experiments. It comes with advanced visualization and enrichment analysis with regard to various biological and pharmacological features. Its circular representation could be potentially useful for the visualization of bipartite networks.

Dataset collections

Last, we present some repositories (databases) that hold numerous biological network datasets, including bipartite ones. Even though some of the datasets mentioned earlier are highly curated and biologically important, here we restrict our attention to collections of datasets and thus we do not list specific datasets. Some of these databases contain various datasets, even of nonbiological origin (such as the Stanford Large Network Dataset [SLND], Colorado Index of Complex Networks [ICON], and Koblenz Network Collection [KONECT]), whereas there are several databases specialized for ecological networks, highlighting the importance of such data in current network research.

SLND accompanies the SNAP library [276], which has been actively developed since 2004 and is organically growing as a result of the Leskovec group's research in analysis of large social and information networks. The datasets available on the website were mostly collected for the research performed by the team, and the website has been active since 2009. [277]. ICON [278] is a comprehensive index of network datasets from all domains of network science, including social, web, biological, ecological, transportation, and technological networks. Each network record is annotated with its graph properties, description, size, and similar information, and many records include links to multiple networks. The contents of ICON are curated by volunteer experts from Professor Aaron Clauset's research group at the University of Colorado–Boulder. KONECT [279] is a project to collect large network datasets of all types in order to perform research in network science, collected by the Institute of Web Science and Technologies at the University of Koblenz–Landau. KONECT contains several hundred network datasets of various types, including directed, undirected, bipartite, weighted, unweighted, signed, and rating networks. The networks of KONECT cover many diverse areas such as social networks, hyperlink networks, authorship networks, physical networks, interaction networks, and communication networks [280].

The ecological databases include the Web of Life, the Interaction Web Database, and the Kelpforest Database. The *Web of Life* [281] provides a graphical user interface, based on Google Maps, for visualization and download of data on ecological networks regarding species interactions. It is designed and implemented in a relational database, allowing sophisticated and user-friendly searches. Data can be downloaded in several common formats, and a web-service for data transmission in JavaScript Object Notation is also provided. The *Interaction Web Database* [282] contains datasets on species interactions from several communities in different parts of the world. Data currently available cover a variety of interaction types, including plant–pollinator, plant–frugivore, plant–herbivore, plant–ant mutualist, and predator–prey interactions. The developers' goal is to expand the database to make it a repository of data on any kind of interactions. The *Kelpforest Database* [283] serves as a repository for the knowledge of identities, life histories, and interactions between the species present in the near shore kelp

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

forest ecosystems of the eastern Pacific Ocean, focusing on central and southern California. The information that it contains could aid in the interpretation of species' spatial and temporal patterns and serve as the basis on which to construct and parameterize mathematical models of these species' rich communities [284].

Conclusions

Network-based approaches have been used routinely during the last decade to analyze the massive amount of biological/biomedical data produced from modern high-throughput experiments. Bipartite networks constitute an important but usually overlooked and difficult-to-analyze class of networks. However, given that natively bipartite structures have many applications in systems biology and medicine, there is an emerging need for specialized methods and software for analyzing such networks. Based on a review of the literature, ecological networks, which are traditionally constructed by collecting large samples of individuals from the field, are usually analyzed as bipartite networks using the native structure. In addition, research on ecological networks has produced many network metrics designed for bipartite graphs. Several studies have introduced new indices to describe network properties, and consequently dozens of indices are currently available to address similar questions [35].

On the other hand, biomedical networks are usually analyzed through projection and analysis of the projected unipartite networks. This is of no surprise since most of the times the biomedical networks connect abstract entities, such as "diseases," "genes," or "symptoms," and, in most cases, the primary goal of the analysis is the direct interactions between members of the same group. Nevertheless, projection of a bipartite network into its unipartite counterparts results in loss of information. Another issue that needs to be investigated is whether and to what extend the different methods of projection proposed in the literature affect the overall results of such analysis. Additionally, it could be particularly useful to determine if any of the natively bipartite methods or metrics that have been developed for ecological analysis (e.g., nestedness, modularity, community detection, flow) can also be applied in the case of molecular or biomedical networks, such as the diseasome.

Convergence of ecology and bioinformatics is expected in the near future. Such convergence has been achieved in the past, with the most prominent applications in phylogenetics, which is considered a vital part of bioinformatics, in microbial ecology, and in metagenomics [285], as well as in other areas of ecology [286]. Of note, in the past decade, molecular methods (e.g., sequencing, metagenomics, barcoding) were used extensively in studies of HPWs to clarify species concepts [287]. Therefore, network science constitutes an interdisciplinary field, where ecologists and molecular biologists are brought together [288]. We have already noted that indices applied to ecological networks could have potential application in the analysis of biomedical and molecular networks as well. In the opposite direction, methods for identifying modules in ecological networks have stimulated much interest. In addition, several robust module-detecting algorithms that have been applied in other disciplines have also been applied in large pollination networks, showing that these networks were modular and that modularity co-occurred with nestedness [85]. In a similar manner, the large arsenal of biclustering methods described in the pertinent machine learning literature can be applied in the study of ecological and other biological networks.

Last, it is worth mentioning that, in several cases, at least in the context of biomedical networks, researchers try to compile tripartite networks in order to model the complex interactions associated with diseases [108,117]. This is of no surprise since most diseases are multifactorial and affected by various genetic, environmental, and lifestyle factors. Thus, due to data accumulation, additional knowledge is expected to be integrated into gene–disease networks. Taking into account the above information, future studies, at least those on biomedical networks, should focus on the development of analytical methods and software tools capable of handling tripartite and multipartite graphs that would enable the simultaneous analysis of information from multiple sources. For instance, instead of the bipartite gene–disease network, it might be more useful to perform network data analysis without projections and analyze, e.g., a multipartite graph that illustrates exposure–gene–symptoms–disease relationships. A potential way of representing such systems would be to extend the network into multiple layers (in a multipartite graph) or to use a generalization of graphs known as hypergraphs. In a simple graph, a link connects only a pair of nodes, whereas the edges of the hypergraph (hyperedges) can connect groups of more than 2 nodes. Toward this end, analytical methods have been developed in order to extend the application of clustering coefficient and subgraph centrality to complex hypernetworks [34].

Availability of supporting data

The original data for the Analysis of the Gene–Disease Network can be found in the publications and the supplements of Kontou et al. [15,204]. The data for generating the example networks of Figures 1–4 can be found in the Supplementary Material.

Abbreviations

DEG: differentially expressed gene; eQTL: expression quantitative traits loci; FDSM: fixed degree sequence model; FW: food web; GAD: Genetic Association Database; GO: Gene Ontology; GRN: GWAS: Genome-Wide Association Studies; HPW: host–parasitoid web; ICA: independent component analysis; ICON: Colorado Index of Complex Networks; KONECT: Koblenz Network Collection; lncRNA: long noncoding RNA; MW: mutualistic web; NAP: Network Analysis Profiler; NCA: network component analysis; NCBI: National Center for Biotechnology Information; NHGRI: National Human Genome Research Institute; NIH: National Institutes of Health; *NODF*: nestedness metric based on overlap and decreasing fill; NP: Nondeterministic Polynomial time; OMIM: Online Mendelian Inheritance in Man; PCA: principal component analysis; PDN: phenotypic disease network; PPI: protein–protein interaction; RG: regulated gene; SDSM: stochastic degree sequence model; SLND: Stanford Large Network Dataset; SNAP: Stanford Network Analysis Platform; SNP: single nucleotide polymorphism; STD: sexually transmitted disease; SVD: singular value decomposition; TAP-MS: tandem affinity-purification/mass spectrometry; TF: Transcription Factor

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the US Department of Energy (DOE) Joint Genome Institute, a DOE Office of Science User Facility, under contract DE-AC02-05CH11231, and used resources of

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

the National Energy Research Scientific Computing Center, supported by the Office of Science of the DOE.

Author contributions

P.G.B. and G.P. conceived the project and organized the work. All authors wrote parts of the manuscript, and all authors have read and approved the final manuscript.

Acknowledgements

The authors thank the associate editor and the reviewers whose comments and constructive criticism helped in improving the quality of the manuscript. All images of Figure 5A were taken from the Pixabay repository and are subject to CC0 Creative Commons license. According to CC0, users can freely copy, modify, and redistribute the images, even for commercial purposes.

References

1. Fields S, Song O. A novel genetic system to detect protein-protein interactions. Nature 1989;**340**(6230):245–6.
2. Zhu H, Bilgin M, Bangham R et al. Global analysis of protein activities using proteome chips. Science 2001;**293**(5537):2101–5.
3. Ito T, Chiba T, Ozawa R et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences 2001;**98**(8):4569–74.
4. Uetz P, Giot L, Cagney G et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 2000;**403**(6770):623–7.
5. McCraith S, Holtzman T, Moss B et al. Genome-wide analysis of vaccinia virus protein-protein interactions. Proceedings of the National Academy of Sciences 2000;**97**(9):4879–84.
6. Davy A, Bello P, Thierry-Mieg N et al. A protein-protein interaction map of the *Caenorhabditis elegans* 26S proteasome. EMBO Reports 2001;**2**(9):821–8.
7. Rain JC, Selig L, De Reuse H et al. The protein-protein interaction map of *Helicobacter pylori*. Nature 2001;**409**(6817):211–5.
8. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;**10**(1):57–63.
9. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;**17**(6):333–51.
10. Pavlopoulos GA, Oulas A, Iacucci E et al. Unraveling genomic variation from next generation sequencing data. BioData Mining 2013;**6**(1):13.
11. Chen IA, Markowitz VM, Chu K et al. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res 2017;**45**(D1):D507–16.
12. Pavlopoulos GA, Secrier M, Moschopoulos CN et al. Using graph theory to analyze biological networks. BioData Mining 2011;**4**(1):10.
13. Yu D, Kim M, Xiao G et al. Review of biological network data and its applications. Genomics Inform 2013;**11**(4):200–10.
14. Burgos E, Ceva H, Hernandez L et al. Two classes of bipartite networks: nested biological and social systems. Phys Rev E 2008;**78**(4):046113.
15. Kontou PI, Pavlopoulou A, Dimou NL et al. Network analysis of genes and their association with diseases. Gene 2016;**590**(1):68–78.
16. Davis A, Gardner BB, Gardner MR. Deep South: A Social Anthropological Study of Caste and Class. University of South Carolina Press. 2009.
17. Watts DJ, Strogatz SH. Collective dynamics of "small-world" networks. Nature 1998;**393**(6684):440–2.
18. Newman ME. Scientific collaboration networks.??I. Network construction and fundamental results. Phys Rev E 2001;**64**(1):016131.
19. Van Noorden R. Online collaboration: scientists and the social network. Nature 2014;**512**(7513):126–9.
20. Newman ME. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 2001;**98**(2):404–9.
21. Conyon MJ, Muldoon MR. The small world of corporate boards. J Bus Fin & Acc 2006;**33**(9-10):1321–43.
22. Ramasco JJ, Dorogovtsev SN, Pastor-Satorras R. Self-organization of collaboration networks. Phys Rev E 2004;**70**(3):036106.
23. Guillaume J-L, Latapy M, Le Blond S. Statistical analysis of a P2P query graph based on degrees and their time-evolution. In: IWDC: 2004. Springer: 126–37.
24. Taylor PJ. The new geography of global civil society: NGOs in the world city network. Globalizations 2004;**1**(2):265–77.
25. Doreian P, Batagelj V, Ferligoj A. Generalized block modeling of two-mode network data. Social Networks 2004;**26**(1):29–53.
26. Fowler JH. Legislative cosponsorship networks in the US House and Senate. Social Networks 2006;**28**(4):454–65.
27. Deng H, Lyu MR, King I. A generalized co-hits algorithm and its application to bipartite graphs. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2009, 239–48.
28. Yanai K. VisualTextualRank: an extension of visualrank to large-scale video shot extraction exploiting tag co-occurrence. IEICE Transactions on Information and Systems 2015;**98**(1):166–72.
29. Cao L, Guo J, Cheng X. Bipartite graph based entity ranking for related entity finding. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on *Web Intelligence and Intelligent Agent Technology - Volume 01*. IEEE. 2011, 130–7.
30. Li X, Chen H. Recommendation as link prediction in bipartite graphs: a graph kernel-based machine learning approach. Decision Support Systems 2013;**54**(2):880–90.
31. Xu K, Wang F, Gu L. Behavior analysis of internet traffic via bipartite graphs and one-mode projections. IEEE/ACM Trans Networking 2014;**22**(3):931–42.
32. Cai J, Liu WX. A new method of detecting network traffic anomalies. In: Applied Mechanics and Materials. Trans Tech Publ, 2013, 912–6.
33. Latapy M, Magnien C, Del Vecchio N. Basic notions for the analysis of large two-mode networks. Social Networks 2008;**30**(1):31–48.
34. Borgatti SP, Everett MG. Network analysis of 2-mode data. Social Networks 1997;**19**(3):243–69.
35. Dormann CF, Fründ J, Blüthgen N et al. Indices, graphs and null models: analyzing bipartite ecological networks. The Open Ecology Journal. 2009;**2**:1.
36. Theodosiou T, Efstathiou G, Papanikolaou N et al. NAP: the network analysis profiler, a web tool for easier topological analysis and comparison of medium-scale biological networks. BMC Res Notes 2017;**10**(1):278.
37. Doncheva NT, Assenov Y, Domingues FS et al. Topological analysis and interactive visualization of biological

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

networks and protein structures. Nat Protoc 2012;**7**(4): 670–85.
38. Leskovec J, Sosic R. SNAP. ACM Trans Intell Syst Technol 2016;**8**(1).
39. igraph; http://igraph.org/c.
40. Daugulis P. A note on a generalization of eigenvector centrality for bipartite graphs and applications. Networks 2012;**59**(2):261–4.
41. Wasserman S, Faust K. Social Network Analysis: Methods and Applications, vol. 8., Cambridge University Press; 1994.
42. Lind PG, González MC, Herrmann HJ. Cycles and clustering in bipartite networks. Phy. Rev E 2005;**72**(5):056127.
43. Opsahl T. Triadic closure in two-mode networks: redefining the global and local clustering coefficients. Social Networks 2013;**35**(2):159–67.
44. Robins G, Alexander M. Small worlds among interlocking directors: network structure and distance in bipartite graphs. Computational & Mathematical Organization Theory 2004;**10**(1):69–94.
45. Zhang P, Wang J, Li X et al. Clustering coefficient and community structure of bipartite networks. Physica A: Statistical Mechanics and Its Applications 2008;**387**(27): 6869–75.
46. Staniczenko PP, Kopp JC, Allesina S. The ghost of nestedness in ecological networks. Nat Comms 2013;**4**:1391.
47. Ulrich W, Almeida-Neto M, Gotelli NJ. A consumer's guide to nestedness analysis. Oikos 2009;**118**(1):3–17.
48. Atmar W, Patterson BD. The measure of order and disorder in the distribution of species in fragmented habitat. Oecologia 1993;**96**(3):373–82.
49. Araujo AI, Corso G, Almeida AM et al. An analytic approach to the measurement of nestedness in bipartite networks. Physica A: Statistical Mechanics and Its Applications 2010;**389**(7):1405–11.
50. Guimarães PR, Guimaraes P. Improving the analyses of nestedness for large sets of matrices. Environmental Modelling & Software 2006;**21**(10):1512–3.
51. Wright DH, Reeves JH. On the meaning and measurement of nestedness of species assemblages. Oecologia 1992;**92**(3):416–28.
52. Bustos S, Gomez C, Hausmann R et al. The dynamics of nestedness predicts the evolution of industrial ecosystems. PLoS One 2012;**7**(11):e49393.
53. Almeida-Neto M, Ulrich W. A straightforward computational approach for measuring nestedness using quantitative matrices. Environmental Modelling & Software 2011;**26**(2):173–8.
54. Galeano J, Pastor JM, Iriondo JM. Weighted-interaction nestedness estimator (WINE): a new estimator to calculate over frequency matrices. Environmental Modelling & Software 2009;**24**(11):1342–6.
55. Newman ME, Girvan M. Finding and evaluating community structure in networks. Phys Rev E 2004;**69**(2):026113.
56. Fortuna MA, Stouffer DB, Olesen JM et al. Nestedness versus modularity in ecological networks: two sides of the same coin? Journal of Animal Ecology 2010;**79**(4):811–7.
57. Allali O, Tabourier L, Magnien C et al. Internal links and pairs as a new tool for the analysis of bipartite complex networks. Soc Netw Anal Min 2013;**3**(1):85–91.
58. Holme P, Liljeros F, Edling CR et al. Network bipartivity. Phys Rev E 2003;**68**(5):056107.
59. Estrada E, Rodríguez-Velázquez JA. Spectral measures of bipartivity in complex networks. Phys Rev E 2005;**72**(4):046105.
60. Pisanski T, Randić M. Use of the szeged index and the revised szeged index for measuring network bipartivity. Discrete Applied Mathematics 2010;**158**(17):1936–44.
61. Blüthgen N, Fründ J, Vázquez DP et al. What do interaction network metrics tell us about specialization and biological traits. Ecology 2008;**89**(12):3387–99.
62. Mukherjee A, Choudhury M, Ganguly N. Understanding how both the partitions of a bipartite network affect its one-mode projection. Physica A: Statistical Mechanics and Its Applications 2011;**390**(20):3602–7.
63. Guillaume J-L, Latapy M. Bipartite graphs as models of complex networks. Physica A: Statistical Mechanics and Its Applications 2006;**371**(2):795–813.
64. Nacher J, Akutsu T. On the degree distribution of projected networks mapped from bipartite networks. Physica A: Statistical Mechanics and Its Applications 2011;**390**(23-24):4636–51.
65. Li M, Fan Y, Chen J et al. Weighted networks of scientific communication: the measurement and topological role of weight. Physica A: Statistical Mechanics and Its Applications 2005;**350**(2-4):643–56.
66. Neal Z. Identifying statistically significant edges in one-mode projections. Soc Netw Anal Min 2013;**3**(4): 915–24.
67. Neal Z. The backbone of bipartite projections: inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. Social Networks 2014;**39**:84–97.
68. Newman ME. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys Rev E 2001;**64**(1):016132.
69. Zhou T, Ren J, Medo M et al. Bipartite network projection and personal recommendation. Phys Rev E 2007;**76**(4):046115.
70. Zweig KA, Kaufmann M. A systematic approach to the one-mode projection of bipartite graphs. Soc Netw Anal Min 2011;**1**(3):187–218.
71. Puebla JG. Spatial structures of network flows: a graph theoretical approach. Transportation Research Part B: Methodological 1987;**21**(6):489–502.
72. Serrano MÁ, Boguná M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. Proceedings of the National Academy of Sciences 2009;**106**(16): 6483–8.
73. Bonacich P. Technique for analyzing overlapping memberships. Sociological Methodology 1972;**4**:176–85.
74. Borgatti SP, Halgin DS. Analyzing affiliation networks. The SAGE Handbook of Social Network Analysis. SAGE 2011:417–33.
75. Scott J. Social network analysis: developments, advances, and prospects. SOCNET 2011;**1**(1):21–26.
76. Tumminello M, Miccichè S, Lillo F et al. Statistically validated networks in bipartite complex systems. PLoS One 2011;**6**(3):e17994.
77. Klecka J. The role of a water bug, *Sigara striata*, in freshwater food webs. Peer J 2014;**2**:e389.
78. Dunne JA, Williams RJ, Martinez ND. Food-web structure and network theory: the role of connectance and size. Proceedings of the National Academy of Sciences 2002;**99**(20):12917–22.
79. Kondoh M, Kato S, Sakato Y. Food webs are built up with nested subwebs. Ecology 2010;**91**(11):3123–30.
80. Hassell M, Waage J. Host-parasitoid population interactions. Annu Rev Entomol 1984;**29**(1):89–114.

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

81. Henri DC, Van Veen F. Body size, life history and the structure of host-parasitoid networks. Advances in Ecological Research 2011;**45**:135–80.
82. Bascompte J, Jordano P. Plant-animal mutualistic networks: the architecture of biodiversity. Annu Rev Ecol Evol Syst 2007;(1):567–93.
83. Dáttilo W, Guimarães PR, Izzo TJ. Spatial structure of ant-plant mutualistic networks. Oikos 2013;**122**(11):1643–8.
84. Bascompte J, Jordano P, Melián CJ et al. The nested assembly of plant-animal mutualistic networks. Proceedings of the National Academy of Sciences 2003;**100**(16): 9383–7.
85. Olesen JM, Bascompte J, Dupont YL et al. The modularity of pollination networks. Proceedings of the National Academy of Sciences 2007;**104**(50):19891–6.
86. James A, Pitchford JW, Plank MJ. Disentangling nestedness from models of ecological complexity. Nature 2012;**487**(7406):227–30.
87. Montoya JM, Solé RV. Small world patterns in food webs. Journal of Theoretical Biology 2002;**214**(3):405–12.
88. Jordano P, Bascompte J, Olesen JM. Invariant properties in coevolutionary networks of plant-animal interactions. Ecology Letters 2003;**6**(1):69–81.
89. Vázquez DP. Degree distribution in plant-animal mutualistic networks: forbidden links or random interactions? Oikos 2005;**108**(2):421–6.
90. Williams RJ. Biology, methodology or chance? The degree distributions of bipartite ecological networks. PLoS One 2011;**6**(3):e17645.
91. Corel E, Lopez P, Méheust R et al. Network-thinking: graphs to analyze microbial complexity and evolution. Trends in Microbiology 2016;**24**(3):224–37.
92. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011;**12**(1):56–68.
93. Goh K-I, Choi I-G. Exploring the human diseasome: the human disease network. Briefings in Functional Genomics 2012:els032.
94. Goh KI, Cusick ME, Valle D et al. The human disease network. Proceedings of the National Academy of Sciences 2007;**1**(21):8685–90.
95. Amberger JS, Bocchini CA, Schiettecatte F et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research 2015;**43**(D1):D789–98.
96. Barrenas F, Chavali S, Holme P et al. Network properties of complex human disease genes identified through genome-wide association studies. PLoS One 2009; **4**(11):e8090.
97. Liu CC, Tseng YT, Li W et al. DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. Nucleic Acids Research 2014;**42**(W1):W137–46.
98. Darabos C, Harmon SH, Moore JH. Using the bipartite human phenotype network to reveal pleiotropy and epistasis beyond the gene. Pacific Symposium on Biocomputing 2014:188–99.
99. Becker KG, Barnes KC, Bright TJ et al. The genetic association database. Nat Genet 2004;**36**(5):431–2.
100. Welter D, MacArthur J, Morales J et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014;**42**(D1):D1001–6.
101. Gray KA, Yates B, Seal RL et al. Genenames.org: the HGNC resources in 2015. Nucleic Acids Research 2015;**43**(D1):D1079–85.
102. Pawson T, Linding R. Network medicine. FEBS Letters 2008;**582**(8):1266–70.
103. Hartwell LH, Hopfield JJ, Leibler S et al. From molecular to modular cell biology. Nature 1999;**402**(6761supp):C47–52.
104. Oti M, Brunner HG. The modular nature of genetic diseases. Clinical Genetics 2007;**71**(1):1–11.
105. Baranzini SE. The genetics of autoimmune diseases: a networked perspective. Current Opinion in Immunology 2009;**21**(6):596–605.
106. Ahmed SS, Ahameethunisa AR, Santosh W et al. Systems biological approach on neurological disorders: a novel molecular connectivity to aging and psychiatric diseases. BMC Syst Biol 2011;**5**(1):6.
107. Chan SY, White K, Loscalzo J. Deciphering the molecular basis of human cardiovascular disease through network biology. Current Opinion in Cardiology 2012;**27**(3):202–9.
108. Hayasaka S, Hugenschmidt CE, Laurienti PJ. A network of genes, genetic disorders, and brain areas. PLoS One 2011;**6**(6):e20907.
109. Kitsak M, Sharma A, Menche J et al. Tissue specificity of human disease module. Sci Rep 2016;**6**(1):35241.
110. Zhou X, Menche J, Barabási A-L et al. Human symptoms-disease network. Nat Comms 2014;**5**:4212.
111. Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. Epidemiologic Reviews 1998;**20**(2):137–47.
112. Darabos C, Grussing ED, Cricco ME et al. A bipartite network approach to inferring interactions between environmental exposures and human diseases. Pacific Symposium on Biocomputing 2015:171–82.
113. Yildirim MA, Goh KI, Cusick ME et al. Drug-target network. Nat Biotechnol 2007;**25**(10):1119–26.
114. Cheng F, Liu C, Jiang J et al. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol 2012;**8**(5):e1002503.
115. Zhang Y, Tao C, He Y et al. Network-based analysis of vaccine-related associations reveals consistent knowledge with the vaccine ontology. J Biomed Sem 2013;**4**(1):33.
116. Campillos M, Kuhn M, Gavin A-C et al. Drug target identification using side-effect similarity. Science 2008;**321**(5886):263–6.
117. Chavan V, Penev L. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics 2011;**12**(Suppl 15):S2.
118. Cai B, Wang H, Zheng H et al. Integrating domain similarity to improve protein complexes identification in TAP-MS data. Proteome Sci 2013;**11**(Suppl 1):S2.
119. Wu M, Li XL, Kwoh CK et al. Discovery of protein complexes with core-attachment structures from tandem affinity purification (TAP) data. Journal of Computational Biology 2012;**19**(9):1027–42.
120. Li Y, Thai MT, Liu Z et al. Protein-protein interaction and group testing in bipartite graphs. IJBRA 2005;**1**(4): 414–9.
121. Ma B. Challenges in computational analysis of mass spectrometry data for proteomics. J Comput Sci Technol 2010;**25**(1):107–23.
122. Junqueira M, Spirin V, Santana Balbuena T et al. Separating the wheat from the chaff: unbiased filtering of background tandem mass spectra improves protein identification. J Proteome Res 2008;**7**(8):3382–95.
123. Nesvizhskii AI. Protein identification by tandem mass spectrometry and sequence database searching. In: Mass

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

Spectrometry Data Analysis in Proteomics. Matthiesen R, ed., Totowa, NJ: Humana Press; 2007:87–119.
124. Rao VS, Srinivas K, Sujini GN et al. Protein-protein interaction detection: methods and analysis. International Journal of Proteomics 2014;**2014**:1–12.
125. Lee J, Lee J. Hidden information revealed by optimal community structure from a protein-complex bipartite network improves protein function prediction. PLoS One 2013;**8**(4):e60372.
126. Royer L, Reimann M, Andreopoulos B et al. Unraveling protein networks with power graph analysis. PLoS Comput Biol 2008;**4**(7):e1000108.
127. Brynildsen MP, Tran LM, Liao JC. Versatility and connectivity efficiency of bipartite transcription networks. Biophysical Journal 2006;**91**(8):2749–59.
128. Liao JC, Boscolo R, Yang Y-L et al. Network component analysis: reconstruction of regulatory signals in biological systems. Proceedings of the National Academy of Sciences 2003;**100**(26):15522–7.
129. Wang X, Alshawaqfeh M, Dang X et al. An overview of NCA-based algorithms for transcriptional regulatory network inference. Microarrays 2015;**4**(4).
130. Ye C, Galbraith SJ, Liao JC, Eskin E Using Network Component Analysis to Dissect Regulatory Networks Mediated by Transcription Factors in Yeast. PLOS Computational Biology 2009;**5**(3):e1000311.
131. Platig J, Castaldi PJ, DeMeo D et al. Bipartite community structure of eQTLs. PLoS Comput Biol 2016;**12**(9):e1005033.
132. Guzmán-Vargas L, Santillán M. Comparative analysis of the complex transcription-factor gene regulatory networks of *E. coli* and *S. cerevisiae*. BMC Syst Biol 2008;**2**(1):13.
133. Horvát E-Á, Zhang JD, Uhlmann S et al. A network-based method to assess the statistical significance of mild co-regulation effects. PLoS One 2013;**8**(9):e73413.
134. Chen L, Wang H, Zhang L et al. Uncovering packaging features of co-regulated modules based on human protein interaction and transcriptional regulatory networks. BMC Bioinformatics 2010;**11**(1):392.
135. Ge M, Li A, Wang M. A bipartite network-based method for prediction of long non-coding RNA-protein interactions. Genomics, Proteomics & Bioinformatics 2016;**14**(1): 62–71.
136. Bryant WA, Sternberg MJ, Pinney JW. AMBIENT: Active Modules for Bipartite Networks—using high-throughput transcriptomic data to dissect metabolic response. BMC Syst Biol 2013;**7**(1):26.
137. He L, Wang Y, Yang Y et al. Identifying the gene signatures from gene-pathway bipartite network guarantees the robust model performance on predicting the cancer prognosis. BioMed Research International 2014;**2014**:1–10.
138. Ahnert SE. Generalised power graph compression reveals dominant relationship patterns in complex networks. Sci Rep 2015;**4**(1):4385.
139. Beyer A, Workman C, Hollunder J et al. Integrated assessment and prediction of transcription factor binding. PLoS Comp Biol 2006;**2**(6):e70.
140. Luke DA, Harris JK. Network analysis in public health: history, methods, and applications. Annu Rev Public Health 2007;**28**(1):69–93.
141. Ergün G. Human sexual contact network as a bipartite graph. Physica A: Statistical Mechanics and Its Applications 2002;**308**(1-4):483–8.
142. Palchykov V, Kaski K, Kertész J et al. Sex differences in intimate relationships. Sci Rep 2012;**2**(1):370.
143. Robinson K, Cohen T, Colijn C. The dynamics of sexual contact networks: effects on disease spread and control. Theoretical Population Biology 2012;**81**(2):89–96.
144. Rocha LEC, Liljeros F, Holme P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. PLoS Comput Biol 2011;**7**(3):e1001109.
145. Niekamp A-M, Mercken LA, Hoebe CJ et al. A sexual affiliation network of swingers, heterosexuals practicing risk behaviours that potentiate the spread of sexually transmitted infections: a two-mode approach. Social Networks 2013;**35**(2):223–36.
146. Hernández DG, Risau-Gusman S. Epidemic thresholds for bipartite networks. Phys Rev E 2013;**88**(5):052801.
147. Bisanzio D, Bertolotti L, Tomassone L et al. Modeling the spread of vector-borne diseases on bipartite networks. PLoS One 2010;**5**(11):e13796.
148. Malik HAM, Mahesar AW, Abid F et al. Two-mode network modeling and analysis of dengue epidemic behavior in Gombak, Malaysia. Applied Mathematical Modelling 2017;**43**:207–20.
149. Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. Nat Rev Genet 2016;**17**(10):615–29.
150. Hidalgo CA, Blumm N, Barabási A-L et al. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol 2009;**5**(4):e1000353.
151. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;**13**(6):395–405.
152. Hsing AW, Ioannidis JP. Nationwide population science. JAMA Intern Med 2015;**175**(9):1527–9.
153. HuDiNe Dataset. http://hudine.neu.edu. Accessed 15 Jan 2018.
154. Bhavnani SK, Dang B, Visweswaran S et al. How comorbidities co-occur in readmitted hip fracture patients: from bipartite networks to insights for post-discharge planning. AMIA Joint Summits on Translational Science Proceedings 2015;**2015**:36–40.
155. Yannakakis M. Node-and edge-deletion NP-complete problems. In: Proceedings of the Tenth Annual ACM Symposium on Theory of Computing; San Diego, California, USA. 804355: ACM, 1978:253–64.
156. Reed B, Smith K, Vetta A. Finding odd cycle transversals. Operations Research Letters 2004;**32**(4):299–301.
157. Hüffner F. Algorithm engineering for optimal graph bipartization. In: Experimental and Efficient Algorithms: 4th International Workshop, WEA 2005; Santorini Island, Greece, May 10–13, 2005 Proceedings. Nikoletseas SE, ed. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005:240–52.
158. Guo J, Gramm J, Hüffner F et al. Compression-based fixed-parameter algorithms for feedback vertex set and edge bipartization. Journal of Computer and System Sciences 2006;**72**(8):1386–96.
159. Hopcroft JE, Karp RM. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. SIAM J Comput 1973;**2**(4):225–31.
160. Fredman ML, Tarjan RE. Fibonacci heaps and their uses in improved network optimization algorithms. J ACM 1987;**34**(3):596–615.
161. Valiant LG. The complexity of enumeration and reliability problems. SIAM J Comput 1979;**8**(3):410–21.
162. Gale D, Shapley LS. College admissions and the stability of marriage. American Mathematical Monthly 1962;**69**(1):9–15.

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

163. Irving RW. Stable marriage and indifference. Discrete Applied Mathematics 1994;**48**(3):261–72.
164. Iwama K, Miyazaki S. A survey of the stable marriage problem and its variants. In: Proceedings of the International Conference on Informatics Education and Research for Knowledge-Circulating Society (icks 2008). 1345763: IEEE Computer Society 2008:131–6.
165. Halford TR, Chugg KM. An algorithm for counting short cycles in bipartite graphs. IEEE Trans Inform Theory 2006;**52**(1):287–92.
166. Steiner G. On the k-path partition of graphs. Theoretical Computer Science 2003;**290**(3):2147–55.
167. Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York, 1979.
168. Peeters R. The maximum edge biclique problem is NP-complete. Discrete Applied Mathematics 2003;**131**(3):651–4.
169. Di Giacomo E, Grilli L, Liotta G. Drawing bipartite graphs on two curves. In: International Symposium on Graph Drawing, Springer, 2006:380–5.
170. Di Battista G, Eades P, Tamassia R et al. Algorithms for drawing graphs: an annotated bibliography. Computational Geometry 1994;**4**(5):235–82.
171. Fößmeier U, Kaufmann M. Nice drawings for planar bipartite graphs. In: Italian Conference on Algorithms and Complexity. Springer, 1997:122–34.
172. Eades P, Whitesides S. Drawing graphs in two layers. Theoretical Computer Science 1994;**131**(2):361–74.
173. Misue K. Drawing bipartite graphs as anchored maps. In: Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60. Australian Computer Society, Inc., 2006:169–77.
174. Misue K. Anchored maps: visualization techniques for drawing bipartite graphs. In: International Conference on Human-Computer Interaction. Springer, 2007:106–14.
175. Shahrokhi F, Sýkora O, Székely LA et al. On bipartite drawings and the linear arrangement problem. SIAM J Comput 2001;**30**(6):1773–89.
176. Eades P, Wormald NC. Edge crossings in drawings of bipartite graphs. Algorithmica 1994;**11**(4):379–403.
177. Zheng L, Song L, Eades P. Crossing minimization problems of drawing bipartite graphs in two clusters. In: Proceedings of the 2005 Asia-Pacific Symposium on Information Visualisation, **Volume 45**. Australian Computer Society, Inc. 2005:33–37.
178. Valls V, Martí R, Lino P. A branch and bound algorithm for minimizing the number of crossing arcs in bipartite graphs. European Journal of Operational Research 1996;**90**(2):303–19.
179. Callaway DS, Newman MEJ, Strogatz SH et al. Network robustness and fragility: percolation on random graphs. Phys Rev Lett 2000;**85**(25):5468–71.
180. Li D, Zhang Q, Zio E et al. Network reliability analysis based on percolation theory. Reliability Engineering & System Safety 2015;**142**:556–62.
181. Hooyberghs H, Van Schaeybroeck B, Indekeu JO. Percolation on bipartite scale-free networks. Physica A: Statistical Mechanics and Its Applications 2010;**389**(15):2920–9.
182. Lü L, Zhou T. Link prediction in complex networks: a survey. Physica A: Statistical Mechanics and Its Applications 2011;**390**(6):1150–70.
183. Kunegis J, De Luca EW, Albayrak S. The link prediction problem in bipartite networks. In: Computational Intelligence for Knowledge-Based Systems Design: 13th International Conference on Information Processing and Management of Uncertainty, IPMU 2010; Dortmund, Germany, June 28 - July 2, 2010 Proceedings. Hüllermeier E, Kruse R, Hoffmann F, eds. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010:380–9.
184. Benchettara N, Kanawati R, Rouveirol C. Supervised machine learning applied to link prediction in bipartite social networks. In: 2010 International Conference on Advances in Social Networks Analysis and Mining. 9–11 Aug. 2010 2010:326–30.
185. Allali O, Magnien C, Latapy M. Link prediction in bipartite graphs using internal links and weighted projection. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). 10–15 April 2011, 2011: 936–41.
186. Agarwal S. Learning to rank on graphs. Mach Learn 2010;**81**(3):333–57.
187. He X, Gao M, Kan M-Y et al. Birank: towards ranking on bipartite graphs. IEEE Trans Knowl Data Eng 2017;**29**(1):57–71.
188. Wang JJ-Y, Bensmail H, Gao X. Multiple graph regularized protein domain ranking. BMC Bioinformatics 2012;**13**(1):307.
189. Phillips CA. Multipartite Graph Algorithms for the Analysis of Heterogeneous Data. University of Tennessee- Knoxville, 2015.
190. Phillips CA, Wang K, Bubier J, et al. Scalable multipartite subgraph enumeration for integrative analysis of heterogeneous experimental functional genomics data. In: Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics. ACM, 2015:626–33.
191. Fortunato S. Community detection in graphs. Physics Reports 2010;**486**(3-5):75–174.
192. Sawardecker EN, Amundsen CA, Sales-Pardo M et al. Comparison of methods for the detection of node group membership in bipartite networks. Eur Phys J B 2009;**72**(4):671–7.
193. Barber MJ. Modularity and community detection in bipartite networks. Phys Rev E 2007;**76**(6):066102.
194. Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of the National Academy of Sciences 2002;**99**(12):7821–6.
195. Larremore DB, Clauset A, Jacobs AZ. Efficiently inferring community structure in bipartite networks. Phys Rev E 2014;**90**(1):012805.
196. Suzuki K, Wakita K. Extracting multi-facet community structure from bipartite networks. In: 2009 International Conference on Computational Science and Engineering. 29–31 Aug. 2009, 2009:312–9.
197. Liu X, Murata T. Community detection in large-scale bipartite networks. In: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. 15–18 Sept. 2009, 2009:50–57.
198. Du N, Wang B, Wu B et al. Overlapping community detection in bipartite networks. In: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 9–12 Dec. 2008, 2008:176–9.
199. Dormann CF, Strauss R. A method for detecting modules in quantitative bipartite networks. Methods Ecol Evol 2014;**5**(1):90–98.
200. Busygin S, Prokopyev O, Pardalos PM. Biclustering in data mining. Computers & Operations Research 2008;**35**(9):2964–87.
201. Prelić A, Bleuler S, Zimmermann P et al. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 2006;**22**(9):1122–9.

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

202. Liu Y-Y, Slotine J-J, Barabasi A-L. Controllability of complex networks. Nature 2011;**473**(7346):167–73.
203. Nacher JC, Akutsu T. Structural controllability of unidirectional bipartite networks. Sci Rep 2013;**3**(1):1647.
204. Kontou PI, Pavlopoulou A, Dimou NL et al. Data and programs in support of network analysis of genes and their association with diseases. Data in Brief 2016;**8**: 1036–9.
205. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Systems 1695. Available at igraph org/ Accessed November 2006;**30**:2015.
206. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems 2006;**1695**(5):1–9.
207. BiLayout - Cytoscape plugin. http://bilayout.bioinf.mpi-inf.mpg.de/index.php. Accessed 15 Jan 2018.
208. PowerClust. http://www.compgen.org/tools/powerclust. Accessed 15 Jan 2018.
209. Pundir S, Martin MJ, O'Donovan C. UniProt protein knowledgebase. Methods in Molecular Biology 2017;**1558**:41–55.
210. Finn RD, Coggill P, Eberhardt RY et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 2016;**44**(D1):D279–85.
211. Baxevanis AD. The importance of biological databases in biological discovery. Current Protocols in Bioinformatics 2011; Chapter 1:Unit 1 1.
212. Ng A, Bursteinas B, Gao Q et al. Resources for integrative systems biology: from data through databases to networks and dynamic system models. Briefings in Bioinformatics 2006;**7**(4):318–30.
213. van Gend C, Snoep JL. Systems biology model databases and resources. Essays Biochem 2008;**45**:223–36.
214. Bebek G, Koyuturk M, Price ND et al. Network biology methods integrating biological data for translational science. Briefings in Bioinformatics 2012;**13**(4):446–59.
215. Altaf-Ul-Amin M, Afendi FM, Kiboi SK et al. Systems biology in the context of big data and networks. BioMed Research International 2014;**2014**:1–11.
216. Wang J, Pang GS, Chong SS et al. SNP web resources and their potential applications in personalized medicine. CDM 2012;**13**(7):978–90.
217. Hoeppner MP, Barquist LE, Gardner PP. An introduction to RNA databases. Methods in Molecular Biology 2014;**1097**:107–23.
218. Bastone A, Diomede L, Parini R et al. Determination of argininosuccinate lyase and arginase activities with an amino acid analyzer. Analytical Biochemistry 1990;**191**(2):384–9.
219. Ooi HS, Schneider G, Chan YL et al. Databases of protein-protein interactions and complexes. Methods in Molecular Biology 2010;**609**:145–59.
220. Ooi HS, Schneider G, Lim TT et al. Biomolecular pathway databases. Methods in Molecular Biology 2010;**609**:129–44.
221. He Y, Xiang Z. Databases and in silico tools for vaccine design. Methods in Molecular Biology 2013;**993**:115–27.
222. Timmers LF, Pauli I, Caceres RA et al. Drug-binding databases. CDT 2008;**9**(12):1092–9.
223. Gozalbes R, Pineda-Lucena A. Small molecule databases and chemical descriptors useful in chemoinformatics: an overview. CCHTS 2011;**14**(6):548–58.
224. Pavlopoulou A, Spandidos DA, Michalopoulos I. Human cancer databases (review). Oncol Rep 2015;**33**(1):3–18.
225. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;**2**(8):e124.
226. Ioannidis JP, Ntzani EE, Trikalinos TA et al. Replication validity of genetic association studies. Nat Genet 2001;**29**(3):306–9.
227. Pavlopoulos GA, Paez-Espino D, Kyrpides NC et al. Empirical comparison of visualization tools for larger-scale network analysis. Advances in Bioinformatics 2017;**2017**:1–8.
228. Martin S, Brown WM, Klavans R et al. OpenOrd: an open-source toolbox for large graph layout. In: *2011. IS&T/SPIE Electronic Imaging* 2011, p.11. SPIE.
229. Hu Y. Efficient, high-quality force-directed graph drawing. Mathematica Journal 2005;**10**(1):37–71.
230. Morris JH, Apeltsin L, Newman AM et al. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics 2011;**12**(1):436.
231. Newman AM, Cooper JB. AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. BMC Bioinformatics 2010;**11**(1):117.
232. Eisen MB, Spellman PT, Brown PO et al. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 1998;**95**(25):14863–8.
233. Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007;**315**(5814):972–6.
234. Newman ME, Girvan M. Finding and evaluating community structure in networks. Phys Rev E 2004;**69**(2):026113.
235. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003;**4**(1):2.
236. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 2002;**30**(7):1575–84.
237. Nepusz T, Sasidharan R, Paccanaro A. SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. BMC Bioinformatics 2010;**11**(1):120.
238. Wittkop T, Emig D, Lange S et al. Partitioning biological data with transitivity clustering. Nat Methods 2010;**7**(6): 419–20.
239. Pavlopoulos GA, Wegener AL, Schneider R. A survey of visualization tools for biological network analysis. BioData Mining 2008;**1**(1):12.
240. Pavlopoulos GA, Malliarakis D, Papanikolaou N et al. Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. GigaSci 2015;**4**(1):38.
241. Gehlenborg N, O'Donoghue SI, Baliga NS et al. Visualization of omics data for systems biology. Nat Methods 2010;**7**(3):S56–68.
242. Pavlopoulos GA, Iacucci E, Iliopoulos I et al. Interpreting the Omics "era" data. In: Multimedia Services in Intelligent Environments: Recommendation Services. Tsihrintzis GA, Virvou M, Jain LC, eds, Heidelberg: Springer International Publishing; 2013:79–100.
243. Shannon P, Markiel A, Ozier O et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 2003;**13**(11):2498–504.
244. Saito R, Smoot ME, Ono K et al. A travel guide to Cytoscape plugins. Nat Methods 2012;**9**(11):1069–76.
245. Piñero J, Bravo À, Queralt-Rosinach N et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research 2017;**45**(D1):D833–9.
246. Batagelj V, Mrvar A. Pajek-program for large network analysis. Connections 1998;**21**(2):47–57.

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021

247. Hagberg A, Schult D, Swart P. Exploring network structure, dynamics, and function using NetworkX. In: SciPy 2008: Proceedings of the 7th Python in Science Conference. 2008:11–15.
248. Borgatti SP, Everett MG, Freeman LC. Ucinet for windows: Software for social network analysis. 2002.
249. Bastian M, Heymann S, Jacomy M. Gephi: an Open Source Software for exploring and manipulating networks. Icwsm; 2009;**8**:361–2.
250. Beckett SJ, Boulton CA, Williams HT. FALCON: a software package for analysis of nestedness in bipartite networks. F1000Research., 2014; 3.
251. Secrier M, Pavlopoulos GA, Aerts J et al. Arena3D: visualizing time-driven phenotypic differences in biological systems. BMC Bioinformatics 2012;**13**(1):45.
252. Pavlopoulos GA, O'Donoghue SI, Satagopam VP et al. Arena3D: visualization of biological networks in 3D. BMC Syst Biol 2008;**2**(1):104.
253. Barkow S, Bleuler S, Preli A et al. BicAT: a biclustering analysis toolbox. Bioinformatics 2006;**22**(10):1282–3.
254. Baker EJ, Jay JJ, Bubier JA et al. GeneWeaver: a web-based system for integrative functional genomics. Nucleic Acids Research 2012;**40**(D1):D1067–76.
255. Zachary N. ONEMODE: Stata module to produce one-mode projections of a bipartite network. S457498 ed.: Boston College Department of Economics, 2012.
256. Krzywinski M, Schein J, Birol I et al. Circos: an information aesthetic for comparative genomics. Genome Research 2009;**19**(9):1639–45.
257. Krzywinski M, Birol I, Jones SJ et al. Hive plots–rational approach to visualizing networks. Briefings in Bioinformatics 2012;**13**(5):627–44.
258. Bikakis N, Liagouris J, Krommyda M, Papastefanatos G, Sellis T GraphVizdb: A scalable platform for interactive large graph visualization. In: *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on* 2016, pp.1342–5. IEEE.
259. Kohler J, Baumbach J, Taubert J et al. Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics 2006;**22**(11):1383–90.
260. Iragne F, Nikolski M, Mathieu B et al. ProViz: protein interaction visualization and exploration. Bioinformatics 2005;**21**(2):272–4.
261. Hu Z, Hung JH, Wang Y et al. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Research 2009;**37**(Suppl˙2):W115–21.
262. Adar E. GUESS: a language and interface for graph exploration. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montreal, Quebec, Canada. 1124889, ACM, 2006:791–800.
263. Borgatti SP, Everett MG, Freeman LC. Ucinet for Windows: software for social network analysis. Harvard, MA: Analytic Technologies 2002.
264. Thimm O, Blasing O, Gibon Y et al. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant Journal: for Cell and Amolecular Biology 2004;**37**(6):914–39.
265. Demir E, Babur O, Dogrusoz U et al. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. Bioinformatics 2002;**18**(7):996–1003.
266. Pavlopoulos GA, Hooper SD, Sifrim A et al. Medusa: a tool for exploring and clustering biological networks. BMC Res Notes 2011;**4**(1):384.
267. Breitkreutz BJ, Stark C, Tyers M. Osprey: a network visualization system. Genome Biol 2003;**4**(3):R22.
268. Theocharidis A, van Dongen S, Enright AJ et al. Network visualization and analysis of gene expression data using BioLayout Express3D. Nat Protoc 2009;**4**(10):1535–50.
269. Admiraal R, Handcock MS. Networksis: a package to simulate bipartite graphs with fixed marginals through sequential importance sampling. J Stat Soft 2008;**24**(8).
270. Borrett SR, Lau MK. enaR: Anr package for ecosystem network analysis. Methods Ecol Evol 2014;**5**(11):1206–13.
271. Dormann CF, Gruber B, Fründ J. Introducing the bipartite package: analysing ecological networks. Interaction 2008;**1**:0.2413793.
272. Seal A, Wild DJ. Netpredictor: R and Shiny package to perform drug-target bipartite network analysis and prediction of missing links. bioRxiv 2016.
273. Opsahl T. Structure and Evolution of Weighted Networks. Queen Mary, University of London; 2009.
274. Gobbi A, Iorio F, Dawson KJ, Wedge DC, Tamborero D, Alexandrov LB et al.. Fast randomization of large genomic datasets while preserving alteration counts. Bioinformatics 2014;**30**(17):i617–i23. 2013.
275. Vrahatis AG, Balomenos P, Tsakalidis AK, Bezerianos A. DEsubs: an R package for flexible identification of differentially expressed subpathways using RNA-seq experiments. Bioinformatics 2016;**32**(24):3844–6.
276. SNAP library. https://snap.stanford.edu/data/index.html. Accessed Jan 2018.
277. Leskovec J, Sosič R. SNAP. ACM Trans Intell Syst Technol 2016;**8**(1):1–20.
278. Colorado Index of Complex Networks (ICON). https://icon.colorado.edu/#!/. Accessed 15 Jan 2018.
279. Koblenz Network Collection (KONECT). http://konect.uni-koblenz.de/. Accessed 15 Jan 2018.
280. Kunegis J. Konect: the koblenz network collection. In: Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013:1343–50.
281. Web of Life. http://www.web-of-life.es/. Accessed 15 Jan 2018.
282. Interaction Web Database. https://www.nceas.ucsb.edu/interactionweb/index.html. Accessed 15 Jan 2018.
283. Kelpforest Database. http://kelpforest.ucsc.edu/visualizations/. Accessed 15 Jan 2018.
284. Beas-Luna R, Novak M, Carr MH et al. An online database for informing ecological network models: http://kelpforest.ucsc.edu. http://kelpforest. ucsc. edu. PLoS One 2014;**9**(10):e109356.
285. Xu J. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Molecular Ecology 2006;**15**(7):1713–31.
286. Jones MB, Schildhauer MP, Reichman OJ et al. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annu Rev Ecol Evol Syst 2006;**37**(1):519–44.
287. Hrček J, Godfray HCJ. What do molecular methods bring to host-parasitoid food webs? Trends in Parasitology, **31**(1):30–35.
288. Proulx SR, Promislow DEL, Phillips PC. Network thinking in ecology and evolution. Trends in Ecology & Evolution, **20**(6):345–53.

Downloaded from https://academic.oup.com/gigascience/article/7/4/giy014/4875933 by Dokuz Eylul University Library (DEU) user on 02 September 2021