• 2 •

Introduction to Network Analysis

Jörg Menche and Albert-László Barabási

Introduction

The mechanisms underlying human disease involve complex interactions across many levels of cellular organization, from protein–DNA interactions to signal transduction and metabolism. Despite the very different nature of the components and the diversity of the interactions between them, they have one important thing in common: they can all be described as networks. In the past decade, the emerging field of network science has established new paradigms and tools to analyze and understand systems of interacting components and their collective properties. In this chapter, we review the basic concepts and tools of network science and illustrate their application to the study of human disease.

Basic Network Properties

Networks are defined as a collection of components and their interactions. The components are called nodes or vertices and their interactions links or edges. Figure 2–1 shows examples of networks encountered in the study of human disease. Protein interaction networks (Rual, Venkatesan, et al. 2005; Stelzl, Worm, et al. 2005; Venkatesan, Rual, et al. 2009) are best described as undirected networks: two proteins are connected by an undirected link if they physically interact with each other. Most commonly, these links are unweighted, representing a yes/no relationship. In weighted networks the nodes and/or links carry an additional weight, representing, for example, the activity of an enzyme (node weight) or the flux of a reaction (link weight) in metabolic networks (Ideker, Thorsson, et al. 2001; Stelling, Klamt, et al. 2002; Forster, Famili, et al. 2003). Gene regulatory networks (Davidson and Levin 2005) are directed networks, as each interaction has a source and a target, for example, "the expression of gene A inhibits the expression of gene B." The regulatory mechanism is mediated by other molecules such as transcription factors or microRNAs. Networks that explicitly include two different types of nodes are

TABLE 2–1 Mathematical Symbols

C_i	Local clustering coefficient of node i
$<C>$	Mean clustering averaged over all nodes
d	Distance between two nodes (i.e., the length of the shortest path between them)
d_{max}	Diameter of a network (i.e., the largest d between all possible node pairs)
d_s	Shortest distance observed between a node and a given group of nodes
$<d>$	Mean distance averaged over all node pairs in the network
$<d_{AA}>$	Average of d_s for a group of nodes A
$<d_{AB}>$	Average of d_s between two groups of nodes A and B
γ	Exponent of the degree distribution in scale-free networks
k	Degree of a node (i.e., the number of links attached to it)
k_s	Number of links that a node has to a given set of seed genes
$<k>$	Average degree of all nodes in a network
$<k^2>$	Second moment of the degree distribution $P(k)$
l	Length of a path in a network
L	Number of links in a network
L_{max}	Maximal possible number of simple, undirected links in a network
m	Number of nodes in a subgraph
N	Number of nodes in a network
N_d	Number of genes associated with a certain disease
p	Probability that two nodes are connected in an Erdős–Rényi graph
p_c	Critical probability at which a giant component emerges
p_c^{bino}	Critical probability at which a giant component emerges in an Erdős–Rényi graph
$P(k)$	Distribution of the degrees of all nodes
r	Reset probability in a random walk on a network
$<S_{rand}^{degree}>$	Mean random expectation for the largest connected component size according to the degree-preserving randomizing method
σ	Standard deviation
s	Number of seed genes in the network
S	Size of the largest connected component
s_{AB}	Network-based separation of two groups of nodes A and B

called bipartite networks. Examples of such bipartite networks are networks in which diseases are connected to their associated genes (Goh, Cusick, et al. 2007) or symptoms (Zhou, Menche, et al. 2015).

Figure 2–2 illustrates the basic concepts and quantities frequently encountered in the characterization of unweighted, undirected networks (Barabási, 2016). Throughout this chapter, we illustrate the introduced concepts using the interactome (Menche, Sharma, et al. 2014) described in Figure 2–3.

Network size. The total number of nodes N is called the size of the network; L denotes the total number of links. In networks without multiple links between two nodes, the maximal possible number of links between N nodes is

$$L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}, \qquad (2\text{–}1)$$

in which case the network is fully connected. Most real networks are sparse, that is, only a small fraction of all possible links is present. The interactome, for example, has $N = 13{,}460$ nodes and $L = 141{,}296$ links, which is less than 0.2% of all possible links.

Degree. The number of links a node has (i.e., the number of its direct neighbors) is called its degree k. The mean or average degree $\langle k \rangle$ in a network is given by

$$\langle k \rangle = \frac{2L}{N}. \qquad (2\text{–}2)$$

The mean degree of the human interactome is $\langle k \rangle \approx 21$.

Network paths. A network path refers to a sequence of links that connect two nodes A and B; its length l is simply given by the number of steps. The minimal number of links necessary to connect A and B is called the shortest path length and gives their network-based distance d. Note

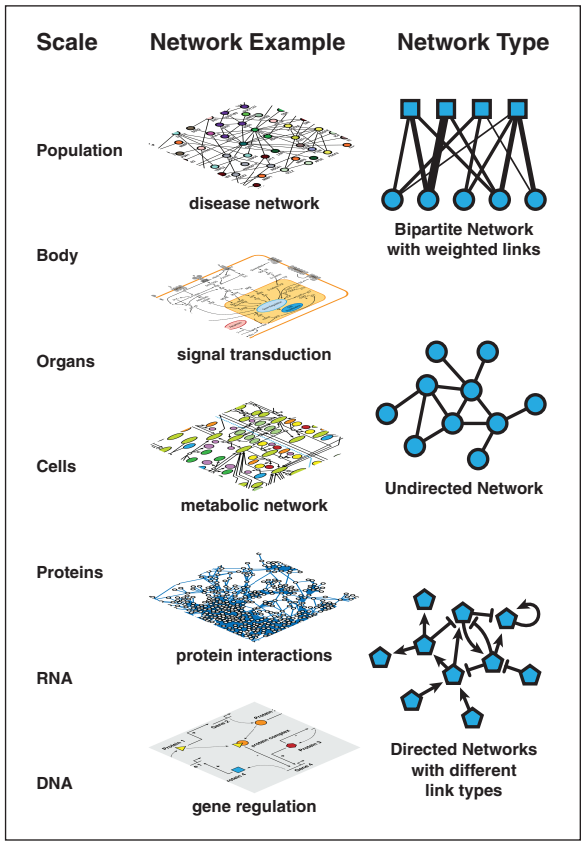

FIGURE 2–1. Networks relevant in human diseases are shown. Network-based approaches to human disease involve different levels of organization. At every level we find systems that are best described in terms of networks, from gene regulatory networks at the molecular scale, to comorbidity networks at the population scale. Depending on the complexity of the system and the desired level of detail in its representation, we can distinguish different network types. The most elementary network types are undirected and unweighted networks. More complex types may include a link directionality or link weights or use different types of nodes, describing the system as bipartite networks.

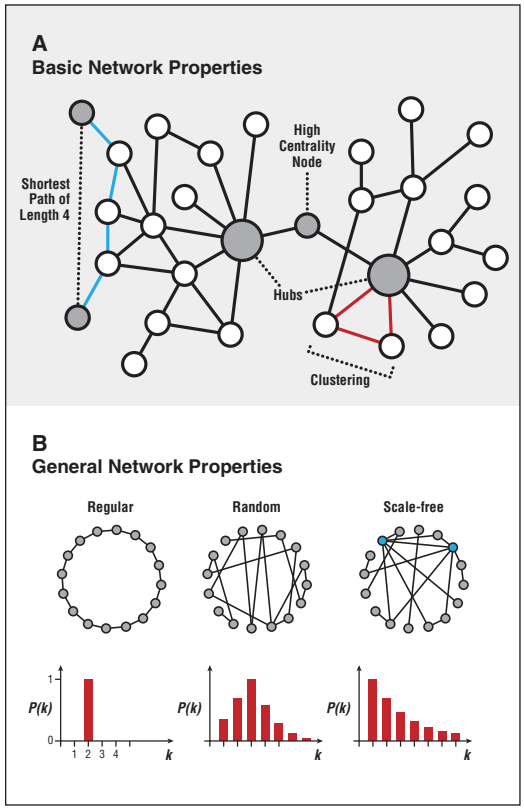

FIGURE 2–2. General network properties. A, Illustration of the fundamental concepts that characterize nodes and their relationship within a network. B, Three important classes of networks and their degree-distribution. In a regular network, all nodes have the same number of links. In a random network, each pair of nodes is connected with a given probability, hence their degree distribution, $P(k)$, follows the binomial distribution (3). In a scale-free network, $P(k) \sim k^{-\gamma}$; its main feature is the presence of highly connected nodes, or *hubs*.

that there is typically a large number of shortest paths connecting most pairs of nodes within a network. Figure 2–3C shows the distribution $P(d)$ of all pairwise distances in the interactome.

Network diameter. The diameter $d_{\max}$ of a network is the longest of all shortest paths between any two nodes. Most real networks have a surprisingly small diameter, a property called the "small world" phenomenon, referring to the popular notion that everyone is connected to everyone else by only a small number of intermediate acquaintances. The diameter of the interactome, for example, is $d_{max}=13$ and the mean distance of all protein pairs is $<d>=3.4$, implying that on average any two proteins are connected via less than four intermediate links (see Figure 2–3B, C).

Degree distribution. The distribution $P(k)$ of the degrees of all nodes in a network can be used to distinguish classes of networks. Historically, the first networks to be studied were regular, like a square lattice, encountered, for example, in crystals. The degree distribution of regular networks typically has a single peak, implying that all nodes have the same number of neighbors (see Figure 2–2B).

Random networks. Many of the fundamental concepts used in modern network science are derived from *random networks* (Erdős 1959; Erdős and Rényi 1960). Consider a network of size N in which each of the possible $L_{\max}$ node pairs is connected by a link with a

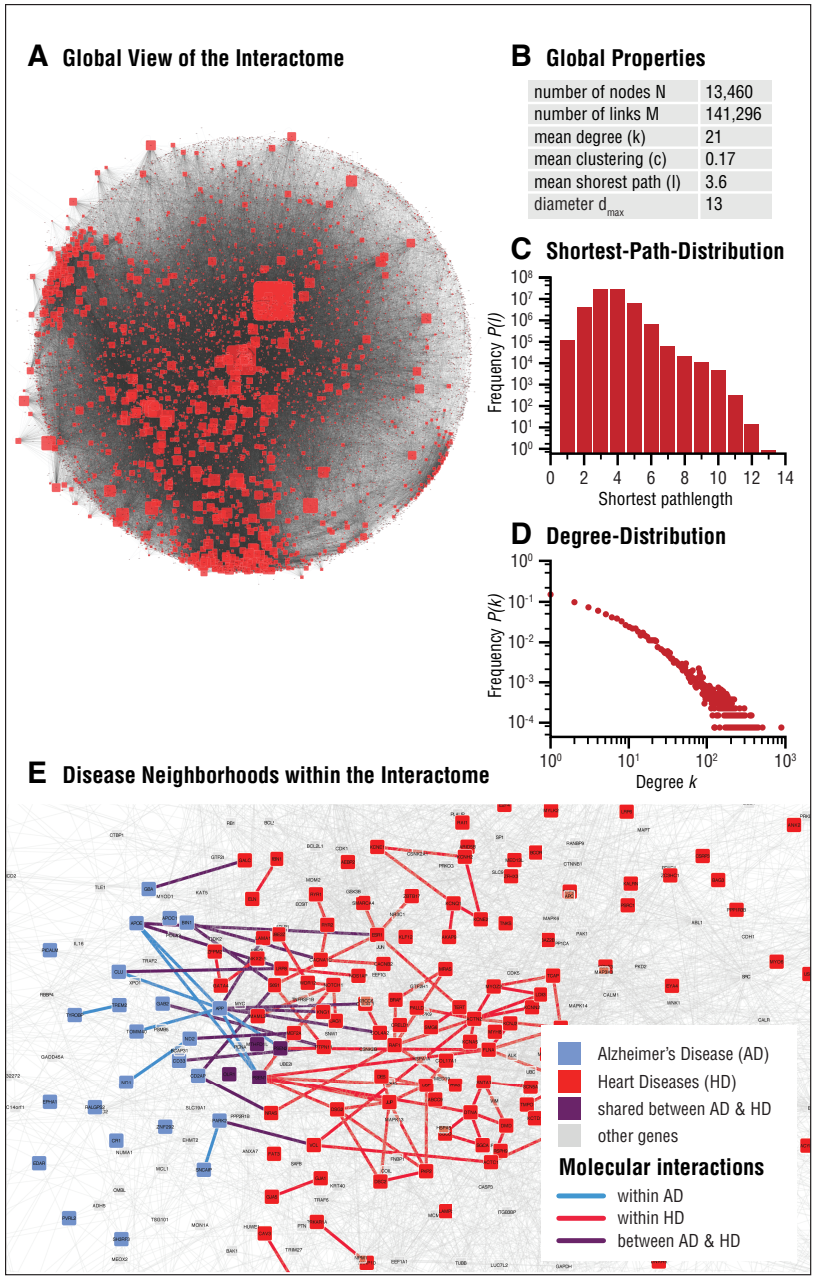

FIGURE 2–3. Overview of the interactome. The interactome represents a comprehensive map of all biologically relevant molecular interactions, for example, binary, regulatory, or signaling interactions. It includes data from high-throughput experiments, such as yeast two-hybrid, as well as literature-curated interactions. A, Global map of the interactome, illustrating its heterogeneity. Node sizes are proportional to their degree, that is, the number of links each node has to other nodes. B, Basic characteristics of the interactome. C, Distribution of the shortest paths within the interactome. The average shortest path is $l = 3.6$. D, The degree distribution of the interactome is approximately scale-free. E, A local neighborhood of the interactome, illustrating the different types of connections and highlighting the proteins associated with two different diseases.

fixed probability p. In a random network, the probability for a node to have exactly k links follows the binomial distribution:

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{2–3}$$

An important property of the degree distribution shown in Eq. (2–3) is that degrees much larger or much smaller than the average are absent, that is, most nodes in the network have a comparable number of links around $<k>$.

Scale-free networks. A key finding of network science is that for most real networks the degree distribution does not follow Eq. (2–3). Instead, as discovered in 1999 (Albert, Jeong, et al. 1999; Barabási and Albert 1999), many real-world networks are scale-free, exhibiting a power-law degree distribution:

$$P(k) \sim k^{-\gamma}. \tag{2–4}$$

A scale-free distribution decays more slowly for large k than does the binomial distribution (2–3). While the vast majority of nodes have only a few connections, there are some nodes in the network with a very large number of links, called hubs. For example, while more than 2000 proteins in the interactome have only a single link, 8 have more than 400 interactions, like *GRB2* ($k = 872$), *YWHAZ* ($k = 502$), and *TP53* ($k = 450$) (see Figure 2–3D). The presence of hubs impacts many network properties. For example, they serve as shortcuts, connecting different parts of a network, making them not just "small," but "ultrasmall" (Cohen and Havlin 2003).

Centrality. A frequently used measure of node importance is its centrality (Freeman 1977; Wasserman and Faust 1994). For example, betweenness centrality measures the number of shortest paths that run through a node. Different centrality measures of a node generally correlate with each other, and hubs tend to have high centrality, as they are likely to lie on many shortest paths. Betweenness may reveal unexpected structural features within a network: low-degree nodes with high betweenness can, for example, hint at an underlying modular network structure (Girvan and Newman 2002).

Clustering coefficient. Clustering describes the tendency for two neighbors of a node to also be connected to each other. In a network, such a relationship is represented by a triangle (see

Figure 2–2A). The local clustering coefficient of node $_i$ (C_i) measures for node i of degree k_i the number of possible triangles present in its neighborhood:

$$C_i = \frac{2L_i}{k_i(k_i - 1)}, \tag{2–5}$$

where L_i denotes the number of all connections between the neighbors of node i. C_i varies between $0 \leq C_i \leq 1$, where $C_i = 0$ indicates that there are no connections between the neighbors of node i, and $C_i = 1$ represents a fully connected subgraph around it. The local clustering coefficient, therefore, measures the local density of a network. The degree of clustering of a network is measured by averaging over all local clustering coefficients $<C_i>$:

$$\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i. \tag{2–6}$$

In a random network, each link is present with the same probability p, regardless of whether or not the two nodes share a neighbor. Hence, the average clustering coefficient is $d = p$. This typically yields values orders of magnitudes below the ones observed in real networks. The interactome depicted in Figure 2–3, for example, exhibits strong clustering with $<C> = 0.17$, whereas the expected value for a random network of the same density is only $<C> = 0.0016$.

Analyzing the Properties of Node Groups

The quantities introduced above capture the global characteristics of networks based on the properties of single nodes or node pairs. Yet, many biological functions and their perturbations in disease states arise from the coordinated action of groups of molecules. Next, we introduce network measures to explore the properties of such node groups.

Motifs. Small recurrent subgraphs in a network are called motifs (Figure 2–4). They are defined as a subgraph that occurs more often in a network than expected by chance under an appropriately chosen null model (Milo, Shen-Orr, et al. 2002). Motifs have attracted considerable attention in gene regulatory networks, where they can be interpreted as molecular building blocks associated with certain functions. For example, a particularly simple motif found in the *Escherichia coli* regulatory network is a single

—-1
—0
—+1

node with an inhibitory self-loop, representing a transcription factor repressing its own expression. This motif has been shown to be beneficial for the dynamics of gene expression, leading to faster response to signals and enhanced stability against noise. Other motifs observed in regulatory networks include feed-forward loops, feedback loops, and oscillators.

The detection of motifs typically relies on network randomization (see section entitled "Randomizing the Network Topology" below") and is computationally challenging, limiting the size of motifs that can be systematically studied to, at most, 10 nodes. At the same time, the functional interpretation of larger motifs is difficult since their interface with the rest of the network increases, thereby impeding their analysis in isolation from the rest of the network.

Communities. Larger topological structures within networks are commonly explored in terms of communities (or modules) (Girvan and Newman 2002; Ravasz, Somera, et al. 2002; Fortunato 2010). A community is loosely defined as a subgraph with high local link density, so that nodes within the community have a higher number of links to each other than to nodes outside the community. A large number of definitions of communities appear in the literature, as well as algorithms to detect them (see Fortunato [2010] for a comprehensive review). Depending on the concrete application, one may, for example, choose between algorithms that allow for overlapping communities or distinct communities, determined by whether a node can belong to several communities at the same time (Palla, Derenyi, et al. 2005; Ahn, Bagrow, et al. 2010). Some algorithms can also reveal hierarchical community structures (Girvan and Newman 2002; Ahn, Bagrow, et al. 2010).

In biological networks, topological communities are often associated with certain biological processes. Using an algorithm from (Ahn, Bagrow, et al. 2010), more than 1112 communities with five or more nodes can be identified in the interactome, with more than half of them being significantly enriched with at least one biological process according to the gene ontology (GO) (Ashburner, Ball, et al. 2000) (see Figure 2–4B, C). Figure 2–4D shows an example of a community of 22 proteins that are connected via 120 links. The community contains all 5 proteins associated with ethanol metabolism, corresponding to highly significant enrichment ($p < 2 \times 10^{-6}$).

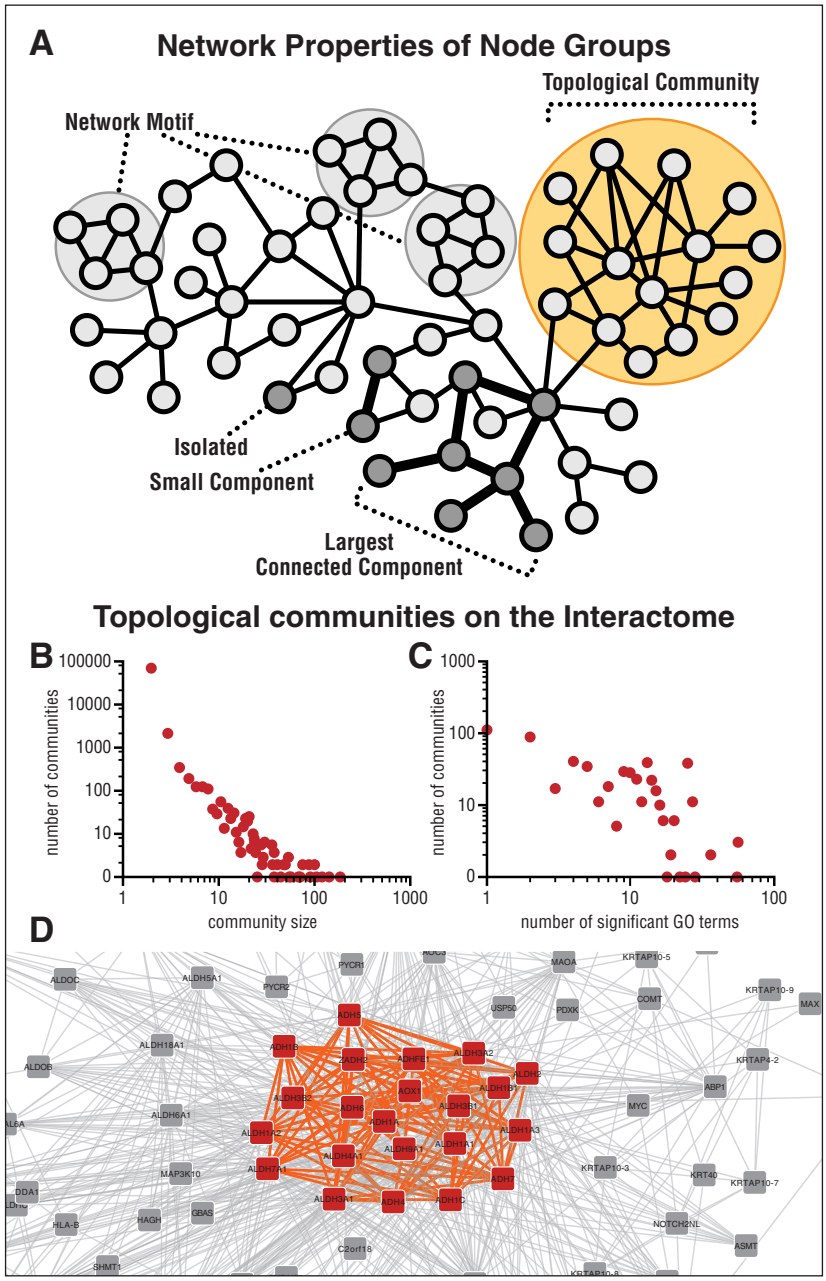

FIGURE 2–4. Properties of node groups. A, Illustration of collective node characteristics: (1) Motifs are small subgraphs that occur more often than expected by chance. (2) Topological communities are local areas of high link density. (3) Connectivity patterns with a given set of nodes, for example, proteins associated with the same disease. They can either be isolated, that is, not interacting with any other nodes of the set, or form connected components of different sizes. B–D, Topological communities on the interactome. The community-finding algorithm of (Ahn, Bagrow, et al. 2010) identified 92,510 communities, of which 1,112 consist of five or more nodes. C, 574 (51%) of these communities are significantly enriched with at least one gene ontology (GO) terms (biological processes); the maximum number is 56 per community. D, Illustration of the community of the 22 densely interconnected genes associated with ethanol metabolism (GO:0006067).

—-1
—0
—+1

Localization of biological function in networks. While in some cases there is a correspondence between topological communities and functional modules in biological networks, there are also important counterexamples: disease modules formed by proteins associated with a particular disease are generally not very densely interconnected within the interactome. This is due in part to the incompleteness of the current interactome and our incomplete list of disease genes (see also, section entitled "Network-Based Disease Gene Discovery" below"). It is also possible, however, that disease modules and functional modules have different topological properties. Regardless of their link density, however, there is evidence that disease modules are highly localized in specific network neighborhoods. Two quantities allow us to measure the degree of network localization of a given set of nodes (Menche, Sharma, 2015):

1. *Size of the largest connected component S,* that is, the number of nodes that form a connected subgraph (see Figure 2–4). Many properties of this quantity can be understood analytically, indicating that its value is relatively sensitive to data incompleteness (see also section entitled "Network-Based Disease Gene Discovery" below"). In extreme cases, a single missing link in the interactome or a single protein whose disease association is not known may destroy the connected component and leave many proteins isolated.

2. *Mean shortest distance.* As a complementary quantity that is less sensitive to network incompleteness, consider the distribution of shortest distances d_s*:* For each disease-associated node we determine the distances d to all other disease-associated nodes. Taking into account only the *shortest* distance d_s among them results in a distribution $P(d_s)$. The mean value $<d>$ can be interpreted as the diameter of the disease module. Note that in contrast to the diameter of the network as introduced above, here *diameter* refers to an average distance, instead of a maximal distance.

In order to interpret the values of S_i and d_s, a comparison with an appropriate random expectation is necessary (see Statistical Tools for Network Analysis). In a comprehensive study of 299 complex diseases, it was shown that proteins associated with 226 diseases exhibit significant localization in the interactome according to both measures (Menche, Sharma, et al. 2015). Furthermore, the more significant the localization of a disease module, the more similar are the molecular functions of the proteins involved in it.

Separation between diseases. The concept of network localization can be further generalized to examine the relation between different

sets of nodes, like proteins associated with two different diseases. The network serves as a map, in which diseases are represented by different neighborhoods. The proximity and degree of overlap of two network neighborhoods has been found to be highly predictive of the pathobiological similarity of the corresponding diseases (Menche, Sharma, et al. 2015).

To quantify the distance of two sets of nodes A and B, we first compute the $P(d_{AB})$ distribution of all shortest distances d_{AB} between nodes A and B and the respective mean distance $<d_{AB}>$ (Figure 2–5). The network-based separation s_{AB} can be obtained by comparing the mean shortest distances $<d_{AA}>$ and $<d_{BB}>$ *within* the respective node sets A and B, to the mean shortest distance $<d_{AB}>$ *between* them (Figure 2–5):

$$s_{AB} = \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle + \langle d_{BB} \rangle}{2} \tag{2–7}$$

A negative s_{AB} indicates topological overlap of the two node sets, whereas a positive s_{AB} indicates topological separation of the two node sets. Rheumatoid arthritis and multiple sclerosis, for example, are two closely related diseases with overlapping disease modules ($s_{AB} = -0.2$), whereas proteins associated with peroxisomal disorders are well separated from the multiple sclerosis proteins ($s_{AB} = 1.3$) (see Figure 2–5B). The network-based separation of disease-associated proteins has been studied for 44,551 pairs among 299 diseases, showing that only 7% of all pairs show network overlap. The degree of this overlap, however, is highly predictive for the pathobiological similarity of diseases: disease pairs with overlapping modules are associated with functionally similar genes that show elevated co-expression, are diseases that have similar symptoms, and are diseases with a high comorbidity. At the same time, nonoverlapping diseases lack any detectable clinical or molecular relationships.

Perturbations and Network Incompleteness

The structural characteristics of a network have important implications for the properties of the dynamic processes they support, like the speed and reliability of signals propagating through them. In general, two nodes in the network can communicate only if there is a path connecting them. An important property of networks is, therefore, their robustness, or resilience, against the breakdown of nodes or links that may break such paths.

—-1
—0
—+1

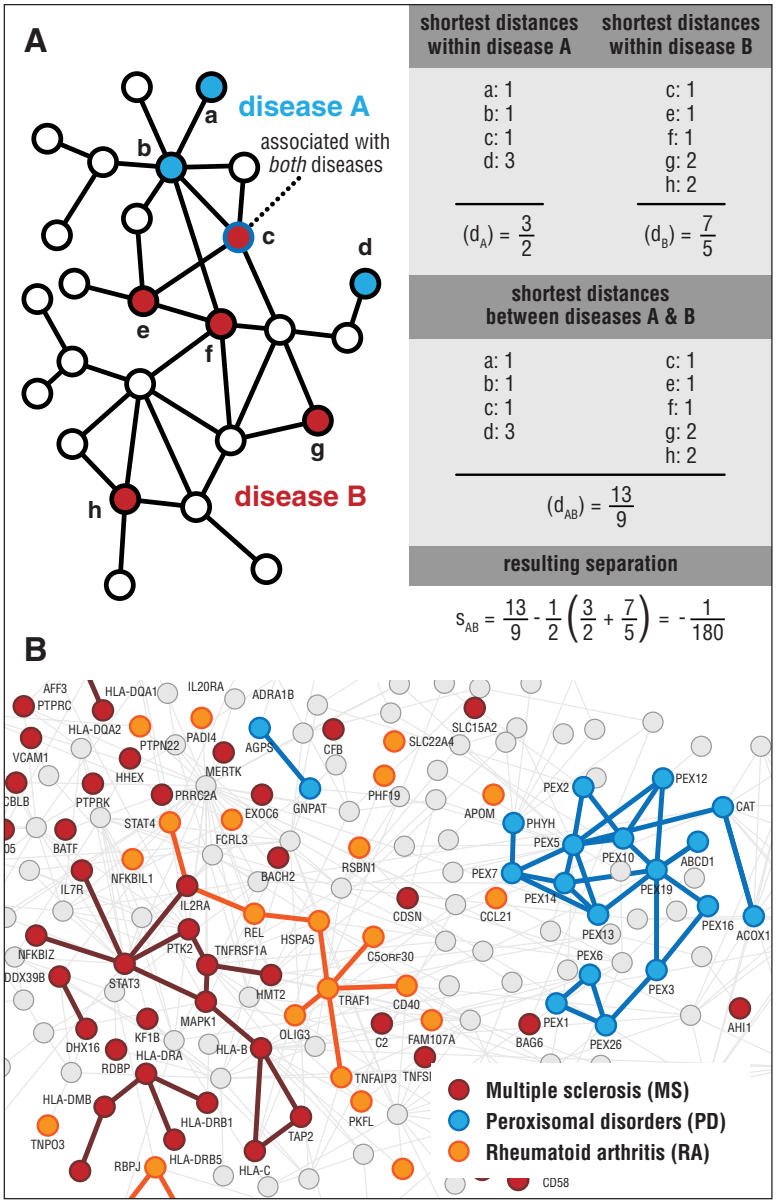

FIGURE 2–5. Network-based separation. A, Illustration of the separation measure s_{AB} for two node sets A (blue) and B (red) with one shared node (c). The tables on the right give the values of the mean shortest distances *within* the sets, $\langle d_{AA} \rangle$ and $\langle d_{BB} \rangle$, as well as the distances for all node pairs *between* them, $\langle d_{AB} \rangle$. Negative values of s_{AB} indicate a topological overlap of the two node sets, whereas positive s_{AB} means that they are topologically separated. In general, s_{AB} is bound by $-d_{max} \leq s_{AB} \leq d_{max}$, where d_{max} denotes the diameter of the network. Since nodes that are shared between sets A and B have $d_{AB} = 0$, the minimal value increases to $-d_{max} + 1$ for sets without common nodes. For sets with at least two nodes, the maximal value is $d_{max} - 1$. B, A subnetwork of the interactome highlighting the network-based relationship between the disease proteins associated with multiple sclerosis, rheumatoid arthritis, and peroxisomal disorders.

-1—
0—
+1—

Network resilience. Biological systems are constantly exposed to external and internal perturbations. Mutations, for example, may affect the ability of a protein to interact with other proteins. A complete loss of function of the protein removes the respective node from the protein-interaction network, whereas link removal corresponds to the case in which only some of its interactions are lost (Zhong, Simonis, et al. 2009).

Networks in which only a fraction of nodes and/or links are present have been studied extensively in the framework of percolation theory (Callaway, Newman, et al. 2000; Cohen, Erez, et al. 2000; Newman, Strogatz, et al. 2001; Dorogovstev 2003). Generally, as long as a certain critical fraction of all N nodes (or L links) is present, the network remains globally connected (Figure 2–6). More precisely, it has a giant component, a connected subgraph that contains most nodes. Below this critical fraction, the giant component disappears and the network breaks into small disconnected components. For random failure, when all nodes (links) have the same probability p of being present in the network, the critical probability p_c, at which the giant component vanishes (called the percolation threshold), is as follows (Callaway, Newman, et al. 2000; Cohen, Erez, et al. 2000; Newman, Strogatz, et al. 2001; Dorogovstev 2003):

$$p_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}, \tag{2–8}$$

where $<k>$ and $<k^2>$ denote the first and second moment of the degree distribution $P(k)$. Using the corresponding expressions for the binomial distribution in Eq. (2–3) we find that the percolation threshold for random graphs is

$$p_c^{\text{bino}} = \frac{1}{p(N-2)}. \tag{2–9}$$

It follows from Eq. (2–9) that connected random networks always have a finite threshold p_c^{bino}, that is, if a critical fraction of nodes/links are removed, the network disintegrates.

Enhanced robustness. Surprisingly, this is not always the case for scale-free networks because a scale-free distribution has a diverging second moment $<k^2> \to \infty$ for $\gamma < 3$, leading to $p_c \to 0$ in Eq. (2–8). This means that in the limit of very large networks, one needs to remove all nodes/links in order to break the network. Strictly speaking, real networks with a finite number of nodes always

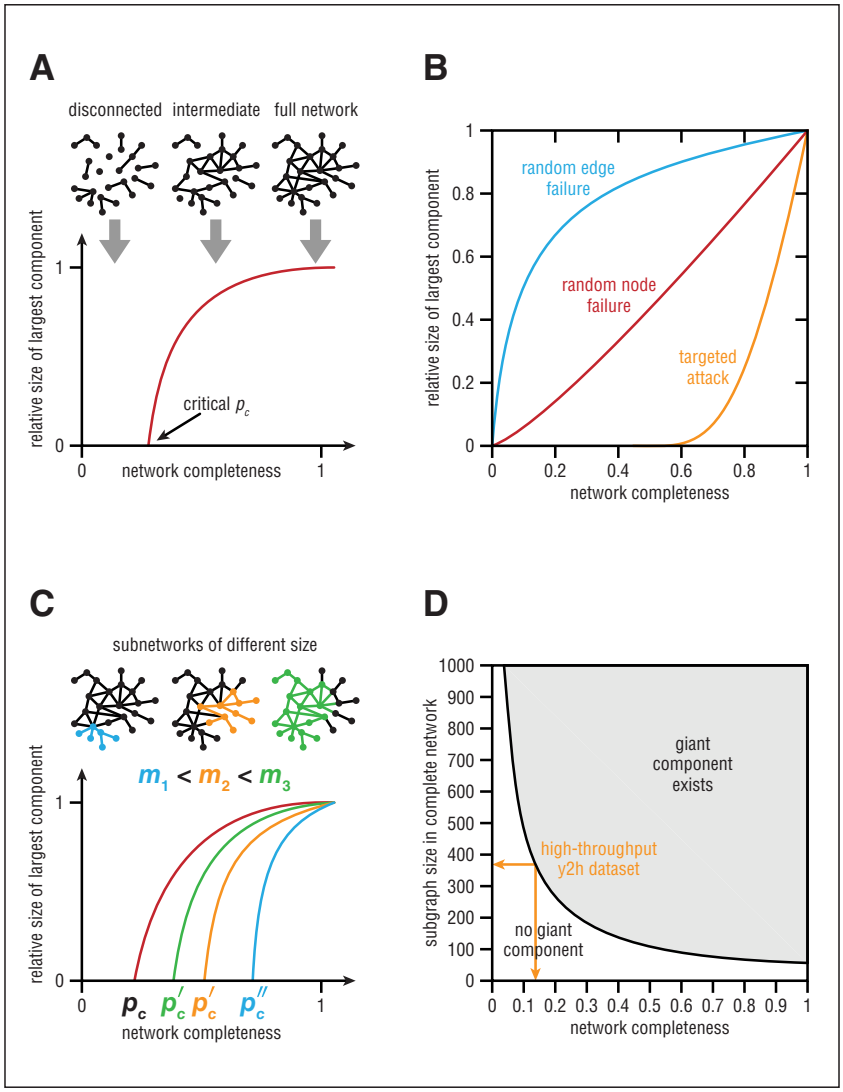

FIGURE 2–6. Percolation theory. A, The behavior of the relative size of the giant connected component as a function of network completeness. Generally, the completeness is given by the product of the observed fractions of all nodes p and all links q. At $pq = 1$, all nodes and links are present and the network consists of one single connected component. As more and more nodes/links are missing, the size of the giant component shrinks until it vanishes at the critical completeness p_c. B, Size of the giant component of the interactome for three different percolation or failure mechanisms: random failure of nodes/edges and targeted removal of the proteins with the highest degrees. C, We observe a similar behavior when subgraphs of size m are considered instead of the whole network. Generally, the percolation threshold $p_c(m)$ will be larger for smaller subgraphs, that is, smaller subgraphs require a higher completeness in order to be observable. D, The phase diagram shows for every level of network completeness how large a module needs to be in the full network in order to exhibit a giant component in the incomplete one, that is, to be observable. (Y2H = yeast two-hybrid.)

have a finite threshold ($p_c > 0$), but its value is negligibly small. For example, for the protein-interaction network discussed above, the threshold is $p_c = 0.01$, so up to 99% of the nodes can be removed before completely fragmenting the network (see Figure 2–6B). Since the vast majority of nodes in scale-free networks have only a few connections, random failure mostly affects such low-degree nodes, which, in turn, have little impact on the overall integrity of the network. Such networks are, therefore, remarkably tolerant against random node removal.

Fragility to attack. This robustness against random failure has also a down side: the networks are particularly vulnerable to a targeted attack that systematically removes the hubs, that is, the nodes in the network with the highest degrees (Albert, Jeong, et al. 2000). The precise fraction of removed hubs under which the network breaks down depends on the details of the degree distribution. For the interactome, we find that removing ~30% of the nodes is sufficient to destroy the network completely (see Figure 2–6B).

Network incompleteness. Percolation theory can also help us understand the implications of the inherent incompleteness of current maps of biological networks. For example, currently available high-throughput human-protein-interaction maps are estimated to cover only around 20% of all true interactions (Venkatesan, Rual, et al. 2009). Using the percolation framework, we can view the current maps as an incomplete sample from the underlying complete network, in which both links and nodes are missing (Stumpf, Wiuf, et al. 2005, 2010; Guimera and Sales-Pardo 2009; Venkatesan, Rual, et al. 2009; Annibale and Coolen 2011). High-throughput interaction maps, as obtained, for example, by yeast two-hybrid (Y2H) assays (Rolland, Tasan, et al. 2014), can be viewed as a uniform subset of the corresponding full interactome: For an unbiased set of proteins, all pairwise interactions have been tested, so the present fraction of all real interactions corresponds to the sensitivity of the experimental protocol. Assuming uniform random sampling, the overall completeness of the obtained network map is given by

$$Pq = (\text{fraction of screened proteins})\ p \times (\text{sensitivity to detect a link})\ q. \qquad (2\text{–}10)$$

—-1
—0
—+1

In the same way that a whole network can fall apart under random node removal or attack, subgraphs inside a network can become disconnected if network incompleteness exceeds some threshold. Disease modules, for example, are expected to form a connected subgraph within the complete interactome. Yet, within the current datasets, only ~20% of the proteins associated with a given disease are part of the giant component. Figure 2–6C illustrates schematically the percolation curves for subgraphs of different sizes of *m*. The percolation threshold is inversely proportional to *m;* that is, smaller subgraphs require a higher network completeness in order to have a giant component. Figure 2–6D shows the minimal subgraph size for which we expect to find a remaining connected component for a given level of network completeness using the Y2H network as input. The yellow arrow indicates the estimated values for the current dataset. We find that the coverage of the current Y2H dataset is still too small to observe significant clustering for the given number of disease-associated genes. Including interactions collected in the literature, however, puts us above the threshold, allowing the systematic identification of multiple disease modules (Menche, Sharma, et al. 2015). We expect that once the ongoing Y2H efforts screen an even larger number of proteins, disease modules can be identified in high-throughput data, as well.

Network-Based Disease Gene Discovery

In addition to missing interactions, we often lack information on important node properties. In particular, for most complex diseases only a fraction of all disease-associated genes are known. As discussed above, disease genes often interact with each other within the same network neighborhood. Building on this observation, and the localization of the disease genes in the same network neighborhood, in recent years a plethora of methods have been developed that exploit the topology of the interactome to infer new disease genes from their connectivity patterns within protein-interaction networks (Tranchevent, Capdevila, et al. 2011). Machine-learning approaches, such as neural networks, support vector machines, or Bayesian networks, typically combine protein-interaction data with other sources of information, such as protein sequence and structure, pathway membership, gene expression, or the genome-wide association study (GWAS) *p* values (Morrison, Breitling, et al. 2005; Aerts, Lambrechts, et al. 2006; Franke, van Bakel, et al. 2006; Hutz, Kraja, et al. 2008). A number of methods aim to identify possible new disease gene candidates relying solely on the position of known diseases genes in the interactome (Krauthammer, Kaufmann, et al. 2004; George, Liu, et al. 2006; Kohler, Bauer, et al.

2008; Dezso, Nikolsky, et al. 2009; Vanunu, Magger, et al. 2010; Bailly-Bechet, Borgs et al. 2011; Erten, Bebek et al. 2011; Guney and Oliva 2012; Sharma, 2015). The already known disease genes that serve as input for these methods are commonly referred to as seed genes. In the following, we briefly review the basic network concepts that underlie these node/gene prioritization methods.

Shortest-path approaches. Based on the observation that seed genes tend to interact with each other, several methods consider the intermediate genes along the shortest paths connecting the seed genes as potential disease gene candidates (George, Liu, et al. 2006; Managbanag, Witten, et al. 2008; Dezso, Nikolsky, et al. 2009). Typically, this results in a large number of candidate genes. In order to identify the most promising candidates, the intermediate genes can be ranked, for example, by the number of shortest paths in which they participate (George, Liu, et al. 2006) and their significance (Dezso, Nikolsky, et al. 2009). An approach introduced by Bailly-Bechet, Borgs, et al. (2011) does not consider all shortest paths, but instead identifies a minimal set of intermediate genes that are sufficient to connect all seed genes into a single subgraph, a so-called Steiner tree.

Dynamic approaches. A different approach for identifying likely disease gene candidates is to propagate known disease associations using dynamic models (Krauthammer, Kaufmann, et al. 2004; Kohler, Bauer, et al. 2008; Vanunu, Magger, et al. 2010; Guney and Oliva 2012). In Köhler, Bauer, et al. (2008), for example, the seed genes serve as sources for a diffusion process that can also be formulated in terms of a random walker that wanders from node to node along the links of the network: at every time step of the iterative algorithm, the walker moves to a randomly selected neighbor of its current position. In order to emphasize the local neighborhood around the seed genes, the walker is reset to a randomly chosen seed gene with a given probability r after every move. The frequency with which the nodes in the network are visited converges after many iterations and can be used to rank the corresponding genes. Genes that are visited more often are considered to be "closer" to the seed genes and, therefore, more likely to be relevant to the disease than those visited less often.

Connectivity-based approaches. Several approaches rank candidate genes based on their number of links to seed genes k_s (Erten, Bebek, et al. 2011; Guney and Oliva 2012; Sharma, 2015). Note that k_s

—-1
—0
—+1

alone is not informative, since hubs (i.e., proteins with many links) are also expected to interact with a large number of seed genes without necessarily implying a disease association. To account for these effects, Menche, Sharma, and colleagues (2015) proposed an algorithm that is based on the significance of k_s. In a network of size N, with s randomly distributed seed genes, the probability that a gene with degree k connects to exactly k_s seed genes is given by the hypergeometric distribution

$$p(X = k_s) = \frac{\binom{s}{k_n}\binom{N-s}{k-k_n}}{\binom{N}{k}}. \tag{2–11}$$

The significance of a given number of connections is, therefore, given by:

$$p\text{-value} = \sum_{n=k_s}^{k} p(X = n), \tag{2–12}$$

which can then be used to iteratively rank all genes in the network (Figure 2–7).

Note that the approaches introduced above can be used for any functional annotation by using suitable protein/gene properties to define seed genes, like pathway membership or differential expression.

Statistical Tools for Network Analysis

In order to assess the statistical significance of a particular network-based finding (e.g., the localization of disease proteins), we need appropriate null models. As many network properties, like the degrees, follow non-Gaussian distributions, we cannot apply standard statistical tests that rely on normality. Network randomization provides a direct way to compare a given measurement with random expectation. Generally, we can randomize the network topology, such as the interaction partners of a particular protein, or randomize the annotation of nodes, like the disease association of proteins. The strategy we apply depends on the statistical feature of the original network we wish to preserve.

Randomizing the Network Topology

Comparison with the random network model. The most basic reference frame is the random network discussed in Basic Network

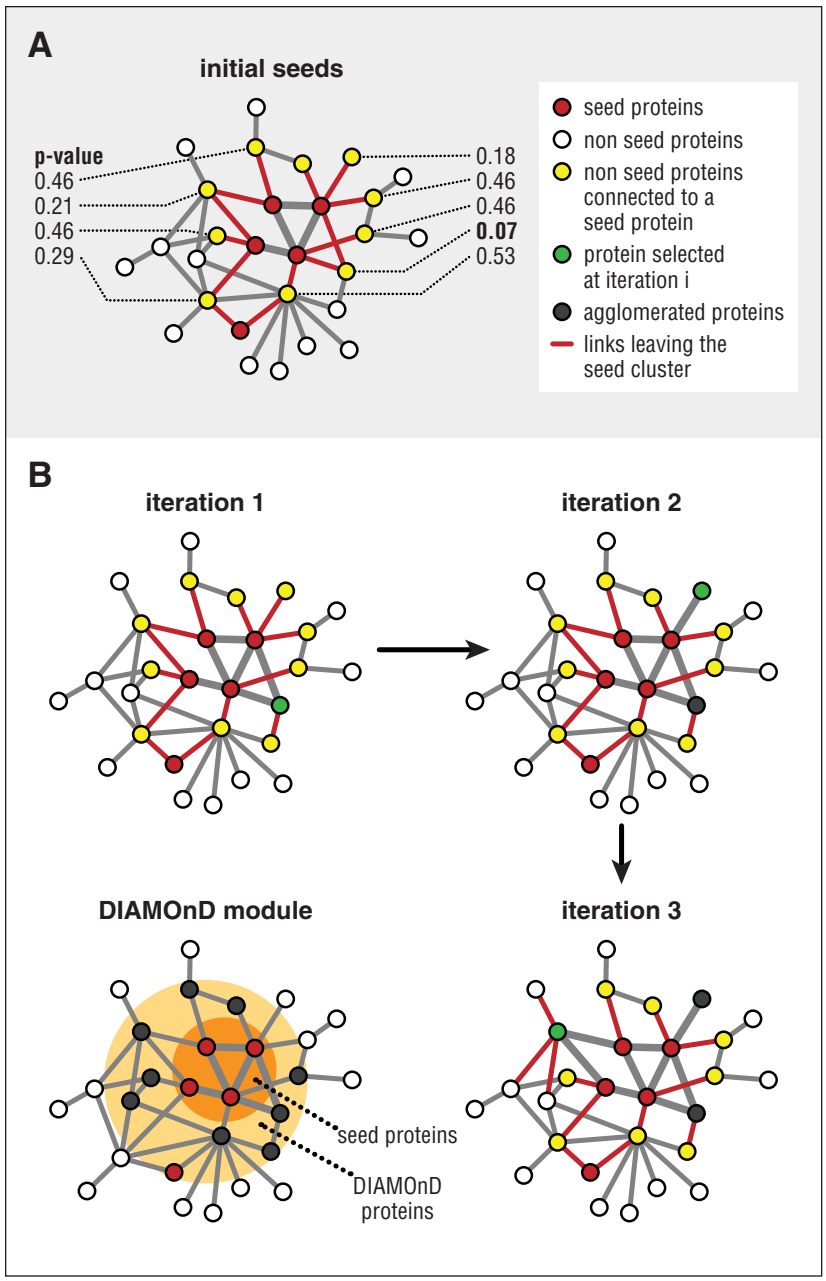

FIGURE 2–7. Network approaches for disease-gene prioritization. Illustration of the connectivity-based DIAMOnD method (Sharma 2015) to construct a full disease module from a set of known disease-associated proteins. A, The seed proteins are placed on the interactome. For every neighboring protein a connectivity *p* value is computed according to Erdős and Rényi (1960). B, At each iteration, the protein with the lowest *p* value is added to the seed cluster. The procedure can be continued until the entire network is selected and added to the module. The order in which the proteins are being pulled in to the module reflects their topological relevance to the disease, resulting in a ranking of all proteins.

—-1
—0
—+1

Properties. For this approach, we randomize a given network, keeping only the number of nodes N and the number of links L constant. Since many properties of a random network can be calculated analytically, we do not need to perform extensive simulations to achieve this goal. The mean clustering coefficient, for example, is simply given by $<C>=p$. For the completely randomized interactome, this yields $<C>=p=L/L_{max}=0.0016$, in excellent agreement with the value obtained from simulations (Figure 2–8C).

Full randomization does not preserve the degree distribution; hence, hubs will no longer be present in the randomized network. As many network properties depend strongly on the degree distribution and the presence of hubs, this method is not suited for most applications.

Degree-preserving randomization. To maintain the degree distribution of a network, we randomize the interaction partners of the nodes while preserving each node's degree. To implement this method, we can use a switching algorithm (Maslov and Sneppen 2002) (see Figure 2–8A). At each step of the algorithm, two links are selected at random and their endpoints are swapped. For example, the links connecting nodes $n_1 \leftrightarrow n_2$ and $n_3 \leftrightarrow n_4$ are exchanged, resulting in two new interactions $n_1 \leftrightarrow n_3$ and $n_2 \leftrightarrow n_4$, respectively. This rewiring can lead to multiple links between a pair of nodes and self-loops. In networks in which such links are not allowed, the original link pairs are restored. Repeating this process a sufficient number of times leads to a network whose topology is randomized, while the degree distribution remains unchanged. In other words, hubs remain hubs. While there is no precise criterion for the necessary number of switches, empirical results suggest that a good randomization is achieved after $100L$ switching attempts (Milo, Nashtan, et al. 2004).

A more efficient approach is the matching algorithm (see Figure 2–8A), based on the configuration model (Bender and Canfield 1978; Bollobas 1979) used to generate networks with a given degree sequence. The algorithm follows these steps: All links are broken at once and then, iteratively, two "half edges" or "stubs" are chosen at random and connected until all links are restored. As before, this process may produce self-loops and multiple links, in which case the respective stubs are not connected and an alternative pair is selected at random. While this method may introduce bias in the ensemble of the generated net-

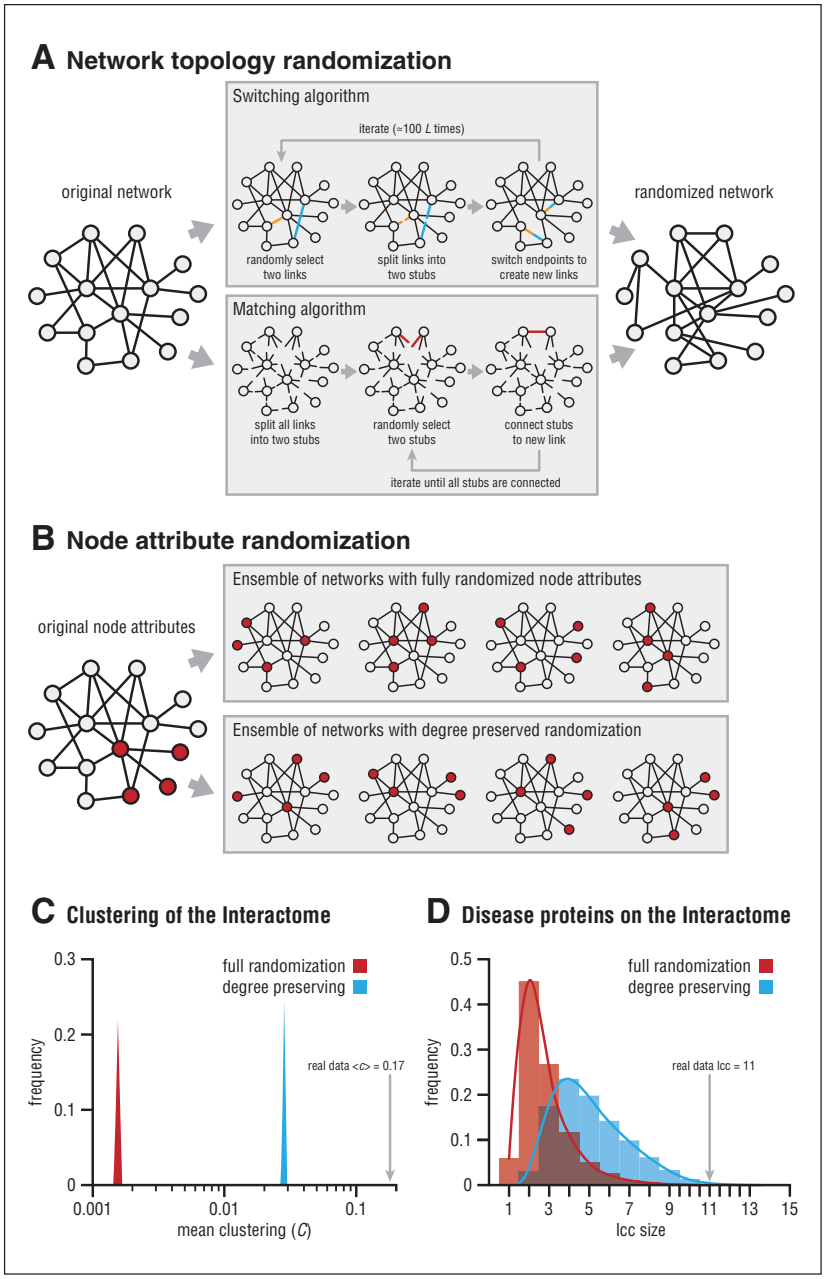

FIGURE 2–8. Network randomization. A, Two algorithms frequently used to randomize the topology of a network while preserving the degrees of the individual nodes. B, Randomizing node attributes, for example, proteins associated with a particular disease with and without preserving the original degrees of the nodes. C, Comparison of the clustering coefficient of the interactome with values obtained by complete randomization and degree-preserving randomization. D, The size of the largest connected component (lcc) of proteins associated with multiple sclerosis in the interactome and compared with the values from randomization according to B.

—-1
—0
—+1

works, this effect can usually be neglected for large N (Milo, Nashtan, et al. 2004). Figure 2–8C shows the distribution of the mean clustering coefficient obtained from 10,000 randomized versions of the interactome. The mean value $<C>=0.03$ is considerably larger than for the fully randomized network, accounting for the influence of the degree distribution, yet it is still smaller than the real value $<C>=0.17$ for the interactome, indicating that the observed high clustering coefficient could not have emerged by chance.

Randomization can also be designed to preserve other topological features of a network. For example, some algorithms generate randomized networks that preserve the clustering coefficient of the original network (Serrano, Boguna, et al. 2005) or the correlations between the degrees of neighboring nodes (Boguna and Pastor-Satorras 2003; Weber and Porto 2007). In metabolic networks, simple link rewiring would generate biochemically unrealistic reactions. Therefore, we need to use more involved procedures that generate only biochemically valid reactions (Basler, Ebenhoh, et al. 2011; Samal and Martin 2011).

Randomizing Node Properties

Randomization of the network topology is primarily used to identify the impact of the network topology on the system's behavior. To explore the network location of a specific group of nodes, we often need to keep the network fixed and randomize the identity or the location of the nodes. We use this method, for example, when we test whether proteins associated with a particular disease have more connections among themselves than expected by chance.

Random label permutation. The simplest approach is to distribute the node attributes of interest randomly on the network (see Figure 2–8B). For instance, to investigate the connectivity patterns of N_d disease proteins, the same number of proteins are selected randomly from the network, and the quantity of interest is measured for this set of randomized nodes. Repeating this procedure will yield a distribution that can be used as a random control against which the statistical significance of the original quantity can be tested. For example, multiple sclerosis has $N_d=69$ known associated proteins in the interactome, forming a largest connected component of size $S=11$. Figure 2–8D shows the size distribution of the largest connected component for 69 randomly chosen proteins obtained from 10,000 simulations. The mean random expectation is $\langle S_{\text{rand}}^{\text{full}} \rangle = 2.9$ with a standard deviation $\sigma = 1.4$.

The statistical significance of the observed size can be quantified using its z-score:

$$\text{z-score} = \frac{S - \langle S_{\text{rand}}^{\text{full}} \rangle}{\sigma}, \tag{2–13}$$

yielding $z = 5.8$. The empirical p value (i.e., the fraction of all random simulations with $\langle S_{\text{rand}}^{\text{full}} \rangle$ S, is $p = 0.003$. As z-scores above 1.65, corresponding to a p value under 0.05 for normal distributions, are considered highly significant, we conclude that the connected component for multiple sclerosis could not have emerged by chance, indicating the potential presence of a disease module.

Degree-preserving label permutation. Similar to the randomization of the network topology, we can introduce additional constraints when reshuffling the node labels. One could argue, for example, that the high number of connections among disease proteins is a result of their relatively high degree. To test this hypothesis, we swap the node labels only between nodes of the same or comparable degree (see Figure 2–8B). Yet, we may have very few or even only one node with a particular high degree. It is, therefore, often useful to relax the requirement of an exact match of the degrees for a label swap and, instead, divide the degrees into bins of different degrees and swap the node characteristics within each bin. Figure 2–8D shows the distribution $\langle S_{\text{rand}}^{\text{degree}} \rangle$ obtained using such binned degree-preserving randomization. The mean value $\langle S_{\text{rand}}^{\text{degree}} \rangle = 5.1$ is increased compared to the full randomization. Yet, the actual value is still significantly higher ($z = 3.1$, empirical p-value $= 0.009$), indicating that the high degree of the disease proteins alone cannot account for the observed large component size.

Perspectives and Further References

In this chapter we could discuss only the most frequently used quantities in network science. For a deeper and broader discussion, we refer the reader to the online Network Science textbook (http://barabasi.com/book/network-science) and other reviews (Albert and Barabási 2002; Newman 2003; Dorogovstev, Goltsev, et al. 2008; Newman 2010; Walhout, Vidal, et al. 2013, Buchanan et al. 2010). Network science is a very active field of research, with new tools emerging

daily. In the following, we highlight a few recent developments that might also provide useful insight for the study of diseases.

Layered networks. As we have seen above, biological systems exhibit different levels of organization, each of which is best described as a separate network (see Figure 2–1). These networks are, however, not independent of each other, but can be considered as networks of networks. Such layered or interdependent networks exhibit a number of interesting phenomena, for example, concerning their stability toward perturbations (Buldyrev, Parshani, et al. 2010; Gao, Buldyrev, et al. 2011). The interdependence between different layers of the network can give rise to cascading failure, where the breakdown of a node in one layer propagates throughout all other layers, leading to a global breakdown.

Temporal networks. The networks we have considered here are essentially static in nature—that is, the nodes and their interactions do not change over time. They capture the biochemical skeleton of all interactions that are chemically possible. This is, of course, a simplified view—for example, proteins are not transcribed at all times and molecular interactions may or may not occur depending on internal or external signals. The rapidly evolving field of temporal networks aims to incorporate these dynamic aspects of networks and to explore the impact of this temporality on its structural and dynamic characteristics (Przytycka, Singh, et al. 2010; Holme and Saramaki 2012). Schulz, Pandit, et al. (2013), for example, used time-sequenced expression data of protein-coding genes and miRNAs to construct a dynamic network to predict the most explanatory factors for changes in expression over time.

Conclusion

An important issue in the study of biological networks boils down to a single question: Can we *control* them? In the past few years there have been a series of rigorous results to address network controllability (Liu, Slotine, et al. 2011, 2013). In systems of biochemical reactions, for example, it has been found that by monitoring a few selected nodes one can infer the complete state of the entire system (Liu, Slotine, et al. 2013). These results could have immediate application in the rational design of biomarkers for disease states, as well as in rational drug

target(s) selection. The ultimate goal is to control these systems, that is, to drive a cell from a disease state to a healthy state (Liu, Slotine, et al. 2011).

References

Aerts, S., D. Lambrechts, et al. (2006). Gene prioritization through genomic data fusion. Nat Biotechnol **24**(5): 537–44.

Ahn, Y. Y., J. P. Bagrow, et al. (2010). Link communities reveal multiscale complexity in networks. Nature **466**(7307): 761–64.

Albert, R., and A. L. Barabasi (2002). Statistical mechanics of complex networks. Rev Mod Phys **74:** 47–97.

Albert, R., H. Jeong, et al. (1999). Internet: diameter of the world-wide web. Nature **401**(6749): 130–31.

Albert, R., H. Jeong, et al. (2000). Error and attack tolerance of complex networks. Nature **406**(6794): 378–82.

Annibale, A., and A. C. Coolen (2011). What you see is not what you get: how sampling affects macroscopic features of biological networks. Interface Focus **1**(6): 836–56.

Ashburner, M., C. A. Ball, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25**(1): 25–29.

Bailly-Bechet, M., C. Borgs, et al. (2011). Finding undetected protein associations in cell signaling by belief propagation. Proc Natl Acad Sci U S A **108**(2): 882–87.

Barabási, A. L., and R. Albert (1999). Emergence of scaling in random networks. Science **286**(5439): 509–12.

Basler, G., O. Ebenhoh, et al. (2011). Mass-balanced randomization of metabolic networks. Bioinformatics **27**(10): 1397–403.

Bender, E. A. and E. R. Canfield (1978). The asymptotic number of labeled graphs with given degree sequences. J Combinat Theory **24, Series A**(3): 296–307.

Boguna, M. and R. Pastor-Satorras (2003). Class of correlated random networks with hidden variables. Phys Rev E Stat Nonlin Soft Matter Phys **68**(3 Pt 2): 036112.

Bollobas, B. (1979). Random graphs. In: B. Bollobas, Graph Theory. New York: Springer, pp. 123–45.

Buchanan, M., Caldarelli, G., De Los Rios, P., Rao, F. and Vendruscolo, M., 2010. Networks in Cell Biology. Networks in Cell Biology, by Mark Buchanan, Guido Caldarelli, Paolo De Los Rios, Francesco Rao, Michele Vendruscolo, Cambridge, UK: Cambridge University Press, 2010, p. 1.

Buldyrev, S. V., R. Parshani, et al. (2010). Catastrophic cascade of failures in interdependent networks. Nature **464**(7291): 1025–28.

Callaway, D. S., M. E. Newman, et al. (2000). Network robustness and fragility: percolation on random graphs. Phys Rev Lett **85**(25): 5468–71.

Cohen, R., K. Erez, et al. (2000). Resilience of the internet to random breakdowns. Phys Rev Lett **85**(21): 4626–28.

Cohen, R., and S. Havlin (2003). Scale-free networks are ultrasmall. Phys Rev Lett **90**(5): 058701.

Davidson, E., and M. Levin (2005). Gene regulatory networks. Proc Natl Acad Sci U S A **102**(14): 4935.

Dezso, Z., Y. Nikolsky, et al. (2009). Identifying disease-specific genes based on their topological significance in protein networks. BMC Syst Biol **3:** 36.

Dorogovstev, S. N. (2003). Evolution of Networks: From Biologial Nets to the Internet and WWW. New York: Oxford University Press.

Dorogovstev, S. N., A. V. Goltsev, et al. (2008). Critical phenomena in complex networks. Rev Mod Phys **80**(4): 1275–335.

Erdős, P. (1959). On random graphs I. Publ Math Debrecen **6:** 290–97.

—-1
—0
—+1

Erdős, P., and A. Rényi (1960). On the evolution of random graphs. Publ Math Inst Hung Acad Sci **5:** 17–61.
Erten, S., G. Bebek, et al. (2011). DADA: degree-aware algorithms for network-based disease gene prioritization. BioData Min **4:** 19.
Forster, J., I. Famili, et al. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res **13**(2): 244–53.
Fortunato, S. (2010). Community detection in graphs. Physics Rep **486**(3): 75–174.
Franke, L., H. van Bakel, et al. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet **78**(6): 1011–25.
Freeman, L. C. (1977). A set of measures of centrality based on betweenness. Sociometry **40:** 35–41.
Gao, J., S. V. Buldyrev, et al. (2011). Networks formed from interdependent networks. Nature Phys **8**(1): 40–48.
George, R. A., J. Y. Liu, et al. (2006). Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res **34**(19): e130.
Girvan, M., and M. E. Newman (2002). Community structure in social and biological networks. Proc Natl Acad Sci U S A **99**(12): 7821–26.
Goh, K. I., M. E. Cusick, et al. (2007). The human disease network. Proc Natl Acad Sci U S A **104**(21): 8685–90.
Guimera, R., and M. Sales-Pardo (2009). Missing and spurious interactions and the reconstruction of complex networks. Proc Natl Acad Sci U S A **106**(52): 22073–78.
Guney, E., and B. Oliva (2012). Analysis of the robustness of network-based disease-gene prioritization methods reveals redundancy in the human interactome and functional diversity of disease-genes. PLoS One **9**(4): e94686.
Holme, P., and J. Saramaki (2012). Temporal networks. Physics Rep **519**(3): 97–125.
Hutz, J. E., A. T. Kraja, et al. (2008). CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genet Epidemiol **32**(8): 779–90.
Ideker, T., V. Thorsson, et al. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science **292**(5518): 929–34.
Köhler, S., S. Bauer, et al. (2008). Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet **82**(4): 949–58.
Krauthammer, M., C. A. Kaufmann, et al. (2004). Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. Proc Natl Acad Sci U S A **101**(42): 15148–53.
Liu, Y. Y., J. J. Slotine, et al. (2011). Controllability of complex networks. Nature **473**(7346): 167–73.
Liu, Y. Y., J. J. Slotine, et al. (2013). Observability of complex systems. Proc Natl Acad Sci U S A **110**(7): 2460–65.
Managbanag, J. R., T. M. Witten, et al. (2008). Shortest-path network analysis is a useful approach toward identifying genetic determinants of longevity. PLoS One **3**(11): e3802.
Maslov, S., and K. Sneppen (2002). Specificity and stability in topology of protein networks. Science **296**(5569): 910–13.
Menche, J., A. Sharma, et al. (2015). Disease networks: uncovering disease-disease relationships through the incomplete interactome. Science **347**(6224): 1257601.
Milo, R., N. Nashtan, et al. (2004). On the uniform generation of random graphs with prescribed degree sequences. [arXiv:cond-mat/0312028 [**cond-mat.stat-mech**]
Milo, R., S. Shen-Orr, et al. (2002). Network motifs: simple building blocks of complex networks. Science **298**(5594): 824–27.
Morrison, J. L., R. Breitling, et al. (2005). GeneRank: using search engine technology for the analysis of microarray experiments. BMC Bioinformatics **6:** 233.
Newman, M. E. (2003). The structure and function of complex networks. SIAM Rev **45**(2): 167–256.
Newman, M. E. (2010). Networks: An Introduction. New York: Oxford University Press.
Newman, M. E., S. H. Strogatz, et al. (2001). Random graphs with arbitrary degree distributions and their applications. Phys Rev E Stat Nonlin Soft Matter Phys **64**(2 Pt 2): 026118.

-1—
0—
+1—

Palla, G., I. Derenyi, et al. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043): 814–18.
Przytycka, T. M., M. Singh, et al. (2010). Toward the dynamic interactome: it's about time. Brief Bioinform **11**(1): 15–29.
Ravasz, E., A. L. Somera, et al. (2002). Hierarchical organization of modularity in metabolic networks. Science **297**(5586): 1551–55.
Rolland, T., M. Tasan, et al. (2014). Expansion of the human interactome landscape by a second-generation proteome-wide map. Cell **159:** 1212–26.
Rual, J. F., K. Venkatesan, et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature **437**(7062): 1173–78.
Samal, A. and O. C. Martin (2011). Randomizing genome-scale metabolic networks. PLoS One **6**(7): e22295.
Schulz, M. H., K. V. Pandit, et al. (2013). Reconstructing dynamic microRNA-regulated interaction networks. Proc Natl Acad Sci U S A **110**(39): 15686–91.
Serrano, M. A., M. Boguna, et al. (2005). Competition and adaptation in an Internet evolution model. Phys Rev Lett **94**(3): 038701.
Sharma, A., J. Menche, et al. (2015). A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. Hum Mol Genet. **24**(11): 3005–20.
Stelling, J., S. Klamt, et al. (2002). Metabolic network structure determines key aspects of functionality and regulation. Nature **420**(6912): 190–93.
Stelzl, U., U. Worm, et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell **122**(6): 957–68.
Stumpf, M. P., and C. Wiuf (2010). Incomplete and noisy network data as a percolation process. J R Soc Interface **7**(51): 1411–19.
Stumpf, M. P., C. Wiuf, et al. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. Proc Natl Acad Sci U S A **102**(12): 4221–24.
Tranchevent, L. C., F. B. Capdevila, et al. (2011). A guide to web tools to prioritize candidate genes. Brief Bioinform **12**(1): 22–32.
Vanunu, O., O. Magger, et al. (2010). Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol **6**(1): e1000641.
Venkatesan, K., J. F. Rual, et al. (2009). An empirical framework for binary interactome mapping. Nat Methods **6**(1): 83–90.
Walhout, M., M. Vidal, et al. (2013). Handbook of Systems Biology: Concepts and Insights. Oxford, U.K.: Elsevier.
Wasserman, S., and K. Faust (1994). Social network analysis. Cambridge, U.K.: Cambridge University Press.
Weber, S., and M. Porto (2007). Generation of arbitrarily two-point-correlated random networks. Phys Rev E Stat Nonlin Soft Matter Phys **76**(4 Pt 2): 046111.
Zhong, Q., N. Simonis, et al. (2009). Edgetic perturbation models of human inherited disorders. Mol Syst Biol **5:** 321.
Zhou, X., J. Menche, et al. (2014). Human symptoms-disease network. Nat Commun **5:** 4212.

—-1
—0
—+1