Applied Network Science

RESEARCH Open Access

# Establish the expected number of induced motifs on unlabeled graphs through analytical models

Emanuele Martorana[1], Giovanni Micale[2], Alfredo Ferro[2] and Alfredo Pulvirenti[2*]

*Correspondence:
alfredo.pulvirenti@unict.it
[2]University of Catania, Dept. of Clinical and Experimental Medicine, Catania, Italy
Full list of author information is available at the end of the article

**Abstract**
Complex networks are usually characterized by the presence of small and recurrent patterns of interactions between nodes, called network motifs. These small modules can help to elucidate the structure and the functioning of complex systems. Assessing the statistical significance of a pattern as a motif in a network *G* is a time consuming task which entails the computation of the expected number of occurrences of the pattern in an ensemble of random graphs preserving some features of *G*, such as the degree distribution. Recently, few models have been devised to analytically compute expectations of the number of non-induced occurrences of a motif. Less attention has been payed to the harder analysis of induced motifs. Here, we illustrate an analytical model to derive the mean number of occurrences of an induced motif in an unlabeled network with respect to a random graph model. A comprehensive experimental analysis shows the effectiveness of our approach for the computation of the expected number of induced motifs up to 10 nodes. Finally, the proposed method is helpful when running subgraph counting algorithms to get the number of occurrences of a topology become unfeasible.

**Keywords:** Induced motifs, Networks, Random graphs, Analytical models

## Introduction

Given a network $G$, a motif $M$ (also referred as subgraph or pattern) of $G$ is defined as a small subgraph of $G$, whose frequency, that is the number of times $M$ occurs in $G$, is statistically significantly higher or lower (under-represented motifs within the network) than expected with respect to a reference null model. The frequency of topologies can be obtained with two different approaches *non-induced* or *induced* according to the structure constraints the users want to keep. When no restriction on the presence or absence of edges in a subset of nodes, we refer to non-induced subgraphs. These motifs are more informative than induced ones since a node could not have significant functions involving all its neighborhood. When we look for motifs in which the missing edges between nodes matter we deal with those named induced. The analysis of induced subgraphs of a network could explain better its structure, since in some classes of networks (e.g. biological) the presence or absence of each edge is important. Moreover, induced subgraphs can be considered as the skeleton of a network for their uniqueness of occurrence. Motif search problem consists in finding all motifs of a given size (i.e. with a given number of nodes) in a network. This problem has several applications ranging from biology to economics and social science (Milo et al. 2002; Chen and Yuan 2006; Squartini and Garlaschelli 2011).

The null model used to establish the statistical significance of a motif is commonly represented as an ensemble of random networks that keep some input network properties, such as the distribution of node degrees. A subgraph is claimed significant as a motif if the expected frequency in the null model is significantly higher or lower than the one observed in the input network.

The most popular strategy to compute the statistical significance of a motif $M$ in a graph $G$ is based on a permutation test which is defined as follows:

1. Compute the frequency of $M$ in $G$;
2. Build a large set of random networks preserving some features of $G$, based on a reference null model;
3. Compute the frequency of $M$ in each generated random network;
4. Calculate a p-value by comparing the frequency of $M$ in $G$ with the average frequency of $M$ in the random networks.

When the average frequency of $M$ in the random networks is substantially lower than the frequency of $M$ in $G$, than $M$ is over-represented in $G$, while if it is considerably higher, then $M$ is under-represented in $G$. Examples of null network models include the Erdös-Renyi (ER) model (Erdös and Renyi 1959), the Fixed degree distribution (FDD) model (Newman et al. 2001), the Expected degree distribution (EDD) model (Chung and Lu 2002; Park and Newman 2003) and the Erdös-Renyi mixture for graphs (ERMG) model (Nowicki and Snijders 2001; Daudin et al. 2008).

The simulation-based method evaluates the statistical significance based on a p-value using a resampling approach (Milo et al. 2002, 2004; Prill et al. 2005; Shen-Orr et al. 2002). Though this method produces reasonable results, it is computationally expensive because it requires generating a large number of random graphs and counting the occurrences of a motif both in the input network and in the ensemble of random graphs. Counting the occurrences of a subgraph involves the subgraph isomorphism which is known to be a NP-complete problem (Cook 1971).

Recently, much research has been spent to avoid such a simulation-based approach. Several models have been developed to assess the significance of a motif without the generation of an ensemble of networks. In (Wernicke 2006), Wernicke proposes an approximated method to estimate the asymptotic normality of the distribution of motif counts. In (Picard et al. 2008), Picard et al. introduced a model for the exact computation of mean and variance of the number of non-induced occurrences of a pattern according to any exchangeable random model. The exchangeability allows to assume that the occurrence probability of a motif does not depend on its location in the network. Thanks to this property every subset of $k$ nodes in the graph could potentially be part of the motif. The authors define equations to compute motif occurrence probability with respect to four exchangeable random models (FDD, ER, EDD and ERMG). Finally, through the Pólya-Aeppli distribution (Johnson et al. 1992) the p-value of motif significance is established. More recently, (Micale et al. 2018; Micale et al. 2019) have extended the work of (Picard et al. 2008) by defining equations for the computation of mean and variance of motifs in node-labeled networks (Micale et al. 2018) and multi-relational networks (Micale et al. 2019) according to the EDD model.

In (Picard et al. 2008) authors provided equations only for computing expectations of non-induced motifs. They also suggest a method to extend the analytical model to induced motifs, by applying the Kocay Lemma (Kocay 1981). The Lemma allows to express the number of induced occurrences of a subgraph as a linear combination of the number of non-induced occurrences of all subgraphs of the same size and vice versa. This result can thus be used to compute the mean and the variance of the count of induced motif. However, computing the coefficients of such a linear combination is a time consuming task, even for motifs of small size (6 or more nodes).

The paper is organized as follows. In "Definitions" section and "Previous work" section we provide definitions and previous works on analytical models for non-induced motifs respectively. In "A novel analytical model for the expectation of induced motifs" section we present a novel analytical model to calculate the mean and the variance of the count of an induced motif according to the EDD random model, without computing the coefficients defined by the Kocay Lemma. This work extends the preliminary work illustrated in (Martorana et al. 2020) by providing an engineering of our Rapid Matrix Elaboration (RaME) algorithm. "Complexity analysis" section presents a detailed theoretical complexity analysis. Our comprehensive experimental analysis clearly shows that our model can compute the mean of count of induced motifs up to 10 nodes in a reasonable running time. On the other hand, the analytical model based on Kocay Lemma becomes impractical starting from motifs of 7 nodes. Although RaME is very efficient for induced mean calculation, it results unfeasible to compute the variance even for small size subgraphs. However, recently, an approach to compute approximated p-values based on the estimation of the variance has been proposed (Micale et al. 2019), and therefore could be taken into account with our approach. Finally, in "Experimental results" section, we describe a case study to show some scenarios where our approach can be used to get the expected counts of induced motifs when establish the actual counts results computationally expensive and therefore unfeasible in resonable time.

## Definitions

In this section we provide some preliminary definitions about networks and motifs.

A *network* (also referred to as graph) is a pair $G = (V, E)$, where $V$ is the set of nodes and $E = \{(a, b) : a, b \in V\}$ is a set of node pairs. The size of $G$ is the number of its nodes $n = |V|$. When $\forall (a, b) \in E, (b, a) \in E$, i.e. all relations between nodes are bidirectional, the graph is called undirected otherwise the it is directed.

A common representation of a graph is the adjacency matrix $A$, which is a $n \times n$ matrix, where $A[\,i, j] = 1$ iff there is an edge between nodes $i$ and $j$, otherwise $A[\,i, j] = 0$.

Two graphs $G = (V, E)$ and $G' = (V', E')$ are isomorphic iff there exists a bijective function $M : V \rightarrow V'$, called isomorphism, such that $\forall a, b \in V : (a, b) \in E \Leftrightarrow (M(a), M(b)) \in E'$. A subgraph *S=(V',E')* of a $G$ is a graph in which $V' \subseteq V$ and $E' \subseteq E$. We refer to $S$ as *induced* when $\forall a, b \in V' : (a, b) \in E \iff (a, b) \in E'$. Otherwise $S$ is called *non-induced*. This means that the definition of induced subgraph is more limiting than the one of non-induced subgraph.

A graph $G' = (V', E')$ is subgraph isomorphic to a graph $G = (V, E)$ when $G'$ is isomorphic to a subgraph of $G$ (Cook 1971). The number of occurrences of $G'$ in $G$ (also called the frequency of $G'$ in $G$) corresponds to the number of subgraph isomorphisms of $G'$ in $G$.

A *motif* $M$ of a graph $G$ is defined as a subgraph of $G$ whose frequency is significantly higher than expected with respect to a null random model. Commonly, the network instance $G$ is thought to be drawn from a universe of random graphs which share some characteristics, such as the degree distribution.

### EDD model

The EDD model generates networks where node degrees follow the degree distribution of a given input network. Let $G$ be a graph with $n$ nodes and let $Deg(G)$ be its node degree distribution. We define an indicator random variable $X_{ij}$ which is equal to 1 iff there is an edge linking nodes $i$ and $j$. In the EDD, the likelihood of getting an edge between two nodes $i$ and $j$, provided the degrees $D_i$ and $D_j$ sampled from the $Deg$ distribution, is:

$$P(X_{ij} = 1 | D_i, D_j) = \min(1, \gamma D_i D_j) \tag{1}$$

with $\gamma = \frac{1}{(n-1)\mathbb{E}[Deg]}$.

## Previous work

### Analytical model for the expectation of non-induced motifs

In (Picard et al. 2008) Picard et al. presented an analytical model to compute the statistical significance of non-induced subgraphs as motifs in both directed and undirected graphs under three different exchangeable random models, i.e. Erdös-Renyi (ER) (Erdös and Renyi 1959), Expected Degree Distribution (EDD) (Chung and Lu 2002; Park and Newman 2003) and Erdös-Renyi Mixture for Graphs (ERMG) (Nowicki and Snijders 2001; Daudin et al. 2008). An exchangeable model is any model in which the occurrence probability of a motif in a network does not depend on the occurrence position. Here we describe equations to establish the significance of non-induced motifs according to the EDD model.

#### *Occurrence probability under the EDD random model*

The occurrence probability of a motif $m$ with $k$ nodes in $G$, given an assignment of expected degrees $D_i$ to the nodes of the motif, is the product of edge probabilities. The overall probability of occurrence of $m$ can be then obtained by summing across all probabilities obtained assigning all the possible degrees $D_i$ present in the input network. In (Picard et al. 2008) authors show that, under the EDD model, the latter probability can be finally expressed as products of some moments of the $Deg$ distribution of $G$:

$$\mu(m) = \gamma^{m_{++}} \prod_{u=1}^{k} \mathbb{E}[Deg]^{m_{u+}} \tag{2}$$

where $m_{++}$ is the number of edges in $m$, $m_{u+}$ is the degree of node $u$ in $m$ and $\mathbb{E}[Deg]^{m_{u+}}$ is the moment of order $m_{u+}$ of $Deg$ distribution.

#### *Mean and variance of the count*

Starting from $\mu(m)$ one can derive equations for computing the mean and the variance of the count of non-induced occurrences of $m$. While the expression of $\mu(m)$ depends on the specific random model used, equations for the mean and the variance hold for any exchangeable random model.

The exchangeability property guarantees that $\mu(m)$ is independent from the position of $m$ in $G$. In other words, all possible $k$-tuples of nodes in $G$ may embed an occurrence of the motif. If $G$ has $n$ nodes, the number of such $k$-tuples is $\binom{n}{k}$. In addition, if $\alpha$ is a $k$-tuple of nodes of network $G$, $m$ can be observed in different configurations, which are given by all $k!$ possible permutations of positions of $\alpha$'s nodes. However, some of these permutations actually yield redundant occurrences, i.e. occurrences of $m$ having the same adjacency matrix. All permutations with distinct adjacency matrices are called Non-Redundant Permutations (NRPs). If we denote the set of NRPs with $R(m)$ and with $\varrho(m) = |R(m)|$ the number of NRPs, the mean number of non-induced occurrences can be expressed by the following equation:

$$\mathbb{E}[N(m)] = \binom{n}{k}\varrho(m)\mu(m) \tag{3}$$

The variance of the count of occurrences of motif $m$ can be computed starting from the expectation of the squared count of non-induced occurrences of $m$, i.e. $\mathbb{E}[N^2(m)]$. The calculation of $\mathbb{E}[N^2(m)]$ can be performed by considering all possible overlaps of nodes and edges of two NRPs of $m$. Given two NRPs $m'$ and $m''$ of $m$. Picard et al. introduced the overlapping operation with $s$ nodes, $m'\Omega_s m''$, whose result is a super-motif with $2k-s$ nodes. The adjacency matrix of this super-motif can be obtained by splitting the adjacency matrices of $m'$ and $m''$ into four blocks of different sizes as follows:

$$m' = \left(\begin{array}{c|c} \begin{matrix} m'_{11} \\ [k-s,k-s] \end{matrix} & \begin{matrix} m'_{12} \\ [k-s,s] \end{matrix} \\ \hline \begin{matrix} m'_{21} \\ [s,k-s] \end{matrix} & \begin{matrix} m'_{22} \\ [s,s] \end{matrix} \end{array}\right) m'' = \left(\begin{array}{c|c} \begin{matrix} m''_{11} \\ [s,s] \end{matrix} & \begin{matrix} m''_{12} \\ [s,k-s] \end{matrix} \\ \hline \begin{matrix} m''_{21} \\ [k-s,s] \end{matrix} & \begin{matrix} m''_{22} \\ [k-s,k-s] \end{matrix} \end{array}\right)$$

The adjacency matrix of the super-motif is then given by:

$$m' \cap_s m'' = \left(\begin{array}{c|c|c} m'_{11} & m'_{12} & 0 \\ \hline m'_{21} & \max(m'_{22}, m''_{11}) & m''_{12} \\ \hline 0 & m''_{21} & m''_{22} \end{array}\right)$$

where the max function in the central term indicates that for the $s$ common nodes of $m'$ and $m''$, all edges of $m'_{22}$ and $m''_{11}$ have to be present. The max function is equivalent to the logical OR. Figure 1 shows all super-motifs generated from the NRPs of the 3-nodes path with $s = 2$.

For the computation of $\mathbb{E}[N^2(m)]$ we need to consider that $m'$ and $m''$ can be disjoint (i.e. they have no nodes in common) or can overlap in one or more nodes. Moreover, if $m'$ and $m''$ overlap in $s$ nodes, with $1 \le s \le k$, we have to take into account all possible ways in which the two NRPs of $m$ can overlap. Therefore, the expected squared count $\mathbb{E}[N^2(m)]$ is given by:

$$\begin{aligned} \mathbb{E}[N^2(m)] = \binom{n}{n-2k,k,k}\left[\sum_{m' \in R(m)} \mu(m')\right]^2 + \\ + \sum_{s=1}^{k}\left[\binom{n}{k-s,s,k-s,n-2k+s} \times \sum_{m',m'' \in R(m)} \mu(m' \cap_s m'')\right] \end{aligned} \tag{4}$$

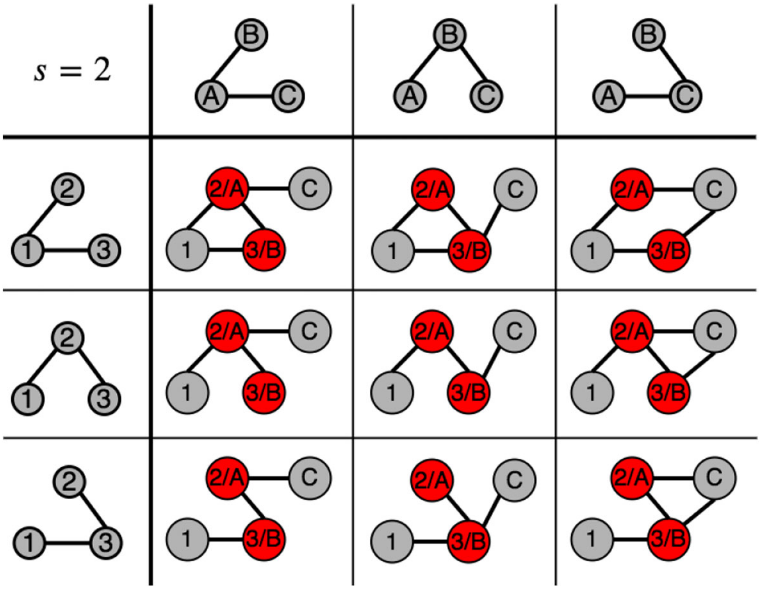


**Fig. 1** Super-motifs resulting from the overlapping between all NRPs of the non-induced 3-nodes path with $s = 2$ nodes. Results are shown in a matrix fashion, where the entry at row *i* and column *j* is the super-motif originating from the NRP at the *i*-th row and the NRP at the *j*-th columnn. Nodes of each super-motif contain node ids of the starting NRPs. Nodes in the overlapping region are colored in red

Finally, the variance is $\mathbb{V}[N(m)] = \mathbb{E}[N^2(m)] - \mathbb{E}[N(m)]^2$.

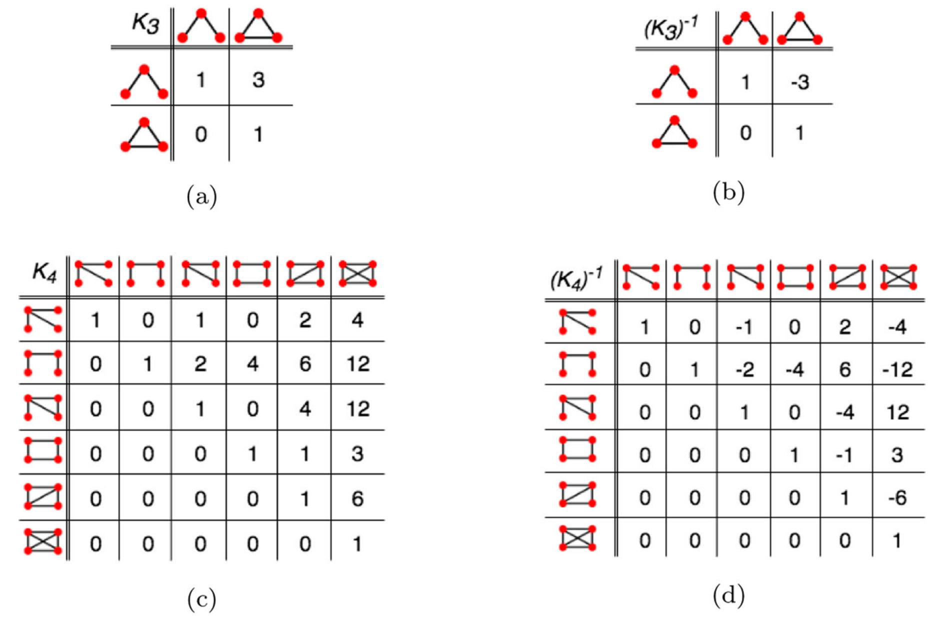
### Analytical model for the expectation of induced motifs

In (Picard et al. 2008) Picard et al. also sketch a possible solution for computing expectations of counts of induced motifs. Authors showed that the induced count of a motif of size $k$ can be always expressed as a linear combination of the non-induced counts of all non-isomorphic subgraphs of size $k$. For instance, the number of induced 3-nodes paths in a graph $G$ is equal to the number of non-induced 3-nodes paths minus the number of non-induced 3-nodes cliques multiplied by 3 (Fig. 3a). The coefficients of such a linear combination are obtained through the Kocay Lemma (Kocay 1981) and correspond to the entries of the inverse of a matrix $K_k$ of size $p \times p$, called Kocay matrix, where $p$ is the number of connected and non- isomorphic topologies of size $k$. The entry $K_k[i,j]$ stores the number of non-induced occurrences of motif $i$ within motif $j$. Figure 2 shows Kocay matrices $K_3$ and $K_4$ and their inverse $K_3^{-1}$ and $K_4^{-1}$. For example, $K_3[1,2] = 3$ because there are 3 occurrences of a 3-nodes path in a 3-nodes clique. Two examples of application of Kocay Lemma are shown in Fig. 3 for the 3-nodes path and the 4-nodes star.

An equivalent relation holds for the induced mean of the count of a $k$-nodes motif $m$, $\mathbb{E}[N_I(m)]$, which can be expressed as a linear combination of the means of non-induced counts of all non-isomorphic subgraphs of size $k$, where the coefficients are again given by the matrix $K_k^{-1}$. Formally:

$$\mathbb{E}[N_I(m)] = \sum_{t \in T_k} K_k^{-1}[m,t]\, N(t) \tag{5}$$

where $T_k$ is the set of all non-isomorphic $k$-nodes subgraphs.



(a)

| $K_3$ | | |
|---|---|---|
| | 1 | 3 |
| | 0 | 1 |

(b)

| $(K_3)^{-1}$ | | |
|---|---|---|
| | 1 | -3 |
| | 0 | 1 |

(c)

| $K_4$ | | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 0 | 1 | 0 | 2 | 4 |
| | 0 | 1 | 2 | 4 | 6 | 12 |
| | 0 | 0 | 1 | 0 | 4 | 12 |
| | 0 | 0 | 0 | 1 | 1 | 3 |
| | 0 | 0 | 0 | 0 | 1 | 6 |
| | 0 | 0 | 0 | 0 | 0 | 1 |

(d)

| $(K_4)^{-1}$ | | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 0 | -1 | 0 | 2 | -4 |
| | 0 | 1 | -2 | -4 | 6 | -12 |
| | 0 | 0 | 1 | 0 | -4 | 12 |
| | 0 | 0 | 0 | 1 | -1 | 3 |
| | 0 | 0 | 0 | 0 | 1 | -6 |
| | 0 | 0 | 0 | 0 | 0 | 1 |

**Fig. 2** Kocay matrices and their inverse: **(a)** The Kocay matrix $K_3$ for 3-nodes subgraphs, **(b)** The Inverse of Kocay matrix $K_3$, **(c)** The Kocay matrix $K_4$ for 4-nodes subgraphs, **(d)** The Inverse of Kocay matrix $K_4$

The same coefficients can be used to estimate the variance of the induced count of $m$, $\mathbb{V}[N_I(m)]$. However, the calculation of variance requires not only the variances of all $k$-nodes subgraphs but also all covariances between them. More formally:

$$\mathbb{V}[N_I(m)] = \sum_{t\in T_k}\left[K_k^{-1}[m,t]\right]^2\mathbb{V}[N(t)] + \sum_{\substack{t',t''\in T_k\\ t'\neq t''}} K_k^{-1}[m,t']\,K_k^{-1}[m,t'']\,Cov\left(N(t'),N(t'')\right) \quad (6)$$

We have that $Cov\left(N(t'),N(t'')\right) = \mathbb{E}[N(t')N(t'')] - \mathbb{E}[N(t')]\,\mathbb{E}[N(t'')]$. The first term of the covariance is obtained like for Eq. 4, considering the non-induced occurrence prob-

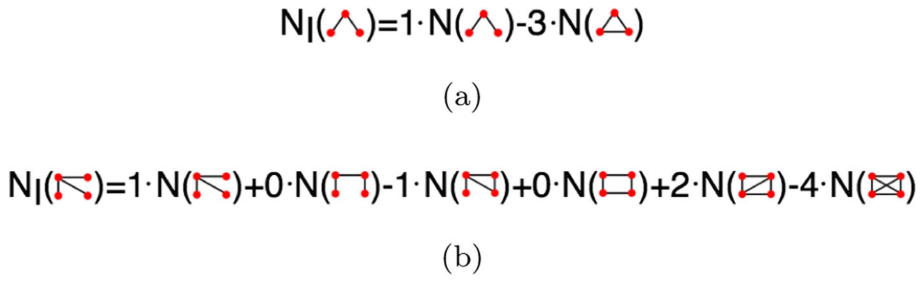


**Fig. 3** Examples of application of Kocay Lemma. **a** Number of induced occurrences of the 3-nodes path as a linear combination of the non-induced counts of the 3-nodes path and the 3-nodes clique: the coefficients of the combination are the entries of the first row of matrix $K_3^{-1}$ (corresponding to the 3-nodes path), i.e. 1 and − 3, respectively. **b** The number of induced occurrences of the 4-nodes star expressed as a linear combination of the non-induced counts of all six 4-nodes subgraphs: the coefficients of the combination are the entries of the first row of matrix $K_4^{-1}$ (corresponding to the 4-nodes star)

abilities of all NRPs of $t'$ and $t''$ and the non-induced occurrence probabilities of the super-motifs generated by the overlapping between a NRP of $t'$ and a NRP of $t''$:

$$\mathbb{E}[N(t^{'})N(t^{''})] = \binom{N}{N-2k, k, k} \sum_{m' \in R(t^{'}), m'' \in R(t^{''})} \mu(m')\mu(m'') + \\ \sum_{s=1}^{k} \binom{N}{k-s, s, k-s, N-2k+s} \sum_{m' \in R(t^{'}), m'' \in R(t^{''})} \mu(m' \cap_s m'') \tag{7}$$

Therefore, the computation of the variance of the induced count using Kocay Lemma is computationally expensive even for motifs of small size.

## A novel analytical model for the expectation of induced motifs

In this section we present a new analytical model to estimate the mean and the variance of the count of induced motifs, that directly estimates the induced occurrence probability and thus avoids the calculation of the Kocay matrix. The analytical model presented here uses the Expected Degree Distribution (EDD) as random graph model and focuses on undirected graphs. However, it can be extended to directed graphs. Furthermore, the equations for induced mean and variance hold for any exchangeable random model.

### Direct estimation of occurrence probability of induced motifs

Let $G$ be a graph and let $Deg$ be its degree distribution. We define as $D_i$ the degree of a node $i$ in $G$.

As described in "EDD model" section, the probability of existence of an edge between two nodes $i$ and $j$ in $G$ is given by Eq. 1. Therefore, the probability of observing no edge between $i$ and $j$ is $P(X_{ij} = 0 | D_i, D_j) = 1 - \gamma D_i D_j$.

To directly compute the occurrence probability of an induced motif under the EDD model we have to consider both the present and the absent edges in the motif. More precisely, the probability is a product of edge probabilities times the probability of absent edges. For example, the occurrence probability of the induced 4-nodes square graph (Fig. 4) can be expressed as:

$$P\ \{\exists(AB, BC, CD, DA), \nexists(AC, BD) | D_A, D_B, D_C, D_D\} = \gamma^4 D_A^2 D_B^2 D_C^2 D_D^2 - \\ -\gamma^5 D_A^3 D_B^2 D_C^3 D_D^2 - \gamma^5 D_A^2 D_B^3 D_C^2 D_D^3 + \gamma^6 D_A^3 D_B^3 D_C^3 D_D^3$$

Each term of the summation corresponds to the occurrence probability of a non-induced motif of the same size and the sign of each term depends on the number of edges of the corresponding motif. Furthermore, the sign in the summation switches when we pass from a motif with $e$ edges to a motif with $e + 1$ edges.

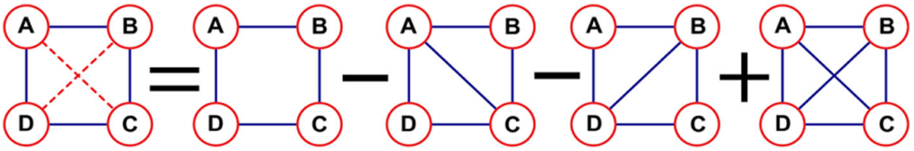


**Fig. 4** Occurrence probability of the induced 4-nodes square motif. The probability is given by a linear combination of occurrence probabilities of non-induced motifs having the same number of nodes. Dashed red edges are those that are not present in the 4-nodes square motif, while blue lines identify edges that are present in the induced motif

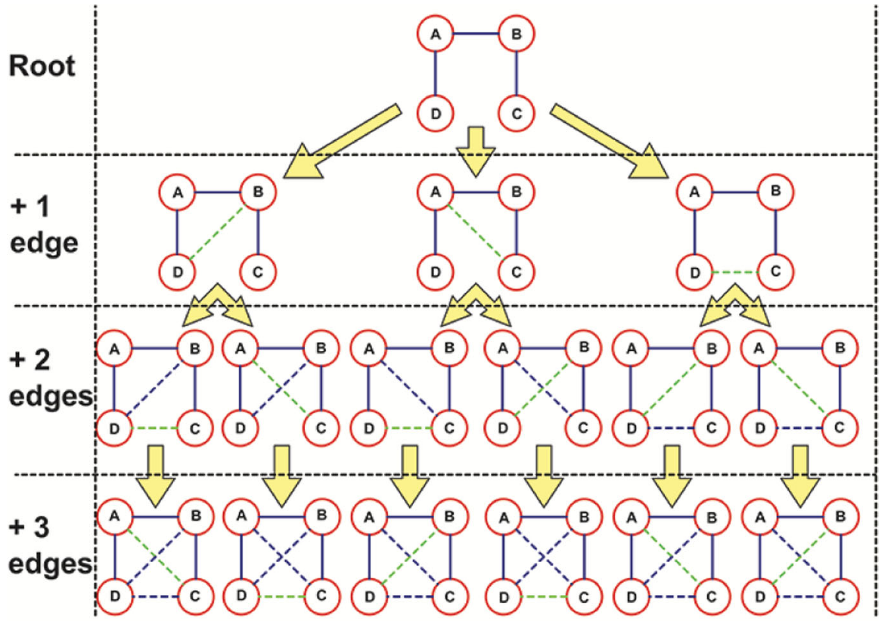


**Fig. 5** The AS of the 4-nodes path. The root is the 4-nodes path. Inner vertices of the AS store all subgraphs that can be obtained from the 4-nodes path by inserting 1,2 or 3 edges. Dashed green edges are the newly added one, while dashed blue edges are those that have been previously added

### Additive Set: an effective data structure for induced probability computation

We introduce a new data structure called Additive Set (AS) that can be used to identify the linear combination of motifs that yields the occurrence probability of an induced motif. The AS is a Directed Acyclic Graph (DAG) where each vertex[1] (except the roots, i.e. vertices with in-degree = 0) corresponds to a subgraph with $k$ nodes that can be obtained from a subgraph of the same size by adding an edge.

An Additive Set is characterized by the following properties:

- Each vertex contains a subgraph with $k$ nodes;
- Vertices at a given level of the AS represent subgraphs with the same number of edges;
- An internal vertex of the AS at level $L$ contains a subgraph obtained from a subgraph at level $L-1$ by adding exactly one edge;
- Levels range from 0 to $r$, where $r$ is the largest number of edges which can be added starting from subgraphs at the root vertices.

An example of Additive Set is shown in Fig. 5. The AS has one root vertex representing the 4-nodes path. Notice that the AS may contain isomorphic subgraphs or the same subgraph multiple times.

A compressed representation of the Additive Set can be built by collapsing all vertices containing automorphic subgraphs into a unique vertex. In the compressed AS, edges are weighted by the number of non-isomorphic occurrences of a subgraph that can be obtained from another subgraph by adding exactly one edge in all possible ways. Figure 6 shows an example of compressed AS containing all 6 non-isomorphic subgraphs with 4 nodes.

[1] To avoid confusion, we use the term "vertex" to indicate the nodes of the AS and the term "node" to denote the nodes of a generic graph or network.

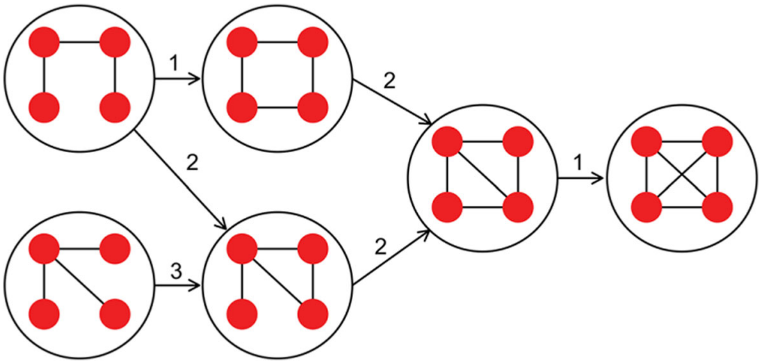


**Fig. 6** Compressed AS containing all the possible 4-nodes non-isomorphic subgraphs. Edges are weighted by the number of occurrences of a subgraph that can be obtained from another subgraph by adding one edge in all possible ways

Starting from a compressed Additive Set $\mathcal{T}$ we introduce the Topology Induced Additive Set (TIAS) of a subgraph $m$. The TIAS of $m$ is a sub-DAG of $\mathcal{T}$ containing $m$ and all subgraphs that can be built from $m$ by adding one or more edges. Figure 7 shows the TIAS of the 4-nodes path extracted from the compressed AS of Fig. 6.

The TIAS provides a simple way to compute the number $N_D$ of non-redundant occurrences of a subgraph $m'$ that can be generated from another subgraph $m$ by adding edges in all possible ways. Consider the TIAS $\mathcal{T}'$ of $m$ and let $P$ a path from $m$ to $m'$ in $\mathcal{T}'$. We first define the weight of $P$, $w(P)$, as the product of the weights of all edges in $P$. By summing the weights of all paths from $m$ to $m'$ in $\mathcal{T}'$, we obtain the number $N$ of all (possibly redundant) occurrences of $m'$ generated from $m$. Finally, the number of non-redundant occurrences of $m'$ is given by $N_D = N/l!$, where $l$ is the level of $m'$ in $\mathcal{T}'$. Figure 7 shows an example of computation of $N_D$.

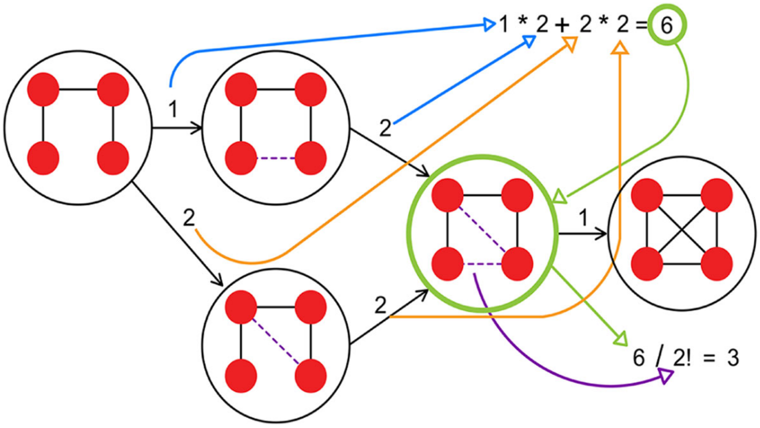


**Fig. 7** TIAS $\mathcal{T}'$ of the 4-nodes path $m$ with an example of computation of the number of non-redundant occurrences of the 4-nodes square with diagonal $m'$ (the subgraph surrounded by a green circle) generated from $m$. There are two paths $P_1$ and $P_2$ between $m$ and $m'$. $P_1$ has weight 2, equal to the product of the weights of its edges (1 and 2). Likewise, $P_2$ has weight 4, given by the product of the weights of its edges (2 and 2). The total number of occurrences of $m'$ generated from $m$ is the sum of the weights of $P_1$ and $P_2$, i.e. 6. The final number of non-redundant occurrences of $m'$ generated from $m$ is 6/2!=3, where 2 is the level of $m'$ in $\mathcal{T}'$

The numbers of non-redundant occurrences of each subgraph in $\mathcal{T}'$ represent the coefficients of the linear combination of non-induced occurrence probabilities yielding the induced occurrence probability of $m$ (see "Direct estimation of occurrence probability of induced motifs" section). The sign of each term in the combination depends on the level of the corresponding subgraph in $\mathcal{T}'$: if the level is even the sign is positive otherwise the sign is negative.

Algorithm 1 summarizes all the steps required to compute the occurrence probability of an induced motif $m$ with $k$ nodes, starting from the AS containing all subgraphs of size $k$.

**Algorithm 1:** Induced Probability

```
InducedProbability($m$);
Input: A motif $m$ with size $k$;
Output: Induced occurrence probability of $m$;
Build the complete AS $\mathcal{T}$ containing all subgraphs with $k$ nodes;
Extract the TIAS of $m$, $\mathcal{T}'$, from $\mathcal{T}$;
Starting from root $m$ in $\mathcal{T}'$ perform a Breadth First Search and for each visited subgraph $m'$ compute the number of non-redundant occurrences of $m'$ that can be obtained from $m$ and multiply it by the non-induced probability of $m'$;
Sum the terms computed in the previous step. The sign of the term relative to a subgraph $m'$ in the summation depends on the level of $m'$ in $\mathcal{T}'$: if the level is even the sign is positive otherwise the sign is negative;
```

Then, the induced occurrence probability of $m$ is given by:

$$\mu_I(m) = \sum_{i=0}^{L_{max}} \frac{(-1)^i}{i!} \left[ \sum_{m \in \mathcal{M}(i)} \beta(m,i)\mu(m) \right]$$

where $L_{max}$ is the maximum level get in the TIAS of $m$, $\mathcal{T}'$ and $\mathcal{M}(i)$ is the set of subgraphs stored of $\mathcal{T}'$ at level $i$. $\beta(m,i)$ is the following recursive function:

$$\beta(m,i) = \begin{cases} 1 & \textit{if } i = 0 \\ \sum_{v \in \mathcal{M}(i-1)} w(v,m)\beta(v,i-1) & \textit{otherwise} \end{cases}$$

with $w(v,m)$ defined as the weight of the edge $(v,m)$ in $\mathcal{T}'$.

### Mean and variance of induced motifs

Starting from the induced occurrence probability of a motif $m$, the mean and the variance of the induced count of $m$ can be computed using equations similar to those presented in "Analytical model for the expectation of non-induced motifs" section. The mean $\mathbb{E}[N_I(m)]$ can be calculated using Eq. 3, by substituting $\mu(m)$ with $\mu_I(m)$:

$$\mathbb{E}[N_I(m)] = \binom{n}{k} \varrho(m)\mu_I(m) \tag{8}$$

Likewise, variance can be computed starting from Eq. 4 and replacing $\mu$ with $\mu_I$. The only important difference is the set of super-motifs resulting from the overlapping of two or more NRPs. To this aim, we first introduce a new overlapping operator, $\cap_s^I$, which takes as input an integer $s$ and two NRPs of $m$, $m'$ and $m''$, and optionally returns a super motif with $2k - s$ nodes, whose adjacency matrix can be written in a block-wise fashion as:

$$m^{'} \cap_s^I m^{''} = \begin{cases} \left(\begin{array}{c|c|c} m^{'}_{11} & m^{'}_{12} & 0 \\ \hline m^{'}_{21} & m^{'}_{22} \vee m^{''}_{11} & m^{''}_{12} \\ \hline 0 & m^{''}_{21} & m^{''}_{22} \end{array}\right) & \text{if } \sum_{i,j} m^{'}_{22_{i,j}} = \sum_{i,j} m^{''}_{11_{i,j}} \\ \emptyset & \text{otherwise} \end{cases} \tag{9}$$

where all blocks are defined as in the case of non-induced motifs.

Equation 9 implies that the overlapping operation for induced motifs actually results in a super-motif only iff all the edges in the overlapping region come from both $m'$ and $m''$. Figure 8 shows all super-motifs resulting from the overlap of two NRPs of an induced 3-nodes path with $s = 2$ nodes.

In order to correctly compute the variance, for each super-motif $S$ resulting from the application of the overlapping operation (Eq. 9), we also need to take into account the set of all motifs that can be built starting from $S$ by adding one or more edges between nodes that are outside the overlapping region in all possible ways. The latter set is called expansion set of $S$ and is denoted as $\mathcal{E}(S)$. Figure 9 shows an example of expansion set of a super-motif.

The expected square count $\mathbb{E}[N_I^2(m)]$ is then computed considering: (i) the induced occurrence probability of all possible pairs of non-overlapping NRPs, (ii) the induced occurrence probability of all super-motifs generated using the overlapping function $\cap_s^I$

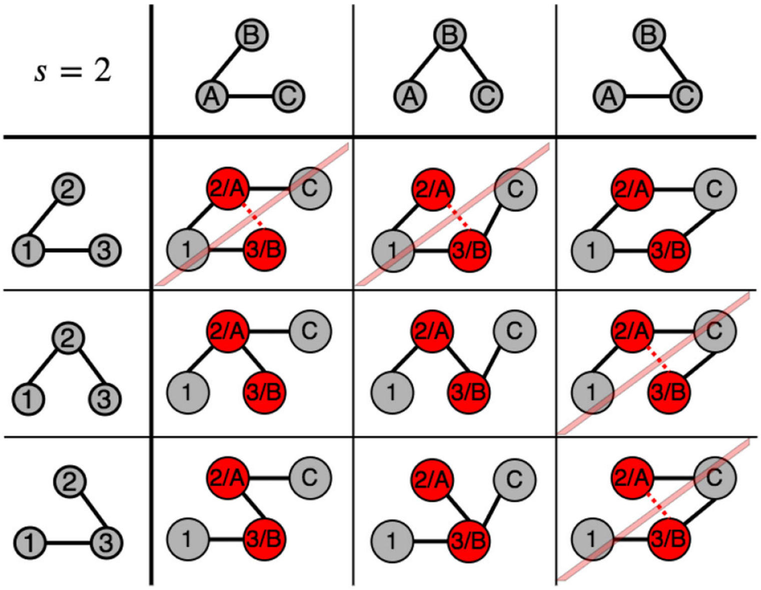


**Fig. 8** Super-motifs resulting from the overlap between all NRPs of the induced 3-nodes path with $s = 2$ nodes. Results are shown in a matrix fashion, where the entry at row $i$ and column $j$ is the super-motif originating from the NRP at the $i$-th row and the NRP at the $j$-th columnn. Nodes of each super-motif contain node ids of the starting NRPs. Nodes in the overlapping region are colored in red. Red dashed lines represent edges between overlapping nodes coming from only one of the two NRPs. All super-motifs containing red dashed edges are barred, meaning that they are not returned by the overlapping operator $\cap_s^I$

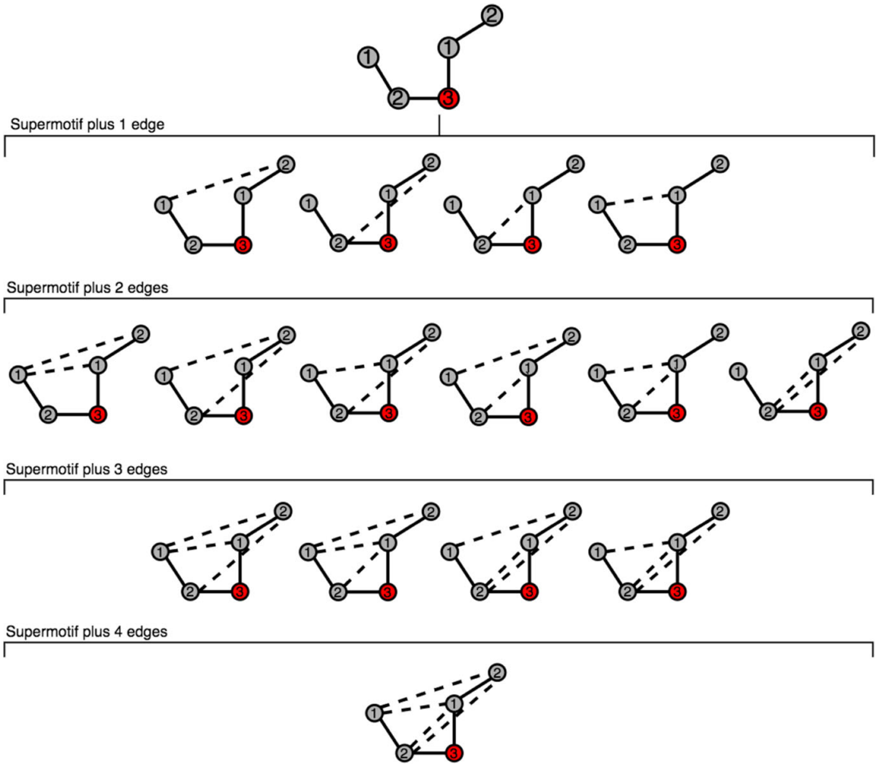


**Fig. 9** Expansion set of the 5-nodes path super-motif resulting from the overlap of two NRPs of the 3-nodes path. Starting from the top, the figure shows at each level *L* all super-motifs that can be built starting from all super-motifs at level *L* − 1, by adding exactly one edge between non-overlapping nodes in all possible ways

and resulting from the overlap of two NRPs with 1 or more nodes and (iii) the induced occurrence probability of all extensions of each super-motif generated from overlapping:

$$\mathbb{E}[N_I^2(m)] = \binom{n}{n-2k,k,k}\left[\sum_{m'\in R(m)} \mu(m')\right]^2 +$$

$$+\sum_{s=1}^{k}\left[\binom{n}{k-s,s,k-s,n-2k+s} \times \sum_{\substack{m',m''\in R(m),\\ S=\mu(m'\cap_s m'')\neq\emptyset}} \; \sum_{S_E\in\mathcal{E}(S)} \mu(S_E)\right] \qquad (10)$$

Finally, the variance $\mathbb{V}[N_I(m)]$ is given by $\mathbb{E}[N_I^2(m)] - \mathbb{E}[N_I(m)]^2$.

## RaME: rapid matrix elaboration

In this section we describe a matrix-based implementation of the proposed analytical model, called RaME (Rapid Matrix Elaboration). RaME uses efficient matrix operations to compute occurrence probabilities of induced motifs. The key idea behind RaME is that the computation of induced probabilities does not require the explicit generation of the Additive Set, but only some moments of the degree distribution of these motifs.

Let $m$ be a motif with $k$ nodes, $\mathcal{T}$ its TIAS and $E_A(m)$ the set of absent edges in $m$. Let $\mathcal{C}$ be the set of all $2^{|E_\mathcal{A}|} - 1$ possible edges combination in $E_A(m)$. Let $M_{\mathcal{CA}}$ be a matrix with $2^{|E_A|} - 1$ rows and $k$ columns, where each entry $M_{\mathcal{CA}}[c,x]$ stores the number of edges of combination $c$ to which node $x$ is incident. By adding to each cell $(i, j)$ of $M_{\mathcal{CA}}$ the degree of node $j$ in $m$, we get a new matrix $M_{\mathcal{D}}$ where $M_{\mathcal{D}}[m', u]$ stores the degree of node $u$ of a motif $m'$ in $\mathcal{T}$. Figure 10 shows an example of computation of $M_{\mathcal{CA}}$ and $M_{\mathcal{D}}$ for the 4-node path.

Starting from node degrees of each motif in $\mathcal{T}$, we can compute a matrix of moments $M_{\mathbb{E}}$ in which $M_{\mathbb{E}}[m', u]$ has the moment of order $M_{\mathcal{D}}[m', u]$ of the degree distribution of motif $m'$.

Finally, we can define two vectors, the gamma vector $V_\gamma$ and the sign vector $V_{\mathcal{S}}$. For each motif $m'$ in $\mathcal{T}$, $V_\gamma[m'] = \gamma^{m'_{++}}$, where $m'_{++}$ is the number of edges of $m'$, and $V_{\mathcal{S}}[m']$ is the sign of $m'$, with $V_{\mathcal{S}}[m'] = 1$ if the level of $m'$ in $\mathcal{T}$ is odd and $V_{\mathcal{S}}[m'] = -1$ if the level is even.

Given $M_{\mathbb{E}}$, $V_\gamma$ and $V_{\mathcal{S}}$, the non-induced probability of $m'$ can be computed as:

$$\mu(m') = V_{\mathcal{S}}[m']\, V_\gamma[m'] \prod_{u=1}^{k} M_{\mathbb{E}}[m', u] \tag{11}$$

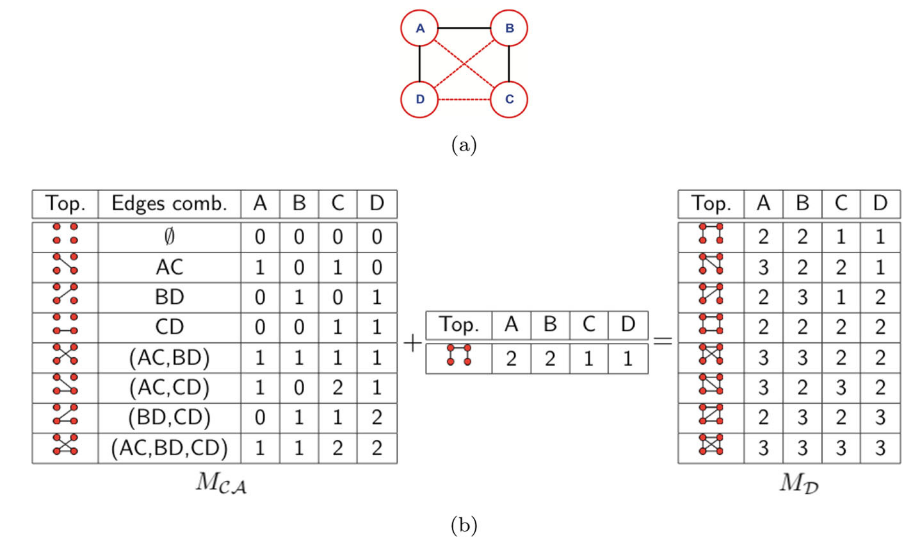
So, the non-induced probability computation can be reduced to a product of matrices and vectors. In Fig. 11 we show an example of non-induced probability computation for all motifs in the TIAS of the 4-node path of Fig. 10a.

Finally, the induced occurrence probability of $m$ is:

$$\mu_I(m) = \sum_{m' \in \mathcal{T}} \mu(m') \tag{12}$$

## Complexity analysis

In this section we analyze the time complexity of RaME and the Kocay Lemma-based method. Let $G$ be a network with $N$ nodes and let $m$ be a motif with $k$ nodes. For both algorithms we assume as computed the moments of the degree distribution of $G$.

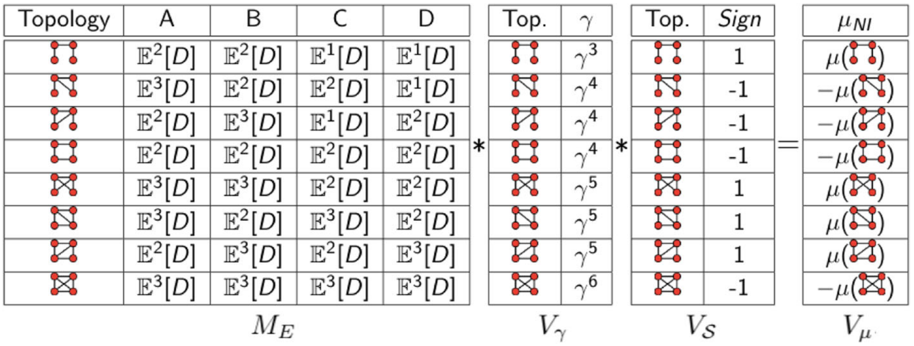


| Top. | Edges comb. | A | B | C | D |
|---|---|---|---|---|---|
| | ∅ | 0 | 0 | 0 | 0 |
| | AC | 1 | 0 | 1 | 0 |
| | BD | 0 | 1 | 0 | 1 |
| | CD | 0 | 0 | 1 | 1 |
| | (AC,BD) | 1 | 1 | 1 | 1 |
| | (AC,CD) | 1 | 0 | 2 | 1 |
| | (BD,CD) | 0 | 1 | 1 | 2 |
| | (AC,BD,CD) | 1 | 1 | 2 | 2 |

| Top. | A | B | C | D |
|---|---|---|---|---|
| | 2 | 2 | 1 | 1 |

| Top. | A | B | C | D |
|---|---|---|---|---|
| | 2 | 2 | 1 | 1 |
| | 3 | 2 | 2 | 1 |
| | 2 | 3 | 1 | 2 |
| | 2 | 2 | 2 | 2 |
| | 3 | 3 | 2 | 2 |
| | 3 | 2 | 3 | 2 |
| | 2 | 3 | 2 | 3 |
| | 3 | 3 | 3 | 3 |

**Fig. 10** **a** A 4-nodes path where present and absent edges are denoted with black and red lines, respectively. **b** From left to right: the matrix $M_{\mathcal{CA}}$ of combinations of absent edges of the 4-nodes path, where $M_{\mathcal{CA}}[c, x]$ is the number of edges in combination $c$ to which node $x$ is incident; node degrees of the 4-node path; the matrix $M_{\mathcal{D}}$ of node degrees of motifs in the TIAS of the 4-nodes path, resulted by adding node degrees of the 4-nodes path $(2, 2, 1, 1)$ to each row of $M_{\mathcal{CA}}$



| Topology | A | B | C | D | Top. | $\gamma$ | Top. | *Sign* | $\mu_{NI}$ |
|---|---|---|---|---|---|---|---|---|---|
| | $\mathbb{E}^2[D]$ | $\mathbb{E}^2[D]$ | $\mathbb{E}^1[D]$ | $\mathbb{E}^1[D]$ | | $\gamma^3$ | | 1 | $\mu()$ |
| | $\mathbb{E}^3[D]$ | $\mathbb{E}^2[D]$ | $\mathbb{E}^2[D]$ | $\mathbb{E}^1[D]$ | | $\gamma^4$ | | -1 | $-\mu()$ |
| | $\mathbb{E}^2[D]$ | $\mathbb{E}^3[D]$ | $\mathbb{E}^1[D]$ | $\mathbb{E}^2[D]$ | | $\gamma^4$ | | -1 | $-\mu()$ |
| | $\mathbb{E}^2[D]$ | $\mathbb{E}^2[D]$ | $\mathbb{E}^2[D]$ | $\mathbb{E}^2[D]$ | | $\gamma^4$ | | -1 | $-\mu()$ |
| | $\mathbb{E}^3[D]$ | $\mathbb{E}^3[D]$ | $\mathbb{E}^2[D]$ | $\mathbb{E}^2[D]$ | | $\gamma^5$ | | 1 | $\mu()$ |
| | $\mathbb{E}^3[D]$ | $\mathbb{E}^2[D]$ | $\mathbb{E}^3[D]$ | $\mathbb{E}^2[D]$ | | $\gamma^5$ | | 1 | $\mu()$ |
| | $\mathbb{E}^2[D]$ | $\mathbb{E}^3[D]$ | $\mathbb{E}^2[D]$ | $\mathbb{E}^3[D]$ | | $\gamma^5$ | | 1 | $\mu()$ |
| | $\mathbb{E}^3[D]$ | $\mathbb{E}^3[D]$ | $\mathbb{E}^3[D]$ | $\mathbb{E}^3[D]$ | | $\gamma^6$ | | -1 | $-\mu()$ |

**Fig. 11** The product between the matrix of moments $M_E$, where the order of moments are given by the corresponding entries in $M_{\mathcal{D}}$ (Fig. 10b), the gamma vector $V_\gamma$ of the powers of $\gamma$, where the exponents are the number of edges of each motif in the TIAS $\mathcal{T}$ of the 4-nodes path Fig. 10a, and the sign vector $V_{\mathcal{S}}$, containing the signs of all motifs in $\mathcal{T}$ gives as result the vector $V_\mu$ of non-induced occurrence probabilities of all motifs in $\mathcal{T}$

### Complexity of RaME

To evaluate the complexity of RaME for computing the mean and the variance of the induced count of a motif we consider the worst case, i.e. a motif with the minimum number of edges.

We start by analyzing the running time for calculating the induced occurrence probability of the motif. RaME requires $\mathcal{O}(k^2)$ to compute the set of absent edges $E_A(m)$, because we need to scan all the elements of the adjacency matrix of $m$. The number of absent edges in $m$ is given by $\alpha = \frac{k*(k-1)}{2} - |E(m)|$. The set $\mathcal{C}$ of all possible combination of absent edges can be generated in $\mathcal{O}(2^\alpha) \approx \mathcal{O}(2^{k^2})$ time, using a recursive algorithm. To calculate $M_{\mathcal{CA}}$ we need to check for each combination of edges $C \in \mathcal{C}$ which nodes are incident to the edges of $C$, so the computation of $M_{\mathcal{CA}}$ costs $\mathcal{O}(2^{k^2} * k)$. The number of rows of $M_{\mathcal{CA}}$ is $2^{k^2} - 1$, which corresponds to the number of motifs in the TIAS $\mathcal{T}$ of $m$ and the size of vectors $V_\gamma$ and $V_\mathcal{S}$. The matrix $M_\mathcal{D}$ of node degrees is obtained by adding degrees of $m$'s nodes to each row of $M_{\mathcal{CA}}$, so this step requires $\mathcal{O}(2^{k^2} * k)$. Computing the matrix $M_E$ of moments of the degree distribution costs $\mathcal{O}(2^{k^2} * k)$ too. The power of $\gamma$ and the sign for each motif in $\mathcal{T}$ can be computed in constant time, therefore $V_\gamma$ and $V_\mathcal{S}$ can be calculated in $\mathcal{O}(2^{k^2})$ time. Non-induced occurrence probabilities of motifs in $\mathcal{T}$ requires $\mathcal{O}(2^{k^2} * k)$ (Eq. 11). Finally, computing the induced probability of $m$ costs $\mathcal{O}(2^{k^2})$ (Eq. 12). Indeed, the overall complexity for the induced occurrence probability of $m$ is $\mathcal{O}(2^{k^2} * k)$.

Once we have the induced occurrence probability of $m$, the mean of the induced count only requires the computation of all NRPs of $m$ (Eq. 8). The latter task can be performed in $\mathcal{O}(k!^2)$ time, because we need to consider all $k!$ permutations of node indexes in $m$ and check each new permutation with the previously generated ones for redundancy. So, calculating the mean of the induced count requires $\mathcal{O}(2^{k^2} * k + k!^2))$, which can be rewritten as $\mathcal{O}(2^{k^2} * k))$.

Concerning the variance of the induced counts, we separately analyze the two terms in the summation of Eq. 10. Calculating all NRPs of $m$ requires $\mathcal{O}(k!^2)$ time and the number of NRPs is at most $k!$. So, the first term in the summation of Eq. 10 requires $\mathcal{O}(k!^2 + k! * 2^{k^2} * k)$, i.e. $\mathcal{O}(k! * 2^{k^2} * k)$ . Regarding the second term in the summation, we first can notice that a super-motif can have at most $2k - 1$ nodes. Computing the induced occurrence probability of a single super-motif $S$ requires $\mathcal{O}(2^{(2k)^2} * 2k)$, which corresponds to $\mathcal{O}(2^{k^2} * k)$. Calculating the expansion set of $S$, $\mathcal{E}(S)$, requires $\mathcal{O}(2^{2k^2})$, i.e. $\mathcal{O}(2^{k^2})$, because it is equivalent to calculating the set of all combination of absent edges in $S$. The number of motifs in $\mathcal{E}(S)$ is at most $2^{(2k)^2} - 1$, i.e. $\mathcal{O}(2^{k^2})$. Building a super-motif costs $O((2k)^2)$, i.e. $O(k^2)$. Therefore, computing the second term in the summation of Eq. 10 requires $\mathcal{O}(k * k!^2 *(k^2 + 2^{k^2} + 2^{k^2} * 2^{k^2} * k))$, which can be simplified as $\mathcal{O}(k^2 * k!^2 *(2^{k^2})^2)$. The latter is also the time complexity of the induced variance in RaME.

### Complexity of kocay lemma-based method

Kocay Lemma-based method relies on the computation of the inverse of the Kocay matrix. This requires first the calculation of the whole matrix. The coefficient for a pair of subgraphs $T'$ and $T''$ is given by the number of non induced occurrence of $T'$ in $T''$. This number can be computed by using any subgraph matching algorithm, whose complexity is $\mathcal{O}(k! *k)$ in the worst case. Since the maximum number of connected subgraphs with $k$ nodes is $2^{k^2}$, computing the inverse of Kocay matrix requires $\mathcal{O}((2^{k^2})^2 * k! *k)$. Calculating the non-induced occurrence probability of a motif requires $O(k)$ because it is just a product of $k$ moments of the degree distribution (Eq. 2). To compute the non-induced mean we also need to calculate all NRPs of $m$. So, calculating the non-induced mean of a single motif costs $\mathcal{O}(k!^2 +k)$, i.e. $\mathcal{O}(k!^2)$. The number of non-induced means needed for the linear combination giving the induced mean of $m$ is at most $2^{k^2}$. So, the complexity for the induced mean is $\mathcal{O}((2^{k^2})^2 * k! *k + 2^{k^2} * k!^2)$, i.e. $\mathcal{O}((2^{k^2})^2 * k! *k)$.

Regarding the induced variance, we first analyze the complexity of the non-induced variance of a motif, which is determined by the computation of the expectation of the squared count (Eq. 4).The first term of the summation of Eq. 4 requires $\mathcal{O}(k!^2 +k! *k)$, i.e. $\mathcal{O}(k!^2)$, while the second term requires $\mathcal{O}(k * k!^2 *(k^2 + 2k))$, i.e. $\mathcal{O}(k^3 * k!^2)$. Therefore, the time complexity for the non-induced variance is $\mathcal{O}(k^3 * k!^2)$. For the complexity of the induced variance (Eq. 6), we separately analyze the two terms in the summation. Calculation of the first term costs $\mathcal{O}(2^{k^2} * k^3 * k!^2)$. The second term of the summation implies the computation of covariance between all pairs of motifs in $\mathcal{T}$ (Eq. 7). For the covariance between two motifs $m'$ and $m''$ we first need to calculate all NRPs of $m'$ and $m''$. This step costs $\mathcal{O}(k!^2)$. The first term of the summation in Eq. 7 can be computed in $\mathcal{O}(k!^2 *k)$. The second term requires $\mathcal{O}(k * k!^2 *(k^2 + 2k))$. So, the calculation of the covariance costs $\mathcal{O}(k^3 * k!^2)$. Since the number of motifs in $\mathcal{T}$ is $2^{k^2} - 1$, computing the second term of the summation in Eq. 6 requires $\mathcal{O}(2^{k^2} * k^3 * k!^2)$. So, the time complexity of the induced variance using the Kocay Lemma-based method is $\mathcal{O}(2^{k^2} * k^3 * k!^2)$.

### Discussion

To sum up, the time complexities for the induced mean using RaME and Kocay Lemma-based method are $\mathcal{O}(2^{k^2} * k))$ and $\mathcal{O}((2^{k^2})^2 * k! *k)$, respectively. The time complexities for computing the induced variances using RaME and Kocay Lemma-based method are $\mathcal{O}(k^2 * k!^2 *(2^{k^2})^2)$ and $\mathcal{O}(2^{k^2} * k^3 * k!^2)$, respectively. Indeed we can conclude that RaME is faster than Kocay Lemma-based method for the calculation of the induced mean and slower for the computation of the induced variance. However, calculating the induced variance remains unfeasible for both algorithms.

**Table 1** Main features of the real network dataset

| Network | Category | Nodes | Edges |
|---|---|---|---|
| Human Protein | Metabolic | 3,133 | 6,726 |
| CAIDA | Computer | 26,475 | 53,381 |
| DBLP | Coauthorship | 317,080 | 1,049,866 |
| LiveJournal | Social | 5,204,176 | 49,174,464 |

## Experimental results

To evaluate the performance of RaME, we collected a dataset of real undirected networks of medium and large size and we compared RaME to the Kocay Lemma-based algorithm described in "Analytical model for the expectation of induced motifs" section, considering subgraphs of different sizes (up to 10 nodes). Performance of the two models have been also evaluated on a case study using a knowledge biological network. In our experimental analysis we focus on the running time for the calculation of the number of induced subgraph occurrences of a motif. Therefore we do not discuss on the importance of a specific motifs to understand the structure of the network. Since calculating variance of the induced count using both methods is computational intensive and therefore unfeasible in a single machine even for small motifs (from size 5 on), we just focused on the calculation of the mean of the induced counts. We implemented RaME in Java language and compared it to a Java implementation of the Kocay Lemma-based model. All experiments were executed on an Intel core i5-7400 processor with 8GB of RAM.

### Dataset

In our experimental analysis we used 4 real undirected networks extracted from KONECT[2]:

- ***Human Protein (Vidal)***: a Protein-Protein Interaction (PPI) network describing the physical interactions between proteins in human;
- ***CAIDA***: Internet communication network between Autonomous Systems (AS);
- ***DBLP***: a co-authorship graph, nodes are the authors and two authors are linked by an edge iff they co-authored at least one published paper;
- ***LiveJournal***: a social network of LiveJournal weblog service where nodes are users and edges are their relationship status.

In Table 1 we report the number of nodes and edges of each network.

### Experiments on the KONECT dataset

In Table 2 we report the running times (in seconds) of both algorithms for the computation of the mean count of all induced motifs of a given size in each real network of the KONECT dataset. Table 3 shows the sum of the expected induced mean for topologies with k-nodes for $k = 3, ..., 7$.

We also measured the execution time (Table 4), expected value (Table 5) and memory consumption (Table 6) of RaME for the computation of the mean induced count of a single motif, i.e. the star topology. Indeed, the $k$-star subgraph is one of the sparsest topologies

[2] http://konect.cc

**Table 2** Running times (seconds) of RaME and Kocay Lemma-based method for the computation of the induced mean of all motifs with 3-7 nodes

| Motif | DBLP | | CAIDA | | Human Protein | | LiveJournal | |
|---|---|---|---|---|---|---|---|---|
| size | *RaME* | *Kocay* | *RaME* | *Kocay* | *RaME* | *Kocay* | *RaME* | *Kocay* |
| 3 | **1.01** | 1.28 | **0.10** | 0.22 | **0.02** | 0.12 | 88.23 | **86.21** |
| 4 | **1.03** | 1.30 | **0.10** | 0.25 | **0.02** | 0.13 | 90.65 | **89.35** |
| 5 | **1.05** | 1.37 | **0.12** | 0.32 | **0.02** | 0.20 | 93.43 | **92.72** |
| 6 | **1.17** | 2.70 | **0.25** | 1.65 | **0.21** | 1.75 | **94.67** | 97.28 |
| 7 | **52.08** | 113.25 | **59.73** | 118.92 | **48.14** | 117.44 | **160.56** | 245.07 |

with $k$ nodes and its TIAS contains the highest number of $k$-subgraphs. Therefore, it represents one of the worst-case scenarios for our method because the execution time for calculating the induced mean count of any motif with $k$ nodes in RaME will be less than or equal to the one of the $k$-star topology.

Concerning Kocay Lemma-based method we only report execution times for all $k$-node motifs (Table 2) since even for a single motif the algorithm has to compute the entire Kocay matrix, its inverse matrix and the non-induced occurrence probabilities of all $k$-node motifs.

Results clearly show that RaME is faster than Kocay Lemma-based method for larger motifs, because calculating the Kocay matrix becomes computationally harder as the motif size increases. Moreover, the size of the network has a negligible impact on the execution time of RaME. In fact, the only information needed by RaMe about the input network is the node degree distribution. Indeed, the small overhead observed in the running time only depends on the computation of the moments of the distribution.

The number of possible subgraphs of size $k$, for $k = 8, 9, 10$ cause a combinatorial explosion in the computation, therefore we just report the mean of the induced count of the star topology (Table 4). In these cases RaME was able to compute induced means in reasonable running time (at most few hours for motifs of size 10) in all networks, while Kocay Lemma-based method didn't manage to complete the task or ran out of memory.

## A case study: Hetionet knowledge graph

In this section we present a case study to demonstrate the applicability of RaME and its ability to find interesting motifs of various size. To this aim we consider the Hetionet knowledge network.

Hetionet[3] is an integrative network of biomedical knowledge that brings together biological data collected in the last 50 years from 29 different databases. It is an heterogeneous information network formed by different type of nodes (such as for example genes, compounds, anatomy, diseases and ontologies) and different types of edges. Figure 12a illustrates the metagraph diagram of Hetionet.

Hetionet contains 47,031 nodes of 11 different types and 2,250,197 edges of 24 different types. For our case study we extracted from Hetionet the maximum connected component of a tripartite network, where nodes are compounds, genes and diseases. The metagraph diagram of the extracted tripartite network is shown in Fig. 12b. The resulting network has 5,485 nodes and 32,210 edges. Since our algorithm works on undirected

[3] http://het.io

**Table 3** Expected number of occurrences for induced counts of all topologies with 3-7 nodes

| Motif size | DBLP | CAIDA | Human Protein | LiveJournal |
|---|---|---|---|---|
| 3 | 2.2603e7 | 3.5139e3 | 3.464e3 | 1.4311e9 |
| 4 | 9.7480e8 | 4.5469e4 | 4.0749e4 | 4.8060e11 |
| 5 | 5.4184e10 | 6.1725e5 | 5.1009e5 | 4.0914e14 |
| 6 | 3.5178e12 | 8.7608e6 | 6.6951e6 | 4.2887e17 |
| 7 | 2.5141e13 | 1.2601e8 | 8.8596e7 | 4.4733e20 |

networks, we treated the graph as undirected. However, this does not bring to an information loss since all edges of the extracted tripartite network always go from one type of node to another one.

In order to test RaME we built five undirected motif graphs of variable size, depicted in Fig. 13. Motifs in Fig. 13a and b consist of a compound and a disease having one and two genes in common, respectively. Motifs in Fig. 13c, d and e represent two sets of genes $G_1$ and $G_2$, each formed by one, two and three genes, respectively. The two sets are connected to unrelated compounds $c_1$ and $c_2$ and unrelated diseases $d_1$ and $d_2$.

Table 7 shows the expected number of induced occurrences found using RaME ($E[N_I(m)]$) and the running time for each query ($T_{E[N_I(m)]}$). We also report the number of induced occurrences ($N_{obs}$) in the input network returned by the GTrie algorithm (Ribeiro and Silva 2014) and its execution time $T_{N_{obs}}$.

As expected, the mean number of induced occurrences according to the EDD model for the smallest motif (Fig. 13a) is high compared to the input network. Even though we don't have the variance and we cannot calculate an exact p-value, we can state that this motif is weakly significant. Motif in Fig. 13b seems to be slightly more significant. For the biggest motifs GTrie failed to return the exact occurrence count. This shows that for some instances subgraph isomorphism results computationally expensive and therefore establish motif significance using permutation test becomes unfeasible. In this regard, analytical methods like RaME can be useful to provide an approximate estimation of the frequency of a subgraph in a graph when the exact subgraph counting becomes computationally unfeasible. Nevertheless, the computation of the variance results still intractable when motif size increases, also with the proposed analytical model. Interestingly, we

**Table 4** Running time (seconds) of RaME for the computation of the mean of induced counts of star topologies with 3-10 nodes

| Star size | DBLP | CAIDA | Human Protein | LiveJournal |
|---|---|---|---|---|
| 3 | 0.87 | 0.04 | 0.01 | 73.46 |
| 4 | 0.87 | 0.05 | 0.01 | 75.81 |
| 5 | 0.88 | 0.05 | 0.02 | 78.62 |
| 6 | 0.93 | 0.05 | 0.03 | 79.31 |
| 7 | 1.15 | 0.07 | 0.06 | 80.36 |
| 8 | 1.62 | 0.43 | 0.50 | 81.54 |
| 9 | 36.38 | 33.52 | 33.87 | 134.60 |
| 10 | 9’009.11 | 8’861,13 | 9’517.49 | 13’152.37 |

**Table 5** Expected number of occurrences with RaME for induced counts of star topologies with 3-10 nodes

| Star size | DBLP | CAIDA | Human Protein | LiveJournal |
|---|---|---|---|---|
| 3 | 2.2601e7 | 3.501e3 | 3.430e3 | 1.4311e9 |
| 4 | 4.5240e8 | 2.7675e4 | 1.8495e4 | 3.8289e11 |
| 5 | 1.2185e10 | 2.0662e5 | 1.0230e5 | 3.2520e14 |
| 6 | 3.7174e11 | 1.3393e6 | 5.1427e5 | 3.1559e17 |
| 7 | 1.1949e13 | 7.5664e6 | 2.2914e6 | 2.8696e20 |
| 8 | 3.8911e14 | 3.7725e7 | 9.0298e6 | 2.3638e23 |
| 9 | 1.2470e16 | 1.6790e8 | 3.1619e7 | 1.7567e26 |
| 10 | 3.8508e17 | 6.7337e8 | 9.9062e7 | 1.1821e29 |

found that the induced mean of the biggest motifs considerably decreases as the motif size increases. This seems to agree with the fact that, the more genes in common two diseases have, the more likely they are similar to each other.

## Conclusions

In this paper we introduced a novel algorithm to compute the expected count of induced motifs in undirected large networks under any exchangeable random model. We also described a matrix-based implementation of our algorithm, called RaME. Compared to the method based on the application of Kocay Lemma and illustrated in (Picard et al. 2008), RaME becomes faster when the size of the motif increases. Our method can be applied to find undirected motifs up to 10 nodes in reasonable time. Its running time results almost independent from the size of input network and can be easily extended to directed networks. RaME shows also a reasonable memory consumption, as reported in Table 6. Most of the memory is used to store the input graph and the set of non redundant permutation topologies drawn from a query graph. In addition, the case study presented here shows that analytical models like RaME can be useful to roughly estimate the number of occurrences of medium and large motifs, whenever subgraph counting algorithms are unable to return the answer in a limited amount of time. We also investigated the second moment calculation with our model. Howerver, currently it results unfeasibile for RaME even for small size subgraphs. We plan to implement RaME on top of SPARK framework to deal with networks having billions of nodes and investigate approximated variants to calculate expectations and variance of larger motifs.

**Table 6** Physical memory consumption (GB) of RaME for the computation of the mean of induced counts of star topologies with 3-10 nodes

| Star size | DBLP | CAIDA | Human Protein | LiveJournal |
|---|---|---|---|---|
| 3 | 0.3 | 0.1 | 0.01 | 5.1 |
| 4 | 0.3 | 0.1 | 0.01 | 5.1 |
| 5 | 0.3 | 0.1 | 0.01 | 5.1 |
| 6 | 0.3 | 0.1 | 0.01 | 5.1 |
| 7 | 0.3 | 0.1 | 0.1 | 5.1 |
| 8 | 0.4 | 0.3 | 0.2 | 5.2 |
| 9 | 0.5 | 0.5 | 0.3 | 5.3 |
| 10 | 1.9 | 1.8 | 1.8 | 6.5 |

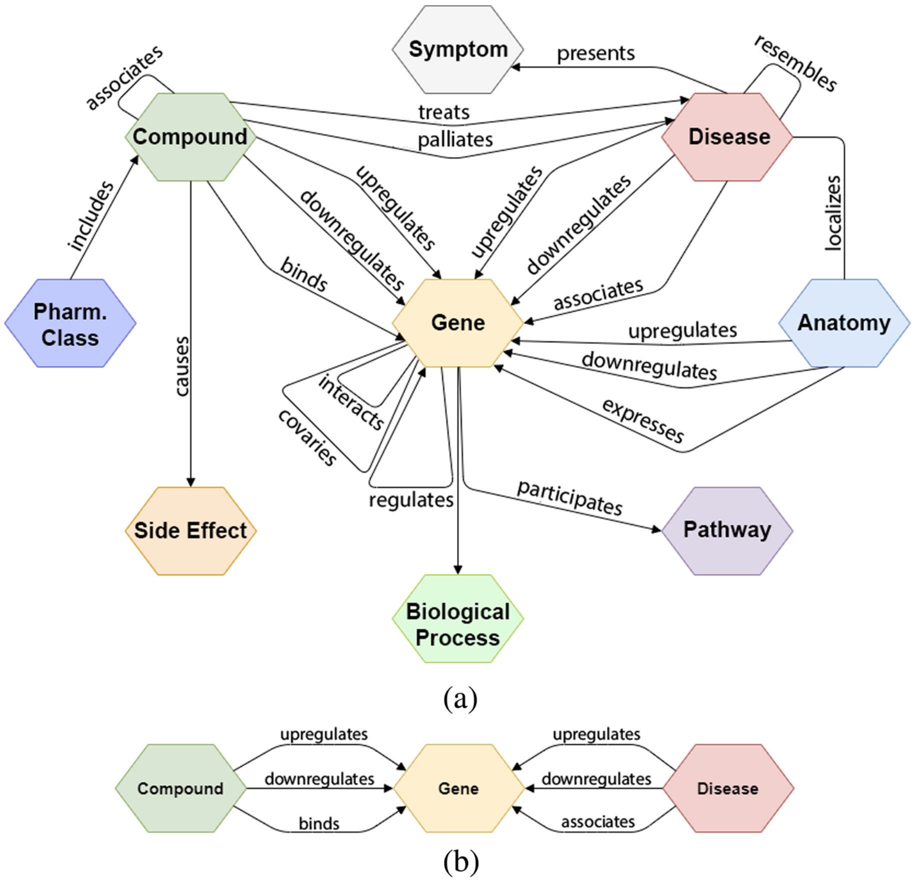


**Fig. 12** **a** A metagraph diagram representing all types of nodes and edges of Hetionet knowledge network. **b** The metagraph diagram of the case study network extracted from Hetionet

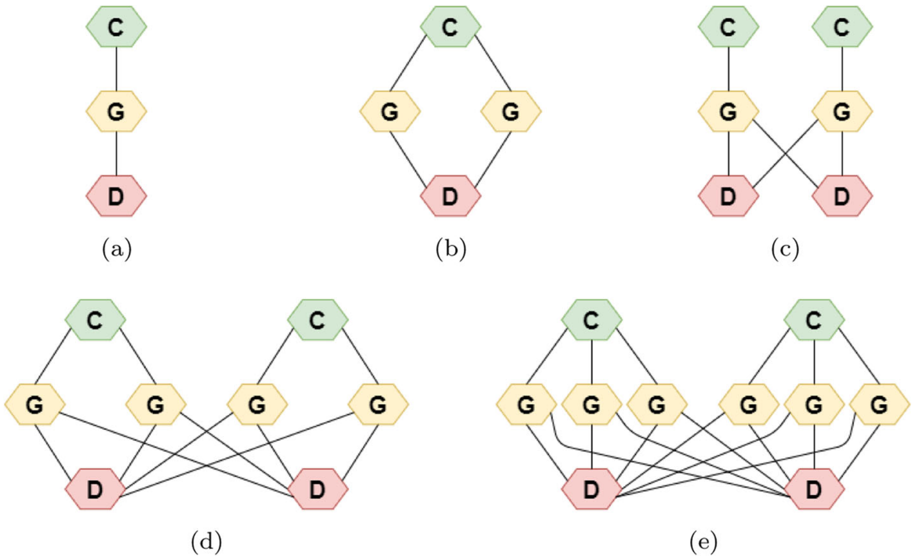
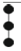
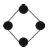
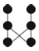
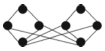
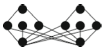


**Fig. 13** Motif graphs of the Hetionet network with 3 nodes (**a**), 4 nodes (**b**), 6 nodes (**c**), 8 nodes (**d**) and 10 nodes (**e**). Node names are *C* for Compounds, *G* for Genes and *D* for Diseases

**Table 7** Induced means $E[N_I(m)]$ according to the EDD model and running times $T_{E[N_I(m)]}$ (in seconds) for the five motif graphs of Fig. 13 using RaME

| Topology | $N_{obs}$ | $T_{N_{obs}}$ | $E[N_I(m)]$ | $T_{E[N_I(m)]}$ |
|---|---|---|---|---|
|  | 5’045’184 | 0.48 | 171’102.35 | 0.21 |
|  | 6’862’764 | 3.6 | 24’376.51 | 0.20 |
|  | - | > 1 day | 889’928’264’336.37 | 0.25 |
|  | - | > 1 day | 76’665.85 | 1.16 |
|  | - | > 1 day | 174.63 | 3’031.21 |

The table also reports the number of occurrences $N_{obs}$ in the input network using GTrie algorithm and the relative execution time $T_{N_{obs}}$ (in seconds)

### Abbreviations

ER: Erdös-Renyi model; FDD: Fixed Degree Distribution model; EDD: Expected Degree Distribution model; ERMG: Erdös-Renyi Mixture for Graphs; NP: Nondeterministic Polynomial; RaME: Rapid Matrix Elaboration algorithm; NRP: Non-Redundant Permutation; AS: Additive Set; DAG: Directed Acyclic Graph; TIAS: Topology Induced Additive Set

### Acknowledgments

Not applicable.

### Authors’ contributions

GM, AF and AP conceived the project. EM, GM and AP developed the analytical model for computing the mean and variance of the count of induced motifs. EM implemented the analytical model and RaME algorithm. EM performed the experiments and the case study. EM, GM and AP wrote the paper. All authors read and approve the final manuscript.

### Funding

GM, AF and AP have been supported by Ministero dell’Istruzione, dell’Università e della Ricerca Grant Number: SCN_00451 “CLARA - CLoud plAtform and smart underground imaging for natural Risk Assessment”.

### Availability of data and materials

Networks of KONECT dataset are publicly available at http://konect.cc/. Hetionet network can be downloaded from https://github.com/hetio/hetionet. Java implementations of RaME and Kocay Lemma-based algorithms are available at https://martorana.email/RaME/.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]University of Catania, Dept. of Physics and Astronomy, Catania, Italy. [2]University of Catania, Dept. of Clinical and Experimental Medicine, Catania, Italy.

Received: 23 March 2020 Accepted: 29 July 2020

Published online: 01 September 2020

### References

Chen J, Yuan B (2006) Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics 22(18):2283–2290

Chung F, Lu L (2002) The average distances in random graphs with given expected degrees. Proc Natl Acad Sci 99(25):15879–15882

Cook SA (1971) The complexity of theorem-proving procedures. In: Proc. 3rd ACM Symposium on Theory of Computing. pp 151–158

Daudin JJ, Picard F, Robin S (2008) A mixture model for random graphs. Stat Comput 18(2):173–183

Erdös P, Renyi A (1959) On random graphs. Publ Math 6:290–297

Johnson NL, Kotz S, Kemp AW (1992) Univariate discrete distributions. Wiley

Kocay W (1981) An extension of Kelly’s lemma to spanning subgraphs. Congr Num 31:109–120

Martorana E, Micale G, Ferro A, Pulvirenti A (2020) Establish the Expected Number of Injective Motifs on Unlabeled Graphs Through Analytical Models, Complex Networks and Their Applications VIII. Springer. pp 255–267

Micale G, Giugno R, Ferro A, Mongiovì M, Shasha D, Pulvirenti A (2018) Fast analytical methods for finding significant labeled graph motifs. Data Min Knowl Disc 32(2):1–28

Micale G, Pulvirenti A, Ferro A, Giugno R, Shasha D (2019) Fast methods for finding significant motifs on labelled multi-relational networks. J Compl Netw 00:1–22

Milo R, Kashtan N, Itzkovitz S, et al. (2004) On the uniform generation of random graphs with prescibed degree sequences. Cond Mat 0312028:1–4

Milo R, Shen-Orr S, Itzkovitz S, et al. (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827
Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 026118:64
Nowicki K, Snijders T (2001) Estimation and prediction for stochastic block structures. J Am Stat Assoc 96:1077–1087
Park J, Newman M (2003) The origin of degree correlations in the internet and other networks. Phys Rev E 68:026112
Picard F, Daudin JJ, Koskas M, et al. (2008) Assessing the exceptionality of network motifs. J Comput Biol 15(1):1–20
Prill R, Iglesias PA, Levchenko A (2005) Dynamic properties of network motifs contribute to biological network organization, Vol. 3
Ribeiro P, Silva S (2014) G-Tries: a data structure for storing and finding subgraphs. Data Min Knowl Disc 28(2):337–377
Shen-Orr SS, Milo R, Mangan S, et al. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31:64–68
Squartini T, Garlaschelli D (2011) Analytical maximum-likelihood method to detect patterns in real networks. New J Phys 13(8):083001
Wernicke S (2006) Efficient detection of network motifs. IEEE/ACM Trans Comput Biol Bioinforma 3(4):347–359

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.