Analytical Methods for Finding Significant Colored Graph Motifs

***Abstract*—**

I. Introduction

II. Related works

The original definition of network motif regards unlabeled patterns of interconnections that arise unexpectedly often in a network, named *topological uncolored motifs*. The basic idea is that subgraphs with the same topology might be functionally similar. Motifs may correspond to conserved patterns that are linked to important cellular functions.

Given a topological pattern m on an input network G, the common approach to determining whether m is a motif consists of following steps: (i) generate a large set of random networks sharing the characteristics (same number of nodes and edges with perhaps more constraints) of N; (ii) find the number of occurrences of m in each of those networks; (iii) their average is the expected count of m.

The first step creates random networks under a specified random reference model that have the same number of nodes and edges of the real network. Examples of reference models include:

- The Erdös-Renyi model (ER model) [] in which the probability of connecting two nodes n_1 and n_2 is the same as the probability of connecting any other two nodes n_3 and n_4, where the probability depends on the network density of G.
- The Fixed degree distribution model (FDD model) [], where the target graph is generated by swapping edges starting from the input network G, implying that each node n in each random graph R has the same degree as n did in G.
- The Expected degree distribution model (EDD model) [] which generates graphs whose node degrees have the same expectation as the input network G.
- The Erdos-Renyi mixture for graphs model (ERMG model) [] which is based on mixture population edges and is used to model heterogeneous connectivity.

To find all motifs, algorithms generate candidates by searching for all subgraphs having k nodes in the input network G and in a set of random variations of G [**?**].

Such a method however presents important computational drowbacks mainly related to the generation of a large number of networks and the application of subgraph isomorphism algorithms to compute the number of occurrences.

The simulation based approach described above yields a measure of the significance of each candidate through the computation of a p-value using a resampling approach [**?**], [**?**], [**?**], [**?**]. Unfortunately, this method requires a large number of random graphs whose analysis turns out to be computationally expensive (far more expensive than analyzing the target network alone).

Over the last decades, researchers have worked on replacing the simulation by analytical methods. For uncolored motifs, approximation methods, based on the Erdös-Renyi (ED) model, have tried to compute the asymptotic normality of the distribution of topology counts [**?**]. Empirically, the Erdos-Renyi random model offers a poor fit for many real-world networks [**?**].

More recently, Picard et al [**?**] proposed a model to exactly compute the mean and variance of the count of a given pattern under any exchangeable random graph model. Exchangeability means that the probability of occurrence of a topology does not depend on its position in the graph (i.e., on the structure of the neighborhood of the pattern). The authors make use of the Pólya-Aeppli distribution (also known as the Poisson Geometric distribution which is a special case of the Poisson-Compound distribution). The Pólya-Aeppli distribution supposes that objects (which are to be counted) occur in clusters, the number of clusters follow a Poisson distribution, while the number of objects per cluster has a geometric distribution[**?**]. This is the case when distinct topologies can share nodes and edges (i.e. clumps) [**?**]. In fact, the authors show that when the number of clumps has a Poisson distribution with mean λ and the sizes of the clumps are independent of each other and have a Geometric distribution $G(1-a)$, the number of observed events X (topologies) has a distribution $P(\lambda, a)$. These results lead to an estimate of the count of occurrences of a given topology. Picard et al [] show that is a good model for the distribution of the counts of subgraph topologies (both induced and non-induced), yielding a more accurate p-value than a Gaussian model for the graphs of many applications.

A. Different characterization of motifs

Focusing only on topologies ignores the meaning of the nodes. Such meaning can be important. For example, in a protein-protein interaction network, topologies having to do with metabolism may be different than topologies having to do with meiosis. We name motifs where the nature of (i.e. information carried on) nodes matters, *colored motifs*. However, to deal with colored motifs we need to generalize its definition according to the constraints that can be defined either on the topology or on the color labels assignment. Thus, we introduce three different definition of motifs which are hierarchically related (see Fig **??**).

In their seminal work, Schabat et al [**?**] define a motif as any connected topology of k nodes having a given multiset of colors M, denoted *multiset colored motif*. For example, a

connected topology consisting of five nodes having three reds and two blues. In that example, five would be the size of the motif. An occurrence of a motif is defined as a connected subgraph whose labels match the motif (see Fig **??**).

The authors [**?**] propose an approach for assessing the exceptionality of this kind of colored motifs which does not require simulations. They established an exact analytical model for the mean and the variance of the count of a colored motif using the Erdös-Rényi (ER) random graph model. In doing so, they assumed that the color assignment to nodes is independent from the topology of the network, and therefore modeled the probability of a multiset of colors as a multinomial distribution. To estimate a p-value associated to the motif, the authors also modeled the complete distribution of the count of a colored motif in an Erdös-Renyi random graph model by making use again of the Pólya-Aeppli distribution.

In some applications, we are interested in both topology and colors (e.g. carbon rings in chemical networks or feed-forward networks of different gene types in protein-protein networks). For that reason, we propose a second definition of motif consisting of a subgraph of k nodes with a given topology having nodes belonging to a multiset of colors M, denoted *topological multiset colored motif*.

A third interesting definition defines motifs consisting of a topology and a specific color assignment to each node in the topology. For example, a star topology of seven nodes in which the center node is blue and the other nodes are red. So, in this case a motif is subgraph of k nodes having fixed colors connected through a given topology, denoted *topological colored motif*.

B. Our view of motifs

In this paper we deal with the last two definitions of motifs, *topological multiset colored motif* and *topological colored motif* for which no analytical model has yet been proposed. Inspired by the work of [**?**], [**?**] we introduce analytical models to establish the significance of colored motifs on directed and undirected graphs, under the EDD random model, and in which colors are either independent or dependent on the degrees of nodes.

III. Definitions

A *colored graph* $G(V, E, C, c)$ is a graph where V is the set of nodes, $E \subseteq (V \times V)$ is the set of edges, C is a set of colors and $c : V \longrightarrow C$ is a function that assigns a color to each node in V. If $(u, v) \in E$, we say that v is a neighbor of u. G is undirected iff $\forall (u, v) \in E$, then $(v, u) \in E$, i.e. u is a neighbor of v and vice versa. If colors are not taken into consideration that $G(V, E)$ is called unlabeled graph.

Intuitively, given a graph G, a topology that occurs "unusually" frequently in G is called a *motif*. The number of occurences of a motif counts only non reduntant occurrences. A motif occurence is redundant if it is an automorphism of anoter occurence. Given a graph $G = (V, E)$ a permutation ξ of the vertex set V, is an authormophism if for all the pair of vertices $u, v \in V$ we have $(u, v) \in E \iff (\xi(u), \xi(v)) \in E$.

To establish the significance of the motifs, the target graph is suppose to be drawn from a set of graphs belonging to a random graph model. Random graphs models allows to generate graphs preserving a certaing carachtersitics. An important properti of random graph model is exchengability. Given two random graphs G^1 and G^2 under a random model R_G. We say that R_G is an exchangable random model when the two different random variables X_1 and X_2 denoting two random edge assignments to G_1 and G_2 have the same distribution.

We define two types of motifs.

Definition 3.1: **(Motif Type I: topological multiset colored motif)** Let $G = (V, E, C, c)$ be a colored graph drawn from a distribution of graphs G^* under a given reference exchangeable random model R_G. Let $m(V_m, E_m, C_m)$ be a subgraph (induced or non-induced) of G where V_m and E_m are the set of motif nodes and motif edges and C_m is the multiset of node colors of the nodes V_m. Let $N_{obs}(m)$ be the number of isomorphic non-redundant occurrences of m in G having the same multiset of colors C_m, and let α be a critical value. We say that m is a motif of G if

$$P\left[N(m) \geq N_{obs}(m)\right] \leq \alpha$$

Where $N(m)$ is a random variable representing the number of non-reduntant occurrences of the motif under the reference model R_G.

Definition 3.2: **(Motif Type II: topological colored motif)** Let $G = (V, E, C, c)$ be a colored graph drawn from a distribution of graphs G^* under a given reference exchangeable random model R_G. Let $m(V_m, E_m, C_m)$ be a subgraph (induced or non-induced) of G where V_m is the sequence of k motif nodes with indexes $1, 2, \cdots, k$, E_m is the set of motif edges and $C_m = (c_1, c_2, \cdots, c_k)$ is an array of node colors, where c_i is the color of the i-th node, for $1 \leq i \leq k$. Let $N_{obs}(m)$ be the number of isomorphic (non redundant) occurrences of m in G (automorphisms of m w.r.t. both topology and colors are considered only once). Let α be a critical value. We say that m is a motif of G if

$$P\left[N(m) \geq N_{obs}(m)\right] \leq \alpha$$

where $N(m)$ is a random variable representing the number of occurrences of the motif under the reference model R_G.

Definition 3.3: **(Motif Type II: topological colored motif)** Let $G = (V, E, C, c)$ be a colored graph drawn from a distribution of graphs G^* under a given reference exchangeable random model R_G. Let $m(V_m, E_m, C_m, c)$ be a subgraph (induced or non-induced) of G where V_m is the set of k motif nodes, E_m is the set of motif edges and C_m is the multiset of node colors. Let $N_{obs}(m)$ be the number of isomorphic non-redundant occurrences of m in G, where $p(V_p, E_p, C_p, c)$ is an occurrence if there is a 1-to-1 onto mapping from E_m to E_p such that for every $u, v \in E_m \exists u', v' \in E_p$ such that $c(u) = c(u')$ and $c(v) = c(v')$ (automorphisms of m w.r.t. both topology and colors are considered only once). Let α be a critical value. We say that m is a motif of G if

$$P\left[N(m) \geq N_{obs}(m)\right] \leq \alpha$$

where $N(m)$ is a random variable representing the number of occurrences of the motif under the reference model R_G.

From now on, we will denote $m(V_m, E_m, C_m)$ as m_c. The significance of a motif is always evaluated with respect to a reference random model, so the aim is to find a good estimation of the distribution of the random variable $N(m_c)$ under a properly selected random graph model.

IV. The Expected Degree Distribution random Model

The Expected Degree Distribution (EDD) model was introduced in [**?**], [**?**]. EDD generates graphs in which node degrees follow a given distribution. We review its definition and give the details of its extension to directed colored graphs, considering the cases where node degrees and colors are (i) independent and (ii) dependent.

A. EED on graphs with independent node labels and node degrees

Dennis thinks we have to be consistent. We call everything a label or a color. I prefer color, but we have to stick to one or the other for clarity.

Given an undirected graph $G = (V, E)$ with $|V| = N$, define a random variable Deg based on the degree distributions of G. Specifically, $P(Deg = d)$ is the probability that a node has degree d in G.

Given the random degree distribution Deg based on the input graph G, intuitively, we can create new graphs $G' = (V', E')$ with $|V'| = |V|$. Assign valences to each node i in V' by sampling according to the discrete distribution Deg by keeping fixed the expectation of the degrees. An edge between nodes i and j, with $i \neq j$, exists with probability:

$$P(i,j) = \frac{D(i) \times D(j)}{\sum_{k=1}^{N} D(k)} \tag{1}$$

Where $D(i)$ be the degree of node i within the input graph.

To guarantee $P(i,j) \leq 1$ we assume that $\max D(i)^2 < \sum_{k=1}^{N} D_{out}(k)$. When this condition is not satisfied $P(i,j)$ is set to 1.

Then, we define the probability of a uncolored topology of k nodes within the graph under the EDD model in the following way. First, we can observe that, given a set of nodes with a given list of degrees, under the EDD model, the probability of a topology is the product of all the edges probabilities.

To compute the probability of a topology within the graph we sum across all the possible combinations of degree assignments to each node. Following [**?**] this can be expressed as:

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

Where $\gamma = 1/\sum_{k=1}^{N} D(k)$, m_{++} is the total number of edges in m, m_{u+} is the number of edges from node u in m and $\mathbb{E}[Deg^{m_{u+}}]$ is the m_{u+}-th moment of distribution Deg. The $i-th$ moment of a random variable X defined as the expectation of X^i which is $\mathbb{E}[X^i] = \sum_{x \in X} P(X = x) x^i$.

Next we define the occurrence probability of a colored motif under the EDD model. Since the color assignment to nodes is independent from its degree, the probability of observing the colored motif m_C of k nodes having a multiset of color C_m is:

$$\sigma(m_C) = \mu(m) \times \nu(C_m)$$

In such a case, the probability to assign colors in C_m to the k nodes of m_C follows a multinomial distribution

$$\nu(C_m) = \frac{k!}{\prod_{c \in C_m} s(c)!} \prod_{c \in C_m} f(c)$$

where $s(c)$ is the multiplicity of color c in C_m and $f(c)$ is the frequency of color c in the graph.

Intuitively, the same probability can be computed in the case of directed graphs by properly adapting the EDD to sample within a space of in-degree and out-degree distributions.

When dealing with directed graphs $G = (V, E)$ with $|V| = N$, we can generate random graphs by defining two random variables D_{out} and D_{in} by sampling from distribution of Deg_{in} and Deg_{out}, which are the random variables of in-degree and out-degree distributions of the graph. Let $D_{out}(i)$ and $D_{in}(i)$ be the out-degree and in-degree of node i in the input graph, respectively. Then EDD random graphs can be created according the following equation:

$$P(i,j) = \frac{D_{out}(i) \times D_{in}(j)}{\sum_{k=1}^{N} D_{out}(k)} \tag{2}$$

To guarantee $P(i,j) \leq 1$ $\max D_{out}(i) \times D_{in}(j) < \sum_{k=1}^{N} D_{out}(k)$. When this condition is not satisfied, $P(i,j)$ is set to 1.

To compute the probability of a motif within the graph we use the following equation:

$$\mu(m) = \gamma^{m_{++}/2} \prod_{u=1}^{k} \mathbb{E}[Deg_{out}^{m_{u+}}] \mathbb{E}[Deg_{in}^{m_{u-}}]$$

Where $\gamma = 1/\sum_{k=1}^{N} D_{out}(k)$, m_{++} is the total number of out-going edges in m, m_{u+} is the number of out-going edges from node u in m and m_{u-} is the number of in-going edges to node u in m. $\mathbb{E}[Deg_{out}^{m_{u+}}]$ and $\mathbb{E}[Deg_{in}^{m_{u-}}]$ represent the moments of order m_{u+} and m_{u-} of distributions Deg_{out} and Deg_{in}, respectively.

B. EED on graphs with dependent node labels and node degrees

Let $G(V, E, C, c)$ be a colored undirected graph with $|V| = N$. When dealing with graphs in which node degrees depend on colors, we define a number of EDD conditioned distributions, one for each color.

Let $Deg_{|c}$ be a random variable defined as the degree distribution for nodes with color c within the input graph G. Let $P(Deg_{|c} = x)$ be the probability of sampling a node in G with a degree x given the color c. Random graphs can be created by defining the probability of adding an edge between two nodes as in the case of undirected graphs with motifs of type I (see equation 1). Where $D(i)$ is the degree of node i according to $Deg_{|c_i}$.

We define the probability of a colored motif m_C of k nodes within the graph under the EDD model in the following way:

$$\sigma(m_C) = \gamma^{m_{++}/2} \prod_{u=1}^{k} P(c_u)^{m_{u+}} \mathbb{E}[Deg_{|c_u}^{m_{u+}}] \quad (3)$$

where $Deg_{|c_u}$ is the degree distribution for nodes of color c_u in the input network, $\gamma = 1/\sum_{k=1}^{N} D(k)$, m_{++} is the total number of out-going edges in m_c, m_{u+} is the number of out-going edges from node u in m_c, $P(c_u)$ is the probability of observing the color c_u within the graph and $\mathbb{E}[Deg_{|c_u}]$ is the $m_{u+} - th$ moment of distribution $Deg_{|c_u}$.

When dealing with directed colored graphs we have to define $2 \times |C|$ distributions. Given a color c we have two conditioned random variables $Deg_{out|c}$ and $Deg_{in|c}$ by making use of both in-degree and out-degree distributions of the input network.

We can then define two random variables D_{out} and D_{in} by sampling from the distributions $Deg_{out|c}$ and $Deg_{in|c}$, respectively. Let $D_{out}(i)$ and $D_{in}(i)$ the expected out-degree and in-degree of node i, respectively. The probability of adding an edge between two nodes is then defined as in the case of directed graphs with motifs of type I (see equation 2. We define the probability of a colored motif m_c of k nodes within the target network under the EDD model in the following way:

$$\sigma(m_c) = \gamma^{m_{++}} \prod_{u=1}^{k} P(c_u)^{m_{u+}+m_{u-}} \mathbb{E}[Deg_{out|c_u}^{m_{u+}}] \mathbb{E}[Deg_{in|c_u}^{m_{u-}}] \quad (4)$$

where $\gamma = 1/\sum_{k=1}^{N} D_{out}(k)$, m_{u+} is the number of out-going edges from node u in m_c and m_{u-} is the number of in-going edges from node u in m_c and $\mathbb{E}[Deg_{out|c_u}^{m_{u+}}]$ and $\mathbb{E}[Deg_{in|c_u}^{m_{u-}}]$ are the moments of order m_{u+} and m_{u-} of $Deg_{out|c_u}$ and $Deg_{in|c_u}$, respectively.

V. Expectation and Variance of Motifs within the target network

A. Motifs of Type I

We describe a method to compute the mean and the variance of the number of non-induced occurrences of a colored motif under any random graph model[],[].

We make the following assumptions: (i) the occurrence probability of a given motif does not depend on the occurrence position; (ii) the disjoint occurrences are independent one to another; and (iii) colors are independent from topologies.

Let m_C be a motif of k nodes, it can occur in different positions within a graph G. Let $\alpha = (i_1, i_2, ..., i_k)$ a k-uple of indexes representing a potential location of m_C in G. The number of such positions is $\binom{N}{k}$. We introduce a random variable $Y_\alpha(m_C)$ which equals one if the topology m_C occurs at position α and 0 otherwise.

Since we assume exchangeability of our random model, the distribution of $Y_\alpha(m_C)$ does not depend on α permutations. $Y_\alpha(m_C)$ is distributed according to a Bernoulli random variable $B(p)$, where $p = \sigma(m_C)$ is the probability of occurrence of motif m_C at any position within G.

Moreover, a motif m_C in a position can occur in different *configurations*, where each configuration corresponds to a permutation of indexes in α. Some permutations of the indexes yield the same motif, so we need to consider only the set of its Non-Redundant Permutations (NRP) which we denote with $R(m_C)$. We also denote with $\rho(m_C) = |R(m_C)|$ the number of Non-Redundant Permutations of m_C.

To compute $R(m_C)$ we generate all possible $k!$ simultaneous permutations of the rows and columns of the adjacency matrix of m. For each permutation, we build the corresponding adjacency matrix and check the latter for redundancy. We then have the following random variable $N(m_C) = \sum_\alpha \sum_{m_C^{'} \in R(m_C)} Y(m_C^{'})$.

For the exchangeability assumption, each permutation of m_C has the same probability of occurrence. The expectation of the count of a colored motif m_C in a graph G with N nodes is

$$\mathbb{E}[N(m_C)] = \binom{N}{k} \times \rho(m) \times \nu(C_m) \times \mu(m) \quad (5)$$

where $\binom{N}{k}$is the number of all possible locations of m in G, $\nu(m_C)$ is the occurrence probability of the multiset of colors C of m_C and $\mu(m)$ is the occurrence of the motif according to the chosen random model.

To compute the variance of the number of occurrences of the colored motif, we have to take into account the overlapping of the occurrences. Two occurrences of a motif overlap if they share at least one node.

We define the concept of super-motif, which is a motif composed by two NRPs of overlapping occurrences of a given motif. Given two NRPs, $m^{'}$ and $m^{''}$ of a motif m, and an integer s, we define the overlapping operation with s common nodes as $m^{'} \Omega_s m^{''}$; The result of the operation is a new motif with $2k - s$ edges (see Figure **??** for an example). Furthermore the notion of super-motif is transitive.

Concerning the colors, a super-motif inherits them from the ancestor motifs. Due to node overlapping, one or more colors can overlap. Colors can overlap in many ways, therefore the same super-motif can be colored with different super-sets of colors. As for the super-motif topology, we can define an overlapping operations for two multi-sets of colors C_1 and C_2 with overlap s: $C_1 \Pi_s C_2$, where $C_1 \Pi_s C_2$ represents the set of all possible intersections of C_1 and C_2 with s elements.

$C \Pi_s C$ consists of the set of all subsets of C with s elements.

Let C a multiset of colors of $m^{'}$ and $m^{''}$, C^* the multiset of colors of the s overlapping nodes of $m^{'}$ and $m^{''}$ and $C^- = C \setminus C^*$, where $\setminus$ is the set difference operator.

The probability of observing the super-multiset of colors $C_1 \Pi_s C_2$ in the graph [**?**], [**?**] is:

$$\nu(C \Pi_s C) = \sum_{C^* \subset C : |C| = s} \frac{\nu(C^*) \left[\nu(C \setminus C^*)\right]^2}{s(C^*)}$$

where $s(C^*)$ is the multiplicity of subset C^* in C. The equation considers the probability of observing the multiset of colors in the intersection of two motifs and the probability of observing the multiset of remaining colors in the non-overlapping region of both motifs. Since the subset of overlapping colors can occur multiple times in C, the probability must be corrected considering $s(C^*)$.

Therefore, the probability of observing a colored supermotif generated from colored motifs is the following:

$$\sigma(m_C^{'}, m_C^{''}, s) = \mu(m_C^{'}\Omega_s m_C^{''}) \times \nu(C\Pi_s C)$$

The computation of variance is based on the expectation of the squared count of a colored motif. The expectation is given by the contribution of two terms, one is related to pairs of disjoint occurrences and one is related to pairs of overlapping occurrences (with different degrees of overlap). In both cases we have to consider: (i) all possible locations of the two occurrences of a motif m_c in the graph; (ii) all possible non-redundant permutations of m_c. The expectation of the squared count is given by the following equation:

$$\mathbb{E}[N^2(m_C)] = \binom{N}{N-2k,k,k}\left[\sum_{m' \in R(m)} \sigma(m'_C)\right]^2 + \sum_{s=1}^{k}\binom{N}{k-s,s,k-s,N-2k+s}\sum_{m',m'' \in R(m)} \sigma(m'_C, m''_C, s) \quad (6)$$

where k is the number of nodes of motif m and N is the number of nodes of the graph, $\binom{N}{N-2k,k,k}$ is the number of all possible combinations of locations of two non-redundant permutations of m with no overlap and $\binom{N}{k-s,s,k-s,N-2k+s}$ is the number of all possible combinations of locations of two non-redundant permutations of m with overlap s. The variance of the count is $\mathbb{V}[N(m_C)] = \mathbb{E}[N^2(m_C)] - \mathbb{E}[N(m_C)]^2$

B. Motifs of Type II

In what follows we describe how to compute the exact mean and variance of the number of non-induced occurrences of a colored motif under the EDD random models to deal with graphs in which there is a dependency between colors and topology. We keep two important assumptions: (i) the occurrence probability of a given motif does not depend on the occurrence position; (ii) disjoint occurrences are independent one another.

In this context we need to extend the original definition of non-redundant permutations of a topology. We introduce the concept of non-redundant colored permutations of a colored motif. A colored permutation of a motif m_C is a colored motif resulting from a permutation of the nodes (and the corresponding colors) of m_C and it is represented by its adjacency matrix plus the array of colors of its nodes. Two colored permutations are non-redundant iff one of the following conditions hold: (i) their adjacency matrices are different; and (ii) their adjacency matrices are equal, but the array of colors are different.

In Figure **??** we give an example of non-redundant colored permutations.

Therefore the expected count of motifs within the target network is computed according to the following equation: $\mathbb{E}[N(m_C)] = \binom{N}{k}\pi(m_C)\sigma(m_C)$, where $\pi(m_C)$ is the number of non-redundant colored permutations of m_C and $\sigma(m_C)$ is the occurrence probability of m_C according to an exchangeable random model. To compute the variance we can use the equation for the motifs of type I (see equation 6) providing the proper probability of the motif. The variance uses the concept of supermotif which in this case have to be implemented carefully. In this case the overlapping has to take into account the node colors. Therefore when two motifs of size k have an overlapping with s nodes, these nodes share also the same colors.

This implies that motifs having nodes of the same colors in not compatible positions will not yield supermotif. In Figure ... we give an example.

VI. Assessing the motifs significance

To establish whether a motif m_c is over represented in a given graph, one needs to calculate the probability $P[N(m_C) \geq N_{obs}(m_C)]$, where $N_{obs}(m_C)$ is the observed number of non-redundant occurrences of m and $N(m_C)$ is a Random Variable representing the number of occurrences of the motif under the chosen reference model. The common approach on the approximation of $P[N(m_C) \geq N_{obs}(m_C)]$ relies on simulation through the usage of permutation test. To avoid such an expensive simulation, a key problem is to identify a proper distribution fitting the number of observations in the the reference random model. In this direction, several attempts have been done. The most successful one has been proposed in [] for uncolored graphs. Authors showed that the Pólya-Aeppli (denoted by PA) distribution (also knonwn as Geometric-Poisson) distribution [] is suitable to describe how the count of events occurring in motifs may vary and can be used as an approximation of the distribution of the count of $N(m_c)$.

Following [], we can notice that, motifs come in clusters since they can overlap, clusters result in several occurrences of a motif with a reduced number of vertices. Hence, given a graph we can observe a certain number of clusters according to the overlapping of the motifs. This number can be modeled as a random variable that we call X_1. On the other hand, suppose we have a set of clusters according to the intersection of pairs of motifs. We can introduce a second random variable called X_2 in which we sample several times a cluster until we observe the size of the cluster we are looking for. We assume that X_1 (modeling the number of clusters) has a Poisson distribution, whereas X_2 (modeling the probability of observing a certain cluster size) has a Geometric distribution. Furthermore, the cluster sizes are independent each other with a common distribution. The PA distribution is obtained when the cluster size has a geometric distribution $G(1-\alpha)$, so the mean size of a cluster is $1/(1-\alpha)$

In this case we have $X \sim PA(\lambda, \alpha)$ is a random variable representing the number of observed events:

$$P(X = x) = \begin{cases} e^{-\lambda}\alpha^x \sum_{c=1\cdots x} \frac{1}{c!}\binom{x-1}{c-1}\left[\frac{\lambda(1-\alpha)}{\alpha}\right]^c & if\ x > 0 \\ e^{-\lambda} & if\ x = 0 \end{cases}$$

The mean and the variance of $PA(\lambda, \alpha)$ are defined as $\frac{\lambda}{1-\alpha}$ and $\frac{\lambda(1+\alpha)}{(1-\alpha)^2}$. By making use of the mean and variance obtained using the exchangeable random graph model we can deduce the parameters of the distribution as $\alpha = \frac{\mathbb{V}[N(m_C)] - \mathbb{E}[N(m_C)]}{\mathbb{V}[N(m_C)] + \mathbb{E}[N(m_C)]}$ and $\lambda = (1-\alpha) \times \mathbb{E}[N(m_C)]$.

VII. Dealing with the Induced case

The equations described in section V-A to compute the mean and the variance of a motif count refer to the non-induced case. The non-induced case is simpler to model since the random variable describing the motif occurrences does not depend on other topological motif occurrences of the same size. This is not the case when we deal with induced occurrences.

We extend the model to compute the mean and the variance of the number of induced occurrences of a motif by relating the number of non-induced occurrences to the number of induced occurrences. In order to do that, we use the theoretical result known as Kocay Lemma []. The Kokay Lemma allows to express the number of non-induced occurences of a motif as a linear combination of the induced occurences of all motifs of the same size.

A. Induced motifs of Type I

Suppose we want to count the number of non-induced occurrences N of a certain subgraph with k nodes. By applying the Kocay Lemma we can express this number as a linear combination of the number of induced occurrences of all possible topologies with k nodes. Therefore, to construct such a relation we find the coefficients of the linear combination. Once such coefficients are found, these can be represented as a matrix. Therefore, we will invert the matrix to find the count of the induced motifs as a linear combination of the non-induced one.

Suppose we have a star topology with k nodes and we wish to find its non-induced occurrences within a target graph G. Suppose we know the number of occurrences of all induced topologies with k nodes. We denote with N the number of induced occurrences. The coefficients of the linear combination can be determined by counting the occurrences of the star topology within each topology of k nodes and this can be done by making use of a subgraph matching algorithm [**?**]. The process above can be repeated for each topology of size k. The coefficient for all topologies can be represented with a matrix notation.

We denote with K_k the Kocay matrix for topologies of size k, where each row refers to a specific topology t. We denote with $K_k(t)$ the corresponding row. $K_k(t_i, t_j)$ is the number of non-induced occurrences of topology t_i in topology t_j. By computing the inverse of a Kocay matrix we can express the number of induced occurrences of a motif as a linear combination of the number of non-induced occurrences of all topologies with k nodes.

In this way we represent the random variable $N'(m_C)$ of the induced counts of colored motif m_C with k nodes as a linear combination of random variables of counts of all non-induces motifs of size k. Let M^k the set of all possible topologies with k nodes. We have:

$$N'(m_C) = \sum_{t \in M^k} K_k^{-1}(m,t) N(m_C)$$

Therefore, we compute mean and variance of such random variable.

The coefficients of the linear combination are the elements of a row of the inverse Kocay matrix, the mean is then given by $\mathbb{E}[N'(m_C)] = \mathbb{E}[\sum_{t \in M^k} K_k^{-1}(m,t) N(m_C)] = \sum_{t \in M^k} K_k^{-1}(m,t) \mathbb{E}[N(m_C)]$.

The variance of $N'(m_C)$ implies the computation of the covariance. We have the following equation:

$$\mathbb{V}[N'(m_C)] = \sum_{t \in M^k} \left[K_k^{-1}(m,t)\right]^2 \mathbb{V}[N(m_C)] +$$

$$+ \sum_{t', t'' \in M^k \mid t' \neq t''} K_k^{-1}(m,t') K_k^{-1}(m,t'') Cov\left(N(t_C'), N(t_C'')\right)$$

We have that $Cov\left(N(t_C'), N(t_C'')\right) = \mathbb{E}[N(t_C') N(t_C'')] - \mathbb{E}[N(t_C')]\mathbb{E}[N(t_C'')]$ where:

$$\mathbb{E}[N(t_C') N(t_C'')] = \binom{N}{N-2k, k, k} \sum_{m' \in R(t'), m'' \in R(t'')} \sigma(m_C') \sigma(m_C'') +$$

$$\sum_{s=1}^{k} \binom{N}{k-s, s, k-s, N-2k+s} \sum_{m' \in R(t'), m'' \in R(t'')} \sigma(m_C', m_C'', s)$$

B. Induced motifs of Type II

Like in the previous case, we can use Kocay matrices to compute the number of non-induced colored motifs as a linear combination of the number of induced colored motifs and vice versa. The main difference is that now we have multiple Kocay matrices for motifs of size k. In fact, if we change the colors of a motif with k nodes, we can obtain different Kocay matrices, and a Kocay matrix becomes function of k and the array C of node colors and will be denote as $K_{k,C}$. Figure ... reports an example of a motif with $k = 3$.

Once we have computed the coefficients we can apply the same equations defined above to compute the mean and the variance of the motif occurrence.

VIII. Experimental analysis