Journal of Theoretical Biology 422 (2017) 18–30

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

Modeling and analysis of modular structure in diverse biological networks

Bader Al-Anzi[a,1,*], Sherif Gerges[e,f,g,1], Noah Olsman[b,1], Christopher Ormerod[c,d,1], Georgios Piliouras[h,1], John Ormerod[i], Kai Zinn[a,*]

[a] Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125 USA
[b] Control and Dynamical Systems Option, Division of Engineering and Applied Sciences, California Institute of Technology Pasadena, CA 91125 USA
[c] Division of Physics, Mathematics, and Astronomy, California Institute of Technology, Pasadena, CA 91125 USA
[d] Department of Mathematics and Statistics, University of Maine, Orono, ME 04469-5752, USA
[e] Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142 USA
[f] Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
[g] Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA
[h] Singapore University of Technology and Design, Engineering Systems and Design (ESD), 8 Somapah Road, 487372 Singapore
[i] School of Mathematics and Statistics F07, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 15 October 2016
Revised 28 March 2017
Accepted 4 April 2017
Available online 8 April 2017

ABSTRACT

Biological networks, like most engineered networks, are not the product of a singular design but rather are the result of a long process of refinement and optimization. Many large real-world networks are comprised of well-defined and meaningful smaller modules. While engineered networks are designed and refined by humans with particular goals in mind, biological networks are created by the selective pressures of evolution. In this paper, we seek to define aspects of network architecture that are shared among different types of evolved biological networks. First, we developed a new mathematical model, the Stochastic Block Model with Path Selection (SBM-PS) that simulates biological network formation based on the selection of edges that increase clustering. SBM-PS can produce modular networks whose properties resemble those of real networks. Second, we analyzed three real networks of very different types, and showed that all three can be fit well by the SBM-PS model. Third, we showed that modular elements within the three networks correspond to meaningful biological structures. The networks chosen for analysis were a proteomic network composed of all proteins required for mitochondrial function in budding yeast, a mesoscale anatomical network composed of axonal connections among regions of the mouse brain, and the connectome of individual neurons in the nematode *C. elegans*. We find that the three networks have common architectural features, and each can be divided into subnetworks with characteristic topologies that control specific phenotypic outputs.

1. Introduction

Complex networks underlie much of biological function. These networks exist at every scale: from the individual interactions of proteins, to the connectivity of neurons, up to entire populations of organisms. While it is clear that networks must encode much of the information that allows for the robustness and diversity found in life, we still lack a detailed understanding of the relationships that link network structure to biological function. Researchers have amassed a vast store of knowledge about individual genes, proteins, and cells, but it is still prohibitively complex to simultaneously analyze more than a handful of these components in any substantive detail. These difficulties have stimulated efforts to analyze biological networks through computational modeling based on graph theory concepts (Barabasi, 2013; Barabasi and Oltvai, 2004; Newman, 2003). These models have provided insights into both the nature of individual interactions and the overall architecture of complex biological systems (Barabasi and Albert, 1999; Clauset et al., 2008; Erdos, 1960; Holland et al., 1983; Watts and Strogatz, 1998).

While many models capture key aspects of networks, such as the distribution of connectivity, the density of clustering, or modular structure, it has proven difficult to devise simple models that accurately recapitulate all of the features of experimental networks (Albert and Barabasi, 2002). As seen in Fig. 1A, networks with, for

* Corresponding authors.
E-mail addresses: baderalanzi13@gmail.com (B. Al-Anzi), zinnk@caltech.edu (K. Zinn).
[1] These authors contributed equally.

http://dx.doi.org/10.1016/j.jtbi.2017.04.005
0022-5193/

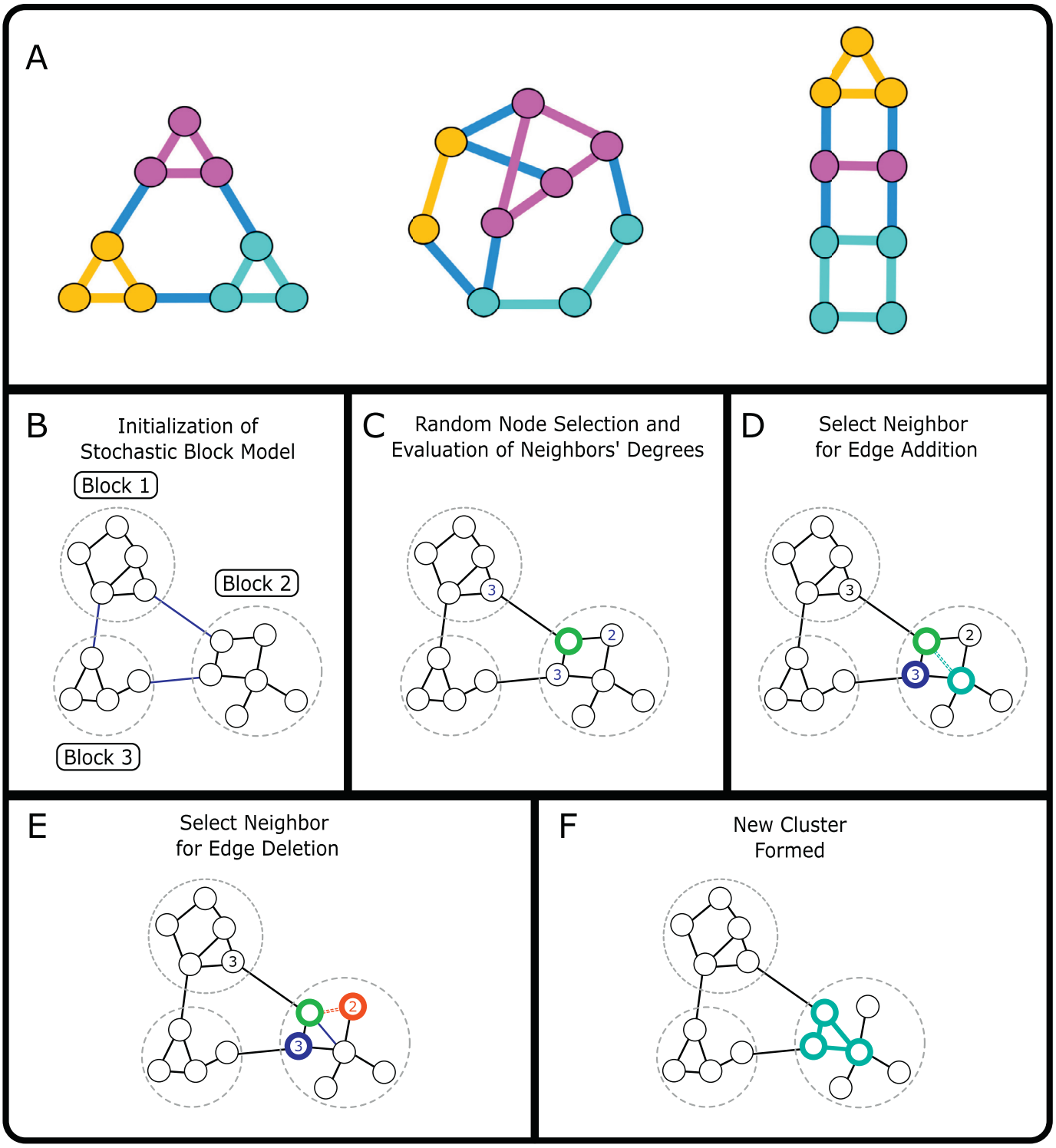

Fig. 1. The SBM-PS model. (A) Three graphs that share an identical degree distribution, but are structurally different. The first has three modules and three clusters, the second has no modules and no clusters, and the third has two modules and one cluster. This illustrates the importance of analyzing multiple properties of a graph to more completely understand its structure. (B-F) Cartoon representation of the PS algorithm. (B) An SBM with three blocks. (C) A random node (green outline) is selected; it is connected to two nodes of degree 3 and one node of degree 2. (D) One of these neighbor nodes (blue outline) is selected with a probability proportional to its degree. Then, a new connection (dotted turquoise line) is created between a random neighbor of the selected node (turquoise outline) and the original node (green outline). (E) Another neighbor node (red outline) is selected with a probability inversely proportional to its degree, and its connection (dotted red line) to the original node (green outline) is deleted. (F) This creates a new cluster (turquoise triangle), but the total number of edges in the network is unchanged.

example, identical degree distributions can differ greatly with regard to other structural properties like clustering and modularity.

In this paper, we propose a new model for generating networks that have the modular structures and clustering characteristic of real biological networks. Our model is appealing because it begins with a simple structure, namely a Stochastic Block Model (SBM) (Holland et al., 1983), and evolves in accordance with a biologically motivated mechanism that we call Path Selection (PS). This selection mechanism relies only on the local fitness of biological components.

We tested the model by using it to simulate real biological networks that were of interest to us. In doing this, we chose three networks of very different types for analysis, so that if SBM-PS was able to accurately simulate all of them we could be confident that the model is of general utility. Further, we sought out networks for which we could be confident that most biologically relevant connections had been experimentally captured, so that the data secured on each network was as faithful a representation of the real-world network as possible.

The three networks we chose that fit these criteria were as follows: the network of physical interactions among proteins involved in mitochondrial function in the budding yeast *Saccharomyces cerevisiae* (this work), the mesoscale network of connections among regions of the mouse brain (Oh et al., 2014), and the network of all connections among individual neurons in *Caenorhabditis elegans* (nematode) (Sulston and Horvitz, 1977; Sulston et al., 1983). Each of these is a highly interconnected network containing several hundred nodes. However, the three networks are functionally and physically disparate, one being molecular and two being anatomical. The two anatomical networks are on very different scales. The mouse network consists of brain areas, while the worm network is composed from single neurons. The networks are derived from organisms separated by at least 500 million years of evolution.

Because these networks are of similar size but very different in character, we hoped that topological attributes found to be common to all three networks might reflect fundamental architectural features of biological network design. The development of a computational model capable of both capturing and reproducing these features should advance our understanding of the processes by which biological networks are generated. The three networks are graphically displayed in Fig. 2, in an organization that displays the communities of which they are comprised. As described below, we tested several well-known computational models for their ability to fit the experimental networks. We found that while other models were capable of achieving good fits to individual topological parameters, the SBM-PS model best matched these experimental networks when accounting for all parameters simultaneously. It will be of interest in the future to determine whether models based on the SBM-PS algorithms can also be used to simulate other types of biological networks, including gene regulatory networks and metabolic networks.

We chose the three specific networks we examine here not only because they are attractive targets for modeling, but also because we have interests in the areas of biology that they represent. We previously published several papers on fat regulation, and analyzed a network of yeast proteins involved in control of fat storage (Al-Anzi et al., 2015). This network is a subset of the mitochondrial network examined here (see below for further discussion). We have also studied networks of interacting neuronal cell surface proteins involved in assembly of neural connectomes (Carrillo et al., 2015; Özkan et al., 2013). In studying these networks, our specific interests lie in understanding how their structures allow them to produce biological outputs (phenotypes). These phenotypes include mitochondrial function, morphology and inheritance (in the case of the yeast mitochondrial network), and specific animal behaviors (in the cases of the mouse and worm neural networks).

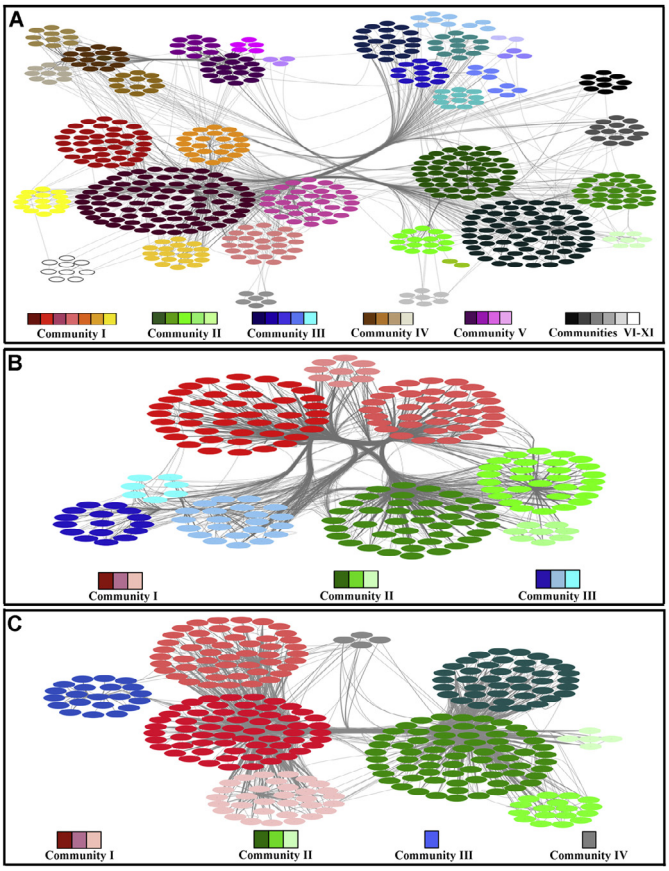

Fig. 2. Modular structures of large biological networks. Here we see graphical representations that highlight the modular structures of the three biological networks considered in this paper: (A) the proteomic network of proteins required for mitochondrial function in budding yeast (B) The mesoscale connectome of the mouse brain, and (C) the network of individual neurons in *C. elegans*. These networks are composed of large communities that contain smaller sub-communities. Community and sub-community boundaries were delineated using a walk-trap algorithm (see Results for details). Connections between nodes (edges) are shown by light gray lines. Nodes that belong to the same large community are represented by different shades of a particular color (see color bars at bottom of each panel), while nodes that belong to the same sub-community have the same shade. The identities of nodes (corresponding to proteins in the yeast mitochondrial network, brain regions in the mouse brain, and single neurons in *C. elegans*) are shown in Fig. 2 supplement 1.

In order to obtain insights into the mechanisms by which the networks generate these diverse outputs, we decided to further characterize the networks in order to examine relationships between network structure and biological function. We used a bioinformatics approach to determine the biological properties of network nodes based on their physical location, the broad biological processes that utilize them, and the impact of experimental alterations in their functions on organismal physiology or behavior. We found that nodes within communities (collections of interconnected modules) tend to share a common location within the same compartment of the cell (for the yeast network) or region of the brain or organism (for the mouse brain and *C. elegans* connectomes). We also constructed subnetworks consisting of protein nodes involved in specific phenotypes (for the yeast network) or neural nodes (brain regions or individual neurons) involved in specific behaviors. We found that these 'phenotypic subnetworks' consist of interconnected nodes distributed among different modules, and they are characterized by clustering coefficients higher than those of random subnetworks of equivalent size. The combination of modeling and analysis in this work provides insights into the

origins of the observed structures in these networks, potentially yielding insights into the ways in which evolution has shaped the structure of proteomic networks and neural connectomes.

2. Results and discussion

2.1. Overview of network models

We begin by discussing several simple graph theoretic models of network generation, primarily focusing on the topological properties that arise from them when they are fit to experimental networks. Since we are trying to examine biological networks that are functionally and physically disparate from one another and perform different functions, it is necessary to focus on properties that are in common. These will be largely structural features selected for their ability to facilitate function. They include: degree distribution *P(k)*, diameter *D* (also known as path length), global clustering coefficient *C*, and modularity *M*. We define *M* as the tendency of a group of nodes to have more connections among themselves than with nodes outside the group. Collectively, these terms encompass the fundamental properties of network architecture and are frequently used to describe real-world networks (Deville et al., 2014; Henderson and Robinson, 2011; Muldoon et al., 2015; Oh et al., 2014). We did not include edge directionality, because most of the protein-protein interactions in one of the experimental networks we examine here are not directional.

While there are many models of network formation, we decide to focus primarily on those that are relatively simple. Because our goal was to generate a model that encompasses many biological processes, we did not want to limit our search to context-specific mechanisms. The models we consider here are the Erdős–Rényi (ER) model (Erdos, 1960), the Watts–Strogatz (WS) model (Watts and Strogatz, 1998), the Barabási–Albert (BA) model (Barabasi and Albert, 1999), the Hierarchical Random Graph (HRG) model (Clauset et al., 2008), and The Stochastic Block Model (SBM) (Holland et al., 1983).

While the ER model is not intended to underpin a plausible mechanism for generating biological networks, the comparison is useful for determining how distinct these networks are from purely random networks. The WS and BA models were designed to replicate two different features of real-world networks not present in ER models. The WS model is a simple model that produces high clustering coefficients like those found in real-world networks (Watts and Strogatz, 1998), and we used this model in earlier work to analyze a proteomic network in budding yeast composed of 94 proteins required for fat storage regulation (Al-Anzi et al., 2015). We had defined this network experimentally by using a genetic screen to identify all nonessential yeast genes required for maintenance of normal fat levels. Because mitochondria are central players in synthesis and metabolism of fat, the mitochondrial network discussed in the present paper includes almost all of the proteins we had previously identified as members of the fat storage regulation network. In our earlier paper (Al-Anzi et al., 2015), we had found that the fat storage regulation network could be fit well by the WS model. However, as discussed below, when we analyzed the much larger mitochondrial network (of which the fat regulation network is a subset), WS and other existing models provided poor fits to some of its key topological parameters.

The BA model uses a preferential attachment mechanism to generate 'hub' nodes with many connections, which are often observed in real-world networks (Barabasi and Albert, 1999). The HRG model does not directly model a particular underlying mechanism, but instead creates a network that matches the hierarchical structure observed in an empirical network (Clauset et al., 2008). The SBM has been used to model the modular structure of real-world networks (Holland et al., 1983).

While these network models are not the most sophisticated or current models available, their popularity and capacity to capture key topological features of our experimental networks makes them useful as benchmarks in testing any new model. For example, we see in Fig. 3 that the ER model can fit the diameter of the experimental networks well, but does poorly on clustering and modularity. Similarly, the WS model can be optimized for clustering and exhibits high modularity, but its degree distribution is not realistic. Alternatively, the BA model may fit the degree distribution of a given experimental network well but does poorly with respect to clustering and modularity. The HRG does even better than the BA at fitting the degree distribution, but also has issues capturing modularity and clustering. The SBM does well at capturing modularity, but does not yield a good fit to the rest of the parameters (Fig. 4).

Here we propose a new model, SBM-PS, that aims to capture all of the statistical properties we observe in real biological networks. We begin with a SBM representation of a modular biological network. We then apply a simple algorithm, called path selection, which models evolutionary selection pressures. With this algorithm, the edges that connect nodes undergo selection, so that edges that increase the connectivity of the graph are preferentially created, while edges that do not contribute to the overall connectivity are likely to decay (Fig. 1B–F). This happens because each edge selection step removes one edge and adds one edge to the graph. The added edge is always part of a triangle, while the removed edge may or may not have been part of a triangle. If the network is sparse (as the networks we are analyzing are), a randomly selected removed edge is relatively unlikely to have been part of a triangle. The edge selection process thus increases the number of triangles in the graph on average. At a local level, minimizing the shortest path between three connected nodes is best achieved when they are completely connected in a closed triangular formation, and this is the basic unit used for *C* measurements. When applied to a highly modular network, the PS algorithm thus produces stronger clustering. It also generates a realistic degree distribution and a propensity to form hubs.

2.2. Generation of SBM-PS networks

Let k be the number of communities in a network G, C_i and p_i respectively be the number of nodes and edge probability for community i, and ω be the intercommunity edge probability. These parameters on their own are sufficient to specify a SBM, which as mentioned earlier does a poor job of directly reproducing most of the topological features observed in the experimental networks, even with a large number of fitted parameters. To remedy this, we now describe the PS edge selection algorithm (Fig. 1B-F).

The primary parameters of the PS algorithm are the number of iterations, denoted T, and a weighting factor, denoted ε, which is a parameter in the range $0 < \varepsilon < 1$. At each iteration of PS, we choose a node g uniformly at random from the network. We then iterate through each neighbor n_i of the node and assign it two weights w_i and d_i,

$$w_i = (1+\varepsilon)^{\deg(n_i)},$$

$$d_i = (1-\varepsilon)^{\deg(n_i)},$$

where $\deg(n_i)$ is the degree of node n_i. These weights can then be normalized to compute probabilities,

$$p_i = \frac{w_i}{\sum w_i},$$

$$q_i = \frac{d_i}{\sum d_i}.$$

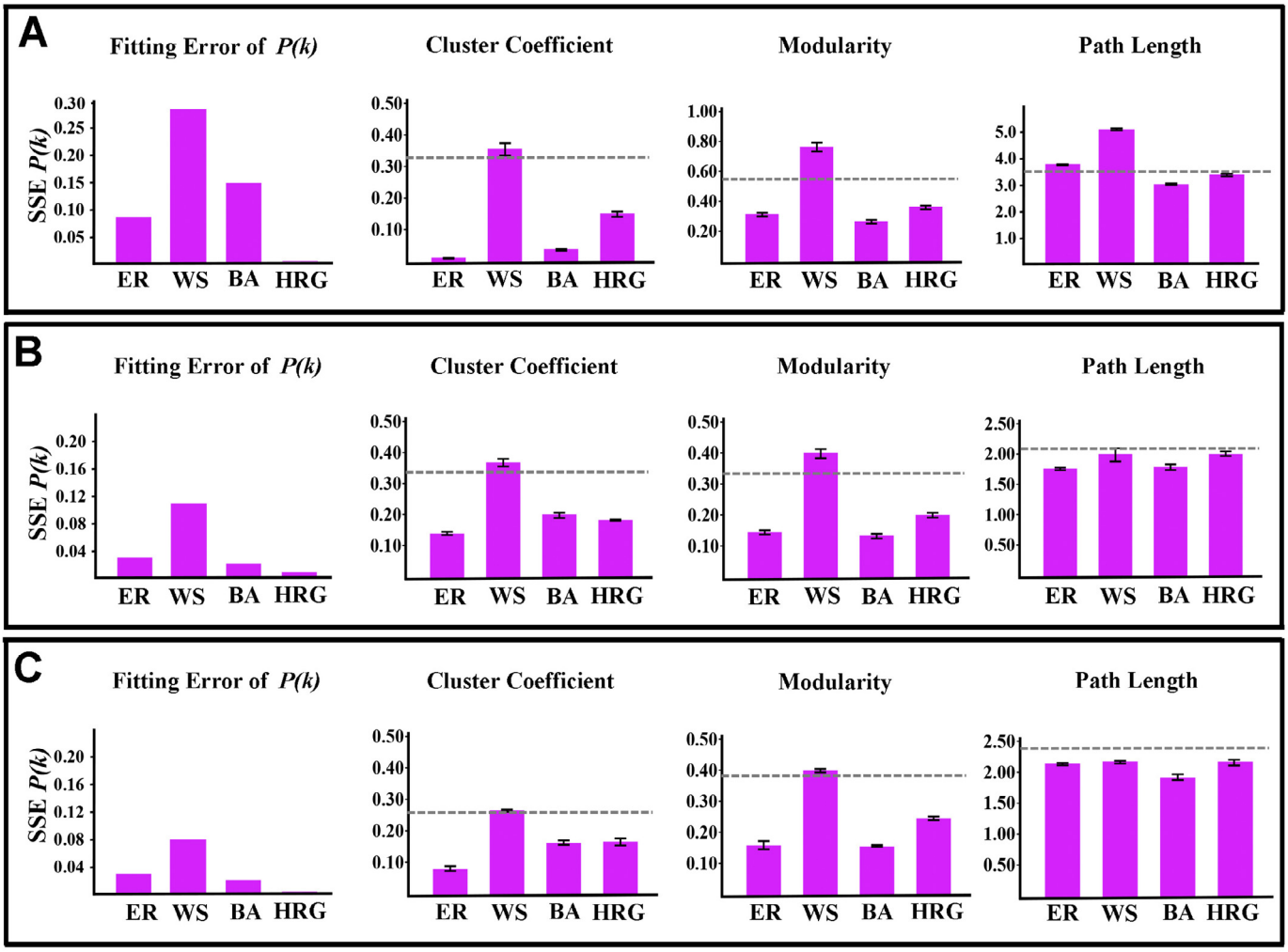

Fig. 3. Best-fit performance of existing network models. In each bar chart we display the performance of the ER, WS, BA, and HRG when fitting a given network model. The dotted lines indicate the measured values of the *C*, *M*, and *D* (path length) parameters for the experimental networks. (A) The yeast mitochondrial network. Only the HRG achieves a near-zero sum-of-squares error (SSE) when fitting to the degree distribution of the yeast network. Only the WS model does well for the clustering coefficient, and none do particularly well to match modularity. ER and HRG provide the best fits to the *D* (path length) parameter. (B) The mesoscale connectome of the mouse brain. All the models provide a better fit to the degree distribution than for the yeast mitochondrial network. Again, HRG provides the best fit. WS does best on modularity and clustering. (C) the *C. elegans* connectome. As in the previous two cases, the HRG does best in fitting the degree distribution and the WS does best on clustering and modularity. Detailed statistical analysis is provided in Supplementary file 5.

Intuitively we can see that, when n_i has a large number of connections relative to the other neighbors of g, p_i is large and q_i is small, and vice versa when n_i has a relatively small number of connections. The first step is to pick a neighbor n_i at random according to the distribution set by the p_i. We then pick a neighbor h uniformly at random from n_i and add an edge between h and g if one does not already exist. If there already is an edge there, we randomly select again until we find an h not already connected to g. If no such h exists, we terminate and move to the next iteration. Since g was already connected to n_i, which is already connected to h, this new connection between g and h must result in a new cluster in the network. Next, we pick another neighbor n_i at random according to q_i and delete a random edge connected to n_i.

Fig. 4A-C show fits of SBM and the SBM-PS to each of the three experimental networks we previously examined. The fitting procedure is described in *Materials and Methods*. While it is expected that the SBM will provide modularity and the PS algorithm will improve clustering, it is not obvious that the combination of the two will not degrade all of the structure in the graph. Surprisingly, when the SBM is evolved under the PS algorithm, clustering is greatly improved with little cost to modularity.

The fitting error for the *P(k)* distribution is also reduced for SBM-PS relative to the original SBM. This is because the three experimental networks have hub-like elements, as reflected by the relatively good fits of their degree distributions to the power-law *P(k)* distribution generated by the BA algorithm (Fig. 3), and the PS algorithm tends to increase the number of connections made by hub-like (highly connected) nodes at the expense of those made by less connected nodes. For the other parameters, *M* and *D*, the SBM already provides a good fit.

We estimated the empirical probabilities that the quantitative properties of the yeast mitochondrial, mouse brain, and *C. elegans* networks are as extreme as or more extreme than the simulated networks for each model. We combined these probabilities using Fisher's and Stouffer's method to produce a single combined probability for each network that reflects the probability that network could be generated by each of the models (Davison and Hinkley, 1997; Fisher, 1925; Wasserman, 2004). Lastly, we considered fitting a multivariate normal distribution to each of the four quantitative

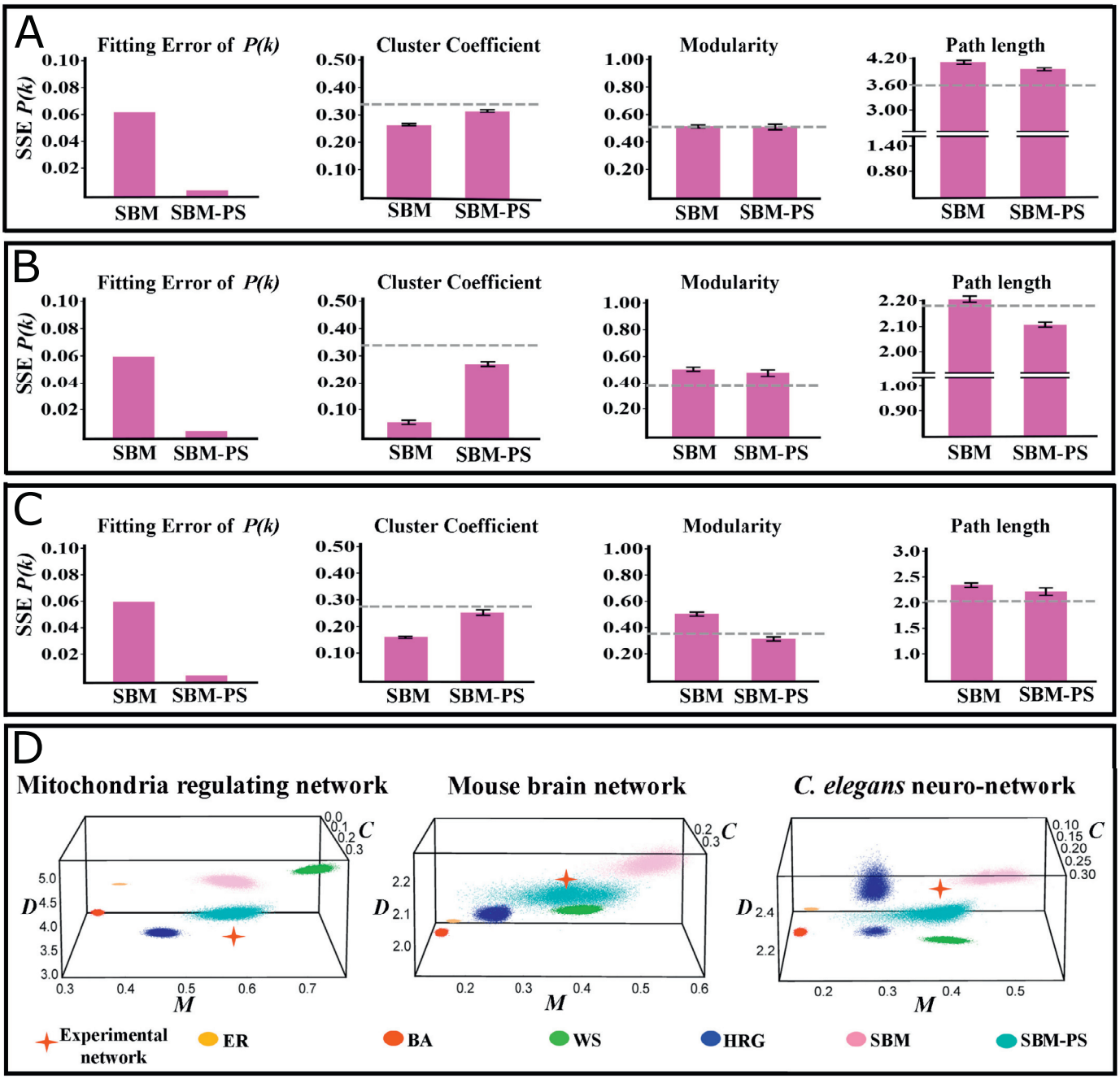

Fig. 4. The SBM-PS model captures the key topological features of all three large biological networks. (A-C) The path selection algorithm greatly improves the fit of an SBM to the three networks: (A) yeast mitochondrial network; (B) mesoscale connectome of the mouse brain; (C) *C. elegans* connectome. (D) A 3-D representation of the distributions of possible values of *C*, *M*, and *D* for the various network models. Each colored region is a cloud of points, representing the possible values of these three parameters that can be produced by each of the six models considered in this paper. The red crosses mark the actual values for the experimental networks. Note that the cross for the mouse brain network is immediately adjacent to the turquoise cloud for the SBM-PS network, indicating this model fits the experimental network excellently. For the mitochondrial network and the *C. elegans* connectome, the cross is closer to the SBM-PS cloud than to any other model clouds. Detailed statistical analysis is provided in Supp. Table 5.

properties. While each of the resulting combined probabilities are low, it is clear that the probability that the three networks could have been generated by the SBM-PS is several orders of magnitude higher than for any competing model. Supp. Table 5 shows the result of this analysis. It demonstrates that SBM-PS produces an excellent global fit to all three experimental networks and outperforms all of the other tested models.

A key aspect of the SBM-PS is that, while the PS algorithm results in changes to the global structure of the network, it relies only on local information from each node. This use of local information is characteristic of evolution in real systems, as no gene knows the global structure of the genome and no region of the brain knows the global structure of the connectome. The PS algorithm is a simple model that can produce networks whose structures resemble those of the real networks examined here, with strong clustering, high modularity, and the correct degree distributions. SBM-PS may be able to produce such networks better than the other models considered here because it focuses on the prop-

erties of edges rather than those of nodes, selecting for edges that increase clustering. The connections in real networks that correspond to such edges should facilitate communication among network nodes and would likely be important for network function. A connection that does not contribute to effective communication within the network would likely correspond to an edge that would be deleted by the PS algorithm, because it would join two nodes of low degree. Thus, while the PS algorithm may not directly model processes that occur during biological evolution, it produces effects like those of natural selection, preserving valuable connections and allowing unimportant ones to decay.

2.3. Overview of the experimental networks

The results from the previous section show that it is possible to generate realistic modular networks from a simple model. We now wished to analyze the biological relevance of the topological parameters that were used as benchmarks for evaluation of the ability of SBM-PS to simulate the three real networks we are examining here. To explain this analysis, we first need to describe the these networks in more detail.

As stated in the Introduction, the three networks have several features that make them suitable targets for modeling. We generated the yeast mitochondrial network using bioinformatics. We first cataloged all *S. cerevisiae* genes for which loss of function (LOF) had been shown to produce abnormal or dysfunctional mitochondria in published studies. We then used global proteomic data to generate a network representing all molecular interactions among the proteins (nodes) encoded by these 883 genes (Al-Anzi et al., 2015; Altmann and Westermann, 2005; Breitkreutz et al., 2010; Gavin et al., 2002; Ho et al., 2002; Ito et al., 2001; Kanki et al., 2011; Merz and Westermann, 2009; Ptacek et al., 2005; Tarassov et al., 2008; Uetz et al., 2000) (Fig. 2A, Supp. Fig. 1). About 85% of these proteins have physical connections to one another. The majority of the connections (edges) in the network (~80%) were identified by co-precipitation methods, which detect stable and abundant protein complexes that exist in vivo (Gavin et al., 2002; Ho et al., 2002). Because the proteins in the network were identified by multiple genetic screens of the entire yeast genome, including essential genes (Altmann and Westermann, 2005), it is unlikely to be missing a large number of nodes. Furthermore, all nodes in the network share the common property that their removal affects mitochondrial morphology or function.

We evaluated the biological significance of the network's edges by examining the ability of randomly selected sets of equal numbers of proteins from the yeast genome to form a network *via* their annotated proteomic interactions. All such networks had less than half as many edges as the mitochondrial network, and they all had low *C* values (Fig. 2, Supp. Fig. 2). Since all of the randomly selected networks have many fewer edges than the actual yeast mitochondrial network and do not have high clustering, these results suggest that the interconnectedness and clustering of the yeast mitochondrial network was probably selected for during evolution of the yeast cell.

Since ~20% of the protein-protein connections (edges) in the mitochondrial network were defined not by co-precipitation, but through yeast two-hybrid or biochemical phosphorylation assays, this raises the possibility that these edges are qualitatively different from the majority of edges in the network and should not be grouped with them in a topological analysis. We thus recalculated *C, D*, and *M* for a network composed only of edges defined by co-precipitation, but did not observe any dramatic changes relative to the parent network. *C* changed from 0.33 to 0.38, while *M* changed from 0.58 to 0.65, and *D* from 3.62 to 3.86. This suggests that it is appropriate to include all biochemically defined protein-protein connections as edges in the network when analyzing its architecture.

It will be of interest in future studies to examine the relative abilities of the models described above to accurately simulate larger-scale protein-protein interaction networks, such as the network of interactions among all *S. cerevisiae* proteins. In this study, we chose to model the smaller mitochondrial network, rather than the larger global network of which it is a subset, for several reasons. First, we were concerned that modules within the global network might not be uniquely defined as specific patterns of connections, since the global network encompasses all the functions of the cell. In a cell, there may be many proteins that utilize different partners and different modules for different biological processes (e.g., protein A might bind to protein B within module 1 to regulate the cell cycle but to protein C in module 2 to regulate fat storage). We hoped that the proteins in the mitochondrial network would be more likely to function within single modules, since the network has a more restricted set of functions. Second, by restricting the size of the analyzed network to several hundred proteins, all of which were known to be involved in mitochondrial function, the analysis of its communities and modules was more tractable. Third, to evaluate the capacities of the models to simulate different types of networks (molecular vs. anatomical), we wanted to use networks of roughly similar sizes (several hundred nodes in each case).

The mouse brain mesoscale connectome was produced by The Allen Institute for Brain Science by injecting a recombinant adeno-associated virus (AAV) expressing EGFP as an axonal anterograde tracer into the mouse brain (Oh et al., 2014) (Fig. 2B, Supp. Fig. 1). This procedure allows mapping of axonal pathways connecting the area injected with the AAV tracer with all other brain areas. The tracer was applied to 295 non-overlapping anatomical regions that span most of the mouse brain (Oh et al., 2014). Only 18 areas in the brain were not labeled due to problems with tracer injections. The original paper on the mouse brain mesoscale connectome examined whether it could be simulated by existing network models such as BA and WS (Oh et al., 2014). A recent paper published while our work was in progress described a generative model for the mesoscale connectome based on a simple algorithm that was designed to capture features such as clustering and degree distribution (Henriksen et al., 2016). This model is based on two principles for formation of edges: source growth and proximal attachment. While the Henriksen et al. model does an excellent job of matching topological parameter values and overall architecture for the mouse brain network, its mechanism is specific for this type of network, in that it is designed to mimic how connections between brain regions may actually form during development, for example through neurotrophic signaling.

The adult *C. elegans* hermaphrodite contains 302 individual neurons. The axons of these neurons and the synaptic connections among them were defined using electron microscopic (EM) reconstruction of entire animals (Altun, 2011; Sulston and Horvitz, 1977; Sulston et al., 1983) (Fig. 2C, Supp. Fig. 1).

2.4. Relationships between network community structure and spatial localization of nodes

In order to analyze the architectures of the yeast mitochondrial, mouse brain and *C. elegans* networks, we partitioned them into communities using a walk-trap algorithm (Newman Fast Community Finding Algorithm), an established method for community detection (Pons, 2006). This method is based on the assumption that short random walks from one node to another will tend to stay in the same community. One of the shortcomings of this approach is that the behavior of the walk-trap algorithm can be sensitive to small changes in the network. That is, the removal of single

nodes could potentially alter the community assignment for a large number of nodes. We addressed this by subjecting the community structure obtained to a stability analysis. Our results indicate that nearly all nodes in all three networks stay in a given community under small changes in the network, indicating a relatively high degree of confidence that all our nodes belong in their assigned community (Supplementary Table 4).

Our analysis indicates that the yeast mitochondrial network is composed of five large communities (each with more than 20 nodes) and seven smaller ones. This organization is graphically displayed in Fig. 2A, which represents the outcome of this community analysis (see also Supplementary Table 1). The mouse brain mesoscale connectome is divided into three communities (Fig. 2B, Supplementary Table 2). Finally, the *C. elegans* connectome contains three large communities (each with more than 20 nodes) and one small community (four nodes) (Fig. 2C, Supplementary Table 3).

In examining the networks, we found that large communities appeared to be preferentially made from nodes that are proximal to one another. We define proximity for proteins as localization within the same subcellular compartment. For brain regions, proximity is defined as localization within the same division of the brain. For individual neurons in *C. elegans*, which lacks a central brain, we define proximity as localization within the same ganglion.

We reached this conclusion by assigning each node to a given cellular or neuronal compartment based on published literature (Altun, 2011; Cherry et al., 2012; Watson et al., 2012). These results are depicted in the diagrams of Fig. 5 and Supp. Fig. 3). For example, in the case of the yeast mitochondrial network, almost all protein nodes in communities II, IV, and V are localized to the mitochondrial compartment, while most proteins in community III are located in the nucleus (Fig. 5A; Supplementary Table 1). In the mouse brain mesoscale connectome, most nodes (brain areas) in community II are in the limbic system. Community I is split between the isocortex and the brain stem, while community III is split between the cerebellum and brain stem (Fig. 5B; Supplementary Table 2). Similar observations were reported by Henriksen et al. in their analysis of the mouse brain mesoscale connectome (Henriksen et al., 2016). In the *C. elegans* connectome, almost all nodes (neuronal cell bodies) in community I are in the head ganglia, while most nodes in community II are in the tail ganglia and the ventral nerve cord (Fig. 5C; Supplementary Table 3).

To determine if the tendency of nodes within the same large community to be located in the same compartment is statistically significant, we calculated the probabilities that the observed distributions of nodes among compartments in the three networks could be generated by chance. Out of 21 discrete structural groups, 17 had p-values based on a χ-squared test of less than 10^{-3}, implying that the community structure is statistically significant.

The large communities in the mouse brain network perform diverse biological tasks and do not belong to a single functional category. Similarly, communities I, II, and III in the yeast mitochondrial network span functional categories. However, communities IV and V correspond to the large and small mitochondrial ribosomal subunit proteins, respectively, so they do have unified functions. For the *C. elegans* connectome, each large community has multiple functions. However, community I has many neurons with sensory functions, while community II is enriched in neurons with motor functions.

When each of the large communities in the three networks were separated from the rest of the network and subjected to further walk-trap analysis, we noted that they consist of smaller sub-communities composed of nodes that are even more proximal to one another either as result of being in the same molecular complexes (in the case of the yeast mitochondrial network), the same brain structures (in the case of the mouse brain network), or the same ganglia (in the case of the *C. elegans* connectome) (Supplementary Tables 2–4). Furthermore, unlike the larger communities, these smaller sub-communities, which we denote as modules, often exhibit unified functions that can be utilized by different biological processes. One example is the mitochondrial sub-community I, G. which is largely composed of the SCF ubiquitin-protein ligase complex and its substrates (Kamura et al., 2002) known to be involved in glucose detection and cell cycle regulation. Another example is the mouse brain sub-community II, B, which is composed of amygdalar structures that are required for memory, decision making, and emotional reactions (Watson et al., 2012). Finally, worm connectome sub-community I, B is composed of amphid chemosensory neurons, which are

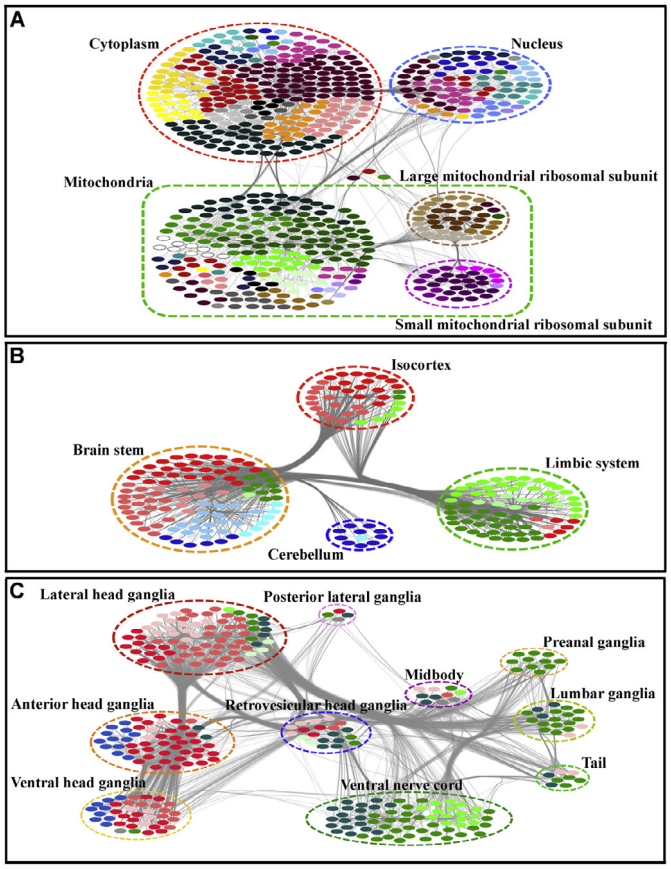

Fig. 5. Modular structure is related to physical location in the network. (A) In the yeast mitochondrial network, we divided proteins according to their known localization patterns within the cell. The large and small mitochondrial ribosomal subunits are within mitochondria (mitochondrial compartments enclosed by green dotted line) but are considered separately from the rest of the mitochondrion because they contain discrete sets of modules that are distinct from other mitochondria protein modules. The five physical locations delineated here correlate strongly with modules identified by the walk-trap algorithm (modules indicated by unique dot colors; see Fig. 2). (B) In the mouse brain connectome, we separated brain regions into four location groups. Each location group is composed mostly of nodes within particular communities and modules identified by the walk-trap algorithm, except for the brain stem, which contains a mixture of nodes from elements from all communities and modules. (C) In the *C. elegans* connectome, we separated neurons by the locations of their cell bodies within different ganglia. again shows strong proximal relationships between nodes within a walk-trap module. Each ganglion is composed mostly of neurons within specific communities and modules identified by the walk-trap algorithm. The color code indicating community and sub-community identity is the same as in Fig. 2. Light gray lines delineate edges. Identities of nodes are indicated in Fig. 5, supplement 1. All location data are shown in sheet 2 s of Supplementary files 2–4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

utilized in aggregation/dispersion behavior and chemosensory responses (Altun, 2011).

2.5. Structures of subnetworks involved in phenotypic outputs

The phenotypic outputs of a node within a protein network can be evaluated by knocking out the gene that encodes it, and examining the resultant phenotype. For example, if the deletion of a particular gene causes loss of the electrical potential gradient across the mitochondrial membrane, we can infer that the protein product of that gene is involved in some way in maintenance of the gradient. The products of genes involved in a phenotypic output usually do not have the same functions, and need not belong to the same module, since many different kinds of proteins are required for each physiological process. Thus, the phenotypic outputs to which a given set of proteins contribute cannot always be predicted based on knowledge of their biochemical functions.

The phenotypic outputs of nodes within neuronal networks can be examined by lesioning neurons or altering their activity and observing the effects of these perturbations on specific behaviors. For mouse brain neurons. this would be done by surgical or pharmacological lesions of a given area, or by using viral vectors to alter neural activity of neurons within a given area. The *C. elegans* connectome nodes correspond to single neurons. These can be laser-ablated, or their activity can be selectively altered using transgenic methods, and the behavior of the perturbed nematodes can be examined. We assembled collections of nodes involved in various phenotypic outputs from the three networks in an unbiased manner by listing all yeast mitochondrial network proteins, mouse brain areas, and worm neurons that had been associated with those outputs in published literature (Al-Anzi et al., 2015; Altmann and Westermann, 2005; Breitkreutz et al., 2010; Gavin et al., 2002; Ho et al., 2002; Ito et al., 2001; Kanki et al., 2011; Merz and Westermann, 2009; Ptacek et al., 2005; Tarassov et al., 2008; Uetz et al., 2000)(Supplementary Tables 2–4).

We observed that many nodes are involved more than one type of phenotypic output. For example, the MRPL9 protein is involved in both voltage generation and inheritance of the mitochondria (Merz and Westermann, 2009). In the mouse brain, the lateral hypothalamus regulates both feeding behavior and predator responses (Watson et al., 2012). In the worm, the RMGL and RMGR neurons function in egg-laying, feeding behavior, and CO_2/O_2 detection responses (Milward et al., 2011).

We rarely observe a situation in which all participating nodes in a phenotypic output are members of the same module. However, we find that nodes involved in the same phenotypic output tend to be connected to one another, and to form subnetworks with specific topological features. These 'phenotypic subnetworks' have high *C* values, as compared with both Erdos-Renyi and randomized subset controls, suggesting that they were generated by evolutionary selection. These comparisons are shown in the bar graphs of Figs. 6–8. However, phenotypic subnetworks often have *M* values that are less than those of one or both of the control subnetworks or of the communities within the network.

In Fig. 6, we show diagrams for four phenotypic subnetworks for the yeast mitochondrial network. Fig. 7 shows four phenotypic subnetworks, representing animal behaviors, for the mouse brain connectome, and Fig. 8 shows four for the *C. elegans* connectome. The figure supplements for each of these figures show the same diagrams, but with the names of components included. These phenotypic subnetworks are of very different sizes and span different numbers of sub-communities (modules). Fig. 6 shows that mitochondrial inheritance involves nodes that are dispersed across 21 sub-communities, while mitophagy uses nodes that are largely, but not exclusively, within a single sub-community. Similarly, in Fig. 8, we see that aggregation/dispersion behavior involves nodes within seven sub-communities, while feeding behavior involves a much smaller number of nodes that are restricted to four sub-communities. Lists of all detected phenotypic subnetworks along with their topological parameters are provided in Supplementary Tables 2–4.

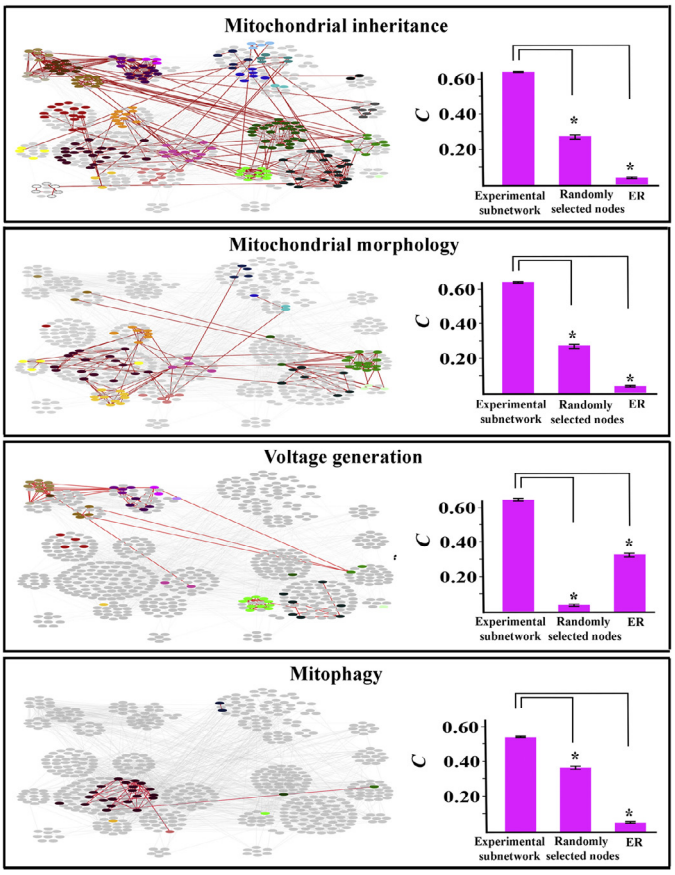

Fig. 6. Example phenotypic subnetworks for the yeast mitochondrial network. Nodes (proteins) implicated in four network functions are superimposed in color on a gray background corresponding to the entire network diagram from Fig. 2A. Proteomic connections (edges) within the phenotypic subnetworks are indicated by red lines. These diagrams show the phenotypic subnetworks of nodes implicated in mitochondrial inheritance, mitochondrial morphology, voltage generation, and mitophagy in published papers (see references in main text). The color code indicating community and sub-community identity is the same as in Fig. 2A. Identities of nodes are indicated in Fig. 6, supplement 1. The histograms on the right side compare *C* for each phenotypic subnetwork to those of networks made by a random selection of the same number of nodes from the entire network, or by an ER model with the same number of nodes. Note that the *C* values for each of the phenotypic networks are much larger than those for the corresponding random or ER networks. $P < 0.0001$ for all indicated comparisons (brackets). Nodes in all subnetworks are listed in Supplementary file 2, sheet 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article

Taken together, these results indicate that, although the walktrap method is not sensitive to edge length, it is capable of capturing proximal modular structure in all of these biological networks. We also find that phenotypic outputs of the networks are products of interplay among nodes in different modules that are connected to form subnetworks with characteristic topological features, such as high *C* values relative to random networks of the same size. The existence of such characteristic features indicates that phenotypic subnetworks are products of selection, and may serve as statistical markers for physiologically relevant functions. However, phenotypic subnetworks cannot yet be defined *via* computational methods alone, and experimental approaches are still required for identification of nodes within these subnetworks.

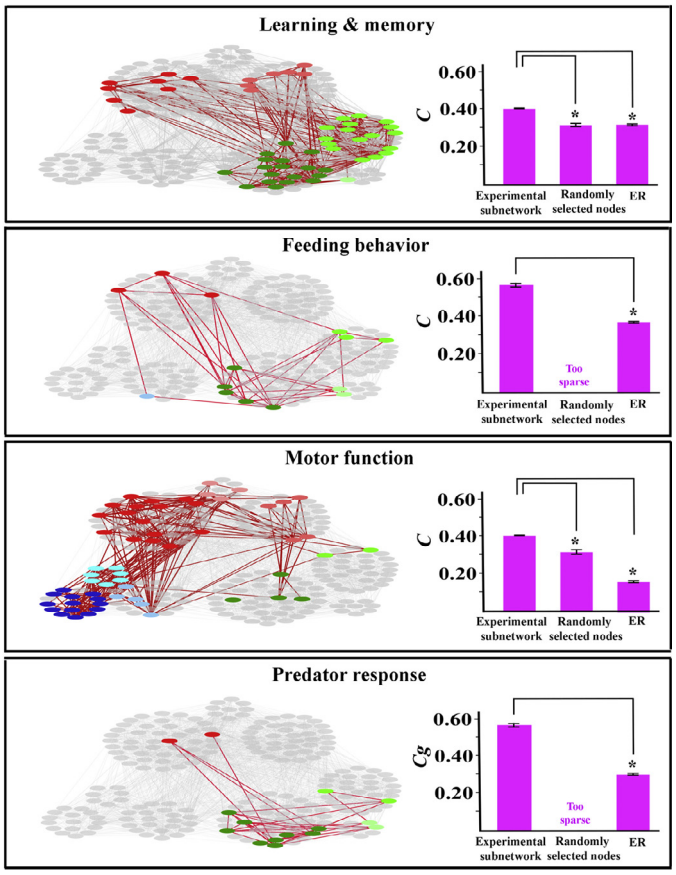

Fig. 7. Example phenotypic subnetworks for the mesoscale connectome of the mouse brain. Nodes (brain regions) implicated in four brain functions are superimposed in color on a gray background corresponding to the entire network diagram from Fig. 2B. Axonal connections (edges) within the phenotypic subnetworks are indicated by red lines. These diagrams show the phenotypic subnetworks of nodes implicated in learning and memory (A), feeding behavior (B), motor function (C), and predator responses (D) in published papers (see references in main text). The color code indicating community and sub-community identity is the same as in Fig. 2B. Identities of nodes are indicated in Fig. 7, supplement 1. The histograms on the right side compare C for each phenotypic subnetwork to those of networks made by a random selection of the same number of nodes from the entire network, or by an ER model with the same number of nodes. "Too sparse" indicates that the randomly selected network had too few connections to allow a calculation of its C value. Note that the C values for each of the phenotypic networks are much larger than those for the corresponding random or ER networks. $p < 0.0001$ for all indicated comparisons (brackets). Nodes in all subnetworks are listed in Supplementary file 3, sheet 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

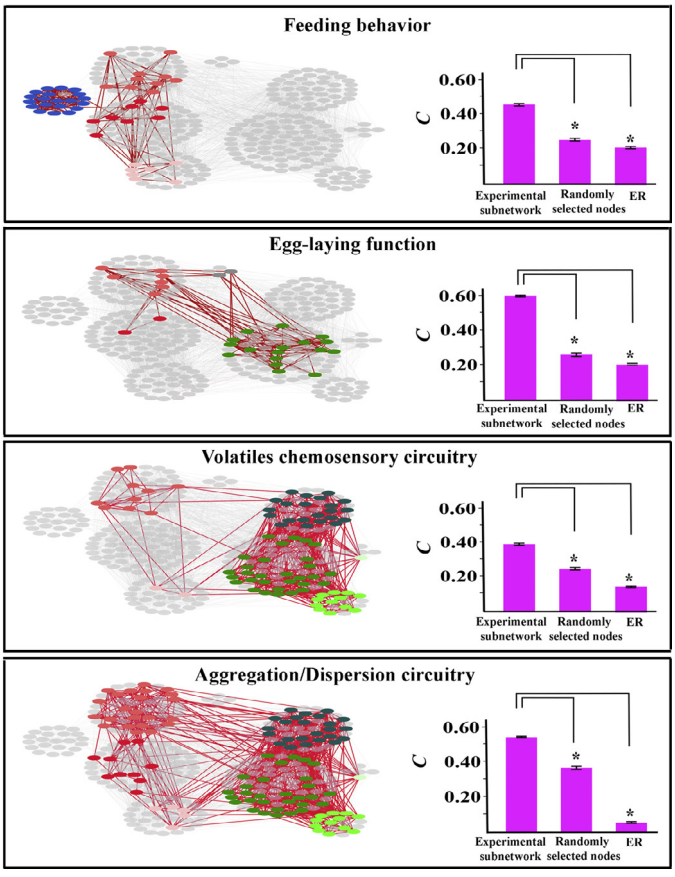

Fig. 8. Example phenotypic subnetworks for the *C. elegans* connectome. Nodes (neurons) implicated in four nervous system functions are superimposed in color on a gray background corresponding to the entire network diagram from Fig. 2C. Axonal/synaptic connections (edges) within the phenotypic subnetworks are indicated by red lines. These diagrams show the phenotypic subnetworks of nodes implicated in feeding behavior (A), egg-laying (B), volatile chemosensation (C), and aggregation and dispersion (D) in published papers (see references in main text). The color code indicating community and sub-community identity is the same as in Fig. 2C. Identities of nodes are indicated in Fig. 8, supplement 1. The histograms on the right side compare C for each phenotypic subnetwork to those of networks made by a random selection of the same number of nodes from the entire network, or by an ER model with the same number of nodes. Note that the C values for each of the phenotypic networks are much larger than those for the corresponding random or ER networks. $p < 0.0001$ for all indicated comparisons (brackets). Nodes in all subnetworks are listed in Supplementary file 4, sheet 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

3. Conclusions

Our work provides two approaches to the analysis of the architecture of biological networks. First, we show that a simple generative model, SBM-PS, is capable of reproducing many of the topological features we find to be common across three divergent networks that evolved through different evolutionary mechanisms. Second, we find that a representation of these complex, dynamic networks as a simple graph is capable of retaining important biological features such as node location and phenotypic subnetworks.

It is of interest that networks with such divergent characteristics can be fit by a common algorithm. The yeast mitochondrial network is a proteomic network of molecules within a single cell. Its nodes are proteins and its edges are noncovalent binding interactions among the proteins. The *C. elegans* connectome is an anatomical network of single neurons, in which edges are axonal/synaptic connections among these neurons. The mesoscale connectome of the mouse brain is also an anatomical neural network, but its nodes are brain regions, each of which contains hundreds or thousands of neurons. Its edges are axonal bundles that connect these brain regions.

We hope that the methods described here can be extended to make them useful for analysis of other types of biological networks, including metabolic, ecological, and gene regulatory networks. Another potential line of future research would be to turn the analysis of phenotypic subnetworks into a diagnostic tool. It might eventually be possible to use computational methods to select candidate phenotypic subnetworks from a larger network. This would be done by searching for collections of genes, neurons, or brain regions that have particular values of key topological parameters. Then, one could search for phenotype(s) controlled by these subnetworks by examining whether there are common effects caused by the removal of any one of its nodes.

4. Materials and methods

4.1. Topological parameters

Here we define the basic metrics we will use to parameterize them. To start, we consider each network as a graph consisting of undirected, unweighted nodes and edges. The biological meaning of the nodes will depend on the context of the network, for example they could represent proteins, cells, or large groups of cells. Edges represent connections between components, e.g. protein-protein interactions or neuronal connectivity. The degree k of a node is defined to be the number of connections associated with it. From this, we can the degree distribution $P(k)$ of a graph, where $P(k)$ is equal to the probability that a node in the graph has degree k. We note that the degree distribution does not uniquely determine a graph, and that it is common for graphs with the same degree distribution to have extremely different topological properties.

Next, we define the diameter of a graph D. Between two nodes in a graph, there are generally many paths that can be taken to reach one from the other, the distance of which is defined as the number of edges traversed in the process. For a given pair of nodes, there will be a (not necessarily unique shortest path. The diameter is defined to be the maximum over the set of all shortest paths in the graph, and is infinite if the graph is not connected. This can be thought of as the worst-case direct traversal time for a given network.

The properties described so far, $P(k)$ and D, give us sense for general properties of a graph, but do not give much insight into the overall structure. To measure connectivity in a more detail way, we now introduce two additional metrics: the global clustering coefficient and the modularity of a graph. At a high level, the global clustering coefficient C measures the density of connection in a graph. Specifically C counts the number of clusters (the number of closed triangles in the graph) relative to the number of connected triplets (sets of three nodes with at least two edges),

$$C = \frac{3 \times number\ of\ triangles}{number\ of\ triplets},$$

where the factor of 3 occurs because each triangle is made up of three distinct triplets. Intuitively, one can see that a graph with very dense connections is more likely to have higher clustering than one that is sparse. For a completely connected graph we have $C = 1$, and for a graph with no closed loops (e.g. a tree graph) we have $C = 0$.

Next we describe the modularity of a graph M. This essentially quantifies the degree to which a graph can be separated into distinct modules, which are groups of nodes that are more highly connected to each other than to other nodes. The modularity for a given community division is related to the difference between that fraction of edges that start and end within a module and the fraction of edges that span modules. The modularity of the entire graph is then determined by the division into modules which maximizes this value.

Network Models

Now we describe three simple models of network generation that are often used as benchmarks against which empirical networks can be tested. Figs. 3 and 4 show that our experimental networks are characterized by high clustering and modularity, despite being sparsely connected overall. This hints that these features are subject to selection, and as we will see it is difficult to achieve both a high clustering and modularity while maintaining a degree distribution that does not correspond to a high density of connections.

Since it is often impossible to actually uncover the precise mechanism that determined structure of a particular network, we instead use models of network generation to see what types of networks generate topological features that match what we observe in biology. The simplest of these models is the Erdős–Rényi (ER) model. The formulation of this model we consider takes in two parameters, the total number of nodes N and the probability p that a given pair of nodes has an edge between them. The ER model has a binomial degree distribution and is a good model for networks with little internal structure as it is not likely to display high modularity or clustering. We see from Fig. 3 that none of our experimental networks are well fit by the ER model.

The Watts–Strogatz (WS) model is more sophisticated than the ER model in that it is designed to yield a graph that is highly structured. It begins with N nodes connected in a ring (i.e. each node has degree 2). Then, for each edge we randomly rewire it with probability p such that one end node is randomly replaced with another. This model was originally designed to generate graphs with a Small World property (i.e. small diameter), essentially looking at what happens to networks that are trying to optimize communication time between nodes. While it accomplishes this using a very elegant and simple mechanism, it is not able to capture the sort of clustering that is observed in many real-world networks (see Fig. 3).

The Barabási–Albert (BA) model aims at capturing a different feature of networks, namely having a scale-free degree distribution, otherwise known as a power law distribution ($P(k) \propto k^{-\beta}$ for some $\beta > 0$. The BA model relies on the mechanism of preferential attachment, where nodes are added to an existing graph and are connect to a node with a probability proportional to that node's degree. This means that highly connected nodes are mostly like to become even more connected. This leads to a few 'hub' nodes that are central to the network, and many lower degree nodes that are not nearly as well connected. While it appears that some real-world networks exhibit a power-law distribution, we actually find that all three of our experimental networks are poorly fit by the BA model (see Fig. 3).

The Stochastic Block Model (SBM) is yet another simple model designed to capture a particular aspect of real-world networks, with the focus now on modularity. The SBM is parametrized by a number r that determines how many communities are in the graph and C_i for $i \in [1, r]$, where C_i is the number of nodes in community i. Each community is treated as an independent ER graph with connectivity determined by the connection probability p_{in} (this can be generalized such that there is a different probability for each community). The communities are then connected to each other by a second connection probability p_{out}. While the SBM performs very well in fitting modularity, it does a poor job of capturing the level of clustering we measure in our experimental networks because it is essentially a collection of ER models, none of which have particularly high clustering on their own (see Fig. 4).

The Hierarchical Random Graph (HRG) model is qualitatively different from the rest of the models discussed so far, in that it is based on an underlying hierarchical structure of the network. Essentially nodes have some notion of relatedness, where more related nodes are more likely to share an edge. This hierarchical structure is directly fitted from the empirical network that the HRG is meant to resemble (Clauset et al., 2008).

Fitting Procedures

Here we describe the procedure for fitting each model to an empirical network. Since many of these models do not necessarily have a unique method for fitting, our approach was to first fit the number of nodes in the graph, then the number of edges, and then if necessary fit to other topological features like clustering coefficient if there were still parameters that had not yet been specified.

ER Model: For the ER model, fitting was performed simply by generating a graph with same number of nodes and edge density as the empirical network.

BA Model: Since the BA model has a process of growth and start with a single node, we simply specify that it terminates after N iterations, where N is the number of nodes in the empirical network. Every time a new node is added, it creates m new edges that link it randomly to the rest of the graph. To match the edge density, we simply set m such that Nm is approximately equal to the number of edges in the empirical network.

WS Model: The WS model begins with a regular lattice, with N nodes each with degree d, that is then randomly rewired. Specifically we iterate through each node and rewire each edge with probability p to a new node chosen uniformly at random from the graph. As before N is set to be equal to the number of nodes in the empirical network and d is set such that $Nd/2$ is approximately equal to the number of edges in the network. Fitting p is somewhat more difficult, as it does not directly correspond to a particular measured parameter of the graph. Because of this, we simulate 10^4 WS models with a given value of p and average their cluster coefficient. We then search over p to find a value that gives the closest match to the empirical networks clustering coefficient.

HRG: The HRG models were fitted to a hierarchy corresponding to a given empirical network. To do this, used a software package called igraph developed by the original authors Clauset et al. to generate random networks. The documentation can be found at http://igraph.org/r/. The R code used for implementation of the HRG model is documented here: https://github.com/sherifgerges/Hierarchical_Random_Graphs

SBM: The number of communities was chosen to coincide with the major communities outputted by the walk-trap algorithm run on the empirical network, while the relative probabilities where chosen so that the expected edge density in each community was equal to the observed edge density of the community, this defines the edge probability p_i for each community and then the probability that edges form between communities was chosen so that the total expected edges was that of our total network.

SBM-PS: Given a fitted SBM, we had two parameters to choose; the number of times we would sample a node, denoted T, and the weighting factor, denoted ε. Firstly, the weighting factor was chosen to be of the order of 0.1. Given the dependence of w_i and d_i on ε is given by $w_i = (1+\varepsilon)^{\deg(n_i)}$ and $d_i = (1-\varepsilon)^{\deg(n_i)}$, and given typical values of deg (n_i) observed in each of our models, a value of ε of the order of 0.1 ensured that the values of w_i and d_i are clearly differentiated by the degrees of the nodes. The number of times we sampled nodes roughly coincided with the number of edges present in the model with some variance based on the experimentally observed clustering. Values of 4000, 5000 and 2000 were used for the neural network of the C. elegans, yeast mitochondrial network and mesoscale mouse brain network models respectively.

Quantitative data for fitting of each network by each of the 6 models is tabulated in Supplementary file 5. Code for SBMs for the three networks is in Supplementary file 6.

Acknowledgments

This work was supported by a grant from the NIH, R21NS083874, to K. Z., and by the Della Martin Foundation. We also acknowledge NVIDIA Corporation for generously donating the NVIDIA GTX980 Graphics Card used in this study. Georgios Piliouras would like to acknowledge SUTD grant SRG ESD 2015 097 and MOE AcRF Tier 2 Grant 2016-T2-1-170. Part of the work was completed while Georgios Piliouras was a Wally Baer and Jeri Weiss postdoctoral scholar in the Department of Computing & Mathematical Sciences at Caltech.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jtbi.2017.04.005.

References

Al-Anzi, B., Arpp, P., Gerges, S., Ormerod, C., Olsman, N., Zinn, K., 2015. Experimental and computational analysis of a large protein network that controls fat storage reveals the design principles of a signaling network. PLoS Computat. Biol. 11, e1004264.
Albert, R., Barabasi, A.L., 2002. Statistical mechanics of complex networks. Rev. Modern Phys. 74, 47–97.
Altmann, K., Westermann, B., 2005. Role of essential genes in mitochondrial morphogenesis in Saccharomyces cerevisiae. Mol. Biol. Cell. 16, 5410–5417.
Altun, Z.F.A.H., D.H (2011). Nervous system, general description. In WormAtlas doi:103908/wormatlas118.
Barabási, A.L., 2013. Network science. Philos. Trans. A Math. Phys. Eng. Sci. 371, 20120375.
Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.
Barabási, A.L., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113.
Breitkreutz, A., Choi, H., Sharom, J.R., Boucher, L., Neduva, V., Larsen, B., Lin, Z.Y., Breitkreutz, B.J., Stark, C., Liu, G., et al., 2010. A global protein kinase and phosphatase interaction network in yeast. Science 328, 1043–1046.
Carrillo, R.A., Ozkan, E., Menon, K.P., Nagarkar-Jaiswal, S., Lee, P.T., Jeon, M., Birnbaum, M.E., Bellen, H.J., Garcia, K.C., Zinn, K., 2015. Control of synaptic connectivity by a network of Drosophila IgSF cell surface proteins. Cell 163, 1770–1782.
Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., et al., 2012. Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res. 40, D700–D705.
Clauset, A., Moore, C., Newman, M.E., 2008. Hierarchical structure and the prediction of missing links in networks. Nature 453, 98–101.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press, Cambridge; New York, NY, USA.
Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V.D., Barabási, A.L., 2014. Career on the move: geography, stratification, and scientific impact. Sci. Rep. 4, 4770.
Erdos, P.R., 1960. On the evolution of random graphs. B. Int. Statist. Inst. 38, 343–347.
Fisher, R.A., 1925. Statistical Methods for Research Workers. Oliver and Boyd), Edinburgh, London.
Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147.
Henderson, J.A., Robinson, P.A., 2011. Geometric effects on complex network structure in the cortex. Phys Rev Lett 107, 018102.
Henriksen, S., Pang, R., Wronkiewicz, M., 2016. A simple generative model of the mouse mesoscale connectome. Elife 5, e12366.
Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al., 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415, 180–183.
Holland, P., Laskey, K., Leinhardt, S, 1983. Stochastic blockmodels: First steps. Social Netw. 5, 109–137.
Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y., 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. U S A 98, 4569–4574.
Kamura, T., Conaway, J.W., Conaway, R.C., 2002. Roles of SCF and VHL ubiquitin ligases in regulation of cell growth. Prog. Mol. Subcellular Biol. 29, 1–15.
Kanki, T., Klionsky, DJ., Okamoto, K., 2011. Mitochondria autophagy in yeast. Antioxid Redox Signal 4, 1989–2001.
Merz, S., Westermann, B., 2009. Genome-wide deletion mutant analysis reveals genes required for respiratory growth, mitochondrial genome maintenance and mitochondrial protein synthesis in Saccharomyces cerevisiae. Genome Biol. 10, R95.
Milward, K., Busch, K.E., Murphy, R.J., de Bono, M., Olofsson, B., 2011. Neuronal and molecular substrates for optimal foraging in Caenorhabditis elegans. Proc. Natl. Acad. Sci. U S A 108, 20672–20677.
Muldoon, S.F.B., E. W., Bassett, D. S. (2015). Small-world propensity in weighted, real-world networks. Cite as arXiv:150502194, 1–13.
Newman, M., 2003. The structure and function of complex networks. SIAM Rev. 45, 89.
Oh, S.W., Harris, J.A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A.M., et al., 2014. A mesoscale connectome of the mouse brain. Nature 508, 207–214.
Özkan, E., Carrillo, R.A., Eastman, C.L., Weiszmann, R., Waghray, D., Johnson, K.G., Zinn, K., Celniker, S.E., Garcia, K.C., 2013. An extracellular interactome of immunoglobulin and LRR proteins reveals receptor-ligand networks. Cell 154, 228–239.
Pons, P., Latapy, M., 2006. Computing communities in large networks using random walks. J. Graph Algorithms Appl. 10, 191–218.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., et al., 2005. Global analysis of protein phosphorylation in yeast. Nature 438, 679–684.

Sulston, J.E., Horvitz, H.R., 1977. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev. Biol. 56, 110–156.

Sulston, J.E., Schierenberg, E., White, J.G., Thomson, J.N., 1983. The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol. 100, 64–119.

Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., Michnick, S.W., 2008. An in vivo map of the yeast protein interactome. Science 320, 1465–1470.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al., 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623–627.

Wasserman, L., 2004. All of Statistics: A Concise Course in Statistical Inference. Springer.

Watson, C., Paxinos, G., Puelles, L. (2012). The Mouse Nervous System, first ed, Academic Press.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442.