



**HAL**  
open science

## Detecting local network motifs

Etienne Birmele

► **To cite this version:**

Etienne Birmele. Detecting local network motifs. *Electronic Journal of Statistics* , 2012, 6, pp.908-933.  
10.1214/12-EJS698 . hal-00726215

**HAL Id: hal-00726215**

**<https://hal.science/hal-00726215>**

Submitted on 29 Aug 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Detecting local network motifs

Etienne Birmelé\*

*Laboratoire Statistique et Génome, CNRS, Université d'Evry,  
F-91037 Evry, France*

*Laboratoire Biométrie et Biologie Evolutive, CNRS, Université Lyon 1,  
F-69100 Villeurbanne, France*

*Equipe BAMBOO, INRIA Grenoble Rhône-Alpes, 655 avenue de l'Europe,  
F-38330 Montbonnot Saint-Martin, France  
e-mail: [etienne.birmele@genopole.cnrs.fr](mailto:etienne.birmele@genopole.cnrs.fr)*

**Abstract:** Studying the structure of so-called *real networks*, that is networks obtained from sociological or biological data for instance, has become a major field of interest in the last decade. One way to deal with it is to consider that networks are partially built from small functional units called *motifs*, which can be found by looking for small subgraphs whose numbers of occurrences in the whole network are surprisingly high. In this article, we propose to define motifs through a local over-representation in the network and develop a statistic to detect them without relying on simulations. We then illustrate the performance of our procedure on simulated and real data, recovering already known biologically relevant motifs. Moreover, we explain how our method gives some information about the respective roles of the vertices in a motif.

**AMS 2000 subject classifications:** Primary 62P10; secondary 05C90.

**Keywords and phrases:** Network motif, Poisson approximation, biological network.

Received September 2011.

## 1. Introduction

One way to reach a better understanding of the structure of networks is to summarize part of the information in the counts of small connected subgraphs. That method is used for decades in social network studies, for example via the triad censuses (Wasserman and Faust, 1994; Watts and Strogatz, 1998). More recent work indicates that biological networks show recurrent small patterns, called *network motifs* and introduced by Milo et al. (2002). They can be thought of as putative small units of given function from which the networks are built. For instance, Alon (2007) describes the regulation role in transcriptional networks of a pattern of three vertices called the feed-forward loop. It is therefore quite natural to ask which are the patterns that are over-represented in a given network.

Looking for over-representation requires a null model to compare the observed network with. The most popular model corresponds to the uniform distribution

---

\*This work has been supported by the CNRS, by the French Agence Nationale de la Recherche under grant NeMo ANR-08-BLAN-0304-01 and by the ERC Advanced Grant SISYPHE.

among all graphs having the same degree distribution as the network of interest. The stub-rewiring algorithm introduced by Milo et al. (2002) is used to sample under this model in several methods for motif detection, including those of Kashani et al. (2009); Kashtan et al. (2004); Wernicke and Rasche (2006), the method based on graph alignments of Berg and Lässig (2004) and the method devoted to labeled graphs of Banks et al. (2008). The earliest implementation of this method by Kashtan et al. (2004) was incorrect as the sampling was non-uniform (Wernicke and Rasche, 2006) but it was subsequently corrected. However, Artzy-Randrup et al. (2004) point out that it does not take into account the preferential links between some vertices and the high local density, which are two major features of biological networks. They show on a toy-example that the use of the stub-rewiring model and of a Z-score to detect motifs selects over-represented patterns in randomly chosen networks when the null model generates spatially clustered networks.

Another way to define the null model is to consider a random graph model defined by a probability distribution. Literature about random graph models and their suitability to real networks is abundant (see e.g. Chung and Lu, 2006). Exponential family models have been developed to take the counting of small subgraphs into account (Holland and Leinhardt, 1970; Hunter and Handcock, 2006), but simulations are needed to compute normalization constants and the law of the motif counts seems to be analytically intractable. Another widely studied family of models are block models, which allow different link probabilities between classes of vertices and thus model the heterogeneity of connection patterns. Moreover, Picard et al. (2008) show that mean and variance calculation for the pattern counts are tractable under such models.

Whatever the choice of the null model, the exact distributions of subgraph counts are unknown. Most of the detection algorithms assume a normal distribution and use the Z-score to compute  $p$ -values. However, Janson (1990) show that even asymptotically this assumption can be wrong. In the case of mixture models, Picard et al. (2008) point out that a Polya-Aeppli approximation for the subgraph counts is better than the normal one, as this distribution is more heavy-tailed than Gaussian. Finally, another approach is to consider the real distribution but to determine only an upper-bound of the  $p$ -value, using for instance concentration of measure inequalities (Boucheron, Lugosi and Massart, 2003; Janson, Oleskiewicz and Rucinski, 2004).

It is important that motifs are defined conditionally on subpattern occurrences, as pointed out by Milo et al.. Indeed, a pattern may appear as over-represented because it contains an over-represented subpattern, which is in fact the biologically relevant structure. The observed network should for instance be compared to random graphs with similar density to decide if triangles are over-represented, and to graphs with similar density and number of triangles to decide if complete graphs of size four are over-represented. This conditioning issue is also taken into account by Banks et al. in the context of labeled graph. Nevertheless, in both cases, the real network is compared to graphs generated by the stub-rewiring procedure. Therefore, to study the patterns of size  $k$ , it is necessary to generate a large number of graphs with the same number of each

type of subgraphs of size ranging between 2 and  $k-1$  as in the observed network. In practice, only the cases  $k = 3$  and  $k = 4$  are implemented to our knowledge (Milo et al., 2002).

Finally, Dobrin et al. (2004) show that the motifs found in the Yeast transcriptional regulatory network aggregate, that is they highly concentrate in some regions of the network, indicating that biologically meaningful mechanisms may not be spread uniformly in the network. Therefore, it seems natural to look for *local* over-representation of patterns.

That phenomenon is also highlighted in the more biologically driven work by Zhang et al. (2005), who suggest to look for motif themes rather than motifs. They define themes as *recurring higher-order interconnection patterns that encompass multiple occurrences of network motifs* and show their biological relevance in the different networks associated to Yeast. In other words, themes are patterns corresponding to several occurrences of a motif that share some of their nodes.

The major contribution of this paper is to propose a definition of a local motif based on the themes of Zhang et al., and a procedure to detect the local motifs of fixed size  $k$  in a network. This procedure builds upon earlier works in motif detection, being to our knowledge the first approach taking into account the conditioning on a subpattern without theoretical size limitation on the considered pattern, as well as the local character of patterns. Moreover, it makes no assumption on the law of the pattern counts. It is composed of four main steps:

- Inference of the parameters of the null model. We consider a model in which each node belongs to a fixed class and each edge is drawn independently from the others under a Bernoulli law whose parameter depends only on the classes of its endvertices.
- Enumeration and localization of all patterns of size  $k$  present in the studied network and of their subpatterns.
- Assignment of a  $p$ -value to each pair (pattern; subpattern) present in the network for testing local over-representation. The key idea of that assignment is to show that the distribution of the number of local occurrences of a pattern is close to a Poisson distribution, allowing us to bound the exact  $p$ -value.
- A filtering procedure which ensures that every emergent local motif conveys some novel information about the network structure when compared to its subpatterns.

The three last steps of this procedure are implemented in the R-package *paloma*, available at the CRAN repository (<http://cran.r-project.org/>).

We define precisely the notion of local over-representation and detail the four steps of our procedure in Section 2. As the obtained  $p$ -value is in fact an upper bound of the exact one, we investigate the tightness of that bound in Section 3. We then show results on both simulated and real data in Section 4.

## 2. Methods

### 2.1. Local network motifs

We consider a network  $G$  of interest on  $n$  vertices. In this work, we consider directed graphs, with possible loops and opposite edges, but all the results can easily be extended to undirected graphs.

A *pattern*  $\mathbf{m}$  of size  $k$  is a directed connected graph on  $k$  vertices, from which we want to know if it is locally over-represented in  $G$ . A *subpattern*  $\mathbf{m}'$  of  $\mathbf{m}$  denotes a pattern on  $k - 1$  vertices obtained by deleting any vertex of  $\mathbf{m}$ . Intuitively, the pattern  $\mathbf{m}$  will be declared as a local motif with respect to  $\mathbf{m}'$  if there exist a surprisingly high number of occurrences of  $\mathbf{m}$  in the network sharing  $k - 1$  vertices, those  $k - 1$  vertices yielding an occurrence of  $\mathbf{m}'$ . Note that this definition is really different from the usual definition of network motifs, defined as patterns which are over-represented in terms of overall count in the network, and which we call global motifs. Indeed, a pattern may be a global motif without being a local one, and vice an versa, as shown in Figure 1.

The intuitive definition of a local motif is however not mathematically precise enough. Indeed, the deletion of two vertices  $a$  and  $b$  in  $\mathbf{m}$  may lead to the same subpattern. Either  $a$  and  $b$  play the same topological role in  $\mathbf{m}$ , as in the bi-fan pattern shown in Figure 2, in which case the deletion of  $a$  and  $b$  should not be treated separately. Or  $a$  and  $b$  may play different topological roles in  $\mathbf{m}$ , as in the feed-forward loop pattern shown in Figure 2, and both deletions should be treated apart one from another even if the subpattern is the same. Therefore, a more formal definition of local motifs has to be given, based on the automorphism classes of  $\mathbf{m}$ .

An automorphism of  $\mathbf{m}$  is a permutation  $\phi$  of its vertices such that, for every pair  $(a, b)$  of vertices,  $\overrightarrow{\phi(a)\phi(b)}$  is an edge if and only if  $\overrightarrow{ab}$  is an edge. Let  $\mathcal{R}$  be the relation defined by  $a\mathcal{R}b$  if  $b$  is the image of  $a$  by an automorphism.  $\mathcal{R}$

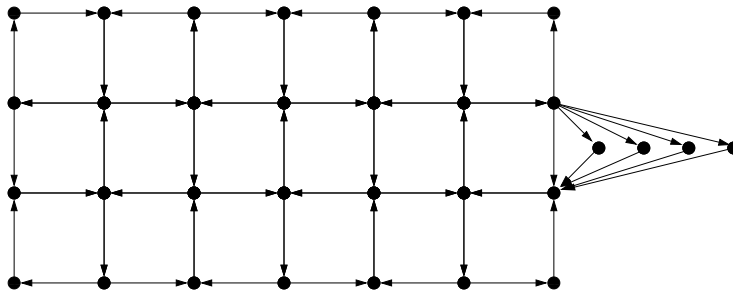


FIG 1. This graph contains 18 occurrences of the directed cycle of length of four, none of them sharing three vertices. On the other hand, it contains only four directed cycles of length three but all of them share an edge. Given the low density of the graph, the directed cycle of length three is a local motif without being a global one, while the one of length four is a global motif without being a local one.

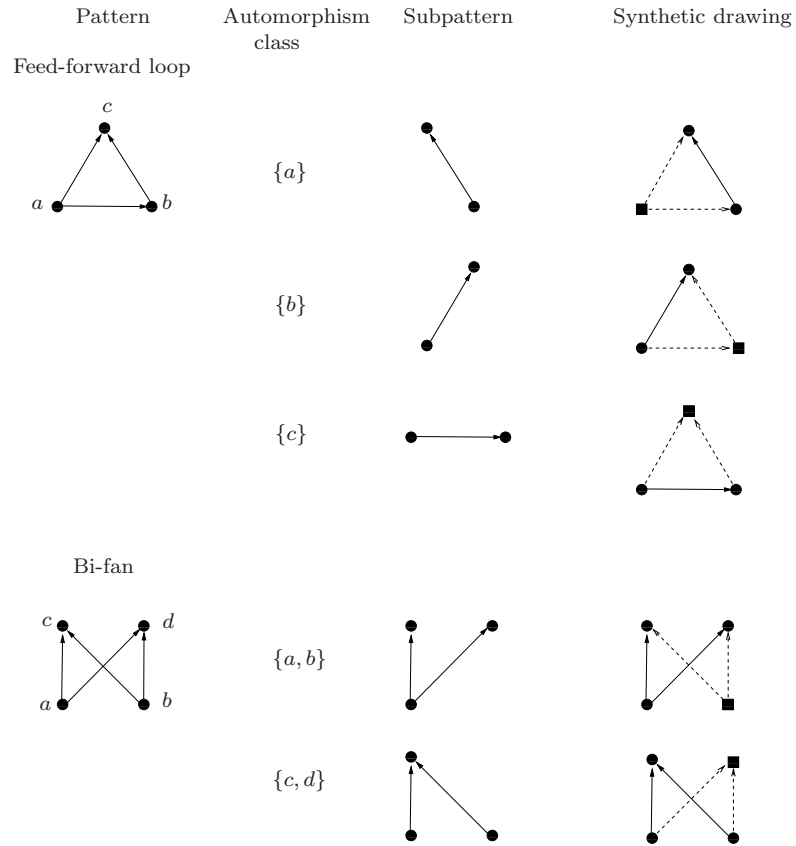


FIG 2. The feed-forward loop and bi-fan patterns with the list of their automorphism classes and subpatterns. The last column shows how we represent a pattern and one of its subpatterns in a single drawing.

is an equivalence relation on the vertices of  $\mathbf{m}$  and those vertices can therefore be partitioned into equivalence classes, which we call *automorphism classes*. For example, in the bi-fan pattern shown in Figure 2, the permutation exchanging  $a$  with  $b$  and  $c$  with  $d$  is an automorphism. However,  $a$  and  $c$  are not equivalent as they have different outdegrees. Thus the bi-fan has two automorphism classes which are  $\{a, b\}$  and  $\{c, d\}$ .

Note that the subpattern obtained by deleting any vertex of an automorphism class is the same. Moreover, vertices having different topological roles will be in different automorphism classes, even if they have a common resulting subpattern. In the following, we adopt the graphical convention shown in the last column of Figure 2 to draw at the same time a pattern  $\mathbf{m}$ , the considered automorphism class  $C$  and the resulting subpattern  $\mathbf{m}'$ . The whole graph represents  $\mathbf{m}$  and the squared vertex whose adjacent edges are dotted is a vertex of  $C$ . The subpattern  $\mathbf{m}'$  is thus obtained by deleting that vertex.

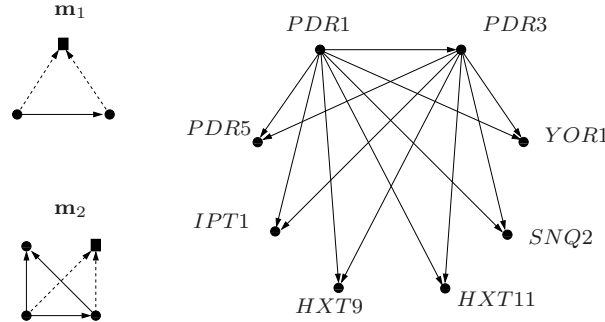


FIG 3. The shown subnetwork of the Yeast regulation network is a  $(\mathbf{m}_1, C_1)$ -theme of order 6 at position  $(\{PDR1, PDR3\})$  and a  $(\mathbf{m}_2, C_2)$ -theme of order 5 at position  $(\{PDR1, PDR3\}, \{PDR5\})$ .  $C_1$  and  $C_2$  are the respective automorphism classes of the squared vertices.

Let  $(C_1, \dots, C_K)$  be the automorphism classes of  $\mathbf{m}$  and  $(i_1, \dots, i_K)$  their respective sizes. A position  $U$  in a network  $G$  for  $\mathbf{m}$  will then denote a list  $(V_1, \dots, V_K)$  of disjoint sets of vertices of  $G$  with respective sizes  $(i_1, \dots, i_K)$ . That position is an occurrence of  $\mathbf{m}$  in  $G$  if the subgraph of  $G$  induced by the vertices of  $U$  is isomorphic to  $\mathbf{m}$ . Writing a position as a list of sets of vertices ensures to count every occurrence of a pattern only once. However, for clarity, we will write positions as lists of vertices throughout the article.

Let  $\mathbf{m}$  be a pattern,  $C$  an automorphism class and  $\mathbf{m}'$  the corresponding subpattern. An occurrence  $U$  of  $\mathbf{m}$  in  $G$  is an extension of an occurrence  $U'$  of  $\mathbf{m}'$  if the vertex set of  $U'$  is a subset of the vertex set of  $U$ . We define the  $(\mathbf{m}, C)$ -theme on  $U'$  as the subgraph of the network induced by the occurrence of  $\mathbf{m}'$  at  $U'$  and all its extensions. The number of those extensions will be the order of the theme (see Figure 3 for an illustration).

We define a potential local motif as a pattern which is locally over-represented with respect to at least one of its automorphism classes. In other words,  $\mathbf{m}$  is a potential local motif with respect to  $C$  if there exist an  $(\mathbf{m}, C)$ -theme whose order is significantly higher than the expected order in a random model to be specified in Section 2.2.

Finally, a potential local motif is a local motif if the information it conveys is not redundant with a smaller local motif, that is if it is not filtered out by the procedure to be described in subsection 2.5.

### 2.2. The random graph model

The detection of local motifs being built on tests about the local structure of networks, it is important to choose a null model showing as much as possible the same local characteristics as real networks. The random generation model we consider is based on blockmodels (White, Boorman and Breiger, 1976), that is that vertices are spread into classes and linked randomly depending on their classes. Such model have been shown to describe well the modular structure

of real networks (Nowicki and Snijders, 2001; Picard et al., 2009) and have two main advantages. First, they take into account the different link densities between different groups of nodes in the network; second, given the vertex classes, all the edges are independent and thus calculations remain tractable.

Those models can be split into two subfamilies depending whether the vertex classes are random or fixed. The models of the former case are mixtures of random graphs and are called *Stochastic Block Models*. Several statistical tools exist to infer their parameters (Daudin, Picard and Robin, 2008; Hofman and Wiggins, 2008; Latouche, Birmelé and Ambroise, 2008; Nowicki and Snijders, 2001). However, the stationarity of those models in terms of degree distribution is in our context a drawback as a high order theme involving vertices of small degree should be more significant than a high order theme involving hubs. Moreover, randomness on the classes would induce correlations between the edges and therefore invalidate the Poisson approximation of Section 2.3.

Therefore, we consider a blockmodel with fixed and known classes on the vertices. It depends on a 4-tuple of parameters  $(n, Q, \mathbf{Z}, \mathbf{\Pi})$ , where:

- $n$  is the number of vertices,
- $Q$  is the number of classes of the model,
- $\mathbf{Z} \in \{1, \dots, Q\}^n$  is a vector giving the class of each vertex,
- $\mathbf{\Pi}$  is a  $Q \times Q$  connectivity matrix. The coefficient  $\Pi_{ql} \in [0, 1]$  of that matrix indicates with which probability a vertex of class  $q$  and a vertex of class  $l$  are linked by an edge.

Under this model, all the edges of the random graph are drawn independently under Bernoulli laws: denoting by  $X_{uv}$  the indicator variable of the edge between vertices  $u$  and  $v$ ,

$$X_{uv} \sim \mathcal{B}(\Pi_{Z(u), Z(v)}).$$

Note that the classical random graph model introduced by Erdős and Rényi (1959) is a model of this family corresponding to the case  $Q = 1$ . This random graph framework also contains the model where  $u$  and  $v$  are linked with probability proportional to the product of their respective observed degrees. Under that model, which we will call *Expected Degree*, the expected degree of each vertex is almost equal to its observed degree (Matias et al., 2006). However, vertices are in the same class if and only if they have the same in- and out-degrees. Therefore, the number of classes may be large on real networks, having a deep impact on the running time of our motif detection procedure.

In applications, the vertex classes and the connectivity matrix  $\mathbf{\Pi}$  are inferred using estimation algorithms for Stochastic Block Models. All such algorithms infer for each vertex a vector of length  $Q$  giving the probability for that vertex to belong to each class, as well as a connectivity matrix. We choose to assign each vertex to the class with highest a posteriori probability, and thus deal with the observed network as if the groups were known and fixed.

Overlapping classes can also be taken into account by using for instance the *Mixed Membership Stochastic Blockmodel* from Airoldi et al. (2008) or the model *Overlapping Stochastic Blockmodel* from Latouche, Birmelé and Ambroise (2011).



**2.3. Local over-representation**

Consider a pattern  $\mathbf{m}$  of size  $k$ , a subpattern  $(C, \mathbf{m}')$  of  $\mathbf{m}$  and a set  $U$  of  $k - 1$  vertices in some graph  $G$ . If  $U$  corresponds to an occurrence of  $\mathbf{m}'$ , we write  $G[U] \sim \mathbf{m}'$ . We use the notation  $U$  rather than  $U'$  in order to simplify the oncoming notations, but this vertex set corresponds to a putative position of the subpattern.

Let  $N_U(\mathbf{m}, \mathbf{m}')$  denote the order of the  $(\mathbf{m}, C)$ -theme located at  $U$ . If  $U$  does not correspond to an occurrence of  $\mathbf{m}'$ , we set  $N_U(\mathbf{m}, \mathbf{m}') = 0$ .

We define  $\lambda_U(\mathbf{m}, \mathbf{m}') = \mathbb{E}(N_U(\mathbf{m}, \mathbf{m}') | G[U] \sim \mathbf{m}')$  and  $\Delta_U(\mathbf{m}, \mathbf{m}')$  as the quantity

$$\Delta_U(\mathbf{m}, \mathbf{m}') = \frac{N_U(\mathbf{m}, \mathbf{m}') - \lambda_U(\mathbf{m}, \mathbf{m}')}{\lambda_U(\mathbf{m}, \mathbf{m}')}.$$

Looking for themes whose order is much larger than expected under our model is then equivalent to look for significantly high values of  $\Delta_U(\mathbf{m}, \mathbf{m}')$ .

In the following, we will omit the reference to  $\mathbf{m}$  and  $\mathbf{m}'$  when there is no ambiguity. Note that the definition of  $\Delta_U$  requires the condition  $\lambda_U \neq 0$ . However, the parameter inference procedures we use (see Section 4.2) ensure that  $\lambda_U$  is non-zero whenever  $N_U$  is non-zero.

Let us consider a set  $U$  corresponding to an occurrence of  $\mathbf{m}'$ . For each vertex  $v \notin U$ , we denote by  $I_U^v$  the indicator random variable which is equal to 1 if adding  $v$  to  $U$  yields an occurrence of  $\mathbf{m}$  in  $G$ . Let  $p_U^v$  be the mean value of  $I_U^v$ . Then  $\lambda_U = \sum_{v \notin U} p_U^v$  and those quantities can easily be deduced from the parameters of the random graph model.

As the indicator random variables  $(I_U^v)_{v \notin U}$  are independent, it is well known that the law of their sum, that is the law of  $N_U$ , can be approximated by a Poisson law (Barbour, Holst and Janson, 1992), by using Chen-Stein's method (Chen, 1975). More precisely, denoting by  $d_{TV}$  the total variation distance between two distributions, we have

$$d_{TV}(\mathcal{L}(N_U), \mathcal{L}(Po(\lambda_U))) \leq \min(1, \lambda_U^{-1}) \sum_{v \notin U} (p_U^v)^2,$$

where  $\mathcal{L}(Po(\lambda_U))$  is the Poisson distribution with parameter  $\lambda_U$ .

This approximation may be used to determine an upper bound for the  $p$ -value of testing if the  $(\mathbf{m}, C)$ -theme order is surprisingly large. In practice, such bounds are quite accurate as the  $p_U^v$ 's are small.

Nevertheless, a better approximation can be obtained by applying again Chen Stein's method for the tail probabilities, as shown in Barbour, Holst and Janson (in the proof of Theorem 2.R, p44).

$$\forall K > 2\lambda_U, \quad \mathbb{P}(N_U \geq K | G[U] \sim \mathbf{m}') \leq \frac{K - \lambda_U}{K - 2\lambda_U} Po(\lambda_U)([K, +\infty)), \quad (2.1)$$

where, for any measurable set  $A$ ,  $Po(\lambda_U)(A)$  is the probability of  $A$  under the Poisson distribution with parameter  $\lambda_U$ .

Setting  $K = \lceil \lambda_U(1+t) \rceil$  for some  $t > 1$  and using elementary bounds and transformations developed in Appendix A, we obtain

$$\forall t > 1, \mathbb{P}(\Delta_U \geq t) \leq \mathbb{P}(G[U] \sim \mathbf{m}') \frac{\sqrt{t+1}}{\sqrt{2\pi\lambda_U}(t-1)} e^{-\lambda_U((1+t)\ln(1+t)-t)} \quad (2.2)$$

For positive values of  $t$  which are smaller than or close to 1, a sharper bound can be obtained by using a concentration inequality on the sum of independent random variables bounded between 0 and 1 (McDiarmid, 1998).

$$\forall t > 0, \quad \mathbb{P}(\Delta_U \geq t) \leq \mathbb{P}(G[U] \sim \mathbf{m}') e^{-\lambda_U((1+t)\ln(1+t)-t)}. \quad (2.3)$$

Moreover, it is straightforward to verify that the function of  $t$  defined on  $]1, +\infty[$  by  $\frac{\sqrt{t+1}}{\sqrt{2\pi\lambda}(t-1)}$  is decreasing and is equal to 1 at

$$t_\lambda = 1 + \frac{1}{4\pi\lambda}(1 + \sqrt{1 + 16\pi\lambda}).$$

Therefore, coupling inequalities (2.2) and (2.3) yields a local bound for the tail probability of the theme order.

**Theorem 2.1.** *Let us define, for any positive  $t$  and  $\lambda$ ,*

$$h(\lambda, t) = \begin{cases} 1 & \text{if } t \leq t_\lambda, \\ \frac{\sqrt{t+1}}{\sqrt{2\pi\lambda}(t-1)} & \text{if } t > t_\lambda. \end{cases}$$

*Then, for any pattern  $\mathbf{m}$ , subpattern  $(C, \mathbf{m}')$ , position  $U$  and positive  $t$ ,*

$$\mathbb{P}(\Delta_U \geq t) \leq \mathbb{P}(G[U] \sim \mathbf{m}') h(\lambda_U, t) e^{-\lambda_U((1+t)\ln(1+t)-t)}.$$

We thus have an exponentially decreasing local bound for the tail probability of the centered and renormalized order of an  $(\mathbf{m}, C)$ -theme at position  $U$ . Moreover, that bound is easily computable from the parameters of the random graph model.

**2.4. A global statistic to detect local motifs**

Theorem 2.1 allows to test whether there is a local over-representation at a given position  $U$  of an  $(m, C)$ -theme. However, the number of possible positions  $U$  is growing as  $n^{k-1}$ . We thus encounter a multiple testing problem. To overcome this issue, we build a statistic characterizing any local over-representation of a pattern somewhere in the graph.

Let us consider the function  $g$  defined for every positive  $\lambda$  and  $t$  by

$$g(\lambda, t) = \lambda((1+t)\log(1+t) - t) - \log(h(\lambda, t)). \quad (2.4)$$

For any positive  $\lambda$ , the function  $g(\lambda, \cdot)$  is a one-to-one increasing function, mapping  $]0, +\infty[$  to itself and which is equivalent to  $\lambda t \log(t)$  as  $t$  tends to

infinity. Thus, the event  $g(\lambda_U, \Delta_U)$  much larger than 1 is equivalent to the event  $\Delta_U$  much larger than 1.

For any positive  $t$ , let us apply Theorem 2.1 to  $y$  such that  $g(\lambda_U, y) = t$ . We then obtain

$$\forall t > 0, \quad \mathbb{P}(g(\lambda_U, \Delta_U) \geq t) \leq \mathbb{P}(G[U] \sim \mathbf{m}')e^{-t}.$$

Noting that the event  $E^t = \{\max_U(g(\lambda_U, \Delta_U)) \geq t\}$  is the union over all the possible positions  $U$  of the events  $E_U^t = \{g(\lambda_U, \Delta_U) \geq t\}$ , and that the exponential term in the upper bound is independent of  $U$ , we obtain our main result, stated in the following theorem.

**Theorem 2.2.** *Let  $g$  be the function defined in Equation (2.4) and  $N(\mathbf{m}')$  the random variable denoting the global number of occurrences of  $\mathbf{m}'$  in  $G$ . Then, for every  $t > 0$ ,*

$$\mathbb{P}\left(\max_U(g(\lambda_U, \Delta_U)) \geq t\right) \leq \mathbb{E}N(\mathbf{m}')e^{-t} \tag{2.5}$$

We thus obtain an upper bound on the global  $p$ -value for detecting a local over-representation of  $\mathbf{m}$  with respect to the subpattern  $(C, \mathbf{m}')$  occurring anywhere in the network.

### 2.5. Motif selection criterion

Consider the two patterns and respective subpatterns of Figure 3. Let  $C_1$  and  $C_2$  denote the respective deletion classes of the subpatterns. Then, as shown by the figure, every  $(\mathbf{m}_2, C_2)$ -theme of order  $K$  is an  $(\mathbf{m}_1, C_1)$ -theme of order  $K + 1$ . In that case, the fact that the  $(\mathbf{m}_2, C_2)$ -theme is of order significantly larger than expected is redundant with the same information for the  $(\mathbf{m}_1, C_1)$ -theme.

To avoid such redundancy in the final motif list, a pattern  $\mathbf{m}$  will be considered as a motif with respect to a subpattern  $(C, \mathbf{m}')$  if the two following conditions hold:

1. The  $p$ -value given by Theorem 2.2 is lower than a fixed threshold, that is  $\mathbf{m}$  is a potential local motif;
2. Let  $\{a\}$  be a vertex of  $C$ . There exist no set  $\mathcal{A}$  of vertices of  $\mathbf{m}$  such that
  - there is no edge between  $a$  and any vertex of  $\mathcal{A}$ ,
  - $\mathbf{m} \setminus \mathcal{A}$  is over-represented with respect to  $(D, \mathbf{m} \setminus (\mathcal{A} \cup \{a\}))$ , where  $D$  is the deletion class of  $\{a\}$  in  $\mathbf{m} \setminus \mathcal{A}$ .

If there exists a set  $\mathcal{A}$  satisfying the two points of the second condition, then the over-representation of  $\mathbf{m}$  with respect to  $(C, \mathbf{m} \setminus \{a\})$  is considered as redundant with the over-representation of  $\mathbf{m} \setminus \mathcal{A}$  with respect to  $(D, \mathbf{m} \setminus (\mathcal{A} \cup \{a\}))$ . Thus, the pair  $(\mathbf{m}, C)$  is filtered out from the local motif list.

Figure 4 illustrates the filtering procedure. Moreover, it shows why the absence of any edge between vertex  $a$  and set  $\mathcal{A}$  is required to filter out a potential local motif. Indeed, consider the first pattern of the figure for which  $\mathcal{A} = \{b\}$  fulfills the previous condition. The corresponding theme doesn't convey any

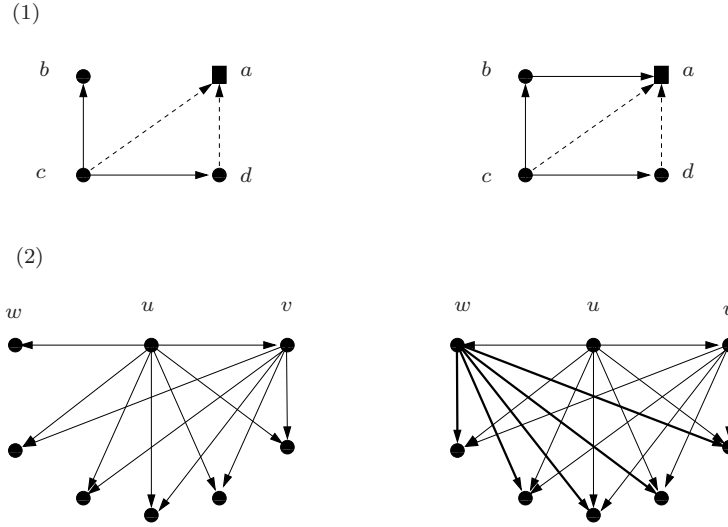


FIG 4. Illustration of the filtering procedure, with the graphical conventions introduced in Section 2.1. (1) Suppose the two patterns shown are potential local motifs with respect to the deletion of  $a$  and that the feed-forward loop was previously declared as a motif with respect to the deletion of  $a$ . Then the left pattern is filtered out by applying the filtering procedure for  $\mathcal{A} = \{b\}$ . However, the second is declared as a motif because of the edge between  $b$  and  $a$ . (2) Themes of order 5 in the network  $G$  for the two previous potential motifs. In the second case, the edge from  $b$  to  $a$  in the pattern implies the presence of 5 additional edges in the theme.

new information compared to the theme of the feed-forward loop obtained by deleting the vertex  $w$ .

On the opposite, a theme of order  $k$  of the second pattern contains  $k$  supplementary edges compared to the feed-forward loop theme. The coefficients of  $\mathbf{\Pi}$  being small in general because of the sparsity of real networks, the presence of those  $k$  edges is informative. Thus, that potential local motif is kept in the list of local motifs.

## 2.6. Algorithmic issues

Given an integer value  $k$ , our procedure first needs to list all the patterns of size  $k$  occurring in the network, as well as their subpatterns. That issue is tackled by using the ESU algorithm of Wernicke (2005). It is important to note that this subgraph count has to be done only once in the observed graph and not a huge number of times, in contrast to simulation-based approaches.

We then apply Theorem 2.2 to each pair (pattern, subpattern). The major cost in terms of computational time of that step is the computation of  $\mathbb{E}(N(\mathbf{m}'))$  in the estimated model. Indeed, it requires to sum a probability of occurrence along all possible positions in the graph. It may be done more efficiently by grouping the nodes belonging to the same class. This approach gives good re-

sults when the number of classes is not too large, which is the case in practice when estimating the classes using Stochastic Block Model algorithms. However, it becomes a real drawback for motifs larger than 4 vertices when using the *Expected Degree* method and observed graph with more than 200 vertices.

Finally, as the algorithm visits every position, it is not time-consuming to keep in memory all the positions and their respective theme orders in order to have a better interpretation of the results.

### 3. Lower bound

The fact that Theorem 2.2 gives an upper-bound of the exact  $p$ -value ensures that the number of false positives is controlled by the threshold used in the procedure. However, the tightness of that bound has to be taken into consideration to tackle the problem of false negatives.

An evaluation of the tightness of the local bound given by Theorem 2.1 can be obtained for moderate deviations, as stated in the following proposition.

**Proposition 3.1.** *Consider any pattern  $\mathbf{m}$  and subpattern  $(C, \mathbf{m}')$ . Let  $U$  be a position corresponding to an occurrence of  $\mathbf{m}'$  in  $G$  and define  $\lambda_{2,U} = \sum_{v \notin U} (p_U^v)^2$ . Denote by  $LB_U(t)$  the local upper bound on  $\mathbb{P}(\Delta_U \geq t)$  given by Theorem 2.1.*

*Suppose that  $\lambda_{2,U} < \frac{1}{4}$ . Then, for every  $t$  such that  $1 < t < \frac{1}{8\sqrt{\lambda_{2,U}}} - 1$ ,*

$$\frac{\mathbb{P}(\Delta_U \geq t)}{LB_U(t)} \geq \left(1 - 52 \frac{\lambda_{2,U}}{\lambda_U} (1+t)\right) \left(1 - \frac{2}{1+t}\right) \left(1 - \frac{1}{10\lambda_U(1+t)}\right)$$

The proof relies on results about Poisson approximations for sums of independent random variables given in Barbour, Holst and Janson and is detailed in Appendix B.1.

This theorem shows that for infinite sequences of real numbers  $t^{(n)}$ , graphs  $G^{(n)}$  and positions  $U^{(n)}$  such that  $t^{(n)}$  and  $(t\lambda_U)^{(n)}$  go to infinity and  $(t\lambda_{2,U}/\lambda_U)^{(n)}$  goes to 0, the bound on  $\mathbb{P}(\Delta_U^{(n)} \geq t^{(n)})$  is asymptotically tight.

For example, let us consider the Erdős-Rényi model with a connection probability  $p^{(n)} = \frac{c}{n}$ . That choice corresponds to a linear growth of the number of edges (Chung and Lu, 2006). Denote by  $k$  and  $n$  the respective sizes of  $\mathbf{m}$  and  $G$  and by  $r$  the number of edges which are in  $\mathbf{m}$  but not in  $\mathbf{m}'$ . Then, for any position  $U$ ,

$$\lambda_U^{(n)} = (n - k + 1)p^r \sim_{n \rightarrow +\infty} \frac{c^r}{n^{r-1}} \tag{3.1}$$

$$\lambda_{2,U}^{(n)} = (n - k + 1)p^{2r} \sim_{n \rightarrow +\infty} \frac{c^{2r}}{n^{2r-1}} \tag{3.2}$$

Thus, for  $r \geq 2$ , choosing  $t^{(n)} \sim n^\alpha$  with  $r - 1 < \alpha < r$  yields a sequence of thresholds for which the bound of Theorem 2.1 is asymptotically tight.

The tightness of the global bound given by Theorem 2.2 is a more intricate issue. Using the notations  $E^t = \{\max_U(g(\lambda_U, \Delta_U)) > t\}$  and  $E_U^t = \{g(\lambda_U, \Delta_U) > t\}$ , the derivation of a lower bound for the event  $E^t$  can be done using

$$\mathbb{P}(E^t) \geq \sum_U \mathbb{P}(E_U^t) - \frac{1}{2} \sum_{U \neq V} \mathbb{P}(E_U^t \cap E_V^t).$$

The term  $\sum_U \mathbb{P}(E_U^t)$  corresponding to the proposed upper bound, it is sufficient to derive tight upper bounds on the intersections  $E_U^t \cap E_V^t$ . Nevertheless, for some patterns, the number of extensions at two overlapping positions  $U$  and  $V$  may be strongly correlated. For instance, consider Figure 3 and in particular the occurrence of the pattern  $\mathbf{m}_2$  at position (PDR1, PDR3, PDR5, IPT1). The number of extensions of  $\mathbf{m}'_2$  at position  $U = (\text{PDR1}, \text{PDR3}, \text{PDR5})$  will be equal to the number of its extensions at position  $V = (\text{PDR1}, \text{PDR3}, \text{IPT1})$ . Therefore, the probability of the intersection is not small with respect to the probabilities of the single events.

However, this approach allows us to show that the global upper bound we propose is tight in the sense that for some patterns and parameters of the Erdős-Rényi model corresponding to sparse graphs, it is asymptotically the best one.

**Proposition 3.2.** *Let  $\mathbf{m}$  be a pattern of size  $k$  admitting some vertex  $a$  linked to every other vertex of  $\mathbf{m}$ . Consider its subpattern  $(C, \mathbf{m}')$  where  $C$  is the deletion class of  $a$ .*

*Consider the Erdős-Rényi model of parameter  $\rho$  with  $\rho = \mathcal{O}(n^{-\frac{1}{2}-\epsilon})$ , with  $\epsilon > \frac{1}{2k-1}$ . Let  $\lambda$  and  $\lambda_2$  denote the values of  $\lambda_U$  and  $\lambda_{2,U}$  which do not depend on  $U$  in that model.*

*Let  $0 < t < n^\epsilon$  and denote by  $GB(t)$  the global bound given by Theorem 2.2. Then, if  $\lambda_2 < \frac{1}{4}$ , and if  $y > 0$  such that  $g(\lambda, y) = t$  satisfies  $1 < y < \frac{1}{8\sqrt{\lambda_2}} - 1$ ,*

$$\frac{\mathbb{P}(E^t)}{GB(t)} \geq (1 - \eta) \left(1 - 52 \frac{\lambda_2}{\lambda} (1 + y)\right) \left(1 - \frac{2}{1 + y}\right) \left(1 - \frac{1}{10\lambda(1 + y)}\right) - kn^{2k} e^{-n^\epsilon + t}$$

where  $\eta = \mathcal{O}(n^{-\epsilon})$ .

Note that the condition on  $\rho$  still allows a growth rate of the number of edges of the order  $\mathcal{O}(n^{\frac{3}{2}-\epsilon})$ , which is faster than the linear growth observed on real data (Chung and Lu, 2006). The detailed proof of the proposition is given in Appendix B.2.

Combining Propositions 3.1 and 3.2 yields that the proposed upper bound is asymptotically tight for some pairs of patterns and subpatterns in a given range of model parameters. Consider for example again the Erdős-Rényi model with a connection probability  $p^{(n)} = \frac{c}{n}$  and a couple (pattern, subpattern) as in Proposition 3.2. For any  $0 < \delta < \frac{1}{2}$  the choice  $t^{(n)} \sim n^\delta \log(n^{k-1+\delta})$  corresponds to  $y^{(n)} \sim n^{k-1+\delta}$ . The first term of the right hand side in Proposition 3.2 then goes to 1 and the second vanishes as  $\delta < \epsilon = \frac{1}{2}$ . The upper bound is thus asymptotically tight in that case.

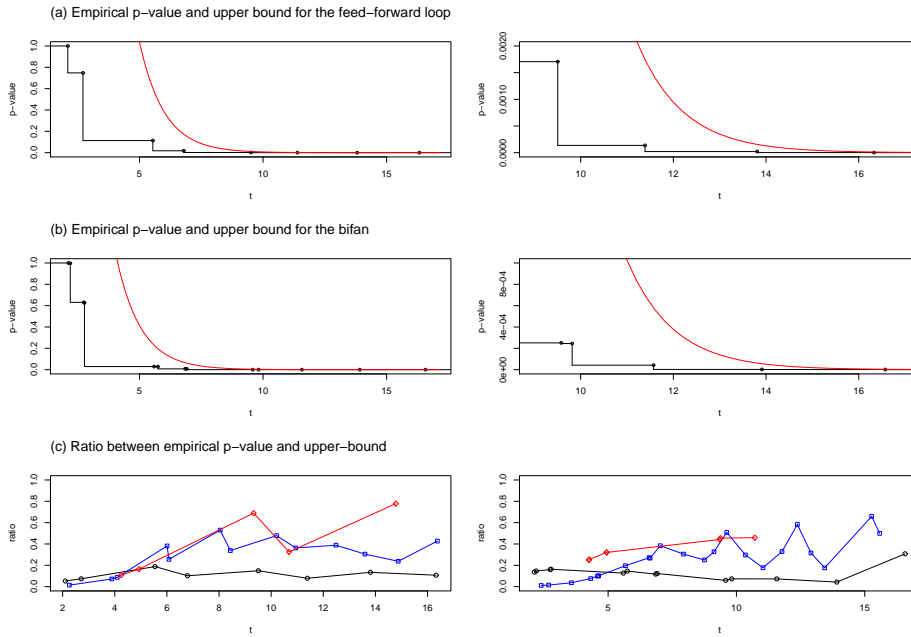


FIG 5. Empirical  $p$ -values and corresponding upper bound. (a) Both curves for the feed-forward loop in the 500,000 reference graphs and zoom at the distribution tail. (b) Same as (a) for the bi-fan pattern. (c) Ratio between the empirical  $p$ -value and the upper bound for the reference graphs (black circles), the dense graphs (blue squares) and the large graphs (red diamonds).

## 4. Illustration of the procedure

### 4.1. Simulated data

500,000 directed graphs with 90 vertices, which we will call the *reference graphs*, were generated under the model with three classes of 30 vertices each and connection probabilities set to 0.04 between vertices of the same class and 0.01 between vertices of different classes. The mean out-degree and in-degree under that model are both equal to 1.76. Our method is illustrated with the feed-forward loop and the bi-fan patterns (see Figure 2). The choice of the subpatterns is done by deleting the only vertex of in-degree 2 for the feed-forward loop and one of the vertices of in-degree 2 for the bi-fan. Nevertheless, that choice plays no role in that particular case, due to the symmetry of the model. Figure 5 (a) and (b) show the empirical tail probabilities and corresponding upper bounds given by Theorem 2.2, as functions of the parameter  $t$ .

To evaluate the quality of the upper bound, the ratio between the empirical  $p$ -values and their upper bounds is shown in Figure 5 (c). The same ratio is also plotted for denser graphs (30,000 graphs sampled with the same number of nodes and classes and connection probabilities five times larger) and for graphs larger than the reference graphs but with comparable density (30,000 graphs

with 360 nodes, three classes of 120 nodes each and connection probabilities 0.01 and 0.0025). For the reference graphs, the ratio is about 0.1 for both patterns, with a minimum of 0.07 for the feed-forward loop and 0.04 for the bi-fan. This ratio increases both for larger graphs and denser graphs, increasing to more than 0.2 for values of  $t$  greater than 5. The simulations therefore indicate that the upper bound proposed is not tight but is still in the right order of magnitude. Moreover, the approximation becomes better when the gap between the expected and the observed count grows.

#### 4.2. Stability with respect to the random model estimation

The  $p$ -value used to decide if a pattern is a motif or not depends on the classes of the vertices and on the inferred connection matrix  $\mathbf{\Pi}$ . It is thus important to evaluate the influence of the method used to infer those parameters of the null model. Let us consider five distinct inference procedures for directed graphs among those listed in 2.2:

**Erdős:** all vertices are in the same class and the connection probability such that the expected number of edges is equal to the observed one. It corresponds to Erdős-Rényi random graph model;

**Mixer** the vertex classes and connectivity matrix of a *Stochastic Block Model* are inferred using the R-package *mixer*. It allows to choose between the classification method of Zanghi, Ambroise and Miele (2008), the variational frequentist method of Daudin, Picard and Robin (2008) and the bayesian procedure of (Latouche, Birmelé and Ambroise, 2008). Those three methods are of increasing precision for the parameter estimation of the model, but also of increasing running time;

**BLOCKS** the vertex classes and connectivity matrix of a *Stochastic Block Model* are inferred using the procedure *BLOCKS* (Nowicki and Snijders, 2001) available in the *STOCNET* software (<http://stat.gamma.rug.nl/stocnet/>).






We run those five inference procedures on a real network to compare the local motifs of size 2 to 4 selected in each case. As *BLOCKS* only supports graphs up to 200 nodes, we consider a subnetwork of the transcriptional Yeast regulation network containing 194 nodes. Table 1 shows all the motifs found using a threshold of 0.001 on the  $p$ -value with at least one of the method. One can see that the list of motifs remains stable.

The Erdős-Rényi model leads to really different  $p$ -values and selects a motif found by no other method, but it is known to poorly describe real networks. However, all methods based on the group structure of the network find roughly the same local motifs with  $p$ -values of the same order of magnitude, excepted the online classification procedure which is rapid but not so precise. The only difference between them in terms of selected local motifs is the one of size two, which interpretation is related to the inferred blocks rather than to the network structure. Indeed, its selection denotes that at least one of the blocks is not



TABLE 1

Local motifs found with different estimation procedures for the parameters of the random model. The table gives the upper-bound on the p-value computed by (2.5). The values within brackets show potential motifs which are filtered out as redundant with a smaller motif

Motif	Erdős	Mixer			BLOCKS
		classification	variational	bayesian	
	5.8 e-43	2.1 e-12		1.6 e-4	4.0 e-4
	1.6 e-58	4.6 e-15	4.6 e-4	1.5 e-7	1.3 e-6
	4.9 e-19	2.6 e-11	3.6 e-6	1.8 e-5	3.9 e-5
	6.0 e-19	3.3e-14	4.5 e-8	1.0 e-5	9.6 e-5
	6.1 e-8				

homogeneous in terms of connection to the other blocks. The filtering procedure however ensures that this phenomenon does not imply irrelevant motifs of higher order.

### 4.3. Comparison with global motif detection methods on real networks

In order to point out the difference between global and local motif detection methods, we run our procedure to determine all local motifs of size 2 to 5 in two standard networks, both studied in Milo et al., and publicly available at <http://weizmann.ac.il/mcb/UriAlon>. Those networks are the transcriptional regulatory networks of Yeast and the electronic circuit *s420* of the *IS-CAS89* benchmark. They were used as a benchmark for most of the motif detection algorithms Kashani et al. (2009); Kashtan et al. (2004); Schreiber and Schwöbermeyer (2005); Wernicke (2005).

From a running time point of view, our method strongly depends on the statistical method chosen to infer the parameters of the model. The choice of a block model as null model forces to spend some time on the parameter estimation. However, the model fitting has to be done only once before looking for motifs of different sizes. This is clearly an advantage compared with the global methods based on network sampling, which have to do hundreds of samples and subgraph counts for each motif size. Table 2 compares the running time on the Yeast data for our procedure with the three possible inference options available in the *mixer* package and the *FANMOD* procedure for global motifs with full or sampled subgraph count. One can see that the running times are basically comparable depending on the level of precision used to run the methods.

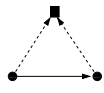
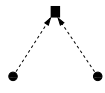
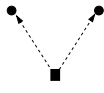
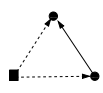
TABLE 2

Comparison in terms of running time in seconds on the Yeast data set (the computer used is a 800MHz single core computer). The local motif procedure is run for three inference methods of increasing precision. The global one is run for full and estimated enumeration of the subgraphs

Method	Option	Estimation	Size 2	Size 3	Size 4	Size 5	Total
paloma	classification	3	.01	.1	3	51	56
	variational	340	.03	.1	3	49	392
	bayesian	2380	.02	.1	3	49	2432
Fanmod	full enumeration			20	399	12160	12579
	sampling			5	29	347	381

TABLE 3

Local motifs of size 3 found in the Yeast regulatory network ranked by increasing  $p$ -value

Local motif				
$p$ -value bound	2.0 e-16	2.3 e-9	4.6 e-4	8.6 e-4
$N_{IJ}^*$	15	38	5	3

Comparing only running times may however be misleading in the sense that global and local motifs are distinct notions and should also be compared in a qualitative way. To that purpose, let us consider all the local motifs of size at most five found in the Yeast regulation network with a threshold of  $1e - 3$ .

Inference of the parameters of the model is done using the Bayesian MixNet approach in order to be able to tackle patterns of size 5 in those graphs. However, that choice implies that hubs may be grouped in the same class, which is relevant from a mixture model point of view but may generate quite inhomogeneous groups in terms of degrees and thus local motifs with poor biological interpretation. For example, the parameter inference on the Yeast network gives rise to a group of two hubs of respective out-degrees 71 and 44. Hence, the expected outdegree in that group is 57.5 and any star pattern will be selected as a local motif when centering the position on the largest hub. To avoid that phenomenon, we first run the whole procedure with the Bayesian MixNet approach and run it again with the Expected Degree approach when the position of the theme leading to a local motif shows that it is selected because of the presence of a vertex of high degree.

The number of local motifs which are not filtered out is limited to four motifs of size 3 shown respectively in Table 3, five motifs of size 4 shown respectively in Table 4, and one motif of size 5.

The first and fourth local motifs of size three are interesting in the sense that they show how the notion of local motif breaks the symmetry between the nodes in a pattern. Indeed, they correspond to the well-known feed-forward loop, composed by a main regulator  $X$  and a gene  $Y$  regulated by  $X$ , both co-regulating a third gene  $Z$ . This pattern is a local motif with respect to the subpattern obtained by deleting  $Z$ . This indicates the existence of places in the

TABLE 4  
Local motifs found of size 4 in the Yeast regulatory network ranked by increasing  $p$ -value

Local motif					
$p$ -value bound	6.5 e-15	3.4 e-6	1.4 e-4	5.6 e-4	9.2 e-4
$N_U^*$	7	2	2	1	1

network where a main regulator  $X$  and a gene  $Y$  regulated by  $X$  both co-regulate a high number of genes  $Z_1, \dots, Z_k$ . The value of  $N_U^*$  indicates that there is at least such a theme of order 15. That phenomenon is already described by Alon (2007), under the denomination *Multi-output feed-forward loops*.

The feed-forward loop is also found to be a local motif with respect to the deletion of the main regulator  $X$ . However, the  $p$ -value upper bound indicates that this local motif carries less information than the previous one. This fact is confirmed by the lower value of  $N_U^*$  which shows that the maximal order of corresponding themes is only 3. Note that the feed-forward loop is not a local motif with respect to the deletion of the intermediate gene  $Y$ , indicating that no main regulator  $X$  regulates a high number of genes  $Y_1 \dots Y_k$  in order to regulate a gene  $Z$ .

The second top local motif corresponds to a pair of regulators co-regulating a gene. Large themes corresponding to that pattern are described by Alon (2007) as *Dense Overlapping Regulons*. What is interesting is that this motif is selected by none of the global motif detection methods. However, all those methods select the motif of size 4 called bi-fan (see Figure 2). That global over-representation of the bi-fan is in fact a consequence of the local motif we detect. Indeed, the three largest themes of our local motif are of respective orders 38, 32 and 18. Thus, their presence imply  $\binom{38}{2} + \binom{32}{2} + \binom{18}{2} = 1352$  occurrences of the bi-fan, explaining its global over-representation. The local character of this motif is confirmed by the fact that those 1352 occurrences represent more than two thirds of the total number of bi-fans in the network.

The method also finds five local motifs of size 4 in the Yeast regulatory network. The first motif appearing in the list is of interest as it corresponds to three regulators co-regulating seven genes with an additional regulation between two of them. Another way to see the theme of that motif is a multi-output feed forward loop of order 7 with a third regulator acting on  $Z_1, \dots, Z_7$ . That motif is also found by the global methods but with a higher  $p$ -value and at the fifth or sixth position among the motifs of size 4.

The other local motifs have a value of  $N_U^*$  lower than 2, suggesting that they appear in the list because of a very low expected value. Note that the bi-fan does not appear in the list as it is filtered out as a redundancy of the second motif of size 3.

Finally, there is only one local motif of size 5. It corresponds to the fourth motif of size 4 with a supplementary edge going out from the squared vertex. Its

value for  $N_U^*$  is 1, showing that its expected value is low, such that it becomes over-represented even at its first occurrence.

The second network, that is the electronic circuit, has no local motif of size between 2 and 5, relying on a threshold of .01. However, one global motif of size 3 and two global motifs of size 4 were found in Milo et al. with  $Z$ -scores larger than 10. This indicates a distinct behavior of the two types of networks, the occurrences of the global motifs of the electronic network being spread in the whole network rather than agglomerated.

## 5. Conclusion

In this work, we propose a new approach to study network motifs, that is to look for locally over-represented patterns. Our framework allows us to take into account the over-representation of a pattern with respect to its subpatterns, for any pattern size. To list the local motifs of a network, we use a model-driven approach to determine a  $p$ -value upper bound for each pair (pattern, subpattern) and then apply a filtering procedure to eliminate redundancy.

Simulated data show that the error made by taking an upper-bound of the exact  $p$ -value is reasonable. The application of our method on standard real data allows us to find information on the role of the vertices of the motifs only by statistical means. Moreover, comparing the lists of local and global motifs highlights a strong structural difference between networks of different nature. In future work, we will investigate non standard data and a deeper understanding of the local motifs which are not global ones.

## Acknowledgements

The author would like to thank Catherine Matias and Gesine Reinert for their remarks and suggestions and Gilles Grasseau for his help while implementing the method.

## Appendix A: Local upper bound

To prove Inequality (2.2), we start from Inequality (2.1) which corresponds to Theorem 2.R in Barbour, Holst and Janson and apply it for  $K = \lceil \lambda_U(1+t) \rceil$ . Then,  $\Delta_U \geq t$  if and only if  $N_U \geq K$ .

For all  $k \geq K$ , let  $u_k = \frac{\lambda_U^k}{k!} e^{-\lambda_U}$ . Then  $Po(\lambda_U)([K, +\infty)) = \sum_{k \geq K} u_k$  and  $\forall k \geq K, \frac{u_{k+1}}{u_k} \leq \frac{1}{t+1}$ .

Thus, using that  $K! \geq \sqrt{2\pi K} \left(\frac{K}{e}\right)^K$ , we get

$$\begin{aligned} Po(\lambda_U)([K, +\infty)) &\leq \frac{1}{1 - \frac{1}{1+t}} \frac{\lambda_U^K}{K!} e^{-\lambda_U} \\ &\leq \frac{t+1}{t} \frac{\lambda_U^K e^K}{\sqrt{2\pi K} K^K} e^{-\lambda_U} \end{aligned}$$

Then, using Inequality (2.1),

$$\mathbb{P}(\Delta_U \geq t | G[U] \sim \mathbf{m}') \leq \frac{t}{t+1} \frac{t+1}{t\sqrt{2\pi\lceil\lambda_U(1+t)\rceil}} \frac{\lambda_U^{\lceil\lambda_U(1+t)\rceil} e^{\lceil\lambda_U(1+t)\rceil}}{(\lceil\lambda_U(1+t)\rceil)^{\lceil\lambda_U(1+t)\rceil}} e^{-\lambda_U}$$

Using its derivative, it is straightforward to check that the function defined by  $f(x) = \frac{\lambda^x e^x}{\sqrt{2\pi x x^x}}$  is decreasing on  $[\lambda, +\infty)$ . The former inequality is therefore still true by ignoring the integer part:

$$\begin{aligned} \mathbb{P}(\Delta_U \geq t | G[U] \sim \mathbf{m}') &\leq \frac{1}{\sqrt{2\pi\lambda_U(1+t)}} \frac{\lambda_U^{\lambda_U(1+t)} e^{\lambda_U(1+t)}}{(\lambda_U(1+t))^{\lambda_U(1+t)}} e^{-\lambda_U} \\ &\leq \frac{1}{\sqrt{2\pi\lambda_U(1+t)}} e^{-\lambda_U((1+t)\log(1+t)-t)} \end{aligned}$$

Writing that

$$\mathbb{P}(\Delta_U \geq t) = \mathbb{P}(\Delta_U \geq t | G[U] \sim \mathbf{m}') \mathbb{P}(G[U] \sim \mathbf{m}')$$

yields Inequality (2.2).

## Appendix B: Lower bound

### B.1. Local lower bound

The first step is to find the best possible bound for the difference between the tail probability of a sum of independent random variables and the tail probability of the corresponding Poisson approximation. This problem is presented and studied in Barbour, Holst and Janson (1992). We use Theorem 9.D presented in that book, namely

**Theorem B.1** (Barbour, Holst, Janson). *Define  $W = \sum_i X_i$ , where  $X_i$  are independent random variables. Set  $\lambda = \sum_i \mathbb{E}(X_i)$  and  $\lambda_2 = \sum_i \mathbb{E}(X_i)^2$ .*

*Let  $K \geq \lambda$  be an integer,  $\xi = \lambda_2/\lambda$  and  $\Gamma = (K - \lambda)/\sqrt{\lambda}$ . Then, uniformly in  $K$  satisfying  $\Gamma \geq 1$ ,  $K \leq \lambda/2\xi$  and  $1 + 4\Gamma^2 \leq (16\xi)^{-1}$ , we have*

$$\mathbb{P}(W \geq K) = Po(\lambda)([K, +\infty))(1 + \mathcal{O}(\xi) + \mathcal{O}(\xi\Gamma^2)).$$

Applying this result in our context for  $W = N_U$  and  $K = \lceil\lambda_U(1+t)\rceil$  for a fixed  $t$  may not be possible because in this case  $\Gamma = t\sqrt{\lambda_U}$  and thus the condition  $\Gamma \geq 1$  may not be satisfied when  $\lambda_U$  is too small with respect to  $t$ .

However, the proof of Theorem B.1 uses the assumption  $\Gamma \geq 1$  only once. Rewriting it without that condition until that step yields:

$$\frac{\mathbb{P}(W \geq K)}{Po(\lambda)([K, +\infty))} = (1 + \eta_1)(1 + \eta_2)$$

with

$$|\eta_1| \leq (1 + 2\Gamma\lambda^{-1/2})^2 2\lambda_2/(K - \lambda) \tag{B.1}$$

and

$$|\eta_2| \leq \sum_{r \geq K} Po(\lambda)(r) |\epsilon_r| / Po(\lambda)([K, +\infty)) \tag{B.2}$$

where  $|\epsilon_r| \leq 8\xi\Gamma^2 + 2(r - K)\xi\Gamma\lambda^{-1/2}$ .

The hypothesis  $\Gamma \geq 1$  is then used to bound the right hand side of Inequality (B.2), which can be alternatively bounded by

$$\begin{aligned} |\eta_2| &\leq 8\xi\Gamma^2 + 2\xi\Gamma\lambda^{-1/2} \sum_{r \geq K} (r - K) \frac{Po(\lambda)(\{r\})}{Po(\lambda)(\{[K, +\infty)\})} \\ &\leq 8\xi\Gamma^2 + 2\xi\Gamma\lambda^{-1/2} \sum_{r \geq K} (r - K) \frac{Po(\lambda)(\{r\})}{Po(\lambda)(\{K\})} \\ &\leq 8\xi\Gamma^2 + 2\xi\Gamma\lambda^{-1/2} \sum_{r \geq K} (r - K) \lambda^{r-K} \frac{K!}{r!} \\ &\leq 8\xi\Gamma^2 + 2\xi\Gamma\lambda^{-1/2} \sum_{r \geq K} (r - K) \lambda^{r-K} \frac{1}{K^{r-K}} \\ &\leq 8\xi\Gamma^2 + 2\xi\Gamma\lambda^{-1/2} \sum_{k \geq 0} k \left(\frac{\lambda}{K}\right)^k \\ &\leq 8\xi\Gamma^2 + 2\xi\Gamma\lambda^{-1/2} \frac{\lambda/K}{(1 - \lambda/K)^2}, \end{aligned}$$

the last inequality deriving from the equality  $\sum_{k \geq 0} kx^{k-1} = \frac{1}{(1-x)^2}$  for  $|x| < 1$ .

Using the additional condition of Proposition 3.1, that is  $K \geq 2\lambda$ , and the fact that it implies  $\Gamma\lambda^{-1/2} \geq 1$  allows us, using elementary bounds, to obtain from Inequalities (B.1) and (B.2) that

$$|\eta_1| \leq 36\xi\Gamma^2 \quad \text{and} \quad |\eta_2| \leq 16\xi\Gamma^2. \tag{B.3}$$

Thus,

$$\frac{\mathbb{P}(W \geq K)}{Po(\lambda)([K, +\infty))} \geq (1 - 36\xi\Gamma^2)(1 - 16\xi\Gamma^2) \geq 1 - 52\xi\Gamma^2$$

As  $\xi\Gamma^2 \leq \frac{K\lambda_2}{\lambda^2}$ , we have

$$\mathbb{P}(W \geq K) \geq Po(\lambda)(\{[K, +\infty)\}) \left(1 - 52\frac{K\lambda_2}{\lambda^2}\right).$$

The second part of the right hand-side of Proposition 3.1 comes from the asymptotic comparison between the tail probability  $Po(\lambda)(\{[\lambda_U(1 + t), +\infty)\})$  and its exponential approximation used in Theorem 2.1. It is derived in a very similar way than in Appendix A, using the following lower bound on  $K!$ , which can be proved from its asymptotic expansion

$$K! \geq \sqrt{2\pi K} \left(\frac{K}{e}\right)^K \left(1 + \frac{1}{10K}\right).$$

**B.2. Global lower bound**

Let  $\mathbf{m}$  be a pattern of size  $k$  admitting some vertex  $a$  linked to every other vertex of  $\mathbf{m}$ . Consider its subpattern  $(C, \mathbf{m}')$  where  $C$  is the deletion class of  $a$ .

Consider the Erdős-Rényi model of parameter  $\rho$  with  $\rho = \mathcal{O}(n^{-\frac{1}{2}-\epsilon})$ , with  $\epsilon > \frac{1}{2k-1}$ .

Under the Erdős-Rényi model with parameter  $\rho$ , the law of the number  $N_U$  of extensions does not depend on the position  $U$ . Thus, the mean number  $\lambda$  is the same for every position and there exist  $y > 0$  and an integer  $K$  be such that, for any position  $U$ ,

$$E_U^t = \{g(\lambda, \Delta_U) \geq t\} = \{\Delta_U \geq y\} = \{N_U \geq K\}$$

Let us also recall the notation  $E^t = \{\max_U(g(\lambda, \Delta_U)) \geq t\}$ .

**Lemma B.1.** *For any  $t < n^\epsilon$ ,*

$$\mathbb{P}(E^t) = (1 - \eta) \sum_U \mathbb{P}(E_U^t) - kn^{2k}\mathbb{P}(E^{n^\epsilon}),$$

where  $\eta = \mathcal{O}(n^{-\epsilon})$ .

*Proof.* We start from the simplest known lower bound for the probability of an union of events, that is:

$$\mathbb{P}(E^t) \geq \sum_U \mathbb{P}(E_U^t) - \frac{1}{2} \sum_U \sum_{V \neq U} \mathbb{P}(E_U^t \cap E_V^t) \tag{B.4}$$

Consider two positions  $U$  and  $V$ , and let  $ext(U)$  be the set of vertices yielding to extensions of  $\mathbf{m}'$  at position  $U$  and  $i = |V \setminus U|$ .

Let  $S$  be any set of vertices not intersecting  $U$ . We decompose  $E_V^t$  as the union of the sets  $\{T \subset ext(V)\}$  for all sets  $T$  of  $K$  vertices. Thus

$$\begin{aligned} \mathbb{P}(E_V^t | ext(U) = S) &\leq \sum_{T, |T|=K} \mathbb{P}(T \subset ext(V) | ext(U) = S) \\ &\leq \sum_{j=0}^{|S|} \sum_{|T|=K, |T \cap S|=j} \mathbb{P}(T \subset ext(V) | ext(U) = S). \end{aligned}$$

Let  $T$  be such that  $|T \cap S| = j$ . As  $\mathbf{m}'$  is connected, at least  $i$  edges need to be present in  $V \setminus U$  to ensure that  $G[V] \sim \mathbf{m}'$ . Moreover, let us recall that  $\mathbf{m}'$  is obtained from  $\mathbf{m}$  by deleting a vertex linked to all other vertices of  $\mathbf{m}$ . Therefore, to ensure that  $T \subset ext(V)$ , all the edges between  $V \setminus U$  and  $T$  and all the edges between  $U \cap V$  and  $T \setminus S$  have to be present, which amounts to a total of at least  $iK + (k-i)(K-j)$  edges.

Therefore,  $\mathbb{P}(T \subset ext(V) | ext(U) = S) \leq \rho^i \rho^{iK+(k-i)(K-j)}$ .

We distinguish between two different cases whether the cardinality of  $S$  is larger or smaller than  $n^\epsilon$ .

- If  $|S| \leq n^\epsilon$ , then

$$\begin{aligned}
 \mathbb{P}(E_V^t | ext(U) = S) &\leq \sum_{j=0}^{\min(|S|, K)} \binom{n}{K-j} \binom{|S|}{j} \rho^{i+K+(k-i)(K-j)} \\
 &\leq \rho^{i+kK} n^K \sum_{j=0}^K \binom{|S|}{j} \left(\frac{1}{n\rho^{k-i}}\right)^j \\
 &\leq \rho^{i+kK} n^K (K+1) \max\left(1, \left(\frac{1}{n\rho^{k-i}}\right)^K\right) \\
 &\leq \rho^i (K+1) \max(n\rho^k, \rho^i |S|)^K \tag{B.5}
 \end{aligned}$$

It is straightforward to check that, for large enough  $n$ , we have  $n\rho^k < \frac{1}{e}$  and  $\rho^i |S| < \frac{1}{e}$ . Therefore, the right hand side of Inequality (B.5) is a decreasing function of  $K$  and

$$\mathbb{P}(E_V^t | ext(U) = S) \leq 2 \max(n\rho^{i+k}, \rho^{2i} |S|)$$

Moreover,  $1 - (i+k)(\frac{1}{2} + \epsilon) \leq -i - \epsilon$  because  $\epsilon \geq \frac{1}{2k-1}$  and  $-2i(\frac{1}{2} + \epsilon) + \epsilon \leq -i - \epsilon$ . Thus, as  $\rho \leq Cn^{-\frac{1}{2}-\epsilon}$  for some constant  $C \geq 1$ , and  $|S| \leq n^\epsilon$ ,

$$\mathbb{P}(E_V^t | ext(U) = S) \leq C^{2k} n^{-i-\epsilon} \tag{B.6}$$

- For  $|S| \geq n^\epsilon$ , we roughly bound  $\mathbb{P}(E_V^t | ext(U) = S)$  by 1. However, let us note that  $g(\lambda, n^\epsilon) > n^\epsilon$  for large enough  $n$  and therefore

$$\begin{aligned}
 \sum_{S, |S| \geq n^\epsilon} \mathbb{P}(ext(U) = S) &= \mathbb{P}(N_U \geq n^\epsilon) \tag{B.7} \\
 &= \mathbb{P}(g(\lambda, N_U) \geq g(\lambda, n^\epsilon)) \\
 &\leq \mathbb{P}(E^{n^\epsilon}). \tag{B.8}
 \end{aligned}$$

Using Inequalities (B.2) and (B.8), we get

$$\begin{aligned}
 \mathbb{P}(E_U^t \cap E_V^t) &= \sum_{S; |S| \geq K} \mathbb{P}(E_V^t | ext(U) = S) \mathbb{P}(ext(U) = S) \\
 &\leq \sum_{K \leq |S| \leq n^\epsilon} C^{2k} n^{-i-\epsilon} \mathbb{P}(ext(U) = S) + \mathbb{P}(E^{n^\epsilon}).
 \end{aligned}$$

As the number of positions  $V$  such that  $|V \setminus U| = i$  is bounded by  $n^i$ , we finally obtain

$$\begin{aligned}
 \sum_{V \neq U} \mathbb{P}(E_U^t \cap E_V^t) &\leq \sum_{i=1}^k n^i (C^{2k} n^{-i-\epsilon} \mathbb{P}(E_U^t) + \mathbb{P}(E^{n^\epsilon})) \\
 &\leq k C^{2k} n^{-\epsilon} + k n^k \mathbb{P}(E^{n^\epsilon}). \tag{B.9}
 \end{aligned}$$

Inequalities (B.4) and (B.9) yield the lemma. □



Let  $t < n^\epsilon$  and  $y$  be such that  $g(\lambda, y) = t$ . By Proposition 3.1, for any position  $U$ ,

$$\mathbb{P}(E_U^t) = \mathbb{P}(\Delta_U \geq y) \geq \left(1 - 52 \frac{\lambda_2}{\lambda} (1+y)\right) \left(1 - \frac{2}{1+y}\right) \left(1 - \frac{1}{10\lambda(1+y)}\right) LB_U(y)$$

Thus, as  $GB(t) = \sum_U LB(y)$ , we obtain that

$$\sum_U \mathbb{P}(E_U^t) \geq \left(1 - 52 \frac{\lambda_2}{\lambda} (1+y)\right) \left(1 - \frac{2}{1+y}\right) \left(1 - \frac{1}{10\lambda(1+y)}\right) GB(t)$$

Using Lemma B.1 and applying Theorem 2.2 to the term  $\mathbb{P}(E^{n^\epsilon})$  yields

$$\begin{aligned} \mathbb{P}(E^t) &\geq (1-\eta) \left(1 - 52 \frac{\lambda_2}{\lambda} (1+y)\right) \left(1 - \frac{2}{1+y}\right) \left(1 - \frac{1}{10\lambda(1+y)}\right) GB(t) \\ &\quad - kn^{2k} \mathbb{E}(N(\mathbf{m}')) e^{-n^\epsilon} \end{aligned}$$

and thus

$$\frac{\mathbb{P}(E^t)}{GB(t)} \geq (1-\eta) \left(1 - 52 \frac{\lambda_2}{\lambda} (1+y)\right) \left(1 - \frac{2}{1+y}\right) \left(1 - \frac{1}{10\lambda(1+y)}\right) - kn^{2k} e^{-n^\epsilon+t}$$

## References

- AIROLDI, E., BLEI, D., FIENBERG, S. and XING, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9** 1981-2014.
- ALON, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics* **8** 450-461.
- ARTZY-RANDRUP, Y., FLEISHMAN, S. J., BEN-TAL, N. and STONE, L. (2004). Comment on “Network Motifs: Simple Building Blocks of Complex Networks” and “Superfamilies of Evolved and Designed Networks”. *Science* **305**.
- BANKS, E., NABIEVA, E., CHAZELLE, B. and SINGH, M. (2008). Organization of Physical Interactomes as Uncovered by Network Schemas. *PLoS Comput. Biol.* **4** e1000203. [MR2457131](#)
- BARBOUR, A. D., HOLST, L. and JANSON, S. (1992). *Poisson approximation*. Oxford University Press. [MR1163825](#)
- BERG, J. and LÄSSIG, M. (2004). Local graph alignment and motif search in biological networks. *Proc. Nat. Acad. Sci.* **101** 14689-14694.
- BOUCHERON, S., LUGOSI, G. and MASSART, P. (2003). Concentration inequalities using the entropy method. *Ann. Probab.* **31** 1583-1614. [MR1989444](#)
- CHEN, L. H. Y. (1975). Poisson approximation for dependant trials. *Ann. Probab.* **3** 534-545. [MR0428387](#)
- CHUNG, F. and LU, L. (2006). *Complex Graphs and Networks (CBMS Regional Conference Series in Mathematics)*. AMS. [MR2248695](#)
- DAUDIN, J. J., PICARD, F. and ROBIN, S. (2008). Mixture model for random graphs. *Stat. Comput.* **18** 173-183. [MR2390817](#)

- DOBRIN, R., BEG, Q. K., BARABÁSI, A. L. and OLTVAI, Z. N. (2004). Aggregation of topological motifs in *Escherichia Coli* transcriptional regulatory network. *BMC Bioinformatics* **5** 10.
- ERDŐS, P. and RÉNYI, A. (1959). On random graphs I. *Publ. Math. Debrecen* **6** 290-297. [MR0120167](#)
- HOFMAN, J. and WIGGINS, C. (2008). Bayesian approach to network modularity. *Phys. Rev. Lett.* **100**.
- HOLLAND, P. W. and LEINHARDT, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology* **76** 492-513.
- HUNTER, D. and HANDCOCK, M. (2006). Inference of curved exponential family models for networks. *Journal of Computational and Graphical Statistics* **15** 565-583. [MR2291264](#)
- JANSON, S. (1990). A functional limit theorem for random graphs with applications to subgraph count statistics. *Random structures & Algorithms* **1** 15-37. [MR1068489](#)
- JANSON, S., OLESKIEWICZ, K. and RUCINSKI, A. (2004). Upper tails for subgraph counts in random graphs. *Israel Journal of Mathematics* **142** 61-92. [MR2085711](#)
- KASHANI, Z. R. M., AHRABIAN, H., ELAHI, E., NOWZARI-DALINI, A., ANSARI, E. S., ASADI, S., MOHAMMADI, S., SCHREIBER, F. and MASOUDI-NEJAD, A. (2009). Kavosh: a new algorithm for finding network motifs. *BMC Bioinformatics* **10**.
- KASHTAN, N., ITZKOVITZ, S., MILO, R. and ALON, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* **20-11** 1746.
- LATOUCHE, P., BIRMELE, E. and AMBROISE, C. (2008). Bayesian methods for graph clustering. *SSB preprint* 17.
- LATOUCHE, P., BIRMELE, E. and AMBROISE, C. (2011). Overlapping Stochastic Block Models with Application to the French Political Blogosphere. *Ann. Appl. Stat.* **5** 309-336. [MR2810399](#)
- MATIAS, C., SCHBATH, S., BIRMELE, E., DAUDIN, J. J. and ROBIN, S. (2006). Network motifs: mean and variance for the count. *REVSTAT* **4** 31-51. [MR2259363](#)
- MCDIARMID, C. (1998). Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics* (J. R.-A. M. HABIB C. MCDIARMID and B. REED, eds.) 195-248. Springer. [MR1678578](#)
- MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. and ALON, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science* **298** 824-827.
- NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *JASA* **96** 1077-87. [MR1947255](#)
- PICARD, F., DAUDIN, J. J., KOSKAS, M., SCHBATH, S. and ROBIN, S. (2008). Assessing the exceptionality of network motifs. *J. Comput. Biol.* **15** 1-20. [MR2383618](#)

- PICARD, F., MIELE, V., DAUDIN, J. J., COTTRET, L. and ROBIN, S. (2009). Deciphering the connectivity structure of biological networks using MixNet. *BMC Bioinformatics* **10** 1-11.
- SCHREIBER, F. and SCHWÖBERMEYER, H. (2005). MAVisto: a tool for the exploration of network motifs. *Bioinformatics* **21** 3572-3574.
- WASSERMAN, S. and FAUST, K. (1994). *Social network analysis: methods and applications*. Cambridge University Press.
- WATTS, D. J. and STROGATZ, S. H. (1998). Collective dynamics of small-world networks. *Nature* **393** 440-442.
- WERNICKE, S. (2005). Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **3(4)** 347-359.
- WERNICKE, S. and RASCHE, F. (2006). FANMOD: a tool for fast network motif detection. *Bioinformatics* **22** 1152-1153.
- WHITE, H. C., BOORMAN, S. A. and BREIGER, R. L. (1976). Social structure from multiple networks I: Blockmodels of roles and positions. *American Journal of Sociology* **81** 730-779.
- ZANGHI, H., AMBROISE, C. and MIELE, V. (2008). Fast online graph clustering via Erdős-Rényi mixture. *Pattern Recognition* **41** 3592-3599.
- ZHANG, L. V., KING, O. D., WONG, S. L., GOLDBERG, D. S., TONG, A. H., LESAGE, G., ANDREWS, B., BUSSEY, H., BOONE, C. and ROTH, F. P. (2005). Motifs, themes and thematic maps of an integrated *Saccharomyces Cerevisiae* interaction network. *J. Biol.* **4**.