

Fast Analytical Methods for Finding Significant Labeled Graph Motifs

Giovanni Micale · Rosalba Giugno ·
Alfredo Ferro · Misael Mongiovì ·
Dennis Shasha · Alfredo Pulvirenti

Received: date / Accepted: date

Abstract Network motif discovery is the problem of finding subgraphs of a network that occur more frequently than expected, according to some reasonable null hypothesis. Such subgraphs may indicate small scale interaction features in genomic interaction networks or intriguing relationships involving actors or a relationship among airlines. When nodes are labeled, they can carry information such as the genomic entity under study or the dominant genre of an actor. For that reason, labeled subgraphs convey information beyond structure and could therefore enjoy more applications. To identify statistically significant motifs in a given network, we propose an analytical method (i.e. simulation-free) that extends the works of Picard et al. 2008 and Schbath et al. 2009 to label-dependent scale-free graph models. We provide an analytical

This work has been partially supported by the U.S. National Science Foundation and National Institutes of Health under grants NSF: MCB-1158273, IOS-1339362, MCB-1412232, MCB-1355462, IOS-0922738, MCB-0929338, and NIH: 2R01GM032877-25A1. This work has been also partially supported by the Italian MIUR projects: PRISMA - CUP E61H12000140005 and CLARA - CUP E64G14000190008.

G. Micale
University of Catania, Department of Mathematics and Computer Science, Catania, 95125,
Italy E-mail: gmicale@dmi.unict.it

R. Giugno
University of Verona, Department of Computer Science, Verona, Italy E-mail: ros-
alba.giugno@univr.it

M. Mongiovì
CNR, Institute of Cognitive Sciences and Technologies, Catania Italy E-mail: mon-
giovì@dmi.unict.it

D. Shasha
New York University, Courant Institute of Mathematical Science, New York, USA E-mail:
shasha@cs.nyu.edu

A. Ferro · A. Pulvirenti
University of Catania, Department of Clinical and Experimental Medicine, Catania, 95125,
Italy E-mail: ferro@dmi.unict.it E-mail: apulvirenti@dmi.unict.it

expression of the mean and variance of the count under the Expected Degree Distribution random graph model. Our model deals with both induced and non-induced motifs. We have tested our methodology on a wide set of graphs ranging from protein-protein interaction networks (PPI) to movie networks. The analytical model is a fast (usually faster by orders of magnitude) alternative to simulation. This advantage increases as graphs grow in size.

Keywords Network Mining · Random Network models · labeled graph motifs · graph algorithms

1 Introduction and Related Work

Many complex systems can be represented as networks of interacting components (e.g. protein-protein interaction networks, social networks, entertainment networks). In biology, frequent sub-networks may represent functional modules [1] or basic building blocks [2] that perform some coordinated activities. These blocks are commonly called network motifs. Network motifs can be defined as consisting of patterns of interconnections (i.e. subgraphs) that arise unexpectedly often in a network. We refer to unlabeled motifs as *topological unlabeled motifs*. The motivation behind finding such motifs is that subgraphs with the same topology might be functionally similar. For example, motifs may correspond to conserved patterns that are linked to important cellular functions.

Before we get into the technical details, we should ask the preliminary question: how do we know when a subgraph appears unusually often? One approach is to declare that if there are more than k instances of a subgraph in the graph, then that subgraph frequently appears and therefore is important. However, without a principled way to choose k , leaving the choice to the user amounts to asking the user to guess.

In principle, statistics offers a better way. Consider the problem of asking whether a coin is fair. Suppose we perform an experiment consisting of flipping the coin 17 times and counting the number of heads. At which point should we consider the coin likely enough to be unfair to warrant an expensive physico-chemical analysis? Intuitively, 8, 9, 10 or 11 heads out of 17 should not raise an eyebrow. 15 or more should. The background knowledge that enables us to choose a threshold is that we start with a null hypothesis (the probability that the coin will land heads is $1/2$) and this gives us a probability distribution of the number of heads under the null hypothesis. Using that distribution we can ask the probability that there would be 12 or more heads (about 0.07) vs 15 or more heads (about 0.001). In the first case, we might conclude that there is no evident need for an expensive test to determine fairness. So, the probability of an outcome (i.e. the p-value) with respect to a reasonable null hypothesis is a principled way to determine unusualness.

In the graph case, we have no simple *a priori* null hypothesis. Should finding a path of five A more than 100 times be unusual in this graph? What about a star of five A? That depends on many properties of the graph. For that reason,

given a topological pattern m on an input network G , the common approach to determining whether m is a motif consists of the following steps: (i) generate a large set of random networks sharing the observable characteristics (roughly, same number of nodes and edges with similar degree distributions) of G ; (ii) find the number of occurrences of m in each of those networks; (iii) estimate the p-value by comparing the number of occurrences in the input network with the numbers in the random networks.

The first step creates random graphs (i.e. networks) under a specified random reference model having the same number of nodes and edges and degree distribution as the input network. Examples of reference models include: (i) The Erdős-Rényi model (ER model) [3] in which the probability of connecting two nodes n_1 and n_2 in a random graph is the same as the probability of connecting any other two nodes n_3 and n_4 and that probability is determined by the network density of G . (ii) The Fixed Degree Distribution model (FDD model) [4], where each random graph is generated by swapping edges starting from the input network G , guaranteeing that each node in each random graph R has the same degree as in G . (iii) The Block Two-Level Erdős-Rényi model (BTER model) [5], which is based on the idea that a graph is composed by subgraphs (“communities”) each of which satisfies ER conditions. These communities are then connected to one another using an EDD model. For this purpose, within each community, those nodes n having degree d'_n lower than the expected one d_n are used. The quantity $d_n - d'_n$ is defined as the excess degree of node n . (iv) The Expected Degree Distribution model (EDD model) typically known as Chung-Lu model [6,7] which generates random graphs whose node degrees have the same expectation as the input network G . (v) The Erdős-Rényi mixture for graphs model (ERMG model) [8,9] which is based on mixture population edges and is used to model heterogeneous connectivity. (vi) The Exponential Random Graph model, in which the input network G is drawn from a family of randomized variants of it, R_G , generated in the following way. Each graph $G_R \in R_G$ has a probability $P(G_R)$. Probabilities $P(G_R)$ are such that a maximally random ensemble of networks is generated, under the constraint that, on average, a set $\{C_a\}$ of desired topological properties is set equal to the values $\{C_a(G)\}$ observed in the input network G . This is achieved as the result of a constrained Shannon-Gibbs entropy maximization [10].

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 variants of G [11]. This baseline method which relies on simulation yields a measure of the significance of each candidate through the computation of a p-value using a resampling approach [2,11–13]. Unfortunately, this method requires a large number (1000 to 10,000) of random graphs whose analysis turns out to be computationally expensive (far more expensive than analyzing the target network alone). Moreover, the expense of simulation increases as the graph size grows.

Over the last decades, researchers have worked on replacing simulation by analytical methods. For unlabeled motifs, approximation methods, based on

the Erdős-Renyi (ER) model, have tried to compute the asymptotic normality of the distribution of topology counts [14]. Unfortunately, empirical evidence suggests that the Erdos-Renyi random model offers a poor fit for many real-world networks [15].

In 2008, Picard et al. [16] 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. 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) [26]. 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 holds when distinct topologies can share nodes and edges (i.e. clumps) [16]. 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 Pólya-Aeppli distribution $P(\lambda, a)$. These results lead to an estimate of the count of occurrences of a given topology. Picard et al. [16] show that this is a good model for the distribution of the counts of subgraph topologies (both induced and non-induced), since the fit is more accurate than a Gaussian model for the graphs of many applications. (An induced subgraph is a subset of the vertices of a graph G together with any edges whose endpoints are both in this subset. In a non-induced subgraph of G , the edges are a subset of those present in the corresponding induced graph over the same nodes.)

More recently, Squartini and Garlaschelli [17] proposed an analytical maximum-likelihood method to detect patterns in real networks, introducing a method that allows one to obtain expected standard deviations of any topological property analytically, for any binary, weighted, directed or undirected network. However authors deal only with topology motifs.

The motif finding problem has attracted a lot of research concerning the design of efficient algorithms for the enumeration of subgraphs. Several methods have been proposed for the identification of induced and non-induced motifs [14, 18–20] of any size. Many tools are capable of dealing with motifs up to $k = 9$ nodes on medium size networks. When the size of the motif is small ($k = 3, 4$), the usage of graphlet decomposition techniques [21] has been proved to be the most efficient solution for unlabeled graphs even with large networks having billions of nodes and edges, because it lends itself to parallelism.

Different characterization of labeled motifs

The above methods give p-values for label-free networks. Focusing only on topology ignores the possible meaning of nodes. Such meaning can lead to important insights. For example, in a protein-protein interaction network,

topologies having to do with the process of metabolism may be different from topologies having to do with another process such as meiosis.

We name motifs where node types matters *labeled motifs*.

To deal with labeled motifs we need to generalize the unlabeled motif definition based on constraints that can be defined on the topology, on the label assignment, or both. This leads us to three different definitions of motifs which are hierarchically related (see Figure 1).

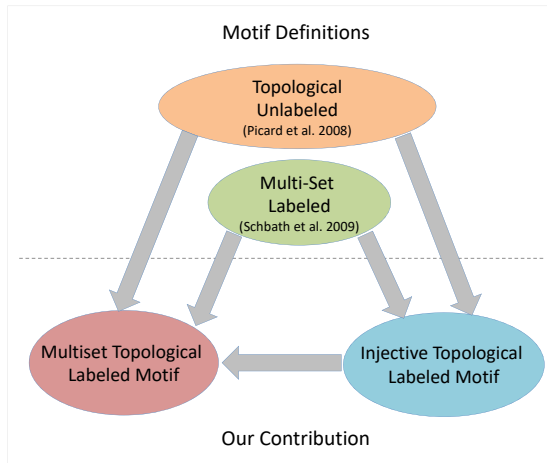


Fig. 1: Motif hierarchy. Four different definitions of motifs. When definition A points to definition B , the set of motifs responding to A is a superset of those responding to B . All the definitions apply to both directed and undirected graphs.

Concerning the generation of networks with real-world structural properties and correlated labels, in [22] the authors introduced the Attributed Graph Model (AGM), which exploits label correlations in connection to generative network models to jointly model network topology and node labels. In [23], authors propose a generative model for labeled graphs called Multiplicative Attribute Graph (MAG) model. MAG generates the network by taking into account the number of vertices, a set of prior probabilities for vertex label values and a set of affinity matrices specifying the probability of an edge conditioned on the vertex labels. In [24], the authors describe AGWAN (Attribute Graphs: Weighted and Numeric), a generative model for random graphs with discrete labels and weighted edges.

In their seminal work on analytical analysis methods for labeled motifs, Schbath et al [25] define a motif as any connected topology of k nodes having a given multiset of labels M . We refer to this kind of motifs as *multiset labeled motif*.

An example of multiset labeled motif is a connected topology consisting of five nodes having one red label and four nodes labeled with blue. In this example, five would be the size of the motif. An occurrence of a motif is defined as any connected topology with exactly one red node and four blue nodes (see Figure 2).

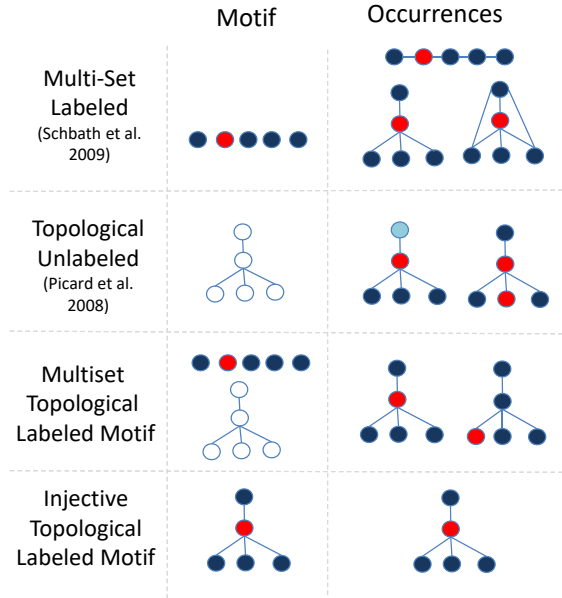


Fig. 2: Example of Motif occurrences within the motif hierarchy. Multi-Set Labeled motifs might consider all connected structures containing four blues and one red to be the same motif. Topological Unlabeled motifs count as a single motif all stars of size five, regardless of labels. Multiset Topological Labeled motifs might count as a single motif all star structures consisting of four blues and one red regardless of which of the five nodes have which labels. An Injective Topological Labeled Motif might count as a motif a star consisting of a single red node in the center surrounded by four blue nodes.

The authors [25] proposed an analytical (simulation-free) approach for assessing the exceptionality of multiset labeled motifs. They established an exact analytical model for the mean and the variance of the count of a labeled motif using the Erdős-Rényi (ER) random graph model. In doing so, they assumed that the label assignment to nodes is independent of the topology of the network, and therefore modeled the probability of a multiset of labels as a multinomial distribution. To estimate a p-value associated to a motif, the authors also modeled the complete distribution of the count of a colored motif

in an Erdős-Renyi random graph model by making use of the Pólya-Aeppli distribution.

In many applications, we are interested in both topology and labels. For that reason, we propose a first new definition of labeled motif consisting of a subgraph of k nodes with a given topology having nodes belonging to a multiset of labels M , denoted *multiset topological labeled motif* (see Figure 1). In this case, the precise assignment of labels to nodes in the topology is unimportant. For example, any star topology of five nodes in which one node is red and the others are blue could be an occurrence of the same topological multiset labeled motif.

A second new definition defines motifs as a topology and a specific label assignment to each node in the topology. For example, a star topology of five nodes in which the center node is red and the other nodes are blue. In this case a motif is a subgraph of k nodes having fixed labels connected through a given topology, so this is called an *injective topological labeled motif* (see Figure 1).

Our view of motifs

In this paper we deal with the two definitions of motifs that constrain both labels and topologies: *multiset topological labeled motif* (or multiset motif) and *injective topological labeled motif* (or injective motif). No analytical model has been proposed yet for either of these. Inspired by the work of [16,25] we introduce analytical models to establish the significance of labeled motifs on directed and undirected graphs, under the EDD random model, and in which labels are either independent or dependent on the degrees of nodes. Finally, our model deals with both induced and non-induced motifs. Thus, we handle two kinds of motifs, directed and undirected graphs, induced and non-induced motifs, and two random models (with color-degree dependency or not). The body of the paper introduces what we consider to be the most useful of these definitions: label-dependent graphs (as in graphs where the label of a node might at least partly determine its degree, e.g. rock musicians have more fans than professors) and the injective motif case for non-induced motifs. The supplementary material (and our software) extends this to both new definitions of motifs, induced graphs, label-independent valence distributions, and directed graphs.

2 Definitions

A *labeled 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 labels and $c : V \rightarrow C$ is a function that assigns a label to each node in V . If $(u, v) \in E$, we say that v is a neighbor of u . G is undirected means that if $\forall (u, v) \in E$, then $(v, u) \in E$, i.e. all neighbor relationships go both ways. If labels are not taken into account $G(V, E)$ is called unlabeled graph.

Intuitively, given a graph G , a topology that occurs “unusually” frequently in G is called a *motif*. The number of occurrences of a motif counts only non-redundant occurrences. A motif occurrence is redundant if it is an automorphism of another occurrence. Given a graph $G = (V, E)$, a permutation ξ of the vertex set V is an automorphism if for each 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 in an input graph G , we imagine that the target graph is drawn from a set of graphs belonging to a random graph model. Random graph models generate graphs that preserve certain characteristics of G . An important property of a random graph model is exchangeability. Given two random graphs G^1 and G^2 under a random model R_G , we say that R_G is a random exchangeable model when the node degree distributions of G^1 and G^2 are the same.

We define two types of motifs.

Definition 21 (Multiset Topological Labeled Motif) *Let $G(V, E, C, c)$ be a labeled graph drawn from a distribution of graphs under a given reference random exchangeable model R_G . Let $m(V_m, E_m, C_m)$ be a subgraph (induced or non-induced) of G having V_m and E_m as sets of nodes and edges and C_m as the multiset of node labels of the nodes V_m . Let $N_{obs}(m)$ be the number of non-redundant occurrences of m in G having the same multiset of labels C_m , and let α be a critical value (provided by the user). We say that m is a motif of G if the probability*

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

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

We discuss the above type of motif further in the appendix. In the body we focus on the following definition of motif, corresponding to the last example of figure 2.

Definition 22 (Injective Topological Labeled Motif) *Let $G(V, E, C, c)$ be a labeled graph drawn from a distribution of graphs under a given reference random exchangeable 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 nodes of m , E_m is the set of edges and C_m is the multiset of node labels. Let $N_{obs}(m)$ be the number of non-redundant occurrences of m in G , where $p(V_p, E_p, C_p, c)$ is an occurrence of m if there is a 1-to-1 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 $c(v) = c(v')$. Let α be a critical value. We say that m is a motif of G if the probability*

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

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.

3 The Expected Degree Distribution Random Model

The Chung-Lu model, also known as Expected Degree Distribution (EDD) model was introduced in [6] for non-labeled graphs. EDD generates graphs in which node degrees follow a given distribution. We review the definition of EDD and extend it to labeled graphs where the degree of nodes depends, at least partly, on labeled. This case would hold, for example, for chemical graphs where most nodes labeled with carbon have degree 4 and all nodes labeled with hydrogen have degree 1.

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

Given f_D , we can generate a new graph $G' = (V', E')$ with $|V'| = |V|$ as follows: assign degrees to each node i in V' by sampling according to the f_D distribution. An edge between two nodes i and j , with $i \neq j$, is generated with probability:

$$P(i, j | D(i), D(j)) = \min(1, \gamma \times D(i) \times D(j)) \quad (1)$$

where $\gamma = 1 / [(N - 1) \times \mathbb{E}[f_D]]$ and $D(i)$ is the degree of node i within the input graph.

According to the exchangeability assumptions, the occurrence probability of a given motif does not depend on the occurrence position; further, disjoint occurrences are independent of one another. Therefore, the conditional occurrence probability of the motif, given an assignment of expected degrees $D(i)$ to the nodes of the motif, can be expressed as the product of the edge probabilities. To compute the occurrence probability of the topology motif m with k nodes, under the EDD model we have to perform a summation over the distributions of degrees $D(i)$. Such a probability, as defined in [16], can be expressed using the following equation:

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

where f_D is the degree distribution for nodes in the input network, m_{++} is twice the total number of edges in m , m_{u+} is the number of out-going edges from node u in m and $\mathbb{E}[f_D^{m_{u+}}]$ is the m_{u+} -th moment of the distribution f_D . Intuitively, the same probability can be computed in the case of directed graphs by adapting the EDD to sample within a space of in-degree and out-degree distributions. When dealing with a directed graph $G = (V, E)$ with $|V| = N$, we can generate random graphs by defining two random variables D_{out} and D_{in} . They are obtained by sampling from distributions of $f_{D_{in}}$ and

$f_{D_{out}}$, which are the random variables of in-degree and out-degree distributions of the input 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 to the following equation:

$$P(i, j) = \min(1, \gamma \times D_{out}(i) \times D_{in}(j)) \quad (3)$$

where $\gamma = 1/[(N-1) \times \mathbb{E}[f_{D_{out}}]]$. To compute the probability of observing an unlabeled topology of k nodes in a directed graph we use the following equation:

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

where 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 . $\mathbb{E}[f_{D_{out}}^{m_{u+}}]$ and $\mathbb{E}[f_{D_{in}}^{m_{u-}}]$ are the moments of order m_{u+} and m_{u-} of distributions $f_{D_{out}}$ and $f_{D_{in}}$, respectively.

In what follows we give the probability of injective motifs. Please refer to the supplementary materials for multiset motifs, for the case in which the degree is independent of labels, and for the analysis of induced motifs.

3.1 Expected degree distribution on labeled graphs

When dealing with graphs in which node degrees depend on labels, we have the f_D distribution of degrees and we can define a number of EDD conditional distributions, one for each label. We extend the model of [6] to labeled graphs as follows.

Let $f_D|c$ be a random variable defined as the degree distribution for nodes with label c within the input graph G . Let $P(f_D = x|c)$ be the probability of sampling a node in G with a degree x given the label c . Random graphs can be created by defining the probability of adding an edge between two nodes as in the case of undirected graphs under the EDD model with label-degree independence (see equation 1), where $D(i)$ is the degree of node i according to $f_D|c_i$.

We define the occurrence probability of the topology of a labeled motif m_C with k nodes, given a label assignment C to the nodes of the motif, within the graph as:

$$\mu(m_C|C) = \gamma^{m_{++}/2} \prod_{u=1}^k \mathbb{E}[f_D^{m_{u+}}|c_u] \quad (4)$$

where $f_D|c_u$ is the degree distribution for nodes with label c_u in the input network, m_{++} is twice the total number of edges in m_C , m_{u+} is the number of out-going edges from node u in m_C and $\mathbb{E}[f_D^{m_{u+}}|c_u]$ is the m_{u+} -th moment of the conditional distribution $f_D|c_u$. Once again, the above equation of

the occurrence probability is obtained by performing a summation over the distributions of degrees.

When dealing with directed labeled graphs we have to define $2 \times |C|$ conditional distributions. Given a label c we create two conditional random variables $f_{D_{out}}|c$ and $f_{D_{in}}|c$ by making use of both the in-degree and out-degree distributions of the input network.

We can then define two random variables $f_{D_{out}}$ and $f_{D_{in}}$ by sampling from the distributions $f_{D_{out}}|c$ and $f_{D_{in}}|c$, respectively. Let $f_{D_{out}}(i)$ and $f_{D_{in}}(i)$ be the out-degree and in-degree of node i , respectively.

We define the occurrence probability of the topology of a labeled motif m_C in a list of k nodes within the directed graph, given a label assignment C , in the following way:

$$\mu(m_C|C) = \gamma^{m_{++}} \prod_{u=1}^k \mathbb{E}[f_{D_{out}}^{m_{u+}}|c_u] \mathbb{E}[f_{D_{in}}^{m_{u-}}|c_u] \quad (5)$$

where m_{u+} is the number of out-going edges from node u in m_C , m_{u-} is the number of in-going edges from node u in m_C and $\mathbb{E}[f_{D_{out}}^{m_{u+}}|c_u]$ and $\mathbb{E}[f_{D_{in}}^{m_{u-}}|c_u]$ are the moments of order m_{u+} and m_{u-} of the conditional distributions $f_{D_{out}}|c_u$ and $f_{D_{in}}|c_u$, respectively.

Finally, the probability of observing the injective labeled motif m_C is:

$$\sigma(m_C) = \mu(m_C|C) \times \nu(C)$$

where $\nu(C)$ is computed as:

$$\nu(C) = \prod_{u=1}^k P(c_u) \quad (6)$$

where $P(c_u)$ is the probability of observing the label c_u of motif node u in the graph. Notice that, in this case the labels in C have an order according to the corresponding label assignment to the nodes of the motif.

Figure 3 presents a toy example showing the computation of the topology occurrence probability, the label probability and the labeled motif occurrence probability for a non-induced injective motif in an undirected graph under an EDD model with label-degree dependence.

4 Expectation and Variance of Non-Induced Motifs

We describe a method to compute the mean and the variance of the number of non-induced occurrences of injective topological motifs under any exchangeable random graph model [16, 25].

Let m_C be a motif of k nodes. It can occur in different positions within a graph G . Let $\alpha = (i_1, i_2, \dots, i_k)$ be a k -tuple of ordered indexes (i.e., $i_1 < i_2 < \dots < i_k$) representing a potential location of m_C in G . The number of such

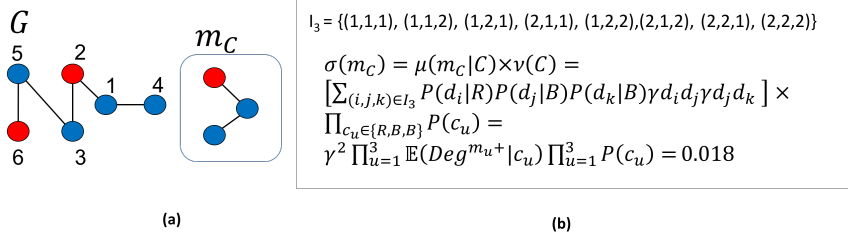


Fig. 3: Occurrence probability for a non-induced injective topological labeled motif under the EDD random model with label-degree dependence on an undirected graph. (a) Input graph $G(V, E)$, input motif m_C . We have two different degrees within G , the f_D distribution assumes values within the set 1, 2, $P(f_D = 1) = \frac{1}{3}$, $P(f_D = 2) = \frac{2}{3}$, $\mathbb{E}[f_D] = \frac{5}{3}$, $\gamma = \frac{3}{25}$. The probability of the two labels are $P(R) = \frac{1}{3}$, $P(B) = \frac{2}{3}$. We have to define the two labels' conditioned degree distributions: the $f_D|R$ distribution assumes values within the set 1, 2, $P(f_D = 1|R) = \frac{1}{2}$, $P(f_D = 2|R) = \frac{1}{2}$, $\mathbb{E}[f_D|R] = \frac{3}{2}$, the $f_D|B$ distribution assumes values within the set 1, 2, $P(f_D = 1|B) = \frac{1}{4}$, $P(f_D = 2|B) = \frac{3}{4}$, $\mathbb{E}[f_D|B] = \frac{7}{4}$, $\mathbb{E}[f_D^2|B] = \frac{13}{4}$. (b) Probability of the motif. Generate the set I_3 containing all degree triples with labels R , B and B . The probability of the motif is given as the sum of all probabilities of each occurrence times the probability of observing such node degrees given the labels times the probabilities of the labels.

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 position α . We deal with overlaps among motifs shortly. $Y_\alpha(m_C)$ 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 α 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 introduce the concept of non-redundant labeled permutations of an injective labeled motif. A labeled permutation of a motif m_C is a labeled motif resulting from a permutation of the nodes (and the corresponding labels) of m_C and the permutation is represented by its adjacency matrix plus the array of labels of its nodes. Two labeled permutations are non-redundant iff one of the following conditions hold: (i) their adjacency matrices are different; or (ii) their adjacency matrices are equal, but the arrays of labels are different.

We also denote with $\pi(m_C) = |R(m_C)|$ the number of Non-Redundant Permutations of m_C . We then have the following random variable: $N(m_C) = \sum_{\alpha} \sum_{m'_C \in R(m_C)} Y_{\alpha}(m'_C)$.

Thanks to the exchangeability assumption, each permutation of m_C has the same probability of occurrence. The expectation of the count of a labeled injective motif m_C with structure m and multiset of labels C in a graph G with N nodes is

$$\mathbb{E}[N(m_C)] = \binom{N}{k} \times \pi(m_C) \times \sigma(m_C) \quad (7)$$

where $\binom{N}{k}$ is the number of all possible locations of m_C in G and $\sigma(m_C)$ is the occurrence probability of the labeled motif m_C , according to the chosen random model.

We compute the variance of the number of occurrences of the labeled motif as $\mathbb{V}[N(m_C)] = \mathbb{E}[N^2(m_C)] - \mathbb{E}[N(m_C)]^2$. The expectation of $N^2(m_C)$ is computed considering that $N^2(m_C)$ can be expressed as:

$$\begin{aligned} N^2(m_C) &= \left(\sum_{\alpha} \sum_{m'_C \in R(m_C)} Y_{\alpha}(m'_C) \right)^2 = \\ &= \sum_{\alpha} \sum_{m'_C \in R(m_C)} \sum_{\alpha'} \sum_{m''_C \in R(m_C)} Y_{\alpha}(m'_C) \cdot Y_{\alpha'}(m''_C) \end{aligned}$$

Therefore, $\mathbb{E}[N^2(m_C)]$ is the sum over all positions of the probabilities of having $Y_{\alpha}(m'_C) \cdot Y_{\alpha'}(m''_C) = 1$. To compute these probabilities, we have to take into account the possibility that occurrences overlap. Two occurrences of a motif overlap if they share at least one node.

As suggested by [16], we define the concept of super-motif, which is a motif composed of 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 4 for an example). A super-motif inherits labels from the ancestor motifs. Due to node overlapping, one or more labels can overlap. Specifically, in the case of injective motifs, the overlapping has to take into account the node labels. Therefore when two motifs of size k overlap on s nodes, these nodes should share the same labels. This implies that motifs having nodes of the same labels in incompatible positions will not yield a super-motif. Figure 4 shows an example.

We can define an overlapping operations for two multi-sets of labels 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 labels assigned to the super-motif.

Therefore, the probability of observing a labeled super-motif generated from labeled motifs is the following:

$$\sigma(m'_C, m''_C, s) = \mu(m'_C \Omega_s m''_C | C \Pi_s C) \times \nu(C \Pi_s C)$$

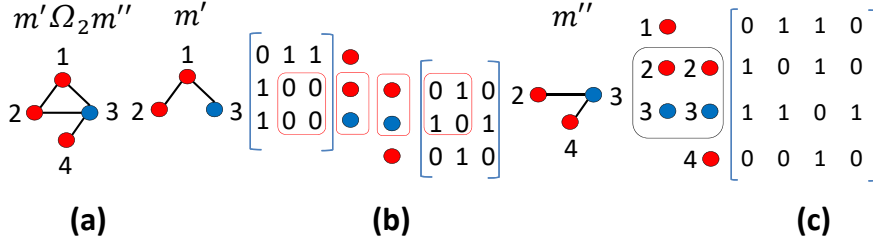


Fig. 4: Labeled Super-motif of a path of 3 nodes with overlap $s = 2$. (a) A super-motif of 4 nodes obtained from the overlapping of two non-redundant labeled motifs of 3 nodes sharing two nodes. (b) Two non-redundant permutations of a path with 3 nodes along with the corresponding adjacency matrices. In this case, overlaps require that the labels of the nodes be compatible. The overlapping involves two nodes, the labels of the last two nodes in the m' motif have to be the same (in an inverted order) of the first two nodes in the motif m'' . The overlapping regions are represented (highlighted in red) by the bottom right sub-matrix of m' and upper left sub-matrix of m'' . (c) The adjacency matrix of the super-motif. The overlapping is applied by using an OR operator on the overlapping entries of the m' and m'' sub-matrices.

where $\nu(C\Pi_s C)$ is computed as in equation 6.

The computation of variance is based on the expectation of the squared count of a labeled 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 amounts 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:

$$\mathbb{E}[N^2(m_C)] = \binom{N}{N-2k, k, k} \rho^2(m_C) \sigma^2(m_C) + \sum_{s=1}^k \binom{N}{k-s, s, k-s, N-2k+s} \sum_{m', m'' \in R(m_C)} \sigma(m'_C, m''_C, s) \quad (8)$$

where k is the number of nodes of motif m_C 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 .

5 Assessing Non-Induced Motif Significance

To establish whether a motif m_C is over-represented in a given graph, one needs to calculate the probability (i.e. p-value) $P[N(m_C) \geq N_{obs}(m_C)]$, where $N_{obs}(m_C)$ is the observed number of non-redundant occurrences of m_C and $N(m_C)$ is a random variable representing the number of occurrences of the motif in a graph generated according to the chosen reference model. The baseline approach to the approximation of $P[N(m_C) \geq N_{obs}(m_C)]$ relies on simulation through the usage of permutation tests. The method consists of generating a certain number of random networks and computing the number of occurrences of the motifs in such networks. The p-value is then approximated as the number of times the occurrences of the motif in the random networks exceeds the number of occurrences of the motif in the target network, divided by the number of random networks. Thus, the reliability of the p-value is strictly related to the number of randomizations performed.

To avoid such an expensive simulation, a key problem is to identify a proper distribution fitting the number of observations in the reference random model. Picard et al. [16] proposed a model for unlabeled graphs in which they showed that the Pólya-Aeppli (denoted by PA) distribution (also known as the Geometric-Poisson distribution) [26] is suitable to describe how the count of motif occurrences may vary and can be used as an approximation of the distribution of the count of $N(m_C)$.

Following [16], we observe that motifs come in clusters because they can overlap. Also, 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 constructed from the overlap 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 obtained from the intersection (i.e. overlap) of pairs of motifs. We can introduce a second random variable called X_2 in which we sample several times a cluster (notice that we can assume we have a distribution of clusters according to the overlap) 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 of each other. The PA distribution is obtained when the cluster size has a geometric distribution $G(1 - \alpha)$, yielding a mean size of a cluster of $1/(1 - \alpha)$.

In this case we have that $X \sim PA(\lambda, \alpha)$ is a random variable representing the number of observed events (i.e. motif occurrences in our case):

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

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 $\lambda = (1 - \alpha) \times \mathbb{E}[N(m_C)]$.

6 Experimental analysis

In this section we analyze the accuracy and speed of the analytical model in the identification of statistically significant motifs under the random EDD model. We make use of directed and undirected graphs of different sizes. To evaluate the quality of results, we compare the analytical p-values with those obtained through a simulation-based permutation test.

In several cases, the Polya-Aeppli (PA) distribution provides a better fit of the empirical distribution of motif counts in a sample of EDD graphs than the Gaussian distribution (see Supplementary materials for the experiments).

In terms of running time, we demonstrate that the analytical model usually vastly outperforms the simulation based method. The speed-up is greatest for non-induced motifs.

Finally, we report some examples of a few small significant and non-significant motifs that have been found in two large networks by using our analytical-based model: (i) an actor-actor collaboration network and (ii) a paper citation network.

In all the experiments we used the GLabTrie algorithm (see The supplementary material for a complete description of the algorithm) to count motifs occurrences in input and random EDD graphs.

The analytical model and GLabTrie have been implemented in Java and integrated in a software called FlashMotif. For purposes of reproducibility and for community use, we provide both source and jar executable of FlashMotif. The executable requires Java 8 (or more) and works on all platforms. The program and the source code are available at <http://alpha.dmi.unict.it/flashMotif/>. Experiments have been performed on an Intel Core i3-3240 CPU with 3.40 Ghz.

6.1 Dataset

We used a dataset of labeled graphs with nine real graphs and two artificial graphs. Table 1 describes their main characteristics.

ROGET graph is taken from the Roget's Thesaurus of 1879 and describes the associations between pairs of words of English that have similar meanings [28]. Nodes are category words and an edge connects two nodes iff they are directly related to each other in the book.

HAMSTERSTER is a graph of friendships between owners of hamsters at the website Hamsterster.com.

OPENFLIGHTS is a graph extracted from Openflights.org and represents all existing air routes between different airports around the world in 2011 [29].

Table 1: Database of colored graphs.

Graph	Orientation	Nodes	Edges	Node colors
ROGET	Undirected	1,010	3,648	6
HAMSTERSTER	Undirected	2,426	16,631	16
OPENFLIGHTS	Undirected	2,939	15,677	5
PPIHUMAN	Undirected	9,506	37,054	11
PPIYEAST	Undirected	2,617	11,855	13
NEURALWORM	Directed	279	2,990	3
POLBLOGS	Directed	1,224	19,022	2
DBLP	Directed	12,591	49,744	8
FOLDOC	Directed	13,356	120,238	14
ARTNETUNDIR	Undirected	2,000	6,000	8
ARTNETDIR	Directed	2,000	8,000	7

PPIHUMAN is a protein-protein interaction (PPI) network in human, taken from the HPRD database [30].

PPIYEAST is a PPI network in yeast, compiled by von Mering et al. [31], combining data taken from different sources (experimental techniques, correlations, genetic interactions and reference sets of known complexes). In order to include trusted and reliable interactions, the PPI network contains only PPIs with "high" and "medium" level of confidence, according to the quality assessment analysis performed by von Mering et al. in [31].

NEURALWORM is the complete neural network in worm and describes the synaptic connections between neuron cells [32]. POLBLOGS is a directed graph of hyperlinks between weblogs on US politics of 2004 [33].

DBLP is the directed citation network of DBLP, a database of scientific publications, where each node in the graph is a publication and an edge goes from A to B iff A cites B [34].

FOLDOC is a directed semantic network taken from the on-line computing dictionary FOLDOC (<http://foldoc.org>), where nodes are computer science terms and edges connect two terms X and Y iff Y is used to explain the meaning of X [35].

Nodes of each real graph have been annotated with the following labels. In the ROGET graph, nodes are labeled according to the existing classification of categories into 6 domains ('abstract relations', 'space', 'matter', 'intellectual faculties', 'voluntary powers' and 'sentiments'). Each domain maps to a specific label. Nodes of HAMSTERSTER have been annotated with 16 different species of hamsters. In OPENFLIGHTS, airports have been associated to one of the five continents.

For the labeling of nodes in PPIHUMAN, we used Gene Ontology (GO) [36], which is a multi-hierarchical dictionary of terms related to biological processes, cellular components and biological functions. GO is commonly used for the analysis of biological networks [37,38]. Each hierarchy of GO terms (e.g. the one for biological processes) is represented by tree data structures. We annotated proteins with GO processes down to the first level of the corresponding tree yielding 11 node labels. Ten of them represent specific kinds of biological processes such as: 'whole-organism process', 'metabolism', 'regulation',

'cellular organization', 'development', 'localization', 'signaling', 'response to stimulus', 'biological adhesion' and 'reproduction'. A special label representing the generic biological process has been associated to proteins for which we did not have GO annotations.

Proteins in PPIYEAST have been annotated using the MIPS functional catalogue FunCat [39]. In particular, we followed the functional annotation scheme designed by von Mering et al. in [31] where each protein is assigned to exactly one of the following categories: 'energy production', 'aminoacid metabolism', 'other metabolism', 'translation', 'transcription', 'transcriptional control', 'protein fate', 'cellular organization', 'transport and sensing', 'stress and defense', 'genome maintenance', 'cellular fate' and 'uncharacterized function'.

Neurons in NEURALWORM are labeled according to their putative roles ('sensory neurons', 'interneurons' and 'motor neurons'). In POLBLOGS nodes have been classified depending on their political leaning ('liberal' and 'conservative'). DBLP nodes has been annotated with different kinds of publications ('articles', 'inproceedings', 'proceedings', 'books', 'incollections', 'phd thesis' and 'master thesis') or 'www' if the node refers to a cited website. Computing terms in FOLDOC have been labeled according to their domains: 'jargon', 'computer science', 'hardware', 'programming', 'graphics' and 'multimedia', 'science', 'people and organizations', 'data', 'networking', 'documentation', 'operating systems', 'languages', 'software' and 'various terms'.

ARTNETUNDIR and ARTNETDIR are artificial graphs where nodes with the same degree have the same label. They have been generated, starting from a lattice graph, by iteratively removing random edges until the desired number of edges is obtained. We include artificial graphs because these graphs allow us to test the model in the color-topology-dependent case.

6.2 Accuracy of the model

The accuracy of the analytical model is determined by its ability to recover the same set of motifs of the simulation-based approach, given a p-value threshold for motif significance. We first computed all labeled subgraphs with 3 and 4 nodes both in the simulation-based approach and in the analytical model. For each real graph we ran eight different experiments, considering all possible definitions of motifs (Multiset and Injective, induced and non-induced) and EDD models (with label-degree dependence and independence) using directed and undirected graphs. As regards artificial graphs, we ran experiments with label-degree dependent only, since in these networks node labels and node degrees are dependent by construction. This led to a total of 160 different tests. The simulation-based p-value was established by using an ensemble of 1,000 random graphs generated under the EDD model. However, we can notice that p-values which are much lower or higher than the critical threshold (i.e. 0.05) are not so affected by the number of networks. Therefore, in principle, we could expect few differences between p-values with 500 or 1000 simulation. On

the other hand, p-values close to the critical threshold are clearly influenced by the number of networks. The choice of this number represents a good trade-off between the accuracy of the simulation-based p-value and the running time of the simulation-based approach.

Labeled motifs present in the input graphs are then classified into 'positives' and 'negatives', depending on whether their simulation-based p-values are lower or higher than a threshold P , respectively. In our experiments, we fixed a p-value P to be 0.05, a common value used to establish the statistical significance of a motif. To obtain the precision/recall curve we fixed the p-value for the simulation and varied the p-values for the analytical computation.

Intuitively, the simulation runs are taken to yield the gold standard and the question is how well do the analytical predictions match that gold standard. We denote as 'true positives' (TP) all the instances of graph motifs according to simulation that have a p-value less than P using the analytical model. We denote as 'true negatives' (TN) all the graph patterns that are not motifs according to simulation and have p-values greater than P in the analytical model. Corresponding definitions hold for both 'false positives' (FP) (analytical assigns low p-value but simulation assigns high p-value, so graph pattern is not a motif according to the simulation gold standard) and 'false negatives' (FN) (analytical assigns high p-value but simulation assigns low p-value).

For each of the 160 tests we plotted the Precision-Recall (PR) curve. In Figs. 5 and 6 we report the PR curves for different kinds of non-induced and induced motifs and different EDD models with 4 nodes in all graphs. PR curves for non-induced and induced motifs with 3 nodes are available as Supplementary Materials.

The results clearly show that the accuracy of the model is very high for all graphs and for all motif and EDD model definitions. The number of false positives and false negatives is generally very low, so there is a strong correlation between the results found using the analytical model as to the gold standard simulation-based approach.

6.3 Running times

We compared the performance of the analytical model to the simulation-based model in all 160 tests.

In Tables 2, 3, 4 and 5 we report, for each graph and for each motif and EDD model definition, the running time for the computation of all non-induced and induced multiset and injective topological labeled motifs of size k (with $k = 3, 4$), respectively.

The results show that the analytical model is faster than the simulation-based algorithm for all non-induced motifs, regardless of motif and EDD model definitions, usually by orders of magnitude. Regarding induced motifs, the analytical model is very fast in the case of undirected graphs but can be slower than the simulation-based algorithm in the case of directed graphs, especially when the number of node labels is high (e.g. FOLDOC). This represents the

Table 2: Running times (secs) of analytical model-based algorithm vs simulation-based algorithm for the computation of non-induced multiset labeled motifs (D=label-degree dependent, I=label-degree independent). Artificial networks have been tested only in the label-degree dependent model.

Graph	k	Simulation-based algorithm		Analytical-based algorithm	
		I-Multiset	D-Multiset	I-Multiset	D-Multiset
ROGET	3	27.92	27.22	0.03	0.03
	4	175.52	206.73	0.28	0.76
HAMSTERSTER	3	283.18	257.05	0.35	0.43
	4	11550.39	9824.35	19.08	30.97
OPENFLIGHTS	3	380.10	339.97	0.48	0.51
	4	16345.06	12922.77	27.24	43.00
PPIHUMAN	3	3413.27	3360.45	2.16	2.75
	4	25736.71	20846.42	34.66	46.69
PPIYEAST	3	173.07	0.73	335.69	1.02
	4	4967.54	6.89	202192.80	235.45
NEURALWORM	3	23.19	25.48	0.05	0.08
	4	782.68	2365.53	4.83	12.18
POLBLOGS	3	374.59	325.71	0.74	1.12
	4	34755.42	30465.09	131.38	234.78
DBLP	3	7639.44	7559.17	4.29	4.41
	4	60142.39	53544.69	111.39	140.00
FOLDOC	3	9352.80	9269.15	5.76	7.38
	4	95923.70	132947.10	300.57	902.32
ARTNETUNDIR	3	-	109.67	-	0.14
	4	-	191.78	-	0.57
ARTNETDIR	3	-	155.13	-	0.13
	4	-	1085.88	-	3.95
Avg performance ratio of anal. vs simul.				724x	588x

worst case for the analytical model, since it needs to compute several Kocay matrices (discussed in the supplementary materials), one for each possible combination of labels, and many terms in the variance and the covariance computation, whose number depends on the count of topologies of a given size. Considering all possible tests, the analytical model is on average 650 times faster than the simulation-based model (last row of Tables 2, 3, 4 and 5).

6.4 Scalability Tests: Actors in Movies and Paper Citations

To test the scalability of our framework we applied FlashMotif on two big networks: an actor collaboration network and a paper citation network.

The actor collaboration network was built from the IMDB database (<http://www.imdb.com>). Nodes are actors and links connect pairs of actors who acted together in at least one movie.

Data about actors, movies and movie genres were downloaded as flat files. We decided to focus only on movies made in the United States. Two actors

Table 3: Running times (secs) of analytical model-based algorithm vs simulation-based algorithm for the computation of non-induced injective labeled motifs (D=label-degree dependent, I=label-degree independent). Artificial networks have been tested only in the label-degree dependent model.

Graph	k	Simulation-based algorithm		Analytical-based algorithm	
		I-Injective	D-Injective	I-Injective	D-Injective
ROGET	3	26.91	26.78	0.03	0.03
	4	206.40	195.36	0.64	0.64
HAMSTERSTER	3	256.22	255.94	0.31	0.35
	4	10081.02	9647.71	20.83	20.87
OPENFLIGHTS	3	342.15	339.95	0.37	0.37
	4	13084.28	12736.22	26.43	26.36
PPIHUMAN	3	3362.08	3364.48	2.55	2.23
	4	21034.03	20590.24	33.38	33.55
PPIYEAST	3	470.30	1.14	366.68	1.19
	4	428612.79	215.62	210876.81	215.50
NEURALWORM	3	26.25	25.10	0.05	0.06
	4	3032.65	2353.22	9.83	9.87
POLBLOGS	3	325.48	327.74	0.70	0.71
	4	30363.26	30263.51	130.06	131.38
DBLP	3	7565.49	7558.59	4.37	4.09
	4	52601.67	53486.89	101.11	101.31
FOLDOC	3	9239.93	9243.77	6.50	6.36
	4	130550.67	129241.64	820.60	824.58
ARTNETUNDIR	3	-	93.57	-	0.11
	4	-	186.16	-	0.47
ARTNETDIR	3	-	154.29	-	0.09
	4	-	1078.60	-	3.83
Avg performance ratio of anal. vs simul.				673x	725x

are linked if they acted together in at least one USA movie. The resulting undirected network consists of 1,283,456 nodes and 54,272,070 edges.

Nodes of the actor collaboration network were then labeled according to the genre of the movies in which the actor mostly acted. For example, an actor was labeled with 'comedy' if he or she mostly acted in comedies. The network contains 29 labels, representing different movie genres, plus one special label for the 'undefined' genre.

The paper citation network has been extracted from the SciMAG 2015 open data set [40,41], which integrates citation data from the Microsoft Academic Graph [41] with paper annotation data coming from the SciMago classification of academic journals (<http://www.scimagojr.com>). In such a directed network, a paper A is linked to a paper B if A cites B.

To build the network, we considered papers published from 2000 to present days and annotated with at least one of the following SciMago categories: 'Computer Science', 'Economics', 'Engineering', 'Mathematics', 'Medicine' and 'Physics and Astronomy'. The resulting network has 9,042,661 nodes and 71,191,166 edges.

We used FlashMotif to count and estimate the significance of all labeled cliques of 3 and 4 nodes in the two networks. For the directed network of paper

Table 4: Running times (secs) of analytical model-based algorithm vs simulation-based algorithm for the computation of induced multiset labeled motifs (D=label-degree dependent, I=label-degree independent). Artificial networks have been tested only in the label-degree dependent model.

Graph	k	Simulation-based algorithm		Analytical-based algorithm	
		I-Multiset	D-Multiset	I-Multiset	D-Multiset
ROGET	3	27.90	26.52	0.02	0.03
	4	175.62	155.15	0.19	3.91
HAMSTERSTER	3	283.50	252.23	0.30	0.41
	4	11427.55	8842.34	14.21	56.63
OPENFLIGHTS	3	379.11	340.70	0.36	0.45
	4	16381.74	12899.35	18.41	24.62
PPIHUMAN	3	3422.87	3362.88	2.58	2.78
	4	25662.37	20225.28	31.66	75.88
PPIYEAST	3	184.31	0.42	189.94	0.81
	4	4342.26	4.07	3806.38	59.93
NEURALWORM	3	23.04	20.41	0.03	0.08
	4	779.08	682.28	14.69	509.97
POLBLOGS	3	375.09	323.70	0.50	0.64
	4	34911.58	29126.04	56.39	183.10
DBLP	3	7639.00	7560.25	4.00	4.60
	4	59834.26	53181.85	118.61	4243.94
FOLDOC	3	9340.76	9174.23	5.21	9.96
	4	96164.98	87260.57	144.02	155903.10
ARTNETUNDIR	3	-	92.26	-	0.08
	4	-	172.69	-	3.52
ARTNETDIR	3	-	150.89	-	0.52
	4	-	395.96	-	7019.75
Avg performance ratio of anal. vs simul.				1034x	472x

citations, a clique is defined as any topology where all the nodes are connected to one another, regardless of the direction of the edges. We considered induced injective motifs with label-degree dependency.

FlashMotif took about 26 minutes to retrieve all 2,787 3-cliques present in the actor network and 24 seconds to compute the analytical p-values. The counting of all 13,649 4-cliques took about 39 hours whereas the analytical p-values were computed in 1.3 minutes. For the paper citation network, FlashMotif took 7.7 minutes and 25.7 minutes to count all 658 3-cliques and 3,979 4-cliques, respectively. Analytical p-values for 3-cliques and 4-cliques were computed in 3.5 seconds and about 169 minutes, respectively.

The complete set of labeled cliques in both networks is available as Supplementary Material. In Tables 6 and 7 we report a few examples of significant and non-significant cliques that we found in our experiments. The table reports, for each clique, the number of occurrences in the input network, the expected number of occurrences in random graphs according to the EDD model and the p-value. We also report the time needed to check the significance of the single labeled motif by using the analytical-based model and the estimation of the time needed by the simulation-based model with 1,000 EDD random graphs. The time for the simulation has been estimated by multiplying the time needed

Table 5: Running times (secs) of analytical model-based algorithm vs simulation-based algorithm for the computation of induced injective labeled motifs (D=label-degree dependent, I=label-degree independent). Artificial networks have been tested only in the label-degree dependent model.

Graph	k	Simulation-based algorithm		Analytical-based algorithm	
		I-Injective	D-Injective	I-Injective	D-Injective
ROGET	3	26.29	26.11	0.03	0.03
	4	145.61	144.87	3.83	3.86
HAMSTERSTER	3	250.70	251.30	0.30	0.34
	4	8787.20	8735.41	51.17	51.31
OPENFLIGHTS	3	341.44	339.43	0.34	0.33
	4	13011.54	12751.47	17.79	17.87
PPIHUMAN	3	3366.29	3363.07	2.55	2.55
	4	20029.44	20459.98	65.08	65.24
PPIYEAST	3	173.55	0.96	184.49	0.99
	4	3790.26	58.02	3638.92	56.83
NEURALWORM	3	20.01	20.15	0.07	0.07
	4	648.55	649.20	506.15	512.08
POLBLOGS	3	324.76	326.32	0.48	0.47
	4	29315.97	29130.96	167.71	169.42
DBLP	3	7555.49	7562.57	4.30	4.27
	4	51616.89	52238.83	4172.64	4212.15
FOLDOC	3	9166.24	9183.45	9.84	9.27
	4	82440.02	83560.30	154798.84	155974.89
ARTNETUNDIR	3	-	91.86	-	0.08
	4	-	167.53	-	3.50
ARTNETDIR	3	-	150.17	-	0.50
	4	-	386.33	-	7017.91
Avg performance ratio of anal. vs simul.				525x	531x

for a single random graph by the number of random graphs (1,000). Notice that such time includes the generation of each random graph and the computation of relative motif frequency. Whereas this operation could be parallelized, it represents a formidable amount of computation.

As expected, actors of the same genre tend to appear together, forming dense communities in the network and a high number of colored 3-cliques and 4-cliques, that are consequently significant as motifs. Likewise, papers of the same (or similar) categories are very frequent and significant as motifs. However, we found many interesting examples of recurrent patterns in which actors of different genres and papers of different disciplines appear together. For instance, drama and comedy actors are very often linked to one another: this can be explained by the fact that i) several actors acting in comedies also act in many dramatic movies and vice versa, ii) 'comedy' and 'drama' are interchangeable genres and many movies can be comedies and dramas at the same time. In the paper citation network, an interesting example is represented by a directed clique with papers published in Medicine and Economics areas. We also found some under-represented motifs. For example, actors of adult movies tend to work together and seldom appear with other actors. In the

Table 6: Examples of significant and non-significant 3-cliques and 4-cliques motifs found in the actor collaboration network. P-values below the significance threshold (0.05) are highlighted in bold. Label map: D=Drama, F=Family, ADU=Adult, CO=Comedy, N=News, W=Western, H=Horror, ADV=Adventure, AC=Action, T=Thriller, CR=Crime, M=Mystery.

Clique labels	Occurrences	Mean EDD	P-value	Analytical time	Simulation time
D, F, ADU	47	131.875	0.955	1.88 secs	\simeq 221 days
CO, N, W	1	10.122	0.982	1.86 secs	\simeq 221 days
H, D, CO	14,404,872	580,375.691	0.0	2.45 secs	\simeq 221 days
ADV, AC, T	13,453	367,417	1.92E-14	1.82 secs	\simeq 221 days
CO, F ADU, W	3	3.391	0.367	1.97 secs	\simeq 221 days
D, ADU ADU, W	181	74.81	0.149	1.90 secs	\simeq 221 days
D, CO CO, CO	26,712,605,349	5,624,541.844	0.0	3.64 secs	\simeq 226 days
CR, M ADV, W	2,337,597	0.102	1.11E-16	1.91 secs	\simeq 221 days

paper citation network, cliques with Medicine and Physics papers appear to be underrepresented.

Once again, the running times clearly show that the analytical model outperforms simulation-based model. In this experiments, the difference between the two approaches is 7 order of magnitude. Since the networks are very big, the generation of a single random graph requires several hours and this impacts the total running time, even though the computation of frequencies is very fast. The analytical model is extremely fast and its time is independent of the size of the network.

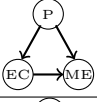
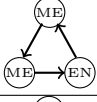
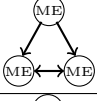
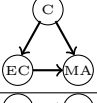
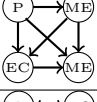
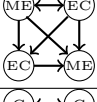
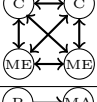
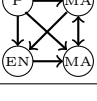
7 Conclusions

The labels of nodes convey meaning just as their relationships do. Node meaning may have to do with political party, physical location, type of actor, or chemical constituent, just to mention a few possibilities. Relationships are characterized by edges.

Finding motifs in such graphs corresponds to finding relationships among labeled nodes that would not be expected at random for similar graphs. The simulation approach to find motifs is to generate random graphs and count. This counting task demands the majority of the time in any motif calculation.

This paper has extended previous works in unlabeled graphs and topology-agnostic labeled graphs [16, 25] to find fast analytical models for labeled motifs that encode topology. Our model handles (i) directed and undirected graphs, (ii) induced and non-induced motifs, (iii) label-dependent valences and label-independent valences, and (iv) two models of labeled motifs. Across a wide variety of real and simulated graphs, our analytical approach is vastly

Table 7: Examples of significant and non-significant 3-cliques and 4-cliques motifs found in the paper citation network. P-values below the significance threshold (0.05) are highlighted in bold. Label map: C=Computer Science, ME=Medicine, EC=Economics, EN=Engineering, MA=Mathematics, P=Physics and Astronomy.

Motif	Occurrences	Mean EDD	P-value	Analytical time	Simulation time
	65	1,568.923	0.859	2.70 secs	\simeq 3167 days
	6	34.103	0.999	2.67 secs	\simeq 3168 days
	176,687	0.266	3.11E-15	2.59 secs	\simeq 3168 days
	1,977	1.881	0.0	2.72 secs	\simeq 3167 days
	17	9.957	0.176	94.24 secs	\simeq 3168 days
	249	4.26E-8	1.11E-16	44.57 secs	\simeq 3167 days
	2,814	2.81E-21	0.0	12.74 secs	\simeq 3167 days
	922	1.36E-7	0.0	60.58 secs	\simeq 3167 days

faster (usually hundreds of times faster) than the simulation approach for non-induced motifs and generally faster for induced subgraphs as well.

Our methods are based on the Expected Degree Distribution Model (EDD), which preserves the degree distribution of the original network. The EDD model has also been showed to be capable to generate the right type of triangles for certain kind of real networks [42]. Although degree-distribution-based approaches are the most widely recognized for measuring significance of network motifs, they do have some limitations when applied as null model in certain kind of real networks [5,42]. In the future we will explore other more realistic graph models in connection with our labeled graph analytical model, like the Block Two-Level Erdős-Rényi (BTER) model and Exponential Random Graph models.

Our software is available at <http://alpha.dmi.unict.it/flashMotif/>.

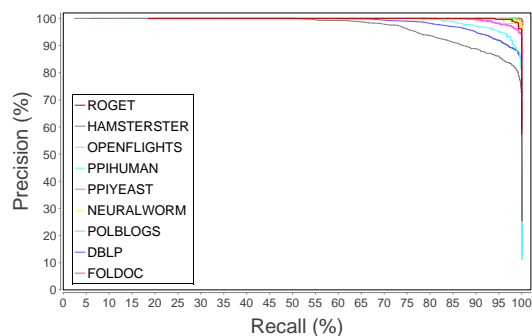
Acknowledgments

The authors would like to thank Simone Severini for insightful discussions.

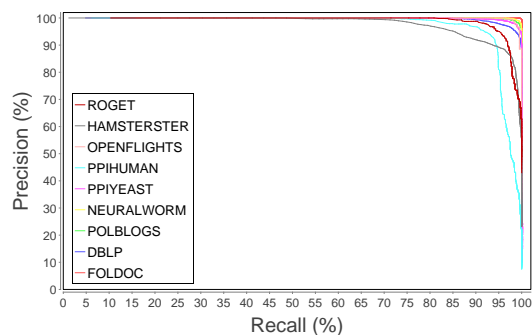
References

1. J. Chen and B. Yuan, "Detecting functional modules in the yeast protein-protein interaction network", *Bioinformatics* 22(18), pp. 2283-2290, 2006.
2. R. Milo, S. Shen-Orr, S. Itzkovitz et al., "Network motifs: simple building blocks of complex networks", *Science* 298(5594), pp. 824-827, 2002.
3. P. Erdos and A. Renyi, "On random graphs", *Publicationes Mathematicae* 6, pp. 290-297, 1959.
4. M. E. J. Newman, S. H. Strogatz and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications", *Phys. Rev. E* 64, 026118, 2001.
5. C. Seshadhri, T. G. Kolda, and A. Pinar, "Community structure and scale-free collections of Erdos-Renyi graphs", *Phys. Rev. E*, 85(5), 2012.
6. F. Chung and L. Lu, "The average distances in random graphs with given expected degrees", *Proc. Natl. Acad. Sci.* 99(25), pp. 15879-15882, 2002.
7. J. Park and M. Newman, "The origin of degree correlations in the internet and other networks", *Phys. Rev. E* 68, 026112, 2003.
8. K. Nowicki and T. Snijders, "Estimation and prediction for stochastic block structures", *J. Am. Statist. Assoc.* 96, pp. 1077-1087, 2001.
9. J. J. Daudin, F. Picard and S. Robin, "A mixture model for random graphs", *Statistics and Computing* 18(2), pp. 173-183, 2008.
10. J. Park, and M.E.J. Newman. "Statistical mechanics of networks", *Physical Review E*, 70(6), 066117, 2004.
11. R. Milo, N. Kashtan, S. Itzkovitz et al., "On the uniform generation of random graphs with prescribed degree sequences", *Cond. Mat.* 0312028, pp. 1-4, 2004.
12. R. Prill, P. A. Iglesias and A. Levchenko, "Dynamic properties of network motifs contribute to biological network organization", *PLoS Biology* 3(11), 2005.
13. S. S. Shen-Orr, R. Milo, S. Mangan et al., "Network motifs in the transcriptional regulation network of *Escherichia coli*", *Nat. Genet.* 31, pp. 64-68, 2002.
14. S. Wernicke, "Efficient detection of network motifs", *IEEE/ACM Trans. Comp. Biology and Bioinformatics* 3(4), pp. 347-359, 2006.
15. A. L. Barabasi and R. Albert, "Emergence of scaling in random networks", *Science* 286(5439), pp. 509-512, 1999.
16. F. Picard, J. J. Daudin, M. Koskas et al., "Assessing the exceptionality of network motifs", *Journal of Comp. Biol.* 15(1), pp. 1-20, 2008.
17. T. Squartini and D. Garlaschelli, "Analytical maximum-likelihood method to detect patterns in real networks", *New Journal of Physics* 13(8), 083001, 2011.
18. P. Ribeiro and F. Silva, "G-Tries: a data structure for storing and finding subgraphs", *Data Mining and Knowledge Discovery* 28(2), pp. 337-377, 2014.
19. L. A. A. Meira, V. R. Maximo, A. L. Fazenda and A. F. D. Conceicao, "Acc-Motif: Accelerated Network Motif Detection", *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 11(5), pp. 853-862, 2014.
20. J. Chen, W. Hsu, M. L. Lee, S. Ng, "NeMoFinder: Dissecting Genome-wide Protein-protein Interactions with Meso-scale Network Motifs", *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 106-115, 2006.
21. N.K. Ahmed, J. Neville, R. A. Rossi, N. G. Duffield and T. L. Willke, "Graphlet decomposition: framework, algorithms, and applications", *Knowledge and Information Systems* 50(3), pp. 689-722, 2017.

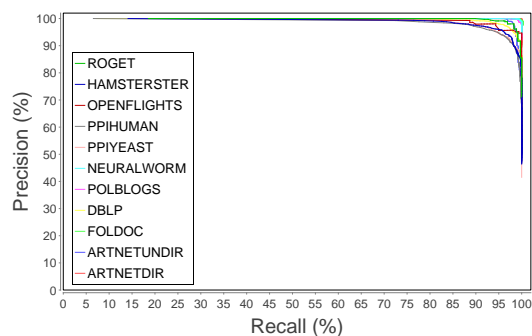
22. J. J. Pfeiffer III, S. Moreno, T. La Fond, J. Neville and B. Gallagher, "Attributed Graph Models: Modeling Network Structure with Correlated Attributes", *Proc. of the 23rd International Conference on World Wide Web*, pp. 831-842, 2014.
23. M. Kim and J. Leskovec, "Modeling Social Networks with Node Attributes Using the Multiplicative Attribute Graph Model", *Proc. of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 400-409, 2011.
24. M. Davis, W. Liu, P. Miller, R. F. Hunter and F. Kee, "Agwan: A Generative Model for Labelled, Weighted Graphs", *New Frontiers in Mining Complex Patterns: Second International Workshop, NFMCP 2013*, pp. 181-200, 2014.
25. S. Schbath, V. Lacroix and M. F. Sagot, "Assessing the exceptionality of coloured motifs in networks", *Journal on Bioinf. Syst. Biol.* 2009(1):616234, 2009.
26. N. L. Johnson, S. Kotz and A. W. Kemp, "Univariate discrete distributions", Second Edition, *Wiley*, New York, 1992.
27. M. Mongiovi, G. Micale, A. Ferro, R. Giugno, A. Pulvirenti and D. Shasha, "GLabTrie: a data structure for motif discovery with constraints", *EDBT Summer School 2015, Graph Data Management*, Springer (in press), 2015.
28. D. E. Knuth, "The Stanford GraphBase: a platform for combinatorial computing", *ACM Press*, New York, 1993.
29. T. Opsahl, "Why anchorage is not (that) important: binary ties and sample selection", <http://toreopsahl.com/2011/08/12>, 2011.
30. T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar et al., "Human protein reference database - 2009 update", *Nucleic Acids Research*, 37(1), pp. D767-D772, 2009.
31. C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields and P. Bork, "Comparative assessment of large-scale data sets of protein-protein interactions", *Nature*, 417, pp. 399-403, 2002.
32. L. R. Varshney, B. L. Chen, E. Paniagua et al., "Structural properties of the Caenorhabditis Elegans neuronal network", *PLoS Comput. Biol.*, 7(2): e1001066, 2011.
33. L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: divided they blog", *Proc. 3rd Int. Workshop on Link Discovery*, pp. 36-43, ACM, New York, 2005.
34. M. Ley, "The DBLP computer science bibliography: evolution, research issues, perspectives", *Proc. Int. Symp. on String Proc. and Inf. Retr.*, 2476, pp. 1-10, 2002.
35. V. Batagelj, M. Mrvar and M. Zavesnik, "Network analysis of dictionaries", *Language Technologies*, pp. 135-142, 2002.
36. M. Ashburner, C. A. Ball, J. A. Blake, et al., "Gene ontology: tool for the unification of biology", *Nature Genetics*, 25(1), pp. 25-29, 2000.
37. S. Maere, K. Heymans and M. Kuiper, "BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks", *Bioinformatics*, 21(16), pp. 3448-3449, 2005.
38. G. Bindea, B. Mlecnik, H. Hackl et al., "ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks", *Bioinformatics*, 25(8), pp. 1091-1093, 2009.
39. A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Gldener, G. Mannhaupt, M. Mnsterktter and H. W. Mewes, "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes", *Nucleic Acids Res*, 32(18), 2004.
40. M. De Domenico, E. Omodei and A. Arenas, "Quantifying the Diaspora of Knowledge in the Last Century", *Applied Network Science* 1, 15, 2016.
41. A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B. Hsu and K. Wang, "An Overview of Microsoft Academic Service (MAS) and Applications", *In Proc. of the 24th International Conference on World Wide Web (WWW 15 Companion)*, pp. 243-246, 2015.
42. N. Durak, A. Pinar, T.G. Kolda, and C. Seshadhri, "Degree relations of triangles in real-world networks and graph models", *In Proc. of the 21st ACM international conference on Information and knowledge management (CIKM'12)*, pp. 1712-1716, 2012.



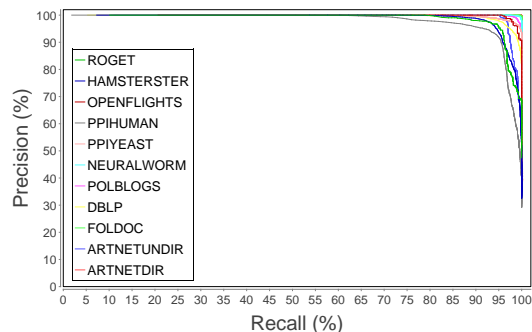
(a) Non-induced multiset motifs in EDD label-degree independent model



(b) Non-induced injective motifs in EDD label-degree independent model

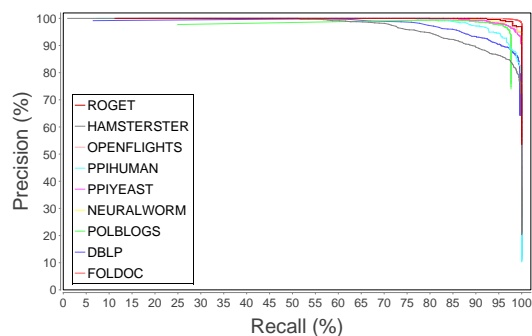


(c) Non-induced multiset motifs in EDD label-degree dependent model

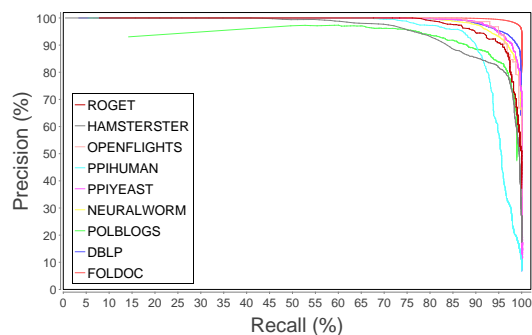


(d) Non-induced injective motifs in EDD label-degree dependent model

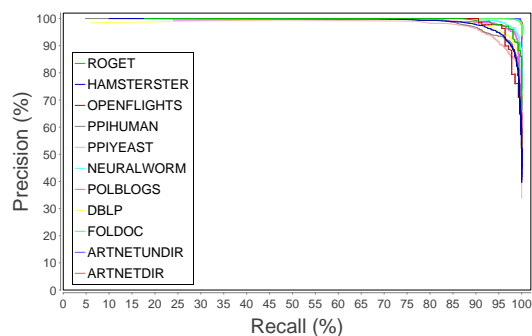
Fig. 5: Precision-Recall curves for simulation-based and analytical p-values for non-induced motifs with 4 nodes. These are nearly perfect curves, showing that across network types and query types, the analytical model yields essentially the same results as the simulation model.



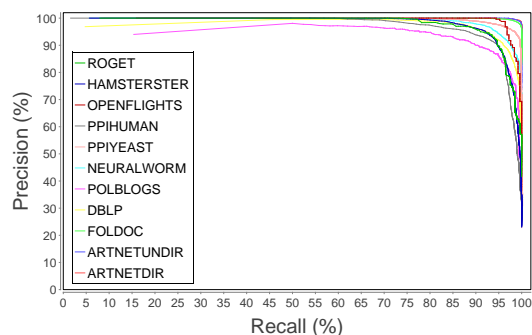
(a) Induced multiset motifs in EDD label-degree independent model



(b) Induced injective motifs in EDD label-degree independent model



(c) Induced multiset motifs in EDD label-degree dependent model



(d) Induced injective motifs in EDD label-degree dependent model

Fig. 6: Precision-Recall curves for simulation-based and analytical p-values for induced motifs with 4 nodes. These are nearly perfect curves, showing that across network types and query types, the analytical model yields essentially the same results as the simulation model.