bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

SOFTWARE

Exploring community structure in biological networks with random graphs

Pratha Sah[1]
, Lisa O. Singh[2]
, Aaron Clauset[3,4,5]
and Shweta Bansal[1,6*]

*Correspondence: sb753@georgetown.edu
[1]Department of Biology, Georgetown University, 20057 Washington DC, USA
Full list of author information is available at the end of the article

Abstract

A modular pattern, also called community structure, is ubiquitous in biological networks. There has been an increased interest in unraveling the community structure of biological systems as it may provide important insights into a system's functional components and the impact of local structures on dynamics at a global scale. Choosing an appropriate community detection algorithm to identify the community structure in an empirical network can be difficult, however, as the many algorithms available are based on a variety of cost functions and are difficult to validate. Even when community structure is identified in an empirical system, disentangling the effect of community structure from other network properties such as clustering coefficient and assortativity can be a challenge. Here, we develop a generative model to produce undirected, simple, connected graphs with a specified degrees and pattern of communities, while maintaining a graph structure that is as random as possible. Additionally, we demonstrate two important applications of our model: (a) to generate networks that can be used to benchmark existing and new algorithms for detecting communities in biological networks; and (b) to generate null models to serve as random controls when investigating the impact of complex network features beyond the byproduct of degree and modularity in empirical biological networks. Our model thus allows for the systematic study of the presence of community structure and its impact on network function and dynamics. This process is a crucial step in unraveling the functional consequences of the structural properties of biological systems and uncovering the mechanisms that drive these systems.

Keywords: biological networks; community structure; random graphs; modularity; benchmark graphs

Background

Network analysis and modeling is a rapidly growing area which is moving forward our understanding of biological processes. Networks are mathematical representations of the interactions among the components of a system. Nodes in a biological network usually represent biological units of interest such as genes, proteins, individuals, or species. Edges indicate interaction between nodes such as regulatory interaction, gene flow, social interactions, or infectious contacts [1]. A basic model for biological networks assumes random mixing between nodes of the network. The network patterns in real biological populations, however, are more heterogeneous than assumed by these simple models [2]. Some of the common non-trivial properties

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

known to exist in biological networks are degree heterogeneity, assortative mixing, clustering coefficient, and community structure (see review by Proulx et al.[1]). Community structure, or the presence of groups of nodes that are highly connected within the group but only loosely connected to other groups, has especially been found to be a central topological property [3, 4]. These groups, called modules or communities, may correspond to well-known or understood partitions of the system, but are based, in the context of network science, only on the configuration of connections between and among these groups, without any *a priori*knowledge of the system's processes.

Although community structure has been found to be a central topological feature of biological networks such as metabolic [5], protein [6, 7], genetic [8], food-web [9, 10] and pollination networks [11], a detailed understanding of its relationship with other network topological properties is still limited. This is perhaps because delineating community structure in empirical networks can prove to be a difficult task, and the numerous community detection algorithms that exist often yield very different results. Another approach has used empirical vertex features or labels to independently estimate the accuracy of detection algorithms [12]. However, such tests can have ambiguous value as a lack of correlation between observed features and the identified communities may simply indicate that the true features driving the network's structure are unobserved, not that the identified communities are incorrect.

A more straightforward method of exploring the structural and functional role of a network property is to generate graphs which are random with respect to other properties except the one of interest. For example, network properties such as degree distribution, assortativity and clustering coefficient have been studied using the configuration model [13], and models for generating random graphs with tunable structural features [14, 15, 2] . These graphs serve to identify the network measures that assume their empirical values in a particular network due to the particular network property of interest. In this work, we propose a model for generating simple, connected random networks that have a specified degree distribution and level of community structure.

Random graphs with tunable strength of community structure can have several purposes such as: (1) serving as benchmarks to test the performance of community detection algorithms; (2) serving as null models for empirical networks to investigate the combined effect of the observed degrees and the latent community structure on the network properties; (3) serving as proxy networks for modeling network dynamics in the absence of empirical network data; and (4) allowing for the systematic study of the impact of community structure on the dynamics that may flow on a network. Among these, the use of random graphs with tunable strength of community structure to serve as benchmarks has received the most attention and several such models have been proposed [16, 17, 18, 19, 20, 21]. A few studies have also looked at the role of community structure in the flow of disease through contact networks [22, 23, 24, 25]. However, the use of modular random graphs, which can be defined as random graphs that have a higher strength of community structure than what is expected at random, is still relatively unexplored in other applications. In this study, therefore, we extend the scope of modular random graphs by using

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

our model to estimate the residual structure in several empirical networks, after controlling for the degree distribution and community structure.

Previous work

In 2002, Girvan and Newman proposed a simple model for generation of random network with a specific configuration and strength of community structure [3]. The model generates modular random graphs of 128 nodes, four modules (and thus average modular size =32) and average degree=16. The connection probability in each module, which corresponds to a Erdős-Rényi random graph, is p_{in} and connection probability between modules is p_{out}. This model is one of the widely-accepted benchmarks to test community-detection algorithms, but has a limited scope due to its several assumptions such as small network size, constant number of modules, and fixed degree of each node in the network, which do not reflect the heterogeneity observed in real network systems. The model proposed by Lancichinetti et al [16] considers heterogeneity in degree and community size, but is restrictive as it assumes only a particular type of degree distribution (power law). A similar model by Bagrow[17] generates modular networks with power law degree distribution and constant community size. Yan et al[23]used a growth and preferential attachment model to generate scale-free networks comprised of communities with power-law degree distribution. Models for special graph types such as hierarchical networks [18], bipartite networks [21], and networks with overlapping modules [20] have also been proposed. These models remain limited due to their assumptions about degree heterogeneity of network and its communities. The stochastic block model is a popular generative model for producing synthetic networks with specified community structure, but like the Girvan and Newman model, it produces unrealistic degree distributions. More recently, Karrer and Newman [26] introduced a "degree-corrected" stochastic block model, which can be used to infer communities from networks with arbitrary degrees, and could, in principle, also generate networks, although this has not been systematically investigated yet. A recently proposed model [19] does generate networks with a broad range of degree distributions, modularity and community sizes, but it is unclear how model parameters map to desired properties (such as degree distribution and modularity).While these models may be sufficient for community-detection benchmarks, the model limitations constrain their use for other purposes.

Our approach

Here, we develop and implement a simple simulation model for generating modular random graphs using only a few intuitive parameters. Our model can generate graphs over a broad range of distributions of network degree and community size. The generated graphs can range from very small (<100) to large (>10^5) network sizes and can be composed of a variable number of communities. In Methods below, we introduce our algorithm for generating modular random graphs. In Results, we show that our null modular graph remains random with respect to other important network properties such as assortativity, clustering coefficient and path length. We next demonstrate the applicability of the generated modular graphs to test the accuracy of extant community detection algorithms. The accuracy of community-detection algorithms depends on several network properties such as the network

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

mean degree and strength of community structure, which is evident in our analysis. Finally, using a few empirical biological networks we demonstrate that our model can be used to generate corresponding null modular graphs under two different models of randomization. We conclude the paper with some thoughts about other applications and present some future directions.

Implementation

We present a model that generates undirected, simple, connected graphs with prescribed degree distributions and a specified level of community structure, while maintaining a graph structure that is otherwise as random (uncorrelated) as possible. Below, we introduce some notation and a metric for measuring community structure, followed by a description of our model and the steps of the algorithm used to generate graphs with this specified structure.

Measure of community structure

We begin with a graph $G = (V, E)$ that is comprised of a set of vertices or nodes $V(G) = \{v_1, \ldots, v_n\}$ and a set of edges $E(G) = \{e_1, \ldots, e_m\}$. G is undirected and simple (i.e. a maximum of one edge is allowed between a pair of distinct nodes, and no "self" edges are allowed). The number of nodes and edges in G is $|V(G)| = n$ and $|E(G)| = m$, respectively. The neighborhood of a node v_i is the set of nodes v_i is connected to, $N(v_i) = \{v_j \mid (v_i, v_j) \in E, v_i \neq v_j, 1 \leq j \leq n\}$. The degree of a node v_i, or the size of the neighborhood connected to v_i, is denoted as $d(v_i) = |N(v_i)|$. A *degree sequence*, D, specifies the set of all node degrees as tuples, such that $D = \{(v_i, d(v_i)\}$ and follows a probability distribution called the *degree distribution* with mean $\overline{d}$.

Each community or *module* C_k is defined as a subset of G that contains both nodes, $V(C_k)$ and edges $E(C_k)$, where both the endpoints of each edge in $E(C_k)$ are contained in $V(C_k)$. Each node v_i of G has a *within-degree*, $d_w(v_i) = |N(v_i) \cap V(C_k)|$, which is the number of *within-edges* connecting v_i to other nodes of the same module C_k; and a *between-degree*, $d_b(v_i) = |N(v_i) - V(C_k)|$, i.e. the number of *between-edges* connecting v_i to nodes in different modules (here, the minus operator represents set difference). The strength of the community structure defined by a partition, $\{C_k\}$, can be measured as *modularity*, Q, and is defined as

$$Q = \sum_{k=1}^{K} \left(e_{kk} - a_k^2\right) \tag{1}$$

where $e_{kk} = \frac{|E(C_k)|}{|E(G)|}$ denotes the proportion of all edges that are within module C_k, and$a_k = [\sum_{v_i \in C_k} d(v_i)]/2m$ represents the fraction of all edges that touch nodes in community C_k. When $Q = 0$, the density of within-community edges is equivalent to what is expected when edges are distributed at random, conditioned on the given degree sequence. Values approaching $Q = 1$, which is the maximum possible value of Q, indicate networks with strong community structure. Typically, values for empirical network modularity fall in the range from about 0.3 to 0.7 [27]. However, in theory Good et al. [28] show that maximum Q values depend on the network size and number of modules.

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

In order to generate a graph with a specified strength of community structure, Q, equation (1) represents our first constraint, which we rewrite below in terms of the expected value of Q, (full derivation in Additional file 1):

$$E[Q] = \sum_{k=1}^{K} \left[\frac{\overline{d_w}.s_k}{m} - \left(\frac{\overline{d}.s_k}{m} \right)^2 \right] \tag{2}$$

where $\overline{d_w}$ and $\overline{d}$ are the average within-degree and average degree, respectively, and s_k is the module size for module k. Thus, equation (2) allows us to specify $\overline{d_w}$ in terms of Q, $\overline{d}$, m and s_k, assuming that the module-specific average degree and average within-degree are equal to $\overline{d}$ and $\overline{d_w}$, respectively. When $s_k = \overline{s}$ for all k, $E[Q]$ reduces to $\frac{\overline{d_w}}{\overline{d}} - \frac{1}{K}$.

We note that as the average within-degree ($\overline{d_w}$) approaches the average degree ($\overline{d}$), the graph, G becomes increasingly modular. Hence, the maximum modularity for G with K modules can be estimated as:

$$Q_{\max} \simeq \sup(Q) = 1 - \frac{1}{K} \tag{3}$$

Model

We present a model and an algorithm that generates undirected, unweighted, simple and connected modular random graphs. The model is specified by a network size (n), degree distribution (p_d, with mean $\overline{d}$), an expected modularity ($E[Q]$), the number of modules (K), and the module size distribution ($P(s)$, with mean $\overline{s}$. Here the size of module C_k is $s = |V(C_k)|$). The algorithm proceeds in four steps:

1. Assign the n network nodes to K modules based on the size distribution $P(s)$.
2. Assign degrees, $d(v_i)$, to each node v_i based on p_d and $\overline{d}$. We next assign within-degrees, $d_w(v_i)$, to each node v_i by assuming that the within-degrees follow the same distribution as p_d with mean $\overline{d_w}$, which is estimated based on equation (2) above (Figure 1a).
3. Connect between-edges based on a modified Havel-Hakimi model and randomize them. (Figure 1b)
4. Connect within-edges based on the Havel-Hakimi model and randomize them. (Figure 1c and 1d)

The generated graph then has a degree distribution that follows p_d with mean $\overline{d}$, K modules with sizes distributed as $P(s)$, and a modularity $Q \approx E[Q]$. We set an arbitrary tolerance of $\epsilon = 0.01$, such that the achieved modularity is $Q = E[Q] \pm \epsilon$. The graph is also as random as possible given the constraints of the degree and community structure, contains no self loops (edges connecting a node to itself), multi-edges (multiple edges between a pair of nodes), isolate nodes (nodes with no edges), or disconnected components. Below, we elaborate on each of the steps of this algorithm.

Assigning Nodes to Modules

We sample module sizes, s_k, for each of the K modules from the specified module size distribution, $P(s)$ so that $\sum s_k = n$. The n nodes are then arbitrarily (without loss of generality) assigned to each module to satisfy the sampled module size sequence.

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

Assigning Degrees

Based on the degree distribution specified, a degree sequence is sampled from the distribution to generate a degree, $d(v_i)$, for each node v_i. To ensure that the degree sequence attains the expected mean of the distribution (within a specified threshold) and is *realizable*, we verify the Handshake Theorem (the requirement that the sum of the degrees be even) and the Erdős-Gallai criterion (which requires that for each subset of the k highest degree nodes, the degrees of these nodes can be "absorbed" within the subset and the remaining degrees) [29], and that no node is assigned a degree of zero.

We assume that the within-degree distribution follows the degree distribution specified, p_d, with mean $\overline{d_w}$ based on equation (2). This assumption is considered reasonable as it holds true for several of the empirical networks we analyze (shown in Figure S1 in Additional File 1). However, our model can be extended for arbitrary within-degree distributions, although the space of feasible within-degree distributions given a specified degree distribution is restricted. Next, we sample a within-degree sequence, $d_w(v_i)$, from this within-degree distribution. (We note that degree and within-degree sequences can be specified instead of a degree distribution as inputs to the algorithm.) Using rejection sampling, we ensure that the within-degree sequence attains the expected overall mean, $\overline{d_w}$ within a tolerance $\epsilon_{d_w} = \epsilon\overline{d}$ (with details in the Additional file 1), and satisfies the following conditions:

- Condition 1: $d(v_i) \geq d_w(v_i)$ for all v_i. To ensure this, we sort the degree sequence and within-degree sequence, independently. If $d(v_i) < d_w(v_i)$ for any v_i in the ordered lists, the condition is not satisfied.
- Condition 2: a realizable within-degree sequence for each module, C_k, as defined by the Handshake Theorem and the Erdos-Gallai criterion.

In addition, to ensure that each module approximately achieves the overall mean within-degree, $\overline{d_w}$, we specify the following constraint: $\min[s_k] \leq \max[\{d_w(v_i)\}_{v_i \in G}]$. If the sampled module sizes do not satisfy this criteria, the module sizes are re-sampled or an error is generated. In Figure S2 of Additional File 1, we discuss the rejection rates for the rejection sampling of both the within-degree sequence as well as the module size distribution.

The between-degree sequence is generated by specifying $d_b(v_i) = d(v_i) - d_w(v_i)$ for each node v_i. To test if the between-degree sequence is realizable, we impose a criterion developed by Chungphaisan[30] (reviewed by Ivanyi [31]) for realizable degree sequences in multigraphs. To do so, we imagine a coarse graph, H, where the modules of G are the nodes of H (i.e. $V(H) = \{C_1, C_2, \ldots C_K\}$), and the between-edges that connect modules of G are the edges of H. We note that H is a multigraph, because G allows multiple between-edges of G to connect each pair of modules. In this case, the degree sequence of H is $D = \{(C_k, d(C_k)) | d(C_k) = \sum_{v_j \in C_k} d_b(v_j), k = 1 \ldots K\}$.

The Chungphaisan criterion then specifies that the multigraph degree sequence $\{d(C_k)\}$ on H is realizable if the following conditions are satisfied:

- Condition 1: the Handshake theorem is satisfied for $\{d(C_k)\}$: $\sum_{k=1}^{K} d(C_k) = \sum_{k=1}^{K} \sum_{v_j \in C_k} d_b(v_j)$ is even
- Condition 2:$\sum_{k=1}^{j} d(C_k) - bj(j-1) \leq \sum_{k=j+1}^{K} \min[jb, d(C_k)]$ for $(j = 1, \ldots, K-1)$.

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

Here, b is defined as the maximum number of edges allowed between a pair of nodes in H; in our case, $b = \max[\{d_b(v_i)\}]$, the maximum between-degree of any node $v_i \in G$.

We also generate network graphs with $Q = 0$ by assuming the network is composed of a single module with no between-edges. Thus, $d_w(v_i) = d(v_i)$ and $d_b(v_i) = 0$ for all $v_i \in G$.

Connecting edges

Based on the within-degree sequence and between-degree sequence specified above, edges are connected in two steps (Figure 1). Nodes that belong to different modules are connected based on their between-degree to form between-edges (Figure 1b) and nodes that belong to the same module are connected according to their within-degree to form within-edges (Figure 1c and 1d).

We connect between-edges using a modified version of the Havel-Hakimi algorithm. The Havel-Hakimi algorithm [32, 33] constructs graphs by sorting nodes according to their degree and successively connecting nodes of highest degree with each other. After each step of connecting the highest degree node, the degree list is resorted and the process continues until all the edges on the graph are connected. Here, we modify this to construct between-edges by sorting nodes by highest between-degree, in order of highest total between-degree for the module to which they belong, and successively connecting the node at the top of the list randomly with other nodes. Connections are only made between nodes if they are not previously connected, belong to different modules, and do not both have within-degree of zero (to avoid disconnected components). After each step the between-degree list is resorted, and the process continues until all between-edges are connected. After all between-edges have been connected, the connections are randomized using a well-known method of rewiring through double-edge swaps [34]. Specifically, two randomly chosen between-edges (u, v) and (x, y) are removed, and replaced by two new edges (u, x) and (v, y), as long as u and x, and v and y belong to different modules, respectively. The swaps are constrained to avoid the formation of self loops and multi-edges. This process is repeated a large number of times to randomize edges.

We then connect within-edges using the standard Havel-Hakimi algorithm, applied to each module independently. Specifically, within-edges of a module are connected by sorting nodes of the module according to their within-degree and successively connecting nodes of highest within-degree with each other. Connections are only made between nodes if they are not previously connected, and do not both have a between-degree of zero (to avoid disconnected components). After each step the within-degree list is resorted and the process continues until all the within-edges of the module are connected. The connections are then randomized by rewiring through double-edge swaps [34]. We do not specify that each module be connected (only that the full graph is connected). However, if this is required, Taylor's algorithm can be used to rewire pairs of edges until the module is connected [35]. Specifically, the algorithm selects two random edges (u, v) and (x, y) that belong to two different disconnected components of the module. As long as (u, x) and (v, y) are not existing edges, the (u, v) and (x, y) edges removed and (u, x) and (v, y)

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

added. If k' is connected, then we perform the rewiring operation and let $k_{t+1} = k'$. Taylor's theorem proves that following such operation any disconnected module can be converted to a connected module with the same degree sequence.

Results and Discussion

Using our simulation algorithm, we were able to generate modular random graphs of variable network size, number of communities, degree distribution, and community size distribution. In Figure 2, we show sample networks of varying levels of modularity, $Q = 0.1, 0.3, 0.6$. We note that a network with three modules can approach a maximum modularity value of 2/3 (from equation 3), and thus Q=0.6 is a relatively high modularity for this particular network type.

Performance & Properties of Generated Graphs

Performance

Our model generates graphs that closely match the expected modularity and degree distribution. The deviation of the observed modularity is less than 0.01 from the expected value, given the specified partition. The modular random graphs with Poisson degree distribution generated by our model are similar to the ones described by Girvan and Newman[3] with linking (p_{in}) and cross-linking probability (p_{out}) equal to $\frac{\overline{d_w}}{\overline{s}-1}$and $\frac{\overline{d}-\overline{d_w}}{\overline{s}(K-1)}$ respectively. However, our model overcomes several limitations of the model proposed by Girvan and Newman[3] and others [16, 17] by considering heterogeneity in total degree, within-module degree distribution, and module sizes. Unlike many of the existing models [19, 18, 21, 20], our model can generate modular random graphs with arbitrary degree distributions, including those obtained from empirical networks.

Structural Properties

There are several other topological properties (besides degree distribution and community structure) that can influence network function and dynamics. The most significant of these properties are degree assortativity (the correlation between a node's degree and its neighbor's degrees), clustering coefficient (the propensity of a node's neighborhood to also have edges among them) and average path length (the typical number of edges between pairs of nodes in the graph). We have developed this model to generate graphs with specified degree distribution and modularity, while minimizing structural byproducts. Thus, it is important to confirm that we have reached this goal with the generative model above.

To evaluate the status of other structural properties due to the generative model, we specify graphs of $n = 2000$ following Poisson ($\lambda^k e^{-\lambda}/k!$), geometric ($p(1-p)^{k-1}$), and power-law ($\frac{k^{-\alpha}}{\zeta(\alpha)}$) degree distributions with $\overline{d} = 10$. We chose these particular types of degree distributions as they have widely studied in the context of biological networks[36, 37, 38]. Each network has $K = 10$ modules and a module size distribution $P(s = 200) = 1$. We generate modular random graphs with these specifications and modularity values that range from $Q = 0$ to $Q = 0.8$, in steps of 0.1. For each level of modularity, we generated 50 such modular random graphs and calculated the degree assortativity (r), clustering coefficient (C), and average path length (L) for each network, which is illustrated in Figure 3. In networks with random community structure (($Q = 0$)), that is random graphs with specified degree

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

distributions (such as those that would be generated by the configuration model [13]), the value of r, C, and L are what are expected at random. In Figure 3, we show that for increasing values of modularity, degree assortativity, clustering coefficient, and average path length remain relatively constant for all three network types (i.e. Poisson, geometric and power-law). At the highest levels of modularity, edge connections are constrained, particularly for the heavy-tailed geometric and power-law degree distributions, leading to an increase in clustering coefficient. Correlations between high clustering coefficient and high modularity have also been observed before [2]. The average path length for all network types also increases at the highest levels of modularity, likely reflecting the lack of many paths between modules, requiring additional steps to reach nodes in different modules. Thus, our model is able to increase levels of modularity in random graphs without altering other topological properties significantly.

Biological networks show remarkable variation in network size, connectivity and community size distribution, with some of them having particularly small network size, high degree, and small module sizes (e.g. food-web networks). We therefore tested the performance of our generated networks under deviations in the network specifications of size, mean degree and module size distribution (results presented in Additional file 1 Figure S3, S4 and S5). We find that the structural properties of our generated modular random graphs remain constant, except for two constraining conditions: a) high average degree ($\bar{d}$ >10) and b) low average module size ($\bar{s}$ <50). At these parameter extremes, the modular random graphs become disassortative and have increased clustering coefficient. A similar observation of network disassortativity has been made in hierarchically modular networks [39]. In these two scenarios, the highest value of within-degree ($d_w(v_i)$) that a node can attain is constrained by the community size, which reduces the number of possible high within-degree nodes . As a consequence high within-degree nodes must connect to low within-degree nodes more than expected, resulting in a disassortative network. In these two cases, modules also become more dense and thus create more triangles resulting in a gradual increase in clustering. Path length, on the other hand, is not affected by these conditions and shows a consistent dependence on network size and mean degree, which is well known [40, 41].

Application: Null Analysis of Empirical Networks

It is crucial to have random controls in the study of biological systems. Our algorithm can be used to generate null models and applied to the detection of structure in empirical biological networks. These null networks can be used to test hypotheses regarding the role of modularity and other topological features of the empirical networks. To do so, one would first use our algorithm to generate an ensemble of networks that match the empirical degree structure and community structure, and then compare the structural, functional, or dynamical properties of the empirical network to those of the generated modular random graphs. We demonstrate this application using four classes of biological networks, namely: a) a food-web, representing the trophic interactions at Little Rock Lake in Wisconsin with a network size of 183 and average degree = 26.8 [42]; b) a protein-protein interaction network in *Saccharomyces cerevisiae* (a yeast) of size= 4713 and average degree = 6.3[43]; c)

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

a metabolic interaction network of *Caenorhabditis elegans* of size = 453 and average degree 9.0 [44]; and d) a network of social interactions in a community of dolphins living off Doubtful Sound, New Zealand of size = 62 and average degree = 5.1 [45] (Figure 4, black bars). Visualizations of the dolphin social interaction network and the food-web trophic interaction network and its modular random counterpart are shown in Figure 5.

For each of these four empirical networks, we generate modular random graphs (Figure 4, light gray bars) with three parameters estimated from the empirical networks: (a) the degree sequence, p_k(b) the modularity, Q and (c) the average community size, $\overline{s}$. We note that as our goal is to construct null models, we assume that communities are of constant size, i.e. $P(\overline{s}) = 1$, and that the within-degree distribution matches the degree sequence (with estimated mean, $\overline{d_w}$). (A second class of null models can be constructed with $P(s)$ and the within-degree distribution estimated from the empirical networks, and we do this in the Additional file 1). Specifically, we generate 25 such random graphs and measure structural properties of the generated graphs including clustering coefficient (C), average path length (L), degree assortativity (r).

We also generate simple random graphs based on the configuration model that have the same degree distribution and average network degree as the empirical network but are random with respect to other network properties for each of the four empirical networks (Figure 4, dark gray bars). Our modular random graph model identifies which network measures assume their empirical values in a particular network because of (i) the observed degrees and (ii) the latent community structure. The configuration model, on the other hand, only specifies (i) and not (ii). Comparison to these configuration model networks thus helps us highlight the utility of our model to identify which empirical patterns in a network are deserving of further investigation. Figure 4 shows the value of each of these properties for the empirical networks as well as the ensemble mean of modular random and simple random graphs.

From Figure 4 it is evident that none of the empirical biological networks have network structure identical to their null counterparts. This suggests that the structure of each of these biological systems is governed by more than what is specified by the degree distribution and community structure. However, the observed network properties of empirical networks are closer to the ensemble means of the modular random graphs, which indicates that modularity is an essential structural component of real biological networks and that it plays an important role in influencing other structural properties of the network. For instance, compartmentalization induced by modularity promotes species persistence and system robustness by containing localized perturbation [11, 46, 47], which might favor their selection during the course of evolution. Our results show that the empirical networks tested have a much higher modularity than the simple random graphs (Figure 4a) and therefore provide evidence for this selection. Out of the three network properties that we tested apart from modularity, we found clustering coefficient of the generated random graphs to be significantly different from each of the empirical counterparts. This may point to a functional role for "triangles" in these biological networks, significantly above or below what is prescribed by the degree and community structure.

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

Little Rock Lake food web interactions (FW): Among the four empirical networks that we tested, the properties of the ensemble mean of null models such as assortativity and path length closely match most of the observed properties of Little Rock food web. The observed clustering coefficient of food web is strikingly lower than either of the random graphs which confirm the observations of low clustering in food web made by earlier studies (Figure 4d). The observed path length of this food web is short (Figure 4c) and only slightly longer than the path lengths of random graphs, which has also been noted before [48, 49, 50]. We note that for this food web, the structural properties of the simple random graphs are quite similar to those of modular random graph counterparts, suggesting that the degree distribution, particularly the high density of edges in the network governs most of the other topological characteristics of this network. Modularity, on the other hand, seems to play a minor role in dictating the structural properties of this network.

Yeast protein-protein interaction network (YP): The empirical yeast protein network is more disassortative than the ensemble mean of null modular graphs (Figure 4b). Disassortative interactions in protein-protein interaction networks are known to reduce interferences between functional modules and thus increase the overall robustness of the network to deleterious perturbations [6], while also allowing for functions to be performed concurrently [51]. The results therefore suggest that disassortative interactions may be selected for in the evolution of biological networks. From Figure 4(d) it is also evident that the yeast protein network has a higher value of clustering coefficient than the expected value predicted by the modular random graphs. A high value of clustering coefficient indicates that there are several alternate interaction paths between two proteins, making the system more robust to perturbation[52].

C.elegans metabolic interaction network (CM): The emphC.elegans metabolic network demonstrates a shorter path length but higher clustering coefficient than both modular and simple random graphs (Figure 4c and 4d). A high clustering coefficient and short path length suggests that the graph has small-world properties, which has been observed in other metabolic networks as well [53]. A highly disassortative degree structure is also well known in metabolic networks, although the mechanism leading to this property is unclear (see review by [36]). As the predicted value of disassortativity of the modular random graphs is closer to the observed value, our results suggest that the strong community structure of the metabolic networks could be one of the factors contributing to high degree disassortativity. (As discussed earlier, community structure leads to significant degree correlations in small networks with long-tailed degree distributions; see Figure S3 in Additional file 1 for an example).

Social interaction network of dolphins network at Doubtful Sound, New Zealand (DS): The empirical social interaction network of dolphins that we investigated demonstrated a negative assortativity (or disassortativity) similar to other real biological networks (Figure 4b). Interestingly, the assortativity value of both null modular and simple random network counterparts of the dolphin network is *lower*than

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

the observed value, which suggests that the network is more assortative than expected. Degree assortativity has also been observed in other animal [54] and human [4] social interaction networks. This result is quite intuitive for a social network and is also referred to as homophily: more gregarious individuals tend to interact with other gregarious individuals while introverted individuals prefer to associate with other introverts[14]. The empirical dolphin network also demonstrated a lower value of clustering coefficient than the expected values of either null model. Low clustering coupled with high degree assortativity indicates that dolphin populations may be more susceptible to the propagation of infection or information, as transmission may occur rapidly through the entire network with such properties [54, 55].

Application: Benchmark graphs for community-detection algorithms

Detecting communities in networks has been an area of intensive research in the past decade [56] with Girvan and Newman's seminal paper on community detection [3]. Extant techniques such as modularity maximization, hierarchical clustering, the clique-based method, the spin glass method etc. aim at achieving high levels of accuracy in detecting the correct partition (for a detailed review see [56]), but have their own set of strengths and weaknesses. Choosing the best algorithm can be a difficult task especially as algorithms often use distinct definitions of communities and perform well within that description. Thus, it is exceedingly important to test community-detection algorithms against a suitable benchmark. We propose our modular random graphs as benchmark graphs for the validation of existing and new algorithms of community detection.

To illustrate this use, we test the performance of six popular community detection algorithms: the Louvain method [57], fast modularity method [58], the spin-glass based method [59], the infoMAP method [60], label propagation [61] and the random-walk based method [62] using our modular random graphs as benchmarks. Specifically, we generate 10 modular random graphs for each level of modularity and used these community detection algorithms to detect their community structure. We also tested the performance of the algorithms on random graphs with no modularity (i.e. Q=0). We limit our discussion to the results of the Louvain[57] and fast modularity method [58] in this section; with the remaining results presented in Additional File 1s (Figures S6, S7 and S8). Figure 6 summarizes the performance of these two algorithms, as measured by the estimated Q, for modular random graphs with three different degree distributions (Poisson, geometric and power-law) and two levels of mean network degree ($\overline{d} = 10$: Fig 6(a) and (c) and $\overline{d} = 20$: Fig 6(b) and (d)).

At $\overline{d} = 10$, both the Louvain method (Figure 6a) and fast modularity algorithm (Figure 6c) overestimate the modularity for networks with weak community structure, and underestimate the modularity for networks of moderate community structure across all three network types. Both algorithms are accurate at the highest strengths of community structure across all network types and improve in accuracy for the more dense network ($\overline{d} = 20$, Figure 6b and 6d). In addition to comparing the estimated values of modularity to the known values in the modular random graphs, we can compare the similarity in the partitions detected by the algorithms to the true partitions. For this comparison, we use the *Jaccard index*, which measures the similarity between two partitions based on the proportion of the union of

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

the partitions that is made up by the intersection of the partitions [63], and present results in the Additional File 1 (Figure S9). As reflected in the results above, we find that partitioning is inaccurate when the true community structure is weak but improves as the Q_{true} value increases. The accuracy also improves with a higher average network degree. Both these observations have also been noted before by Lacichinetti and Fortunato [64].

Conclusions

In summary, the model that we propose in this study generates modular random graphs over a broad range of degree distribution and modularity values, as well as module size distributions. We highlight that our model is specifically designed to generate networks which have modularity evenly divided across its modules, modulo the impact of module size. This means that we are mitigating the resolution limit effect and indeed generating networks with the maximum modularity partition. We also confirm that structural properties of our generated modular graphs such as assortativity, clustering and path length remain unperturbed for a broad range of parameter values. This important feature allows these graphs to act as benchmark and control graphs to explicitly test hypotheses regarding the function and evolution of modularity in biological systems.

Compartmentalization of biological networks has been an area of great interest to biologists. What we refer to as community structure in this work is any segregation of a biological system into smaller subunits inter-connected by only a few connections. It has been suggested that modularity in a system promotes system robustness and enhances species persistence by containing localized perturbations [11, 47]. Metabolic networks of organisms living in a variable environment have indeed been found to be more modular [46]. Maintaining and selecting for modularity in biological networks, however, comes at a great cost of reducing system complexity [65], longer developmental time and cost of complete module replacement in case of failure [66]. It is therefore unclear why modularity would be strongly selected for as a structural feature of biological systems. There is also a lack of evidence to prove that the functional localization of sub-goals overlaps with the structural segregation of the network into community structure. Our work provides a tool for the systematic study of network structure (through benchmark graphs) and of the impact of connectivity and compartmentalization on system function and dynamics (through control graphs).

The detection of community structure plays a crucial role in our topological understanding of complex networks. Currently the performance of community detection methods is usually evaluated based on ground-truth from real networks. However, determining reference communities in real networks is often a difficult task. Also, ground truth data on empirical network partitions do not necessarily identify system features based on network topology and thus may create a bias when analyzing community structure. A more convenient technique of evaluating community detection method is to use artificial random graphs, but has been limited as most of the models fail to incorporate degree heterogeneity of real networks. By providing a systematic method to generate benchmark graphs, our model can aid in the development of more robust community detection algorithms, and therefore improve our topological understanding of empirical networks.

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

Beyond identifying the topological presence of community structure is the understanding of the evolution of community structure as well as the functional and dynamical role of community structure. We believe this process can be facilitated by using an appropriate class of control or null graphs. As a model for generating null networks, our method joins a suite of random graph models, each contributing to a hierarchy of null models. The simplest model for generating random graphs (based on only a single parameter) is the Erdős-Rényi random graph model, which produces graphs that are completely defined by their average degree and are random in all other respects. A slightly more complex and general model is one that generates graphs with a specified degree distribution (or degree sequence) and but are random in all other respects [67, 68, 13]. These models can be extended to sequentially include additional independent structural constraints, such as degree distribution and clustering coefficient [2], or degree structure and community structure, as we have demonstrated here. A further extension to this work will be designing models that generate random graphs with multiple structural constraints. For example, our model can be combined with the one proposed by Bansal [2009] to generate random graphs with specified degree distribution as well as tunable strength of modularity and clustering coefficient.

Availability and requirements

- *Project name:* Modular random graph generator
- *Project home page:* http://github.com/bansallab
- *Operating system(s):* Platform independent
- *Programming language:* Python 2.7
- *Other requirements:* Networkx Python package
- *License:* BSD-style
- *Any restrictions to use by non-academics:* None

Competing interests
The authors declare that they have no competing interests.

Author's contributions
PS and SB contributed to algorithm design and implementation. PS, LOS, AC and SB contributed to manuscript writing. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by NSF award DEB-1216054.

Author details
[1]Department of Biology, Georgetown University, 20057 Washington DC, USA. [2]Department of Computer Science, Georgetown University, 20057 Washington DC, USA. [3]Department of Computer Science, University of Colorado, 80309 Boulder, CO, USA. [4] BioFrontiers Institute, University of Colorado, 80303 Boulder, CO, USA. [5]Santa Fe Institute, 1399 Hyde Park Rd. 87501 Santa Fe, NM, USA. [6]Fogarty International Center, National Institutes of Health, 20892 Bethesda, MD, USA.

References

1. Proulx, S.R., Promislow, D.E.L., Phillips, P.C.: Network thinking in ecology and evolution. Trends in ecology & evolution **20**(6), 345–53 (2005)
2. Bansal, S., Khandelwal, S., Meyers, L.A.: Exploring biological network structure with clustered random networks. BMC bioinformatics **10**, 405 (2009)
3. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America **99**(12), 7821–6 (2002)
4. Newman, M.: Mixing patterns in networks. Physical Review E **67**(2), 026126 (2003)
5. Ravasz, E., Somera, a.L., Mongru, D.a., Oltvai, Z.N., Barabási, a.L.: Hierarchical organization of modularity in metabolic networks. Science (New York, N.Y.) **297**(5586), 1551–5 (2002)
6. Maslov, S., Sneppen, K.: Specificity and stability in topology of protein networks. Science (New York, N.Y.) **296**(5569), 910–3 (2002)

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

7. Han, J.-D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P., Vidal, M.: Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature **430**(6995), 88–93 (2004)
8. Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulation network of Escherichia coli. Nature genetics **31**(1), 64–8 (2002)
9. Krause, A.E., Frank, K.a., Mason, D.M., Ulanowicz, R.E., Taylor, W.W.: Compartments revealed in food-web structure. Nature **426**(6964), 282–5 (2003)
10. Guimerà, R., Stouffer, D.B., Sales-Pardo, M., Leicht, E.a., Newman, M.E.J., Amaral, L.a.N.: Origin of compartmentalization in food webs. Ecology **91**(10), 2941–51 (2010)
11. Olesen, J.M., Bascompte, J., Dupont, Y.L., Jordano, P.: The modularity of pollination networks. Proceedings of the National Academy of Sciences of the United States of America **104**(50), 19891–6 (2007)
12. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics - MDS '12, 1–8 (2012)
13. Molloy, M., Reed, B.: A Critical Point for Random Graphs With a Given Degree Sequence. Random Structures and Algorithms **6**(2-3), 161–180 (1995)
14. Newman, M.: Assortative Mixing in Networks. Physical Review Letters **89**(20), 208701 (2002)
15. Xulvi-Brunet, R., Sokolov, I.: Reshuffling scale-free networks: From random to assortative. Physical Review E **70**(6), 066102 (2004)
16. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Physical Review E, 1–6 (2008)
17. Bagrow, J.P.: Evaluating local community methods in networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(05), 05001 (2008)
18. Arenas, A., Díaz-Guilera, A., Pérez-Vicente, C.: Synchronization Reveals Topological Scales in Complex Networks. Physical Review Letters **96**(11), 114102 (2006)
19. Hintze, A., Adami, C.: Modularity and anti-modularity in networks with arbitrary degree distribution. Biology direct **5**, 32 (2010)
20. Sawardecker, E.N., Sales-Pardo, M., a.N. Amaral, L.: Detection of node group membership in networks with group overlap. The European Physical Journal B **67**(3), 277–284 (2008)
21. Guimerà, R., Sales-Pardo, M., Amaral, L.a.N.: Module identification in bipartite and directed networks. Physical review. E, Statistical, nonlinear, and soft matter physics **76**(3 Pt 2), 036102 (2007)
22. Zhao, H., Gao, Z.Y.: Modular effects on epidemic dynamics in small-world networks. Europhysics Letters (EPL) **79**(3), 38002 (2007)
23. Yan, G., Fu, Z.-Q., Ren, J., Wang, W.-X.: Collective synchronization induced by epidemic dynamics on complex networks with communities. Physical Review E **75**(1), 016108 (2007)
24. Chu, X., Guan, J., Zhang, Z., Zhou, S.: Epidemic spreading in weighted scale-free networks with community structure. Journal of Statistical Mechanics: Theory and Experiment **2009**(07), 07043 (2009)
25. Salathe, M., Jones, J.H.: Dynamics and Control of Diseases in Networks with Community Structure. PLoS Computational Biology **6**(4), 1–11 (2010)
26. Karrer, B., Newman, M.: Stochastic blockmodels and community structure in networks. Physical Review E **83**, 1–11 (2011)
27. Newman, M.E.J.: Detecting community structure in networks. The European Physical Journal B - Condensed Matter **38**(2), 321–330 (2004)
28. Good, B.H., de Montjoye, Y.-A., Clauset, A.: Performance of modularity maximization in practical contexts. Physical Review E **81**(4), 046106 (2010)
29. Zverovich, I.E., Zverovich, V.E.: Contributions to the theory of graphic sequences. Discrete Mathematics **105**, 293–303 (1992)
30. Chungphaisan, V.: Conditions For Sequences to be r_Graphic. Discrete Mathematics **7**, 31–39 (1974)
31. Iványi, A.: Degree sequences of multigraphs. Annales Univ. Sci. Budapest., Sect. Comp **37**, 195–214 (2012)
32. Havel, V.: A remark on the existence of finite graphs. Casopis Pest. Mat. **80**, 477–480 (1955)
33. Hakimi, S.: On realizability of a set of integers as degrees of the vertices of a linear graph. I. Journal of the Society for Industrial & Applied . . . **10**(3), 496–506 (1962)
34. Gkantsidis, C., Mihail, M., Zegura, E.: The Markov Chain Simulation Method for Generating Connected Power Law Random Graphs. In: Ladner, R.E. (ed.) Proceedings of the Fifth Workshop on Algorithm Engineering and Experiments, pp. 16–25. SIAM, 2003, ??? (2003)
35. Taylor, R.: Constrained Switchings in Graphs, pp. 314–336. Springer, Berlin Heidlberg (1981)
36. Barabási, A.-L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nature reviews. Genetics **5**(2), 101–13 (2004)
37. Przulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics **23**(2), 177–83 (2007)
38. Tanaka, R.: Scale-Rich Metabolic Networks. Physical Review Letters **94**(16), 168101 (2005)
39. Jing, Z., Lin, T., Hong, Y., Jian-Hua, L.: The effects of degree correlations on network topologies and robustness. Chinese . . . **16**(12), 3571–3580 (2007)
40. Dorogovtsev, S., Mendes, J., Oliveira, J.: Degree-dependent intervertex separation in complex networks. Physical Review E **73**(5), 056122 (2006)
41. Hołyst, J., Sienkiewicz, J., Fronczak, A., Fronczak, P., Suchecki, K.: Universal scaling of distances in complex networks. Physical Review E **72**(2), 026108 (2005)
42. Martinez, N.: Artifacts or attributes? Effects of resolution on the Little Rock Lake food web. Ecological Monographs **61**(4), 367–392 (1991)
43. Colizza, V., Flammini, A., Maritan, A., Vespignani, A.: Characterization and modeling of protein–protein interaction networks. Physica A: Statistical Mechanics and its Applications **352**(1), 1–27 (2005)
44. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, a.L.: The large-scale organization of metabolic networks. Nature **407**(6804), 651–4 (2000)

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

45. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology **54**(4), 396–405 (2003)
46. Parter, M., Kashtan, N., Alon, U.: Environmental variability and modularity of bacterial metabolic networks. BMC evolutionary biology **7**, 169 (2007)
47. Stouffer, D.B., Bascompte, J.: Compartmentalization increases food-web persistence. Proceedings of the National Academy of Sciences of the United States of America **108**(9), 3648–52 (2011)
48. Williams, R.J., Berlow, E.L., Dunne, J.a., Barabási, A.-L., Martinez, N.D.: Two degrees of separation in complex food webs. Proceedings of the National Academy of Sciences of the United States of America **99**(20), 12913–6 (2002)
49. Montoya, J.M., Sole, R.V.: Small world patterns in food webs. Journal of theoretical biology **214**(3), 405–12 (2002)
50. Dunne, J.a., Williams, R.J., Martinez, N.D.: Food-web structure and network theory: The role of connectance and size. Proceedings of the National Academy of Sciences of the United States of America **99**(20), 12917–22 (2002)
51. Khor, S.: Concurrency and network disassortativity. Artificial life **16**(3), 225–32 (2010)
52. Wuchty, S., Barabási, A.-L., Ferdig, M.T.: Stable evolutionary signal in a yeast protein interaction network. BMC evolutionary biology **6**, 8 (2006)
53. Wagner, a., Fell, D.a.: The small world inside large metabolic networks. Proceedings. Biological sciences / The Royal Society **268**(1478), 1803–10 (2001)
54. Croft, D.P., James, R., Ward, A.J.W., Botham, M.S., Mawdsley, D., Krause, J.: Assortaive interactions and social networks in fish. Oecologia **143**, 211–219 (2005)
55. Newman, M.: Properties of highly clustered networks. Physical Review E **68**(2), 026121 (2003)
56. Newman, M.E.J.: Communities, modules and large-scale structure in networks. Nature Physics **8**(1), 25–31 (2011)
57. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008 (2008)
58. Clauset, A., Newman, M., Moore, C.: Finding community structure in very large networks. Physical Review E **70**(6), 066111 (2004)
59. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Physical Review E **74**, 1–16 (2006)
60. Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. The European Physical Journal Special Topics **178**(1), 13–23 (2010)
61. Raghavan, U., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E **76**(3), 036106 (2007)
62. Pons, P., Latapy, M.: Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications **10**(2), 191–218 (2006)
63. Downton, M., Brennan, T.: Comparing classifications: An evaluation of several coefficients of partition agreement. In: Classification Society, Boulder, CO (1980)
64. Lancichinetti, A., Fortunato, S.: Community detection algorithms: A comparative analysis. Physical Review E **80**(5), 056117 (2009)
65. Welch, J.J., Waxman, D.: Modularity and the cost of complexity. Evolution; international journal of organic evolution **57**(8), 1723–34 (2003)
66. Krohs, U.: The cost of modularity. In: Functions in Biological and Artificial Worlds: Comparative Philosophical Perspectives, pp. 259–276 (2009)
67. Aiello, W., Chung, F., Lu, L.: A random graph model for massive graphs. Proceedings of the thirty-second annual ACM symposium on Theory of computing - STOC '00, 171–180 (2000)
68. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. Physical Review E **64**(2), 026118 (2001)

Figures

Figure 1 Schematic representation of the steps of our algorithm. The algorithm assigns a within-degree and between-degree to each node, which are represented here as half-within-edges and half-between-edges respectively (Figure 1a). The half-between-edges are then connected using a modified version of the Havel-Hakimi algorithm, and to remove degree correlations, the between-edges are randomized (Figure 1b). Finally, the half-within-edges are connected using the standard Havel-Hakimi algorithm for each module (Figure 1c) and the within-edges are randomized to remove degree correlations (Figure 1d).

Figure 2 Modular random graphs with $n = 150, m = 375, K = 3, P(s = 50) = 1$and p_k is power law with modularity values of : a) Q=0.1; b) Q=0.3; and c) Q=0.6. As the modularity increases, the ratio of the total number of edges within modules to the number of edges in the network increases (i.e. $overlined_w$ increases), while the remaining parameter values (degree distribution, network mean degree, number of modules) are held constant.

bioRxiv preprint doi: https://doi.org/10.1101/001545; this version posted December 22, 2013.

Figure 3 Values of (a) Assortativity, r, (b) clustering coefficient,C, and (c) path length, L in modular random graphs do not vary significantly with increasing modularity (Q). Each graph has $n = 2000$ nodes, a mean degree $\overline{d} = 10$ and $K = 10$ modules with $P(s = 200) = 1$. The data points represent the average value of 50 random graphs. Standard deviations are plotted as error bars.

Figure 4 Comparisons of empirical, null modular networks and simple random graphs: The Figure summarizes network statistics of the empirical network as well as the ensemble mean of two types of random graphs in terms of (a) Modularity, Q ; (b) Assortativity, r; (c) Path length, L and (d) Clustering coefficient, C. The path length value for the empirical Yeast-Protein interaction network is missing as the network contains disconnected components. Error bars denote standard deviation from the ensemble mean of the generated random graphs. Errors bars for modular random graphs in Figure 5(a) have been omitted as the value of modularity (Q) match the empirical networks perfectly. FW= Little Rock food web, YP= Yeast protein interaction network, CM= *C.elegans* metabolic network and DS= Dolphin social network. of the empirical graphs.

Figure 5 Visualization of empirical and random graphs of social interaction of dolphins and food-web trophic interactions at the Little Rock Lake in Wisconsin. Figure (a) is the empirical network of Dolphin social network, (b) its modular random graph, and (c) its simple random graph. Figure (d) is the empirical network for the food-web trophic interaction at Little Rock Lake in Wisconsin, (e) is its modular random graph and (f) its simple random graph counterpart. Modular random graphs have generated to match the overall degree distribution, network mean degree, the level of modularity and the number of modules of the empirical graphs. Simple random graphs have been generated to match the overall degree distribution and network mean degree of the empirical graphs.

Figure 6 Performance of the Louvain method (top row)[57] and fast modularity algorithm (bottom row)[58] in networks with mean degree 10 (left column) and mean degree 20 (right column). Fill circles, open circles and triangles represent networks with Poisson, geometric and power-law degree distributions, respectively. Each data point represents the average over ten modular random graphs. Error bars have been omitted as they were smaller than the symbols. The solid line is the reference line where estimated modularity is equal to the input modularity.

Additional Files

Additional file 1 — Supplementary Analysis
Additional analysis of algorithm with figures