Subgraph fluctuations in random graphs

Christoph Fretter and Matthias Müller-Hannemann
Institut für Informatik, Martin-Luther-Universität Halle-Wittenberg, 06120 Halle, Germany

Marc-Thorsten Hütt
School of Engineering and Science, Jacobs University, 28759 Bremen, Germany
(Received 13 September 2011; published 29 May 2012)

The pattern of over- and under-representations of three-node subgraphs has become a standard method of characterizing complex networks and evaluating how this intermediate level of organization contributes to network function. Understanding statistical properties of subgraph counts in random graphs, their fluctuations, and their interdependences with other topological attributes is an important prerequisite for such investigations. Here we introduce a formalism for predicting subgraph fluctuations induced by perturbations of unidirectional and bidirectional edge densities. On this basis we predict the over- and under-representation of subgraphs arising from a density mismatch between a network and the corresponding pool of randomized graphs serving as a null model. Such mismatches occur, for example, in modular and hierarchical graphs.

DOI: 10.1103/PhysRevE.85.056119 PACS number(s): 89.75.Hc, 02.10.Ox

I. INTRODUCTION

Network science, i.e., the discipline studying and interpreting a broad range of complex systems from a network perspective, has an enormous impact on how we perceive (and analytically approach) social, biological, and technical systems. One of the most fascinating theoretical challenges of network science is the interdependence of network properties observed at different scales: clustering depends on modularity, heavy-tailed degree sequences can induce degree-degree correlations [1,2], a modular structure influences our expectations of betweenness centralities, and other edge- or node-based properties. The severest impact of these interdependences probably occurs when attempting to interpret the composition of a network in terms of few-node subgraphs. On this level, we can expect a very strong influence of global network properties unless we adjust our null model (i.e., the set of random expectations) to match these global properties. It is therefore essential to understand this interplay from first principles. Here we discuss two types of correlations between network properties: (i) how single-edge fluctuations influence fluctuations in three-node subgraph frequencies and (ii) how global network properties affect three-node subgraphs frequencies.

Network motifs were introduced as a method for analyzing transcriptional regulatory systems [3]. A comparison of the transcriptional regulatory network of the bacterium *E. coli* with random graphs has revealed that three characteristic local node and link patterns appear substantially more frequently than expected at random [4]: feed-forward loops (FFLs), single-input modules (SIMs), and densely overlapping regulons. The benefit from an identification of over-represented node and link patterns is twofold: (i) One can formulate models of the dynamics encoded by such few-node devices and (ii) one can discuss selected examples of such motif occurrences in detail. In this way, feed-forward loops and single-input modules could, in subsequent work [4–6], be linked to specific dynamical functions [such as noise buffering (FFL) and the implementation of temporal programs (SIM)].

To a certain extent, the analysis of such node and link patterns is a balance between an automatized statistical view on a complex network and the discussion of individual cases. An interesting example of this balance is the discussion of various types of feed-forward loops in transcriptional regulatory networks. Once the statistical over-representation of this node and link pattern had been established [4,5], the specific forms of feed-forward loops occurring in the networks could be further analyzed. One classification scheme is to enumerate all distributions of signs on the links (activating and inhibitory) and see whether the two paths (directly and via the third, intermediate node) from the top-level node to the bottom-level node in the feed-forward loop both provide the same signal (both activating or both inhibiting) (coherent FFL) or conflicting signals (incoherent FFL). Surprisingly, not all variants of these coherent and incoherent FFLs seem to occur in equal proportions in transcriptional regulatory networks. Instead there seems to be a strong bias toward only one type of coherent FFL and one type of incoherent FFL [7].

An important debate in the study of biological systems from a network perspective is the biological relevance of statistical signals derived from graph representations (see also Ref. [8]). In addressing this question it is interesting to explore the consistency of large-scale biological data sets with graph abstractions of biological networks. This has been done in particular for the gene regulatory network and the metabolic network of yeast and *E. coli*. Luscombe *et al.* [9] showed that the topology of subnetwork structures in yeast is specific for cellular programs triggered by environmental conditions: Slow programs (e.g., cell cycles) employ a densely interconnected subnetwork structure, while programs required to act rapidly (e.g., DNA repair) employ networks with shorter path lengths and less complex motif content. The arrangement of genes on the genome and their correspondence to the gene regulatory network have been analyzed using methods from point process statistics [10]. The agreement of active metabolic networks (as predicted by flux-balance analysis) and gene expression data [12] has been studied using the method of control strengths derived from effective networks [11]. The interplay between feed-forward loops and larger-scale structures (subsets formed by all nodes topologically downstream of a reference node) in gene

regulatory networks has been explored [13] with the aim of better understanding the validity of the motif perspective. The rationale of this analysis has been to explore the interplay of two scales within the transcriptional regulatory network of *E. coli*. In particular, in Ref. [13] it was shown that when one scale dominates (high subnet usage) few regulatory devices on the smaller scale are found (low feed-forward loop occurrence).

A strong step toward an automatized statistical view on network motifs has been the work in Ref. [14], where the over- and under-representation of three-node subgraphs [the motif signature or triad significance profile (TSP)] compared to randomized networks is analyzed. This analysis of the TSP has been applied to a wide range of complex networks [15–19] and has become synonymous with a motif analysis. Many of the TSPs of real networks either show no significant over- or under-representation of three-node subgraphs or follow one of the four patterns (or superfamilies) discussed in Ref. [14].

A very promising development over the past few years has been that some features of such motif signatures are found to be related to the robustness of the system (see, e.g., Refs. [16,20]). Avetisov *et al.* [21] analyzed the motif signatures of graphs obtained from a block-hierarchical adjacency matrix. By introducing randomness (i.e., random flips $1 \leftrightarrow 0$) in the adjacency matrix, the authors were also able to study the robustness of the motif pattern. They found that the motif signature persists under small amounts of such topological noise. The work of Ref. [21] is one of the few examples (together with the comment on spatial networks in Ref. [22]) of motif signatures arising from global organizational features (in this case, the block-hierarchical structure of the adjacency matrix) of the network. Remarkably, the motif signature is quite similar to one of the superfamilies from Ref. [14].

It is therefore of great interest to better understand the crosstalk between local and global network properties as well as the interdependences between the different few-node subgraphs. In order to investigate this crosstalk, it is helpful to distinguish between two different kinds of global network properties: (i) a property visible only when one looks at the whole network (e.g., modularity) and (ii) a property present at every place of the network (e.g., assortativity).

For the special case of Erdős-Rényi (ER) random graphs we formulate a simple statistical description of expectation values for subgraph frequencies. Similar approaches were formulated in Refs. [23,24]. Subgraph counts in the ER model are a well-investigated topic. In Ref. [25] an analytical framework for computing significant over- and under-representations of (in that case, noninduced) subgraphs in undirected graphs was proposed as an alternative to the usual switch randomization and thus not requiring the simulation of a pool of null model graphs (see also Ref. [26]).

Here our question is different. We want to understand the systematic crosstalk between subgraph statistics and more global network properties. To this end, we require a simple statistical description for the influence, e.g., of link density on subgraph fluctuations. The three-step approach (variations in link density, variations in template counts, and variations in subgraph counts) described in the Sec. III allows us to explore the crosstalk, e.g., between modularity and the triad significance profile. By grouping the possible three-node subgraphs into categories, we are able to understand differences between subgraph counts arising on purely combinatorial grounds. This formulation can be used to analyze potential artifacts in motif signatures arising, e.g., from fluctuations in the number of unidirectional and bidirectional edges.

Both the expectation values and the standard deviations of subgraph counts enter the computation of subgraph z scores, which are frequently employed to quantify the statistical over- and under-representation of subgraphs in real networks. We can thus employ our method to the computation of a motif signature (or TSP) in all cases, where fluctuations in the edge density induce a nonzero motif signature for an otherwise random graph. Modular graphs, as the most important case of this category, are discussed as an application.

II. SUBGRAPH STATISTICS

In this section we introduce a simple model for the emergence of templates and motifs in random networks. We discuss the expectation values of motif counts and the corresponding fluctuations. This yields insight into the correlations from single-node properties to motifs.

A. Subgraph categories

Throughout this paper we will discuss only simple directed graphs with a node number N and an edge number M. Simple means that parallel edges pointing in the same direction and self-links are forbidden. Because of these conditions the graph is complete when it contains $M = N(N-1)$ edges. When two nodes a and b are connected by a single edge they are connected by a unidirectional edge u;when the opposing edge is also present they are connected by a bidirectional edge b. Finally, two nodes can be unconnected. Formally, this can be described by a nonedge n (see Fig. 1).

Between global network properties on the one hand and single-node properties on the other, motifs can be used to understand networks on a mesoscopic scale. For directed networks, most studies use three-node subgraphs and we will do the same here, although the formalism can easily be extended to higher motif sizes. Throughout this paper we discuss induced subgraphs, as they constitute the objects of interest in the vast literature on network motifs (see, e.g., Ref. [5]). A graph H is called an induced subgraph of G if it has exactly the edges that appear in G over the same vertex set. In other words, for any pair of vertices v and w of H, (v,w) is an edge of H if and only if it is an edge of G. Many analytical descriptions have focused on noninduced subgraphs, where the subgraph H contains an arbitrary subset of edges of G, restricted to the vertices of H (e.g., Ref. [25]). Note that the interpretation of motifs as noninduced subgraphs leads to different subgraph counts.

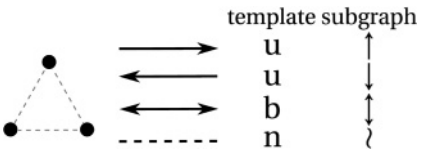

FIG. 1. The three different placements for the three possible types of edges. Shown on the right are the different types of edges that can occupy the placements and their notation for templates and subgraphs.

TABLE I. The 7 templates together with the 13 subgraphs they can form; subgraph numbers in parentheses correspond to the standard ordering from Ref. [14].

Template	Subgraph (a)	Subgraph (b)	Subgraph (c)
uun	↑↓≀ (1)	↓↑≀ (2)	↓↓≀,↑↑≀ (3)
ubn	↓↕≀ (4)	↑↕≀ (5)	
bbn	↕↕≀ (6)		
uuu	↓↓↑ ,↑↑↓ (7)	↓↓↓ ,↑↑↑ (8)	
uub	↑↓↕ (9)	↓↑↕ (10)	↓↓↕,↑↑↕ (11)
ubb	↓↕↕,↑↕↕ (12)		
bbb	↕↕↕ (13)		

In this paper we distinguish between templates and subgraphs. Templates are sets of edges with an undefined position relative to each other. Some templates have two (*uu*, *ub*, and *bb*) and others three (*uuu*, *uub*, *ubb*, and *bbb*) edges. These 7 templates can form 13 different (induced) subgraphs, as illustrated in Table I. A subgraph is obtained by defining the relative orientations of the edges within a template. Orientations are defined using ↑ for clockwise and ↓ for counterclockwise directions of the edge. Using this shorthand notation, we can summarize the template-subgraph relationships in a tabular form (see Table II). As only relative positions and orientations matter, ↓↓↓ is indistinguishable from ↑↑↑; however, ↑↓≀ is a different subgraph from ↓↑≀.

TABLE II. Illustration of how templates are distributed among their constituting subgraphs on the example of the template *uun*.

Template *uun*		Subgraph		r_m
↓↑≀	⇒		⇒	$\frac{1}{4}$
↑↓≀	⇒		⇒	$\frac{1}{4}$
↓↓≀	⇒		⇒	$\frac{1}{2}$
↑↑≀	⇒		⇒	

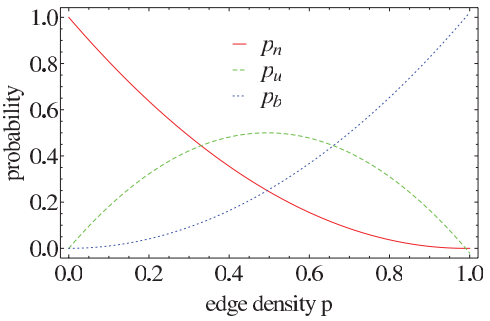

FIG. 2. (Color online) Probabilities p_n (red solid line), p_u (blue dotted line), and p_b (green dashed line) as functions of the edge density p.

B. Edge counts

Here we work with the ER model of random graphs, where a graph is characterized by the number of nodes N and the edge probability p. A graph represented by N and p contains on average $M = pN(N-1)$ edges. We specify a certain number of edges M and then use the corresponding edge density (or connectivity) $p = M/N(N-1)$ to characterize the network in our statistical assessment. For random networks we can estimate some basic probabilities and counts.

In a directed network model two edges that connect the same two nodes pointing in opposite directions form a bidirectional edge. The number of bidirectional edges can be estimated by the probability that a single position is selected twice p^2 times the number of possible slots $N(N-1)/2$:

$$M_{\text{bi}} = \frac{M^2}{2N(N-1)}.$$

The number of unidirectional edges is then given by

$$M_{\text{uni}} = M - \frac{M^2}{N(N-1)}.$$

C. Subgraph counts

In order to obtain expectation values for the subgraph counts c_m, we formulate a simple model of subgraphs where each of the three positions between the three nodes can be in one of three states: the unidirectional edge, the bidirectional edge, and no edge. The edge density of the graph is defined as $p = \frac{M}{N(N-1)}$ and the probabilities for the three states are then given by

$$p_u(p) = 2(p - p^2),$$
$$p_b(p) = p^2,$$
$$p_n(p) = 1 - p_u - p_b = (1-p)^2.$$

Figure 2 shows probabilities over p. By denoting the numbers of unidirectional edges u_m, bidirectional edges b_m, and nonedges n_m for every subgraph m, we can write the expected number c_m of type m as

$$c_m = p_l\, p_u(p)^{u_m}\, p_b(p)^{b_m}\, p_n(p)^{n_m}\, s_m,$$

where the number of possible placements for a subgraph is $p_l = \binom{N}{3}3!$ and s_m are symmetry factors

$$s_m = \xi_t / r_m,$$

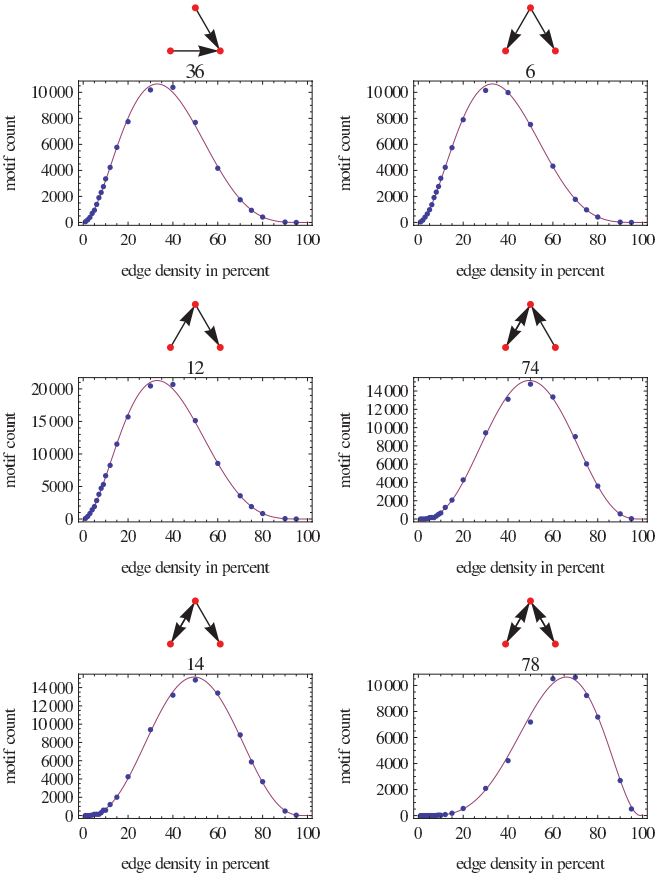

FIG. 3. (Color online) To illustrate the prediction quality of the subgraph counts we show the predicted (full curve) and numerically observed (dots) counts of the first six three-node subgraphs in a random network with $N = 100$. The connectivity is varied over the whole range ($p = 0{-}100\,\%$). The numerical points are obtained by averaging over 100 random graphs. The subgraphs, together with their IDs, are indicated above each plot. The figures for the full 13 motifs are available in Ref. [27].

where ξ_t accounts for the symmetries of the template t and r_m represents the ratio by which the template is split up into subgraphs. These predictions are compared to numerical experiments in Fig. 3.

The symmetry factor ξ_t is the ratio of possible three-symbol permutations of the distinct permutations obtained in a template. A template containing three distinct symbols exhausts the full possible six permutations, yielding $\xi_t = 1$, while a template with two distinct symbols allows for three permutations, yielding $\xi_t = \frac{6}{3} = 2$. If all three symbols in the template are equal, we have $\xi_t = 6$. As the symmetry factor r_m accounts only for the distribution of the templates on the subgraphs (see Table III), there must be $\sum \frac{1}{r_m} = 1$ for every template, where the sum is over all subgraphs m in the template t.

D. Edge fluctuations

Changing any one of p_u, p_b or p_n by a small probability Δ at fixed edge density p results in the change of the other two probabilities according to

$$\hat{p}_u(p) = p_u(p) + 2\Delta = 2(p - p^2) + 2\Delta,$$

TABLE III. Number of unidirectional edges u_m, bidirectional edges b_m, and nonedges n_m and the symmetry factors in all subgraphs.

m	1	2	3	4	5	6	7	8	9	10	11	12	13
t	1	1	1	2	2	3	4	4	5	5	5	6	7
u_m	2	2	2	1	1	0	3	3	2	2	2	1	0
b_m	0	0	0	1	1	2	0	0	1	1	1	2	3
n_m	1	1	1	1	1	1	0	0	0	0	0	0	0
r_m	4	4	2	2	2	1	$\frac{4}{3}$	4	4	4	2	1	1
ξ_t	2	2	2	1	1	2	6	6	2	2	2	2	6
s_m	8	8	4	2	2	2	8	24	8	8	4	2	6

$$\hat{p}_b(p) = p_b(p) - \Delta = p^2 - \Delta,$$
$$\hat{p}_n(p) = p_n(p) - \Delta = 1 - p_u - p_b - \Delta.$$

This is because the creation of a bidirectional edge needs two unidirectional edges and frees one place. To be able to infer the fluctuations of subgraphs it is useful to first derive equations for the fluctuation of bidirectional edges. The expectation value of the number of bidirectional edges is

$$M_{\mathrm{bi}} = \frac{p^2 N(N-1)}{2} = \frac{M^2}{2N(N-1)}.$$

In order to estimate the fluctuations in the number of bidirectional edges at low edge densities (and large numbers of nodes) we take the expectation value and variance $\lambda = nP$ for the Poissonian distribution, but replace the event number n with the number of possible sites for bidirectional edges $\binom{N}{2}$ and the event probability P with the probability of two edges p^2. For the standard deviation we thus obtain $\sigma_{bl} = \sqrt{M_{\mathrm{bi}}}$ at low densities.

When the number of edges approaches its maximum value $M = N(N-1)$ these fluctuations decrease again, which is due to the decreasing number of places that are not yet occupied by single edges that one would have to hit to not create another bidirectional edge. In this case, the event probability is replaced by $(1-p)^2$. It is also clear that σ_{bh} must be symmetric around $p = 0.5$. So at high densities we get

$$\sigma_{bh} = \sqrt{\frac{N^2}{2}} - \sqrt{M_{\mathrm{bi}}}.$$

As both fluctuations are mutually exclusive, their reciprocal sum yields an analytical expression for the total fluctuations of bidirectional edges as a function of the edge density p, i.e.,

$$\sigma_b = \frac{1}{\frac{1}{\sqrt{\sigma_{bl}}} + \frac{1}{\sqrt{\sigma_{bh}}}}.$$

This situation is summarized in Fig. 4. These fluctuations directly transfer to fluctuations of the unidirectional edges and nonedges. As every additional bidirectional edge means two uniedges less, there is a factor of 2 between their fluctuations (see Fig. 4).

E. Subgraph fluctuations

In order to reduce trivial contributions to the subgraph fluctuations, we keep the number M of edges in the ER graph fixed (which makes the edge density p a secondary quantity, as described above). Otherwise, the fluctuations in the number of edges at a given p would partially mask the conceptually more

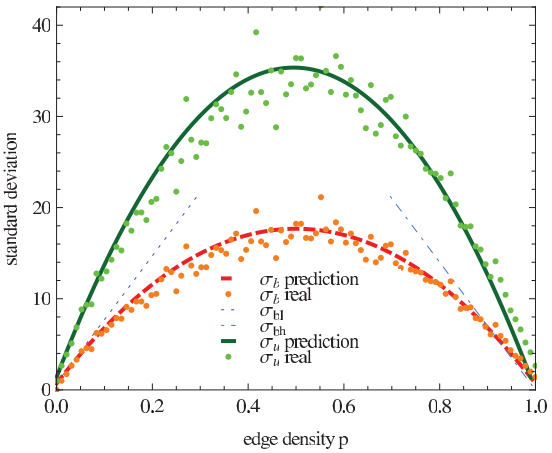

FIG. 4. (Color online) Fluctuations of the number of unidirectional and bidirectional edges in a random graph with $N = 100$ nodes and varying edge density. We show the numerical results together with the corresponding predictions.

important (and less trivial) contribution from fluctuations of unidirectional and bidirectional edges at fixed M.

There are two reasons for the fluctuation of a subgraph count: (i) fluctuations in the number of bidirectional edges and (ii) fluctuations due to the subdivision of templates (i.e., subgraphs with the same number of unidirectional, bidirectional, and nonedges) into subgraphs. This subdivision depends on the direction of the unidirectional edges, as discussed in Tables I and II.

Contribution (i). The fluctuation of the number of bidirectional edges can be translated into the fluctuation of a subgraph count by processing the normal number of subgraphs and subtracting that from the number of subgraphs one gets by changing the probabilities for unidirectional, bidirectional, and no edges by one standard deviation:

$$\sigma_{bm} = c_m(\hat{p}_u, \hat{p}_b, \hat{p}_n) - c_m(p_u, p_b, p_n).$$

Contribution (ii). The other sources of fluctuations are the fluctuations in the combination of unidirectional and bidirectional edges to templates and the distribution of the templates among the subgraphs. Together they can be estimated by the square root of the subgraph count:

$$\sigma_{mm} = \sqrt{c_m}.$$

These sources of fluctuations need to be combined in a Pythagorean sum:

$$\sigma_m = \sqrt{\sigma_{mm}^2 + \sigma_{bm}^2}.$$

The resulting fluctuations are shown in Fig. 5.

III. APPLICATION

Here we will show as a simple example how the theory presented above can be used to better understand properties of subgraph signatures. The subgraph signature of a network (or, more specifically, for three-node subgraphs, the TSP) is the pattern of over- and under-representations of few-node subgraphs in this network. It has become a standard method of analyzing complex networks. More formally, it is the

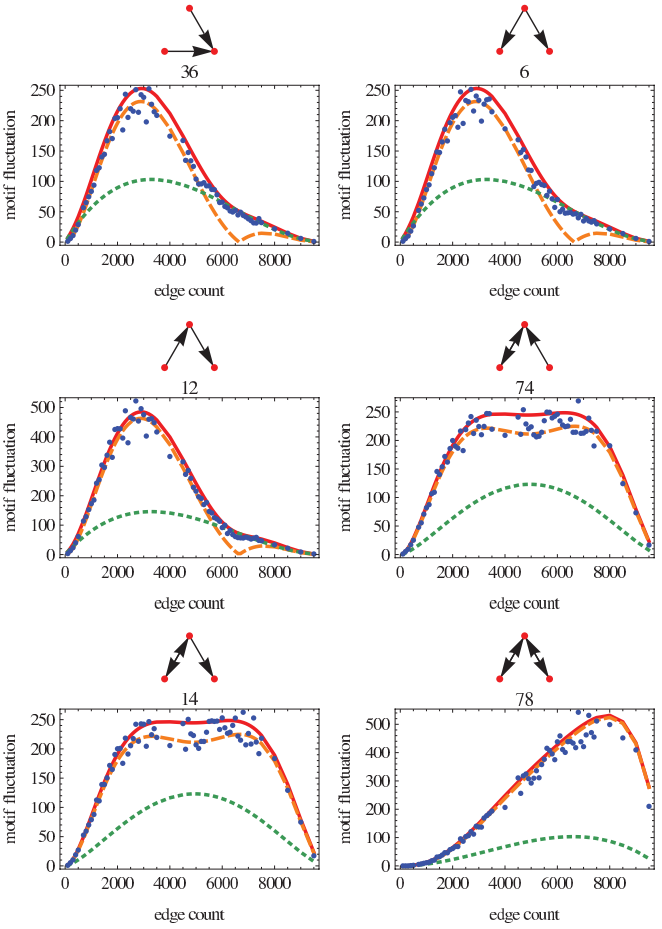

FIG. 5. (Color online) Fluctuations of the first six three-node subgraphs in a random network with $N = 100$. The connectivity is varied over the whole range ($p = 0-100\,\%$) and contribution (i) (orange dashed line), contribution (ii) (green dotted line), the total expected value (red solid line), and numerical results (blue dots) are shown. The subgraphs, together with their IDs, are indicated above each plot. The figures for the full 13 motifs are available in Ref. [27].

(normalized) z score of the subgraph counts. The z score is defined as

$$Z_m = \frac{c_m - \mu_m}{\sigma_m},$$

where for every subgraph c_m is the subgraph count in the original network, μ_m is the expectation value of c_m in an ensemble of randomized networks, and σ_m is the standard deviation of c_m in the randomized networks.

To obtain the ensemble of randomized networks a randomization scheme is repeatedly applied, where typically the in and out degree of each node (i.e., the degree sequence of the graph) is conserved during the randomization process, as well as the number of bidirectional edges at each node. The aim of the randomization procedure is to remove any nonrandom property (beyond the degree sequence). In this way deviations of the subgraph counts (in the real network) from randomness can be detected and functionally interpreted.

Apart from the case where some kind of selective process in the evolution of a network or some other functional requirement is enriching specific subgraphs, which is the most

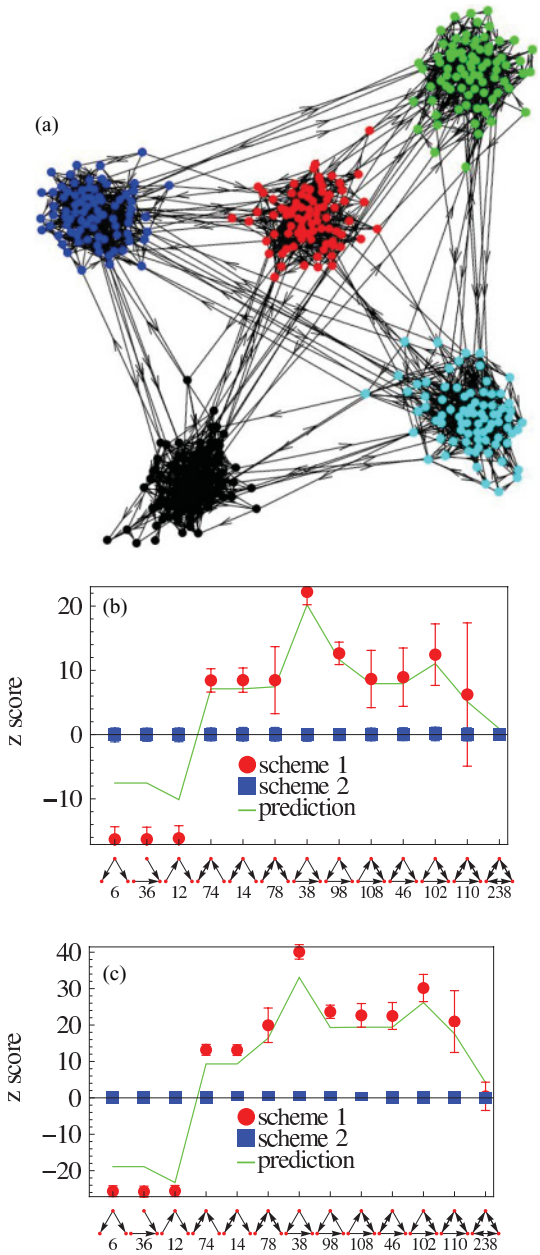

FIG. 6. (Color online) Motif z scores for a network that is composed of five strong modules. (a) Example of such a random modular graph. Two different randomization schemes are applied: (1) simple flipping of two-edge end points and (2) flipping while preserving the module structure. As the analyzed network is random apart from its modularity the z-score using the correct randomization scheme must be 0. This is shown for different densities: (b) $N = 500$, $M = 2000$, and $\rho = 0.08$ and (c) $N = 500$, $M = 8000$, and $\rho = 0.16$.

interesting case, there are many other reasons for a nonzero z score. Here we discuss modularity as one such possible reason. An example of a modular network is depicted in Fig. 6. If the modular structure is not taken into account during the randomization process and thereby conserved in the pool of randomized networks (i.e., eliminated from its effect on expected subgraph numbers), a false nonzero z score appears.

We use a random graph that is composed of five strong modules, where each module is an ER network. Additionally a certain amount of intermodule edges is introduced. As the base networks as well as the intermodule edges are constructed in a motif-blind way, correct randomization should yield a flat motif signature with z scores close to zero. The result of the application of standard, module-blind randomization techniques can be seen in Fig. 6. We also show the result of a module-aware randomization that mixes edges inside the modules and intermodule edges separately. In real world networks the modular structure of a network is generally not known and it is therefore necessary to detect the modules first, before adjusting the randomization scheme accordingly.

To better understand the error made by the standard randomization scheme we will analytically predict the error signature using the formalism introduced above. To this end, it is essential to notice that when the modules are destroyed the effective local density of the network is reduced by a factor of 5, the number of modules in this example. This is because a network with $N^* = 2N$ and $M^* = 2M$ has a density of

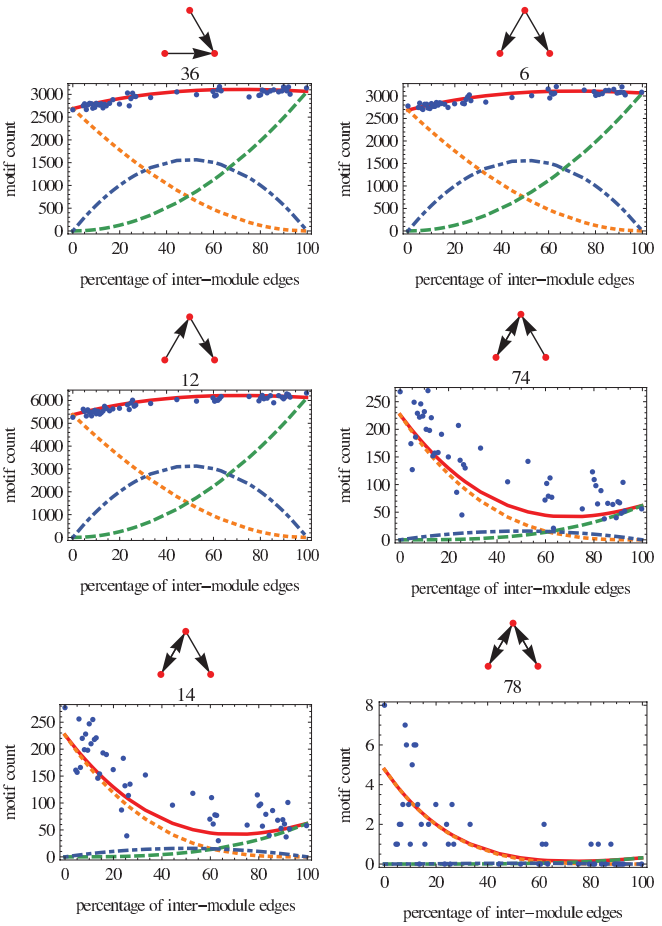

FIG. 7. (Color online) Composition of the first six three-node subgraphs in a modular network. The ratio of intermodular edges is varied over the whole range and the contribution of intramodule subgraphs (orange dotted line), intermodule subgraphs (green dashed line), mixed subgraphs (blue dot-dashed line), the total expected value (red solid line), and numerical results (blue dots) are shown ($N = 500$ and $M = 2000$). The subgraphs, together with their IDs, are indicated above each plot. The figures for the full 13 motifs are available in Ref. [27].

$d^* = \frac{M^*}{N^{*2}} = \frac{d}{2}$. A network with double the size has to have four times the number of edges to have the same density. Let N and M be the number of nodes and edges of the whole modular network. Then c_m can be estimated by $5c_m(N/5, M/5)$ and $\mu_m = c_m(N,M)$ and σ_m by $\sigma_m(N,M)$. The general form when a graph consists of k modules with node counts $n_1, n_2, \ldots, n_k$ is

$$c_m = \sum_{i=1}^{k} c_m\left(n_i, \frac{Mn_i}{N}\right).$$

The ratio of intermodule edges ρ can easily be taken into account by adding motifs from an additional global network, consisting of the intermodule edges:

$$c_m = \sum_{i=1}^{k} c_m\left(n_i, \frac{(1-\rho)Mn_i}{N}\right) + c_m(N, \rho M).$$

This simplification does not acknowledge subgraph instances that contain intramodule and intermodule edges. These are relevant mostly for the two-edge subgraphs. We evaluate the number of these mixed subgraphs d_m by taking into account the different edge densities in the module and between the modules. We therefore introduce probabilities for unidirectional and bidirectional edges in the components p_{uc} and p_{bc} and outside the components p_{uo} and p_{bo}. Using these probabilities we can write the expectation value for the additional subgraphs as

$$d_m = N^3 \begin{cases} p_{uc}p_{uo} + p_{uo}p_{uc} & \text{for } u_m = 2 \wedge b_m = 0 \\ p_{uc}p_{bo} + p_{uo}p_{bc} & \text{for } u_m = 1 \wedge b_m = 1 \\ p_{bc}p_{bo} + p_{bo}p_{bc} & \text{for } u_m = 0 \wedge b_m = 2 \\ 0 & \text{otherwise.} \end{cases}$$

These additional subgraphs are added to the intermodule and intramodule subgraphs to obtain the total subgraph counts. Figure 7 shows the quality of the prediction of the subgraph counts. To verify the quality of the predictions of the z score of a modular network, we compute the geometric mean of the difference of the z score when applying the appropriate module-aware randomization scheme and the simple randomization scheme. This quantity can both be computed numerically and analytically, as in Fig. 8.

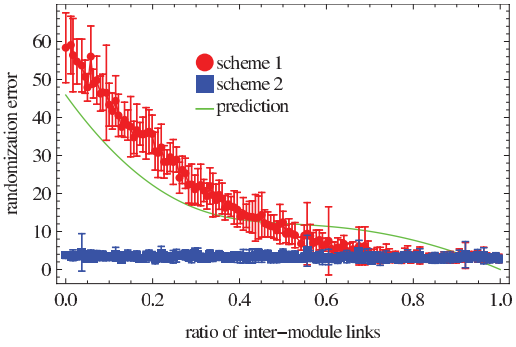

FIG. 8. (Color online) Sum over the squared errors that occurs from applying different randomization schemes over the ratio of intermodule edges ($N = 500$ and $M = 2000$). We show the numerical results for two different randomization schemes together with our prediction.

IV. CONCLUSION

Statistical properties of random graphs have been studied for decades in several disciplines and with a wide range of applications in mind. Here we have focused on a topic that in spite of its practical importance has received comparatively little attention so far, namely, the statistical fluctuations of few-node subgraphs induced by lower-level fluctuations in the numbers of unidirectional and bidirectional edges.

In the discussion of the practical relevance of our findings, one needs to distinguish three elements: (i) The crosstalk between a global graph property and the perceived motif signature, (ii) the qualitative parts of the prediction (in particular the grouping of subgraphs into seven templates, as introduced in Table I), and (iii) the quantitative prediction of subgraph fluctuations. The qualitative part of the prediction is valid in general: Fluctuations in subgraph statistics induced by other network properties (link density, fluctuations in the number of bidirectional links, degree correlations, etc.) will affect the templates and then split up into fluctuations of individual subgraph counts within those templates according to the combinatorial factors summarized in Table III. The quantitative part, where this framework is applied to density-induced subgraph fluctuations, is in the present form restricted to the case of ER graphs. An extension of the formalism to arbitrary degree distributions, which is possible by explicitly including the degree dependence in the probabilities p_i, $i = u,b,n$, introduced in Sec. II C and then averaging over all degrees,

$$p_i(p) \to \sum_k p_i(p,k)P(k),$$

where p is the link density and $P(k)$ is the degree distribution of the graph, is beyond the scope of this paper. A directed graph would require a double sum over the in degrees and the out degrees.

Here the quantitative prediction serves as a proof of principle that we have correctly identified all individual contributions to subgraph fluctuations. In this way we can quantitatively understand some of the crosstalk between global and local network properties. As an example of such a crosstalk we have here presented the motif signature arising from the modularity of the graph. Using our analytical description of subgraph fluctuations, we can precisely predict the artifactual motif signature of this otherwise random graph. By mixing intermodule edges and the different sets of intramodule edges independently, we can additionally show that the full motif signature is a sole consequence of the modular graph structure.

Beyond a better understanding of such artifacts, we believe that the classification of three-node subgraphs into the categories introduced in Sec. II has the potential to unravel the theoretical background behind the empirical observation that only four variants (or superfamilies) of three-node motif signatures are observed across a vast range of complex networks [14]. It is clear that all subgraphs within the same category will display synchronous fluctuations distributed among the participants of a category according to few well-understood combinatorial factors. This approach may constitute a solid

basis for understanding correlated subgraph fluctuations and motif-motif covariations.

Note added. Recently the work by Reichardt *et al*. [28] was pointed out to us, where a similar question is addressed from a different perspective, namely, the construction of a refined random graph model. Combining our concept of motif templates with their random graph model may provide additional insight into the superfamilies from Ref. [14].

ACKNOWLEDGMENT

This work was supported by Volkswagen Foundation Grants No. I/82717 and No. I/83435.

[1] J. Lee, K. Goh, B. Kahng, and D. Kim, Eur. Phys. J. B **49**, 231 (2006).
[2] J. Park and M. E. J. Newman, Phys. Rev. E **68**, 026112 (2003).
[3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
[4] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Nature Genet. **31**, 64 (2002).
[5] U. Alon, Nature Rev. Genet. **8**, 450 (2007).
[6] S. Mangan and U. Alon, Proc. Natl. Acad. Sci. USA **100**, 11980 (2003).
[7] S. Kaplan, A. Bren, E. Dekel, and U. Alon, Mol. Syst. Biol. **4**, 203 (2008).
[8] R. Montañez, M. A. Medina, R. V. Solé, and C. Rodríguez-Caso, BioEssays **32**, 246 (2010).
[9] N. M. Luscombe, M. Madan Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, Nature (London) **431**, 308 (2004).
[10] N. Sonnenschein, M.-T. Hütt, H. Stoyan, and D. Stoyan, BMC Syst. Biol. **3**, 119 (2009).
[11] C. Marr, M. Geertz, M. Hütt, and G. Muskhelishvili, BMC Syst. Biol. **2**, 18 (2008).
[12] N. Sonnenschein, M. Geertz, G. Muskhelishvili, and M.-T. Hütt, BMC Syst. Biol. **5**, 40 (2011).
[13] C. Marr, F. J. Theis, L. S. Liebovitch, and M.-T. Hütt, PLoS Comput. Biol. **6**, e1000836 (2010).
[14] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, Science **303**, 1538 (2004).
[15] P. Kaluza, M. Vingron, and A. S. Mikhailov, Chaos **18**, 026113 (2008).
[16] K. Klemm and S. Bornholdt, Proc. Natl. Acad. Sci. USA **102**, 18414 (2005).
[17] L. Krumov, C. Fretter, M. Müller-Hannemann, K. Weihe, and M. Hütt, Eur. Phys. J. B **84**, 535 (2011).
[18] B. Mirzasoleiman and M. Jalili, PLoS ONE **6**, e20512 (2011).
[19] O. Sporns and R. Kötter, PLoS Biol. **2**, e369 (2004).
[20] P. Kaluza, M. Ipsen, M. Vingron, and A. S. Mikhailov, Phys. Rev. E **75**, 015101 (2007).
[21] V. Avetisov, S. Nechaev, and A. Shkarin, Physica A **389**, 5895 (2010).
[22] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone, Science **305**, 1107c (2004).
[23] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon, Phys. Rev. E **68**, 026127 (2003).
[24] B. Karrer and M. E. J. Newman, Phys. Rev. E **82**, 066118 (2010).
[25] F. Picard, J. Daudin, M. Koskas, S. Schbath, and S. Robin, J. Comput. Biol. **15**, 1 (2008).
[26] C. Matias, S. Schbath, E. Birmelé, J. Daudin, and S. Robin, REVSTAT **4**, 31 (2006).
[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.85.056119 for figures for the full 13 motifs.
[28] J. Reichardt, R. Alamino, and D. Saad, PLoS ONE **6**, e21282 (2011).