Link and subgraph likelihoods in random undirected networks with fixed and partially fixed degree sequences

Jacob G. Foster,[1] David V. Foster,[2] Peter Grassberger,[1,2] and Maya Paczuski[1]

[1]*Complexity Science Group, Department of Physics and Astronomy, University of Calgary, Calgary, Canada T2N 1N4*
[2]*Institute for Biocomplexity and Bioinformatics, University of Calgary, Calgary, Canada T2N 1N4*

The simplest null models for networks, used to distinguish significant features of a particular network from *a priori* expected features, are random ensembles with the degree sequence fixed by the specific network of interest. These "fixed degree sequence" (FDS) ensembles are, however, famously resistant to analytic attack. In this paper we introduce ensembles with partially-fixed degree sequences (PFDS) and compare analytic results obtained for them with Monte Carlo results for the FDS ensemble. These results include link likelihoods, subgraph likelihoods, and degree correlations. We find that local structural features in the FDS ensemble can be reasonably well estimated by simultaneously fixing only the degrees of a few nodes, in addition to the total number of nodes and links. As test cases we use two protein interaction networks (*Escherichia coli, Saccharomyces cerevisiae*), the internet on the autonomous system (AS) level, and the World Wide Web. Fixing just the degrees of two nodes gives the mean neighbor degree as a function of node degree, $\langle k' \rangle_k$, in agreement with results explicitly obtained from rewiring. For power law degree distributions, we derive the disassortativity analytically. In the PFDS ensemble the partition function can be expanded diagrammatically. We obtain an explicit expression for the link likelihood to lowest order, which reduces in the limit of large, sparse undirected networks with L links and with $k_{max} \ll L$ to the simple formula $P(k,k') = kk'/(2L + kk')$. In a similar limit, the probability for three nodes to be linked into a triangle reduces to the factorized expression $P_\Delta(k_1, k_2, k_3) = P(k_1, k_2)P(k_1, k_3)P(k_2, k_3)$.

PACS number(s): 89.75.Hc, 02.50.Cw, 02.70.Uu, 05.20.Gg

I. INTRODUCTION

A pivotal question of empiricism is the degree to which the results of an observation are expected. In ideal cases, either predictions based on these expectations remain valid in view of new measurements, or the expectations have to be changed. But this clear distinction is often blurred by uncertainties resulting from measurement errors, imprecision of model parameters, or the impossibility of extracting exact predictions from complicated models. Whether or not the problem at hand is a typical instance of a wider class of problems that are already understood is a question of statistical inference. In rare cases, the consequences of the expectations (or the model) can be derived analytically prior to observation. If this is not feasible, a widely used strategy is to construct a large number of "surrogates" [1], or instances of a well-defined null model encapsulating the expectations, and to compare the actual observations to this artificial data.

Constructing surrogates is equivalent to simulating a statistical ensemble. In choosing weights for the ensemble of surrogates one often uses Occam's razor—no outcome compatible with the null hypothesis should be preferred, and all such outcomes are equally likely. This is similar to Jaynes' construction of statistical mechanics by maximizing Shannon entropy with physically meaningful constraints. Consequently, the numerical construction of surrogates often uses Monte Carlo methods [2] similar to those used in statistical mechanics.

This paper addresses properties of ensembles used as null models for complex networks. Predictions based on the null models fix expectations, and thereby determine whether or not deviations in the properties of an actual network are functionally or historically significant. While the numerical construction of surrogates of these ensembles has received attention in the recent literature [3–5], much less is known about analytic methods (see discussion below).

Nowadays networks attract enormous interest as representations of complex systems. They take various guises in biological, social, technological, and physical contexts. The nodes designate distinct degrees of freedom (e.g., agents, species, genes, magnetic concentrations in the solar photosphere, or earthquakes) and the links identify primary interactions or relationships between pairs of nodes (e.g., coauthorship, predator-prey relations, gene regulation, magnetic flux tubes, or seismic correlations). For examples see [6–13]. The ubiquity of networks and their relatively easy visualization as graphs, together with notions of universality prevalent in the physics community, have driven speculations that the structure of networks can shed light on fundamental principles of social or biological organization, such as political behavior, ecosystem dynamics, brain function or the regulated homeostasis of organisms.

At the simplest level, networks are purely static entities, with each pair of distinct nodes connected by no more than one edge (or "link"). If, in addition, the interaction *strength* is disregarded (which often is a very useful simplification), the adjacency matrix $\mathcal{M}$ for the graph is a square (0,1) matrix. If $M_{ij}=1$ then an edge points from node i to node j; if $M_{ij}=0$ then the edge is absent. Without self-interactions, $M_{ii}=0$. For undirected networks, the adjacency matrix is symmetric, $M_{ij}=M_{ji}$. The *degree* k_i of node i is then defined as the number of edges incident on it, $k_i = \Sigma_j M_{ji}$. Several reviews may be found in Refs. [6,14,15].

Section II defines more precisely the network ensembles (or null models) we consider in this paper. Our analytical

methods focus on ensembles where the total number of links and nodes in the network is specified as well as the degrees of a small subset of nodes. These are called ensembles of "partially fixed degree sequence" (PFDS). Analytic predictions based on the PFDS ensembles can be compared with numerical results from a "rewiring" algorithm for ensembles with fixed degree sequence (FDS), where the number of links attached to *every* node in the network is simultaneously specified. Section III mainly recalls previous results. We review Monte Carlo methods for sampling the FDS ensemble. Then we discuss how such null models can be used, and we conclude by recalling previous analytic approaches. Section IV discusses some results derived later in Sec. V, namely analytic estimates of the link likelihood p_{ij} (the likelihood for a link to connect nodes i and j). It uses them to make predictions for the average nearest neighbor degree $\langle k' \rangle_k$ and for disassortativity. The calculation in PFDS ensembles simultaneously fixing the degrees of two nodes at a time gives an excellent description of $\langle k' \rangle_k$ for large k, e.g., for an *Escherichia coli* protein interaction network and an AS level map of the Internet. We also compute $\langle k' \rangle_k$ analytically for the case where the degree distribution is a power law, using Eq. (1) given below. In that case, the naive approximation for $P(k,k')$ could give divergent or ill-defined results.

Section V contains our main analytic results. In order to keep the notational clutter of this section to a minimum and to emphasize the intuitive nature of the results, most intermediate steps are moved to Appendix B. In the limit of large sparse networks with L links and with the maximal degree much less than L, we find that the link likelihood depends only on L and on the degrees k and k' of the two nodes, and is given by

$$P(k,k') \approx kk'/(2L + kk'). \tag{1}$$

This improves substantially over the widely used "naive" approximation $P(k,k') \approx kk'/2L$. Finally, we find an expression for the likelihood of a triangle, which factorizes in the same limit of large sparse networks (and when all three degrees are much larger than 1) to $P_\Delta(k_1,k_2,k_3) = P(k_1,k_2)P(k_1,k_3)P(k_2,k_3)$, with $P(k,k')$ given again by Eq. (1). The paper ends in Sec. VI with a discussion and an outlook to further problems.

II. NULL MODELS FOR NETWORKS

A. Erdös-Renyi (undirected) graphs

The simplest null hypothesis is that a given network is completely random, not even the number of links being specified. The only constraint is on the number of nodes, which is assumed to be N. Each pair of nodes may be joined with at most one link. Hence, the number of labeled undirected graphs with fixed N is $Z_0(N) = 2^{N(N-1)/2}$. This quantity is the number of ways undirected links may be placed in $\binom{N}{2} = N(N-1)/2$ possible positions. A statistical ensemble is obtained by assigning weights to each graph. The most natural choice is to weight each graph with L links by a factor p^L, where p is the probability that two given nodes are connected by a link. This gives the average number of links as $\bar{L}$

$= pN(N-1)/2$. The average degree of a node, i.e., the average number of links attached to it, is then $\bar{k} = 2\bar{L}/N = p(N-1)$, and the degree distribution is binomial. In the limit of sparse networks, where $p \to 0$ for $N \to \infty$ such that $pN \to$ const, the degree distribution [or the probability $P(k)$ that a node has k links] becomes Poissonian,

$$P(k) = \frac{\bar{k}^k \exp^{-\bar{k}}}{k!}. \tag{2}$$

While this ensemble can be viewed as a "grand canonical" version of the Erdös-Renyi ensemble [16] since the particle fugacity is fixed, it is more customary to associate Erdös-Renyi graphs with a different ensemble where the total number L of links is fixed, rather than just the average $\bar{L}$. Park and Newman refer to the ensemble with fixed L as "canonical" [17,18], making an analogy between the number of links and the number of particles in traditional statistical mechanics. However, we shall refer to this ensemble, and ensembles with similar hard degree constraints, as microcanonical.

Excluding self-connections as well as multiple edges between any pair of nodes gives

$$Z_1(N,L) = \sum_{\{\text{lg},N\}}^{\{L\}} 1 = \left(\begin{array}{c} \binom{N}{2} \\ L \end{array} \right) \tag{3}$$

distinct labeled, undirected graphs with fixed N and L [19]. The subscript "$\{\text{lg},N\}$" indicates a sum over labeled graphs with N nodes, while the superscript $\{L\}$ on the sum indicates, as in later formulas, the constraints on the edges. The subscript "1" on Z indicates that Z_1 is the number of undirected graphs with one (global) hard constraint on the links, just as $Z_0(N)$ is the number of undirected graphs with zero constraints on the links.

With no further knowledge or constraints on the network, Occam's razor suggests assigning equal weight to each labeled graph satisfying all the hard constraints. This corresponds exactly to the construction of microcanonical ensembles in statistical mechanics. For Z_1, each node has equal probability to be connected to any other node. It is easy to show [20] that the distribution for the number of links attached to each node is again Poissonian for sparse networks with large N, where the grand canonical and microcanonical ensembles become equivalent.

In contrast, observations of real networks reveal fat-tailed degree distributions, which differ starkly from the situation where each node has equal likelihood to be connected to any other node. The most salient consequence is that the average degree $\bar{k}$ fails to characterize the connectivity of the nodes; in particular it cannot account for the dominant nodes or "hubs" with many links, which would not typically appear in the Erdös-Renyi ensemble.

B. Ensembles with fixed degree sequences

As a result, attention has moved to ensembles that build additional information into the null hypothesis about the

"distinguishability" or diversity of the nodes. Although many different and equally plausible ways to account for diversity can be imagined, to begin we focus on the most popular contemporary method. This uses the random ensemble of labeled graphs with fixed degree sequence (FDS) as the relevant null model. The complete degree sequence simultaneously fixes *all* the one-node properties for each member of the ensemble, without reference to their relationships (or links) in the network. Obviously, it is straightforward to obtain the degree sequence for any network, and there exist numerical methods to estimate characteristic properties of the corresponding FDS ensemble (see Sec. III).

The microcanonical FDS ensemble is specified by assigning a specific degree (=number of links) to each node, k_i for $i = 1, \ldots, N$, and giving equal weight to each graph with this degree sequence, while giving zero weight to all those graphs which have a different degree sequence. The null hypothesis for any observable pertinent to a specific graph $\mathcal{G}$ with adjacency matrix $\mathcal{M}_\mathcal{G}$ is then obtained by taking its expectation value in the FDS ensemble with the same degree sequence. For undirected graphs excluding self-interactions, the FDS partition sum is the number of symmetric (0,1) matrices with zeroes on the diagonal and with fixed marginal sums, which can be written according to our previous convention as

$$Z_N(N, L, k_1, k_2, \ldots, k_{N-1}) = \sum_{\{lg, N\}}^{\{k_1, \ldots, k_N\}} 1 \qquad (4)$$

with $k_1 + k_2 + \cdots + k_N = 2L$. For most networks of physical interest, Z_N is astronomically large compared to one, but vanishingly small compared to Z_1. For instance, Chen *et al.* [5] numerically estimate the number of 12×12 (0,1) matrices with each row and column sum equal to 2 (and with no restrictions of symmetry or vanishing diagonal) to be $\approx 2.196 \times 10^{16}$, which agrees well with the exact number found by Wang and Zhang [21]. This number is much smaller than the number of all 12×12 (0,1) matrices with 24 ones, which is $\binom{12^2}{24} \approx 1.69 \times 10^{29}$. Despite efforts by these and other mathematicians over decades [21,22], no well-developed, exact analytical approaches are known for these combinatorial problems, but advanced computational methods exist, as described in Sec. III.

C. Partially fixed degree sequences

On the one hand, the difficulty of enumerating the number of graphs in the FDS ensemble suggests strong correlations in the graphs, since similar problems in systems lacking correlations can often be solved exactly. Indeed, the FDS ensemble makes very different predictions from the Erdös-Renyi (ER) ensemble, showing that taking into account *some* information about the nodes' degrees is crucial.

On the other hand, it might be the case that not all the constraints in the FDS ensemble must be taken into account simultaneously. After all it is the simultaneous fixing of *all* the marginal sums in the matrix that makes the calculation of Z_N difficult. Perhaps taking into account all the constraints, but not requiring them to be simultaneously enforced, is al-

ready sufficient to capture some nontrivial aspects of the FDS ensemble. If this is possible, then we must also identify which specific small subset of the nodes' degrees gives the most reliable estimate of various expectation values in the FDS ensemble.

Here, we study ensembles where the degrees of a very small subset of the nodes are simultaneously fixed—the other degrees being arbitrary up to the constraint on the total number of links. We also demand that no more than one link may connect any two nodes in the network, and disallow self-connections. All graphs satisfying these constraints have equal weight. Those not satisfying these constraints are given zero weight. These ensembles can all be viewed as subensembles of the ER ensemble. For each possible degree subset, we calculate the different partition functions corresponding to all possible subgraphs of a certain size. From these we can approximate expectation values of various quantities in the FDS ensemble.

In the following we shall always label the nodes such that the first m degrees, $k_1, \ldots, k_m$, are fixed. We call the resulting ensembles PFDS($m+1$): partially fixed degree sequence with $m+1$ constraints (the final constraint comes from fixing the number of links, L). Clearly, putting more constraints on the ensemble of labeled graphs diminishes its size until, when each and every edge is specified, the ensemble contains just one member—the real network $\mathcal{G}$ being studied. For ensembles with increasing numbers of link constraints this implies that Z_{m+1} decreases monotonically with $m+1$, and

$$1 \equiv Z_\mathcal{G} \ll Z_N(N, L, k_1, k_2, \ldots, k_{N-1})$$
$$\ll Z_{m+1}(N, L, k_1, \ldots, k_m) \ll Z_1(N, L) \qquad (5)$$

for $1 < m < N-1$.

We find that fixing only the degrees of the nodes participating in the small subgraph (e.g., link or triangle) under consideration, with explicit exclusion of self-connection and multiple connections between any nodes, already gives a good approximation to the disassortativity (and to other properties) in the FDS ensemble. As noted above, this uses information about the whole degree sequence, but in each contribution corresponding to one specific (labeled) subgraph only part of this information is used.

The information stored in the degree sequence is most important when its distribution is very wide. Even for networks exhibiting broad degree distributions, such as protein interaction networks or autonomous system maps of the Internet, it is sufficient to fix the degrees of the node pairs directly involved (as well as the total number of links in the network) to obtain a good estimate of the mean neighbor degree, $\langle k' \rangle_k$, and of the disassortativity (i.e., the tendency of nodes with low degree to connect to nodes of high degree, and *vice versa* [23]). The mean neighbor degree $\langle k' \rangle_k$ is defined as

$$\langle k' \rangle_k = \frac{\sum_{k'} k' N(k, k')}{\sum_{k'} N(k, k')}, \qquad (6)$$

where $N(k, k')$ is the average number of links between nodes with degree k and nodes with degree k' in a given ensemble

(see Sec. IV for a more detailed discussion). In order to estimate the number of triangles (i.e., the clustering), one has to fix the degrees of node triples. If we fix in addition the degrees of (some) hubs, this slightly improves the approximations. It is much easier to make analytical calculations for small m, with the smallest meaningful m being the size of the subgraph being considered. Hence for link likelihoods and for $\langle k' \rangle_k$ this minimum is $m=2$, while for triangle likelihoods it is $m=3$. By comparing analytic properties of the PFDS$(m+1)$ ensemble with numerical estimates of the FDS ensemble, we can assess to what extent the correlations in the PFDS ensembles resemble those in the FDS and in the ER ensembles.

To begin, we shall focus in Sec. IV on the link likelihood, p_{ij}, which is the probability that a randomly chosen graph from the ensemble contains an edge from node i to node j. From this microscopic quantity one can calculate the standard degree-degree correlations that are commonly compared with real-world networks to identify statistically significant features. Details of the calculation of p_{ij} are deferred to Sec. V (and to Appendix B). There we will also treat the generalization to $m \geq 3$ which is needed in order to estimate the frequencies of higher order quantities such as motifs [24]. As an example of a motif calculation, we include an estimate of the number of triangles.

III. BACKGROUND

A. Monte Carlo algorithms to estimate the FDS ensemble

As for many other problems where one wants to sample complex instances from some well-defined ensemble, here two basic strategies predominate: Markov chain Monte Carlo and sequential sampling [4,25,26]. For the present case, the most obvious and popular Markov chain algorithm is the *rewiring algorithm* [3,27–30]. We describe it here for directed graphs; the generalization to undirected graphs is immediate. We start by making an initial network with N nodes, no self-connections, and the desired degree sequence, but without paying attention to multiple links between pairs of nodes [31]. The Monte Carlo algorithm proper consists of a sequence of moves, randomly chosen from a move set, which continues until equilibrium (i.e., uniformity of sampling) is reached with sufficient accuracy. A move is initiated by choosing randomly four different nodes i, j, l, and m with $M_{ij}>0$ and $M_{lm}>0$. If either $M_{im}>0$ or $M_{lj}>0$, a null move is performed (the graph is left as it is). If neither of the pairs $\{i,m\}$ and $\{l,j\}$ were already connected by a link, M_{ij} and M_{lm} are each decreased by 1, while M_{im} and M_{lj} are increased by 1. This corresponds to swapping one pair of links.

It can be shown easily that this algorithm eventually leads to a graph without multiple links (provided such a graph consistent with the fixed degree sequence exists). After this happens, the algorithm satisfies detailed balance (any sequence of moves is equally likely to be chosen as its reversed sequence) and is ergodic (each graph with the same degree sequences can be reached by a suitable move sequence.) As shown in [28], ergodicity is not strictly satisfied, but the few

exceptions can be taken into account by including three-link exchanges in the move set.

Sequential sampling proceeds, in contrast, by repeatedly building a new graph from scratch. For this we start with an empty adjacency matrix and fill its entries randomly. In the simplest version, this is done without paying any attention to the degree sequence, to the absence of self-loops, or to the exclusion of multiple links. Instead, the candidate graph is discarded if any of these constraints are violated. In this way the uniformity of the sampling is guaranteed, but the *attrition* (i.e., the chance to reach an illegal configuration) is overwhelming, rendering the algorithm useless. But there are more sophisticated options for sequential sampling. The most efficient algorithm studied in the literature [5] uses detailed mathematical results for the structure of legal adjacency matrices [32] to bias the matchings in a much more clever way.

B. Uses of null models

Statistically significant deviations between a null model and a real network point to organizing laws or historical accidents that are not accounted for by the null hypothesis. On the other hand, finding no statistically significant deviations would promote the belief that the entire structure of the network could be accounted for by the model, e.g., by the complete degree sequence in case of the FDS ensemble. This process of building null models can, in principle, be iterated to understand the full set of organizing principles or physical constraints on a network: one builds a null model, tests for significant deviations, and then builds a new null model with richer structure to try to reduce any significant deviations to typicality.

Through an application of this discriminatory method, Maslov *et al.* [33] showed that a significant part of the disassortativity [23] observed in the Internet could be attributed to the broad degree distribution together with the restriction of no multiple links between any pair of nodes. For a scale free network of N nodes with a degree distribution $P(k) \sim k^{-\gamma}$, the maximum expected degree $k_c(N)$ scales as $k_c(N) \sim N^{1/(\gamma-1)}$. In a random network with no constraints on edge multiplicities, the expected number of edges between the two largest hubs would then scale as $k_c(N)^2/N \sim N^{2/(\gamma-1)-1}$. For $\gamma < 3$, this number diverges with N. If the constraint of no multiple edges is imposed, these links must be distributed so that they connect the hubs to other nodes. This creates fewer links between hubs than naively expected, and more links between hubs and nodes with small degree; it also leads to a suppression of links between nodes with small degrees (as the degrees of these nodes are "used up" by connecting to hubs). The net effect is that fixing a broad degree sequence decreases assortativity (the preference for nodes with similar degrees to be connected to each other) [33].

On the other hand, Milo *et al.* [24] have discovered subgraphs or motifs that are significantly more frequent in actual networks than in the corresponding FDS ensemble. Identifying these motifs allows a classification of networks that share the same motifs. For instance feed-forward loops are overrepresented in gene regulation networks and in some electronic circuits, while fully connected triangles are most overrepresented in the World Wide Web.

So, on the one hand we see that the FDS ensemble, together with nontrivial (power law) degree distributions, allows both discrimination between features of the network and comparison with other networks; on the other hand, the ensemble itself exhibits strong correlations. To explain how these correlations are related to each other and to the degree sequence it is useful to have an analytic approach. This is also important if one wants to develop more refined null models, or study very large networks for which rewiring is prohibitive. While most authors have considered the FDS ensemble as the most natural null model for networks, there have been attempts to generalize to more complex ensembles. Maybe the most interesting is due to Mahadevan *et al.* [34].

C. Previous analytic approaches

The present paper builds on papers by Burda *et al.* [20] and Berg and Lässig [35]. An alternative strategy to incorporate information on degree distributions was proposed by Park and Newman [17,18]. While we fix the degree sequence $\{k(m), m=1,2,\dots\}$ *exactly*, Park and Newman constrain only the ensemble averages $\langle k(m) \rangle$. Thus, while our approach is microcanonical, the one of [17,18] is grand canonical. As in statistical mechanics, calculations are often simpler in the grand canonical ensemble, but they are feasible and not too difficult for the PFDS($m+1$) ensemble considered in this paper, with m small. Note that for finite sized networks, the two ensembles are not equivalent. Further, for a given network, physical arguments may suggest that one ensemble is more explanatory than another.

IV. LINK LIKELIHOODS p_{ij} AND DISASSORTATIVITY IN NULL MODELS

For undirected networks, all pairs of nodes with the same degree have the same likelihood to be connected in the FDS ensemble. For directed networks the likelihood to form a link from node i with out-degree $k_{out,i}$ to node j with in-degree $k_{in,j}$ also depends on $k_{in,i}$ and $k_{out,j}$. This is demonstrated in Fig. 1, where the actual link likelihood estimated using a rewiring algorithm is plotted vs the naive analytic estimate of the link likelihood $k_{out,i}k_{in,j}/L$ (for pairs in the directed networks) or $k_i k_j/2L$ (for pairs in the undirected network). In particular, directed networks exhibit a high degree of scatter for the same values of the connected (out- and in-) degrees, showing the importance of the other degrees associated with the pair (in- and out-, respectively). Further, the likelihood does not approach the naive estimate for $k_i k_j \ll L$. This is due to the constraint of one link between two nodes and to the presence of hubs, which thus have to distribute their links to different nodes.

For any ensemble A, let $N_A(k,k')$ be the average number of links between nodes with degree k and nodes with degree k'. In terms of the link likelihood,

$$N_A(k,k') = \sum_{i,j} p_{ij,A}\,\delta(k_i - k)\,\delta(k_j - k'), \qquad (7)$$

where the sum over i,j indicates a sum over all pairs of nodes, and $p_{ij,A}$ is the link likelihood for ensemble A. If the
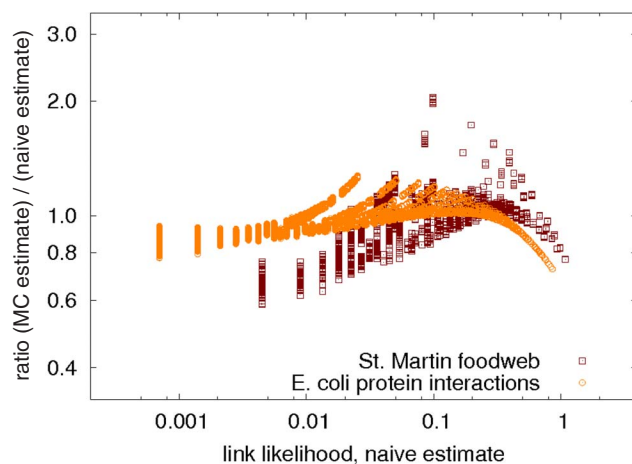

FIG. 1. (Color online) Log-log scatter plot of the naive analytic estimate of link likelihood $k_{out,i}k_{in,j}/L$ (directed network, St. Martin foodweb [37]) or $k_i k_j/2L$ (undirected network, *Escherichia coli* protein interaction network [38]) versus the ratio of the Monte Carlo rewiring estimate to the naive estimate for the corresponding nodes. Note that the directed network has considerably more scatter for given $k_{out,i}, k_{in,j}$.

ensemble A is the trivial ensemble consisting of just one network, namely the experimentally observed graph $\mathcal{G}$ with adjacency matrix $\mathcal{M}_\mathcal{G}$, then $p_{ij,A} = M_{\mathcal{G},ij}$.

Recall the average degree $\langle k' \rangle_k$ of neighbors of nodes with degree k, as given by Eq. (6). This quantity can be related to the (dis)assortativity, i.e., the tendency of nodes to connect (less) preferentially to nodes with similar degree. The assortativity was formally introduced by Newman as the Pearson correlation coefficient for the degrees of any two nodes connected by an edge [23]. Intuitively, when the average degree $\langle k' \rangle_k$ is an increasing function of k then the network shows assortative mixing, i.e., nodes of low degree tend to connect to nodes of low degree and nodes of high degree tend to connect to nodes of high degree. When $\langle k' \rangle_k$ is flat, the network shows no assortativity, and when $\langle k' \rangle_k$ is a decreasing function of k then the network shows disassortative mixing [36].

We can compute $\langle k' \rangle_k$ in several PFDS ensembles. The ensemble $Z_3(k_1,k_2)$ consists of uncorrelated random graphs with N nodes, L edges, and no multiple or self-connections, where we fix the degrees of one pair of nodes. Evidently, we choose the pair whose link likelihood is being evaluated. Equation (6) is then calculated by averaging over all pairs of nodes in the network. This clearly allows us to take the whole degree sequence into consideration, although only pairs of node degrees are considered simultaneously. To include the presence of a hub, we work in $Z_4(k_1,k_2,k_{max})$, the ensemble of uncorrelated random graphs with N nodes, L edges, and no multiple or self-connections, where we fix the degree of the pair of nodes whose link likelihood is being evaluated, as well as the degree k_{max} of the strongest hub.

As shown in Sec. V and in Appendix B, we can compute $\langle k' \rangle_k$ exactly in $Z_3(k_1,k_2)$ and $Z_4(k_1,k_2,k_{max})$, as well as in the approximate $Z_3(k_1,k_2)$ ensemble with p_{ij} given by Eq. (1) [see also Eq. (22) below], which becomes exact in the limit
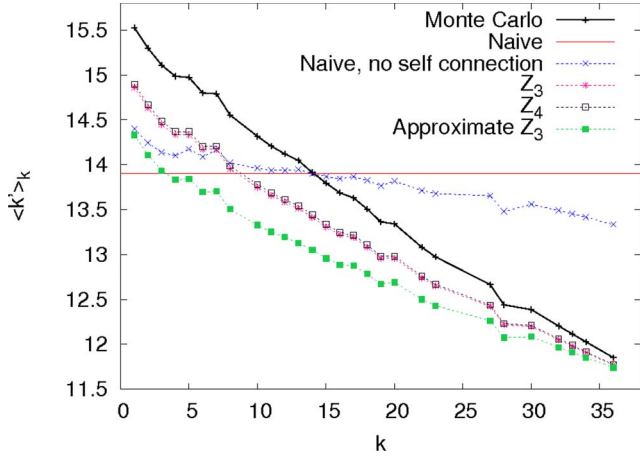
FIG. 2. (Color online) Average degree of the neighbor $\langle k' \rangle_k$ vs node degree k for an *Escherichia coli* protein interaction network [38] in several ensembles. The FDS ensemble, sampled by Monte Carlo rewiring, shows disassortativity as $\langle k' \rangle_k$ is a decreasing function of k. For the naive estimate $p_{ij} = k_i k_j / 2L$ and using the exact degree sequence, there is no disassortativity [while using a power law as in Eq. (10) leads to divergence]. However, using the naive estimate but forbidding self-connection results in slight disassortativity. Approximate Z_3, exact Z_3, and $Z_4(k_1, k_2, k_{max})$ are increasingly refined estimates of the FDS ensemble. Notice that the latter two can hardly be distinguished.

of large N for sparse networks, and for $k_{max} \ll L$. In Fig. 2 we plot $\langle k' \rangle_k$ versus k for an *Escherichia coli* protein interaction network [38]. The FDS ensemble, as sampled by the Monte Carlo rewiring procedure, is clearly disassortative, while the estimate of $\langle k' \rangle_k$ using the standard naive estimate $p_{ij} = k_i k_j / 2L$ shows no disassortativity or assortativity, as expected. We note that forbidding self-connection but otherwise using the naive estimate for p_{ij} results in slight disassortativity, while approximate Z_3, exact $Z_3(k_1, k_2)$, and $Z_4(k_1, k_2, k_{max})$ are increasingly refined estimates of the FDS behavior (see Fig. 2). Finally, $Z_4(k_1, k_2, k_{max})$ gives only a slight improvement over Z_3, indicating that hubs *per se* are less important to global properties such as disassortativity than constraints such as no self-or multiple connections, which are already implemented at the level of Z_3, along with information about the whole degree sequence, taken in degree pairs.

The approximate Z_3 ensemble is of further interest because $\langle k' \rangle_k$ can be calculated analytically. Note that

$$N(k,k') = \sum_{i,j} p_{ij}^{model} \delta(k_i - k) \delta(k_j - k') = P(k,k') N(k) N(k'). \quad (8)$$

For a power law degree distribution,

$$N(k) = (\gamma - 1) N k^{-\gamma}, \quad (9)$$

where N is the number of nodes in the network and the degree distribution of the network is a power law with exponent γ. In this case the sums over k' in Eq. (6) can be approximated by integrals, yielding
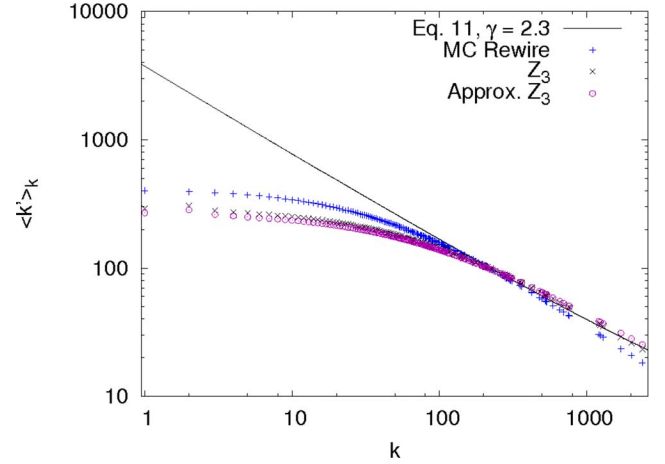

FIG. 3. (Color online) $\langle k' \rangle_k$ versus k for the Internet at the AS level from [39]. We plot the analytic estimate of Eq. (11) for $\gamma = 2.3$; for comparison we plot the results in the approximate Z_3 and Z_3 ensembles, computed directly from the degree sequence of [39], as well as Monte Carlo rewiring estimates of the FDS ensemble.

$$\langle k' \rangle_k = \frac{\int_1^\infty k' P(k,k') k'^{-\gamma} dk'}{\int_1^\infty P(k,k') k'^{-\gamma} dk'}. \quad (10)$$

If we approximate $P(k,k')$ by Eq. (1), these integrals can be solved in terms of hypergeometric functions:

$$\langle k' \rangle_k = \frac{\gamma - 1}{\gamma - 2} \frac{{}_2F_1\left(1, \gamma - 2; \gamma - 1; \frac{-2L}{k}\right)}{{}_2F_1\left(1, \gamma - 1; \gamma; \frac{-2L}{k}\right)}. \quad (11)$$

We note that for $\gamma = 2.5$ the hypergeometric functions can be expressed in terms of elementary functions:

$$\langle k' \rangle_k = \frac{\arcsin\left(\sqrt{\frac{2L}{2L+k}}\right)}{\left(\frac{k}{2L}\right)^{1/2} - \left(\frac{k}{2L}\right)\arctan\left(\left(\frac{k}{2L}\right)^{-1/2}\right)}. \quad (12)$$

To test the validity of Eq. (11), we turn to a large network, specifically Newman's recent AS level Internet data [39], for which $L = 48436$. In [17] it is reported that $\gamma \approx 2.2 \pm 0.3$ for the Internet; we estimate $\gamma \approx 2.1 \pm 0.3$ for Ref. [39].

In Fig. 3 we plot the analytic estimates of $\langle k' \rangle_k$ given by Eq. (11) for $\gamma = 2.3$. This value for γ gives the best fit for $\langle k' \rangle_k$ and is within the uncertainty of the direct degree distribution measurement of γ. For comparison we also show the results in the approximate Z_3 and Z_3 ensembles, computed directly from the degree sequence of [39], as well as Monte Carlo rewiring estimates of the FDS ensemble. Note the strong similarities between the Z_3 results and the Monte Carlo estimates of the FDS ensemble; in particular, we ob-

serve a flattening of $\langle k' \rangle_k$ for small k both in the Z_3 ensemble and in the Monte Carlo rewiring. This is consistent with the observations of [17].

Also note the similar scaling of the various estimates and of the Monte Carlo results at large k. For the Internet, $\langle k' \rangle_k$ has been reported to scale with k as a power law, $\langle k' \rangle_k \approx k^{-\nu}$ with $\nu \approx 0.5$ [17]. Our Monte Carlo results for the FDS ensemble, using the degree sequence of Ref. [39], show indeed such a power law for large k, but with $\nu \approx 0.75$. The exact and approximate Z_3 calculations, obtained with the exact degree sequence, give $\nu \approx 0.7$ resp. 0.62. When using a power law degree sequence and Eq. (11), the scaling depends on γ. But in this case, one can verify that scaling does not hold in the large k limit, but in the limit $k \ll L$. The curvature of the continuous line visible in Fig. 3 results entirely from the fact that k is not much less than L. Thus it is the slope of the continuous line at small k which should be used for extracting ν. With this, one finds that ν varies from ≈ 0.79 for $\gamma = 2.2$ to 0.5 for $\gamma = 2.5$. Note that $\nu = 0.5$ for $\gamma = 2.5$ is an exact result that can be obtained analytically by taking the limit $2L/k \to \infty$ in Eq. (12).

In contrast to the approach of [17], all of our results can be computed directly from the degree distribution or the degree sequence, omitting the intermediate step of constructing a fugacity distribution to match the statistics of the degree distribution and then extracting $\langle k' \rangle_k$ from the fugacities. The fact that the disassortativity properties of the Internet can be studied so directly in the simple approximate Z_3 ensemble suggests that Eq. (1) should replace the naive estimate $P(k,k') = kk'/2L$ in other applications, for example in the study of motifs. This is explored in Sec. V B [see Eq. (25)].

For undirected networks and any null model A, Maslov *et al.* [33] defined a quantity called the correlation profile $R(k,k') = N_{\mathcal{G}}(k,k')/N_A(k,k')$ and the $\mathcal{Z}$-score $\mathcal{Z}(k,k') = [N_{\mathcal{G}}(k,k') - N_A(k,k')]/\sigma_A(k,k')$, where $N_A(k,k')$ is defined as in Eq. (7), $N_{\mathcal{G}}(k,k')$ is the analogous quantity for the trivial ensemble (with p_{ij} replaced by $M_{\mathcal{G},ij}$), and $\sigma_A^2(k,k')$ is the variance of the number of links connecting nodes with degrees k and k' in ensemble A [remember that $N_A(k,k')$ was the average of that number]. The specific null model studied in [33] was the FDS ensemble. As shown in Appendix A, a similar analysis can be done comparing different null models to each other. Results are also discussed in Appendix A.

V. ANALYTIC ESTIMATES OF THE FDS ENSEMBLE FOR UNDIRECTED NETWORKS

A. Notation and basic identities

We now derive our principal analytic results. Our central object is the partition function Z, which counts the number of graphs in the ensemble. The elementary constraints on the network (N nodes, no multiple or self-connections) imply that the adjacency matrix $\mathcal{M}$ is $N \times N$, is symmetric (for undirected networks), has zeroes along the diagonal, and consists solely of 0's and 1's. If we add the constraint of L links, the partition function can be written as
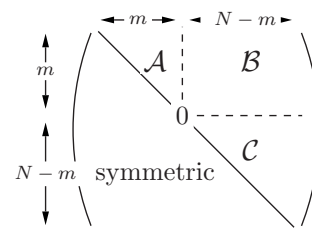

FIG. 4. Schematic decomposition of the adjacency matrix $\mathcal{M}$ into the three submatrices $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ discussed in the text. Note that $\mathcal{M}$ is symmetric for undirected networks.

$$Z_1(N,L) = \sum_{\substack{\{M_{ij}=0,1\} \\ i<j}} \delta\left(L - \sum_{i<j} M_{ij}\right), \qquad (13)$$

where the sum is over the upper triangle of $\mathcal{M}$ due to symmetry. A simple computation gives Eq. (3) for the number of ways to distribute L links among $\binom{N}{2}$ possible pairs of nodes.

Now let us specify the degrees of m of the nodes. We refer to m as the "order" of a calculation. The partition function becomes

$$Z_{m+1}(N,L,k_1,\ldots,k_m) = \sum_{\substack{\{M_{ij}\} \\ i<j}} \delta\left(L - \sum_{i<j} M_{ij}\right)$$
$$\times \prod_{l=1}^{m} \delta\left(k_l - \sum_{j=1}^{l-1} M_{jl} - \sum_{j=l+1}^{N} M_{lj}\right), \qquad (14)$$

where we use the symmetry of the adjacency matrix to write the degree constraints in terms of the variables M_{ij} with $i < j$.

It will be helpful at this point to introduce some further notation to assist us in organizing this calculation. We split the matrix $\mathcal{M}$ into four pieces: $\mathcal{A}$, the square submatrix controlling the edges linking the m nodes with fixed degree to each other; $\mathcal{B}$, the rectangular matrix encoding the connections of the m nodes of fixed degree with the rest of the nodes in the graph, and its transpose $\mathcal{B}^T$; $\mathcal{C}$, the square submatrix encoding the edges among the $N-m$ remaining nodes (whose degrees are not specified). Due to the symmetry of $\mathcal{M}$, only $\mathcal{B}$ and the upper triangular parts of $\mathcal{A}$ and $\mathcal{C}$ are independent. In Fig. 4 we present a schematic decomposition of $\mathcal{M}$.

The sum over all adjacency matrices decomposes into a sum over the $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ submatrices, with suitable constraints. In particular, we write the symbol $\Sigma_{\{\mathcal{A}\}}$ for the sum over all possible values $\{0,1\}$ of the matrix elements of the submatrix $\mathcal{A}$. Each term of this sum corresponds to a particular (possibly disconnected) labeled subgraph involving m nodes of fixed degree. This is analogous to a diagrammatic expansion of the partition function, where the partition function is now written as the sum over all possible subgraphs involving nodes 1 through m, and each subgraph is weighted by a degeneracy factor resulting from the summations over $\mathcal{B}$ and $\mathcal{C}$. This degeneracy counts the number of possible graphs

in the ensemble containing that particular subgraph (respectively submatrix) $\mathcal{A}$. The partition function is written in this notation as

$$Z_{m+1} = \sum_{\{\mathcal{A}\}} Z_{m+1}(\mathcal{A}), \tag{15}$$

where $Z_{m+1}(\mathcal{A})$ is the partition function or degeneracy factor for a given fixed submatrix (equivalently, subgraph) $\mathcal{A}$. For each $Z_{m+1}(\mathcal{A})$, the nodes with fixed degrees (i.e., the first m nodes) are connected in a specified way. Thus, for example, the probability of some particular $m \times m$ subgraph $\mathcal{A}$ occurring would be

$$\text{Prob}(\mathcal{A}) = \frac{Z_{m+1}(\mathcal{A})}{Z_{m+1}}. \tag{16}$$

As shown in Appendix B, the degeneracy of a given subgraph $Z_{m+1}(\mathcal{A})$ can be written as

$$Z_{m+1}(\mathcal{A}) = \begin{pmatrix} \binom{N-m}{2} \\ L + \sum_{i<j,=2}^{j=m} A_{ij} - \sum_{l=1}^{m} k_l \end{pmatrix}$$
$$\times \prod_{l=1}^{m} \begin{pmatrix} N-m \\ k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj} \end{pmatrix}. \tag{17}$$

The first term on the right-hand side is the degeneracy associated with the upper half triangle of the square submatrix $\mathcal{C}$. Recall that the submatrix $\mathcal{C}$ defines connections between all $N-m$ nodes not in $\mathcal{A}$, i.e., all nodes with free degree. This matrix has $\binom{N-m}{2}$ independent places to put a specified number of 1's. The number of 1's in $\mathcal{C}$ depends on L, the total number of 1's in the entire (upper triangular) adjacency matrix, minus the number of 1's from edges that have at least one end on a node of fixed degree. By definition, those 1's cannot appear in $\mathcal{C}$. Due to the symmetry of the entire matrix, the number of 1's to be placed in the upper triangular half of $\mathcal{C}$ is $L + \sum_{i<j,j=2}^{j=m} A_{ij} - \sum_{l=1}^{m} k_l$.

Each factor in the product $\prod_{l=1}^{m}$ of Eq. (17) is the degeneracy associated with a row in the matrix $\mathcal{B}$. For every such row there are $N-m$ places to put 1's, and the number of 1's that must be placed in the l-th row is the degree of the node k_l minus the number of 1's in the corresponding row of the entire matrix $\mathcal{A}$. This latter number is the degree of the node within the subgraph $\mathcal{A}$.

B. Calculation of link likelihoods for undirected networks

As a first application we compute the link likelihood for two nodes in this framework. At lowest order we can just specify the degrees of the two nodes under consideration, giving the ensemble Z_3 with $m=2$. For convenience, the two nodes are labeled 1 and 2. The two possible configurations of this subsystem, where the nodes are either connected or disconnected, must be weighted by the appropriate degeneracy factors according to Eq. (17). Let us denote these two con-

figurations by $\aleph$ and $\aleph'$. In $\aleph$ an edge connects the two nodes, so $\aleph_{12} = 1$. We denote this "connected" part of the total partition function $Z_3^{conn} = Z_3(\aleph)$. In $\aleph'$ there is no edge between the two nodes, so $\aleph'_{12} = 0$. We denote this "disconnected" part of the total partition function $Z_3^{disc} = Z_3(\aleph')$. The total partition function is thus $Z_3^{total} = Z_3^{conn} + Z_3^{disc} = Z_3(\aleph) + Z_3(\aleph')$, so

$$p_{12}^{(m=2)} = \frac{Z_3^{conn}}{Z_3^{total}} = \frac{Z_3(\aleph)}{Z_3(\aleph) + Z_3(\aleph')}. \tag{18}$$

From Eq. (17), the explicit expressions are

$$Z_3(\aleph) = \begin{pmatrix} \binom{N-2}{2} \\ L + 1 - k_1 - k_2 \end{pmatrix} \prod_{l=1}^{2} \binom{N-2}{k_l - 1}, \tag{19}$$

$$Z_3(\aleph') = \begin{pmatrix} \binom{N-2}{2} \\ L - k_1 - k_2 \end{pmatrix} \prod_{l=1}^{2} \binom{N-2}{k_l}. \tag{20}$$

Straightforward calculation gives

$$p_{12}^{(m=2)} = \left[1 + \frac{(L+1-k_1-k_2)(N-1-k_1)(N-1-k_2)}{k_1 k_2 ((N-2)(N-3)/2 - L + k_1 + k_2)} \right]^{-1} \tag{21}$$

and in the limit $L \to \infty$, $L/N^2 \to 0$, and $k_i \ll L$ this reduces to

$$p_{12}^{(m=2)} \approx \frac{k_1 k_2}{2L + k_1 k_2} = P(k_1, k_2). \tag{22}$$

Thus we have derived the approximate link likelihood given by Eq. (1). In the limit $k_1 k_2 \ll L$ we recover from Eq. (22) the naive estimate $k_1 k_2 / 2L$ used by most authors. This naive estimate is a bad approximation if either node 1 or node 2 is a hub. In general, the full expression given in Eq. (21) is a better approximation, although as we have shown in Sec. IV the approximate Z_3 ensemble given by Eq. (22) is both analytically tractable and significantly better than the naive estimate.

The presence of any hubs in the network reduces the link likelihood between two nodes, particularly nodes of low degree, as their links are "stolen" by the hubs. This effect already appears in the calculation of the disassortativity, as shown in Figs. 2 and 3. We can refine the preceding computation by including constraints on the degree sequence coming from the hubs. We incorporate these constraints from the hubs by considering the nodes with the highest degrees first. The partially fixed degree sequence ensemble $Z_4(k_1, k_2, k_{max})$, then, includes the two nodes k_1, k_2 whose link likelihood we compute (free to vary over all node pairs) and the largest hub in the network, with degree (k_{max}). In the case that $k_{max} = k_1$ or k_2 we use the next largest hub degree for the third constraint. In the submatrices $\mathcal{A}$, all possible ways to connect the three nodes are enumerated. To compute the link likelihood $p_{ij}^{(m=3)}$ we divide the partition function into connected Z_4^{conn} and disconnected Z_4^{disc} parts, where connected means an edge connects node 1 and node 2, i.e., $A_{12} = 1$. The diagrammatic expansion for the connected sub-
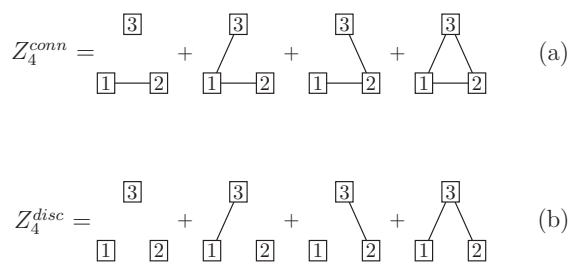
$$Z_4^{conn} = \; \boxed{3} \quad + \quad \boxed{3} \quad + \quad \boxed{3} \quad + \quad \boxed{3} \qquad \text{(a)}$$

$$Z_4^{disc} = \; \boxed{3} \quad + \quad \boxed{3} \quad + \quad \boxed{3} \quad + \quad \boxed{3} \qquad \text{(b)}$$

FIG. 5. Diagrammatic expansion for the connected (a) and disconnected (b) contributions to the partition function Z_4 for p_{12}.

partition function is shown in Fig. 5(a), and the disconnected subpartition function is shown in Fig. 5(b).

So we can compute the link likelihood as

$$p_{12}^{(m=3)} = \frac{Z_4^{conn}}{Z_4^{conn} + Z_4^{disc}} = \frac{Z_4(\aleph_1) + Z_4(\aleph_2) + Z_4(\aleph_3) + Z_4(\aleph_4)}{\sum_{i=1}^{8} Z_4(\aleph_i)},$$

$$(23)$$

where $\aleph_1$ to $\aleph_8$ denote the adjacency matrices for the eight graphs shown in Fig. 5. In Fig. 6 we compare the link likelihood p_{ij} for an undirected network in the FDS ensemble (obtained numerically by the rewiring method) to the analytic results in the ensembles $Z_3(k_1, k_2)$ and $Z_4(k_1, k_2, k_{max})$. For each pair (i, j) we plot the naive estimate $\frac{k_i k_j}{2L}$ on the horizontal axis; the vertical coordinate is the ratio of the numerical (Monte Carlo) estimate to the naive, no hub, and one hub estimates for that pair. The Z_3 estimate is already a substantial improvement over the naive analytic estimate, with slight further refinement coming as expected from the inclusion of hubs in all the diagrams.
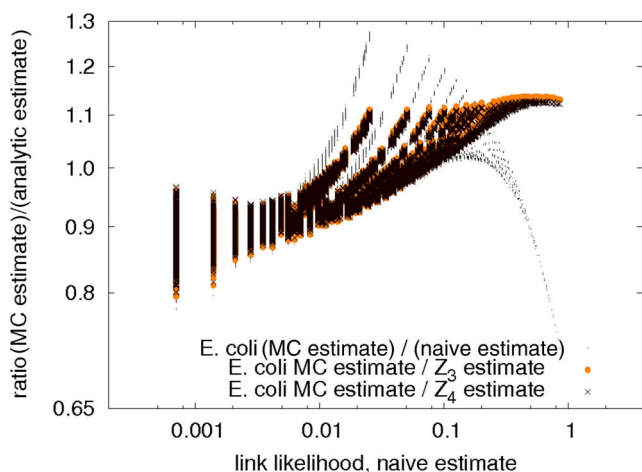

FIG. 6. (Color online) Scatter plot comparing estimates of the link likelihood in an *Escherichia coli* protein interaction network. On the horizontal logarithmic axis we plot the naive estimate $\frac{k_i k_j}{2L}$; on the vertical logarithmic axis we plot for all corresponding nodes the ratio of the Monte Carlo rewiring estimate to the naive, Z_3 (no hub), and Z_4 (one hub) estimates. The latter two nearly coincide.

C. Calculation of subgraph likelihoods for undirected networks

Estimating the likelihoods of larger subgraphs can be done along exactly the same lines. As an example, we estimate the number of triangles in an undirected network and test this estimate on an *Escherichia coli* protein interaction network [38], a yeast (*Saccharomyces cerevisiae*) protein interaction network [40], two artificial yeast protein interaction networks created by modifying the degree sequence of [40] by hand to make it narrower or broader, the Newman AS level map of the Internet [39], and a symmetrized snapshot of the World Wide Web [41]. Working in an ensemble with four constraints, $Z_4(k_1, k_2, k_3)$, we consider all permitted triples of nodes (i, j, k), forbidding self- and multiple connection. Note that node 3 is no longer fixed as the largest hub, but allowed to range over all nodes. Given fixed nodes 1, 2, and 3, we can compute the likelihood of a triangle between them as

$$p_\Delta^{(m=3)} = \frac{Z_4(\aleph_4)}{\sum_{i=1}^{8} Z_4(\aleph_i)}, \qquad (24)$$

where $\aleph_4$ corresponds to the fully connected subgraph, i.e., the last term in the sum of Fig. 5(a). The resulting combinatorial expression is quite unwieldy. However, in the limit $L \to \infty$, $L/N^2 \to 0$, and $1 \ll k_i \ll L$, i.e., the large, sparse network limit used in deriving Eq. (22) with the additional assumption of $1 \ll k_i$, we find a remarkable simplification. The expression factorizes to

$$p_\Delta^{(m=3)} \approx P(k_1, k_2) P(k_1, k_3) P(k_2, k_3), \qquad (25)$$

where $P(k_i, k_j)$ is given by Eq. (22).

We now test these formulas against the various trial networks. The results are shown in Table I, where "MC" represents the averaged triangle count for many Monte Carlo rewirings, i.e., a numerical estimate of the average number of triangles in the FDS ensemble. The only noticeable trend is the decrease in the absolute value of the % Error (defined as [Eq. (24) − Eq. (25)]/ Eq. (24)) between the simple factorized expression of Eq. (25) and the elaborate expression of Eq. (24) as N increases. This verifies the approximation made in deriving Eq. (25). The time required to make the "complex" estimate given by Eq. (24) is roughly equal to the time required to count the number of triangles for a single Monte Carlo instance; the simple estimate given by Eq. (25) is much faster, running on the WWW data in roughly 90 sec (on an Intel core duo processor) without any optimization for speed. We note that in many cases the % Error is rather large for both approximations [Eq. (24) and Eq. (25)], although of roughly the same size as the % Error of Ref. [42] for many of the "triangle" subgraphs in directed networks. In some cases Eq. (24) appears to be the better approximation, and in others Eq. (25). Unlike the case for disassortativity, neither approximation scheme seems to provide a completely satisfactory estimate of the number of triangles, and further work is needed to develop approximations that are better able to capture this quantity.

TABLE I. Comparing estimates of the total number of triangles in various networks with N nodes. MC refers to Monte Carlo rewiring estimates of the FDS ensemble. As expected, the results of Eq. (24) approach the asymptotic result, Eq. (25), for large, sparse networks.

Network	N	Eq. (25)	Eq. (24)	MC	% Error Eq. (25) vs Eq. (24)	% Error Eq. (25) vs MC	% Error Eq. (24) vs MC
E. coli	230	215.82	289.65	322.14	25.49	33.01	10.09
Yeast(narrow)	1373	302.77	247.65	339.10	−22.26	10.71	26.97
Yeast	1373	651.07	592.59	1160.39	−9.87	43.89	48.93
Yeast (broad)	1373	1553.94	1667.70	2813.37	6.82	44.77	40.72
AS Internet	22963	29157.38	31840.23	37810.68	8.43	22.89	15.79
WWW	325729	379371.15	379706.63	274926.89	0.09	−37.99	−38.11

It is tedious to verify similar factorizations for larger motifs. However, the convergence to Eq. (25) in the predicted limit provides further evidence for the replacement of the naive estimate $P(k,k')=kk'/2L$ used, e.g., in [42], where factorizations as in Eq. (25) were assumed, with Eq. (22). This will be the topic of future work. Such extensions of our method can also be used to study larger motifs in complex networks [43,44], or to study large networks where computational time for rewiring grows prohibitively, but the approximation underlying Eqs. (22) and (25) should still be valid.

VI. CONCLUSION

Detecting and describing local structure is an important frontier in the study of complex networks, as many of the features distinguishing real-world networks from their random analogs or null models are local: degree-degree correlations, motifs, and so forth. One of the major obstacles to this project is the lack of analytical techniques to study the fixed degree sequence ensemble, which is the most common null model for complex networks associated with the rewiring method. In this paper we have reviewed the numerical tools for studying the FDS ensemble and discussed some of the practical uses (e.g., disassortativity, motif calculation, correlation profiles) to which knowledge of local structure can be put. Through a careful study of the partition function of the FDS ensemble and the PFDS ensembles containing it, we derive simple and general combinatorial expressions that improve naive estimates of the link likelihood by explicitly including important constraints from the FDS ensemble (the exclusion of multiple edges and self-connections, and the appearance of a broad range of degrees) in a "Gaussian" type approximation where the set of degree constraints are treated minimally but nontrivially.

In particular, for undirected networks we have developed the analytically tractable approximate Z_3 ensemble, where the link likelihood $P(k,k')=kk'/(2L+kk')$ [Eq. (1)] gives clear disassortativity, while the naive estimate $kk'/2L$ does not. We have also introduced a diagrammatic expansion of the PFDS partition function, which organizes the combinatorial calculations usefully and leads to a simple, approximate factorized formula for the estimated number of undirected triangles between three nodes, $P_\Delta^{(m=3)}(k_1,k_2,k_3) = P(k_1,k_2)P(k_1,k_3)P(k_2,k_3)$ [Eq. (25)] where k_1, k_2, k_3 are

the degrees of the three nodes. The factorization suggests the application of Eq. (22) to extended local structures such as motifs.

It should be emphasized that these analytic results are not merely useful for the null model they have been explicitly developed to approximate (the FDS ensemble). They also provide guidance in developing more complicated null models that incorporate higher-level constraints. The astronomical $\mathcal{Z}$-scores observed in work on extended motifs [43,44] dramatize the need for such extensions, which might constrain, for example, the number of triangles in addition to the number of nodes, number of links, and a few degrees. Further work will explore applications of the approximate Z_3 and other PFDS ensembles to motif estimates, as well as the incorporation of higher-level constraints in the PFDS ensemble to improve likelihood estimates of extended motifs. The extension of the results of this paper to directed networks is in preparation.

ACKNOWLEDGMENTS

APPENDIX A: COMPARING NULL MODELS VIA THE CORRELATION PROFILE

We can study how the real network deviates from various null hypotheses by calculating $R(k,k')$ with respect to various null hypotheses. This provides an overall measure of how close the ensembles are to each other and helps establish the relevant features that distinguish the real network from the different ensembles.

In general, we can define correlation profiles and $\mathcal{Z}$-scores for any pair (A,B) of ensembles:

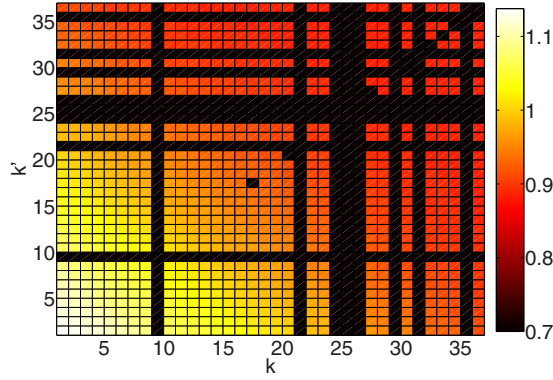$$R_{A|B}(k,k') = \frac{N_A(k,k')}{N_B(k,k')} \tag{A1}$$

and

FIG. 7. (Color online) Correlation profile $R_{Z_3|\text{FDS}}(k,k')$ for the *Escherichia coli* protein interaction network [38]. Note that the dark regions (blue online) are artificially set to the value .7; they correspond to values of (k,k') for which no data exist.

$$\mathcal{Z}_{A|B}(k,k') = \frac{N_A(k,k') - N_B(k,k')}{\sqrt{\sigma_A^2 + \sigma_B^2}}. \tag{A2}$$

In particular, we can calculate the correlation profile for the numerically sampled FDS ensemble and the ensemble $Z_3(k_1,k_2)$. Recall that the ensemble Z_3 consists of uncorrelated random graphs with N nodes, L edges, and no multiple or self-connections, where we fix the degree of the pair of nodes whose link likelihood is being evaluated. We might also compare the FDS ensemble to $Z_4(k_1,k_2,k_{\max})$, the ensemble of uncorrelated random graphs with N nodes, L edges, and no multiple or self-connections, where we fix the degree of the pair of nodes whose link likelihood is being evaluated, as well as the degree of the largest node in the network, $k_{\max}$. In Z_3 and Z_4, p_{ij} can be calculated exactly (see Sec. V). We plot $R_{Z_3|\text{FDS}}(k,k')$ for the *Escherichia coli* protein interaction network in Fig. 7. It is clear that the Z_3 ensemble captures many of the features of the FDS ensemble, as $R_{Z_3|\text{FDS}}(k,k')$ is close to 1 for all k, k'. $R_{Z_4|\text{FDS}}(k,k')$ exhibits slight improvement over Z_3, as expected (data not shown). The correlation profile allows us to identify correlations due to the degrees of the other nodes in the network and provides a test of our hypothesis that the PFDS ensemble captures much of the structure of the FDS ensemble.

APPENDIX B: DERIVATION OF Z_{m+1} FOR UNDIRECTED NETWORKS

We note that the number of possible undirected subgraphs of m nodes is $2^{(m^2-m)/2}$. So we write

$$Z_{m+1}(N,L,k_1,\ldots,k_m)$$

$$= \sum_{\{A_{ij}\}} \sum_{\{B_{ij}\}} \sum_{\{C_{ij}\}} \delta\left(L - \left(\sum_{\substack{i<j\\j=2}}^{j=m} A_{ij}\right) - \left(\sum_{\substack{i\leq m\\j=m+1}}^{j=N} B_{ij}\right)\right.$$

$$\left. - \left(\sum_{\substack{i<j\\i=m+1}}^{i=N} C_{ij}\right)\right) \prod_{l=1}^{m} \delta\left(k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj} - \sum_{j=m+1}^{N} B_{lj}\right). \tag{B1}$$

For concision we henceforth write the sum more compactly, as in the next equation.

We now Fourier transform the delta functions:

$$Z_{m+1}(N,L,k_1,\ldots,k_m)$$

$$= \sum_{\{ABC\}} \frac{1}{(2\pi)^{m+1}} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} dz\,da_1\ldots da_m\, e^{\imath z(L - \sum_{\substack{i<j\\j=2}}^{j=m} A_{ij})}$$

$$\times \prod_{l=1}^{m} e^{\imath a_l(k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj})} \prod_{l=1}^{m} \prod_{j=m+1}^{N} e^{-\imath(z+a_l)B_{lj}}$$

$$\times \prod_{\substack{i<j\\i=m+1}}^{N} e^{-\imath z C_{ij}}. \tag{B2}$$

We can do the sum over B_{ij}, yielding

$$Z_{m+1}(N,L,k_1,\ldots,k_m)$$

$$= \sum_{\{AC\}} \frac{1}{(2\pi)^{m+1}} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} dz\,da_1\ldots da_m\, e^{\imath z(L - \sum_{\substack{i<j\\j=2}}^{j=m} A_{ij})}$$

$$\times \prod_{l=1}^{m} e^{\imath a_l(k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj})} \prod_{l=1}^{m} \left(1 + e^{-\imath(z+a_l)}\right)^{N-m}$$

$$\times \prod_{\substack{i<j\\i=m+1}}^{N} e^{-\imath z C_{ij}}. \tag{B3}$$

Performing the standard binomial expansion yields

$$Z_{m+1}(N,L,k_1,\ldots,k_m)$$

$$= \sum_{\{AC\}} \frac{1}{(2\pi)^{m+1}} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} dz\,da_1\ldots da_m\, e^{\imath z(L - \sum_{\substack{i<j\\j=2}}^{j=m} A_{ij})}$$

$$\times \prod_{l=1}^{m} e^{\imath a_l(k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj})} \prod_{l=1}^{m} \sum_{n_l=0}^{N-m} \binom{N-m}{n_l} e^{-\imath n_l(z+a_l)}$$

$$\times \prod_{\substack{i<j\\i=m+1}}^{N} e^{-\imath z C_{ij}}. \tag{B4}$$

Integrating over $a_1 \cdots a_m$ sets $n_l = k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj}$, so we have

$$Z_{m+1}(N,L,k_1,\ldots,k_m)$$

$$= \sum_{\{AC\}} \frac{1}{(2\pi)} \int_{-\pi}^{\pi} dz\, e^{\imath z(L - \sum_{\substack{i<j\\j=2}}^{j=m} A_{ij})}$$

$$\times \prod_{l=1}^{m} e^{-\imath z(k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj})}$$

$$\times \binom{N-m}{k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj}} \prod_{\substack{i<j\\i=m+1}}^{N} e^{-\imath z C_{ij}}. \tag{B5}$$

We now sum over C_{ij} and perform the binomial expansion of the resulting quantity:

$$Z_{m+1}(N,L,k_1,\ldots,k_m)$$

$$= \sum_{\{\mathcal{A}\}} \frac{1}{(2\pi)} \int_{-\pi}^{\pi} dz\, e^{\imath z(L+\sum_{\substack{i<j\\j=2}}^{j=m} A_{ij} - \sum_{l=1}^{m} k_l)}$$

$$\times \prod_{l=1}^{m} \left(\begin{array}{c} N-m \\ k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj} \end{array} \right) \sum_{n=0}^{\binom{N-m}{2}} \left(\begin{array}{c} \binom{N-m}{2} \\ n \end{array} \right) e^{-\imath z n},$$

$$(B6)$$

where we have used the fact that

$$\prod_{l=1}^{m} e^{\imath z(\sum_{j=1}^{l-1} A_{jl} + \sum_{j=l+1}^{m} A_{lj})} = e^{\imath z(\sum_{l=1}^{m}(\sum_{j=1}^{l-1} A_{jl} + \sum_{j=l+1}^{m} A_{lj}))} = e^{2\imath z\sum_{\substack{l<j\\j=2}}^{j=m} A_{lj}}$$

by the symmetry of the adjacency matrix. We may now integrate over z to give the result:

$$Z_{m+1}(N,L,k_1,\ldots,k_m)$$

$$= \sum_{\{\mathcal{A}\}} \left(\begin{array}{c} \binom{N-m}{2} \\ \left(L + \sum_{i<j,j=2}^{j=m} A_{ij} - \sum_{l=1}^{m} k_l \right) \end{array} \right)$$

$$\times \prod_{l=1}^{m} \left(\begin{array}{c} N-m \\ k_l - \sum_{j=1}^{l-1} A_{jl} - \sum_{j=l+1}^{m} A_{lj} \end{array} \right). \qquad (B7)$$

This can be written as a sum over subpartition functions $Z_{m+1}(\mathcal{A})$, each of which is given by Eq. (17). Thus we recover the result for fixed $\mathcal{A}$ given in Sec. V.

[1] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).
[2] T. Schreiber and A. Schmitz, Physica D **142**, 346 (2000).
[3] S. Maslov and K. Sneppen, Science **296**, 910 (2002).
[4] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon, e-print arXiv:cond-mat/0312028v2.
[5] Y. G. Chen, P. Diaconis, S. R. Holmes, and J. S. Liu, J. Am. Stat. Assoc. **100**, 109 (2005).
[6] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[7] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).
[8] S. Gupta, R. M. Anderson, and R. M. May, AIDS **3**, 807 (1989).
[9] M. Kretzschmar, Y. T. H. P. van Duynhoven, and A. J. Severijnen, Am. J. Epidemiol. **144**, 306 (1996).
[10] J. Camacho, R. Guimerà, and L. A. Nunes Amaral, Phys. Rev. Lett. **88**, 228102 (2002).
[11] S. A. Kauffman, in *Current Topics in Developmental Biology 6*, edited by A. Moscana and A. Monroy (Academic Press, New York, 1971), pp. 145–182.
[12] J. Davidsen, P. Grassberger, and M. Paczuski, Geophys. Res. Lett. **33**, L11304 (2006).
[13] M. Paczuski, "Networks as Renormalized Models for Emergent Behavior in Physical Systems," in *Complexity, Metastability and Nonextensivity*, edited by C. Beck, G. Benedek, A. Rapisarda, and C. Tsallis (World Scientific, London, 2005).
[14] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[15] A. S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[16] B. Bollobas, *Random Graphs* (Cambridge Univ. Press, 2001).
[17] J. Park and M. E. J. Newman, Phys. Rev. E **68**, 026112 (2003).
[18] J. Park and M. E. J. Newman, Phys. Rev. E **70**, 066117 (2004).
[19] This is the number of ways of placing L identical objects (in this case "1"s) in $N(N-1)/2$ places.
[20] L. Bogacz, Z. Burda, and B. Waclaw, Physica A **366**, 587 (2006).
[21] B. Y. Wang and F. Zhang, Discrete Math. **187**, 211 (1998).
[22] E. A. Bender and E. R. Canfield, J. Comb. Theory, Ser. A **24**, 296 (1978).
[23] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002); Phys. Rev. E **67**, 026126 (2003); S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon, Phys. Rev. E **68**, 026127 (2003).
[24] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
[25] S. J. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2001).
[26] P. Grassberger, Comput. Phys. Commun. **147**, 64 (2002).
[27] J. Besag and P. Clifford, Biometrika **76**, 633 (1989).
[28] A. R. Rao, R. Jana, and S. Bandyopadhya, Sankhya, Ser. A **58**, 225 (1996).
[29] J. M. Roberts, Jr., Soc. Networks **22**, 273 (2000).
[30] G. W. Cobb and Y.-P. Chen, Am. Math. Monthly **110**, 265 (2003).
[31] In many cases one wants to construct a surrogate ensemble with the same degree sequence as a given graph $\mathcal{G}$. If a given graph $\mathcal{G}$ is available, one does not need this first step, and starts from $\mathcal{G}$ directly.
[32] H. J. Ryser, Can. J. Math. **9**, 371 (1957).
[33] S. Maslov, K. Sneppen, and A. Zaliznyak, Physica A **333**, 529 (2004).
[34] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, e-print arXiv:cs.NI/0605007v2.
[35] J. Berg and M. Lässig, Phys. Rev. Lett. **89**, 228701 (2002).
[36] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).
[37] L. Goldwasser and J. Roughgarden, Ecology **74**, 1216 (1993).
[38] G. Butland *et al.*, Nature (London) **433**, 531 (2005).
[39] M. Newman, data at www-personal.umich.edu/~mejn/netdata/
[40] A. C. Gavin *et al.*, Nature (London) **440**, 631 (2006); data received from G. Musso (University of Toronto).
[41] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999); data at www.nd.edu/~networks/resources.htm
[42] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon, Phys. Rev. E **68**, 026127 (2003).
[43] K. Baskerville and M. Paczuski, Phys. Rev. E **74**, 051903 (2006).
[44] K. Baskerville, P. Grassberger, and M. Paczuski, Phys. Rev. E **76**, 036107(2007).