

Fast Analytical Methods for Finding Significant Colored Graph Motifs

Dennis Shasha

Joint work with Alfredo Ferro, Rosalba Giugno, Misael
Mongiovi, Giovanni Micale, and Alfredo Pulvirenti

July, 2016

Problem and Example Applications

- 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 a subgraph may indicate:
 - a small scale interaction feature in genomic interaction networks
 - an intriguing relationship involving rock musicians and fans
 - connections among airports.

The Simulation Approach to Evaluating p-values

Given a topological pattern m on an input network G , the common approach to determining whether m is a motif works as follows:

- ① Generate a large set of random networks sharing the characteristics of G .
- ② Find the number of occurrences of m in each of those networks.
- ③ Estimate the p-value by comparing the number of occurrences in the input network with the numbers in the random networks.

Pro and Con of Simulation Method

- Simulation-based method 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, much more expensive say than just finding subgraphs that appear more than 5 times in the target graph which is the data mining approach.
- Conclusion: Worthwhile to find analytical methods that can be hundreds of times faster.

Take a step back: Random models

- The significance of a motif is always evaluated with respect to a reference random model.
- Random graphs are generated such that they preserve some characteristics of a network, typically the degree distribution.

Examples of Random Models

- The Erdős-Renyi model (ER model):
 - 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 .
- The Fixed degree distribution model (FDD model):
 - random graphs are generated by swapping edges starting from the input network G , guaranteeing that each node in each random graph has the same degree as in G .
- The Expected degree distribution model (EDD model):
 - node degree distribution of random graphs have the same expectation as the input network G (more tractable than FDD).

Research directions

- Over the last few decades, researchers have worked on replacing simulation by analytical methods.
- For uncolored/unlabel motifs, approximation methods, based on the Erdős-Renyi (ER) model, have computed the asymptotic normality of the distribution of topology counts.
- Unfortunately, empirical evidence suggests that the Erdos-Renyi random model offers a poor fit for many real-world networks.

Seminal work on topological motifs

- Picard et al. proposed a model to exactly compute the mean and variance of the count of a given pattern of unlabeled/uncolored nodes under any exchangeable random graph model.
- *Exchangeability*: 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 to deal with objects occurring in clusters which makes additional assumptions (which should be justified empirically):
 - the number of clusters follow a Poisson distribution
 - the number of objects per cluster has a geometric distribution.
- However, topology without labels/colors is limited. Labels can be important (male/female, transcription factor/ metabolite). Finding colored/labeled motifs can lead to important insights.

Types of colored motifs

- To deal with colored motifs we need to generalize the uncolored motif definition based on constraints that can be defined on the topology, on the color label assignment, or both.
- This leads us to three different definitions of motifs which are hierarchically related.

Multiset colored motifs

- Schbath et al. define a motif as any connected topology of k nodes having a given multiset of colors C .
- The authors proposed an analytical approach for assessing the exceptionality of multiset colored motifs.
- Exact analytical model for the mean and the variance of the count of a colored motif using the Erdős-Rényi (ER) random graph model.

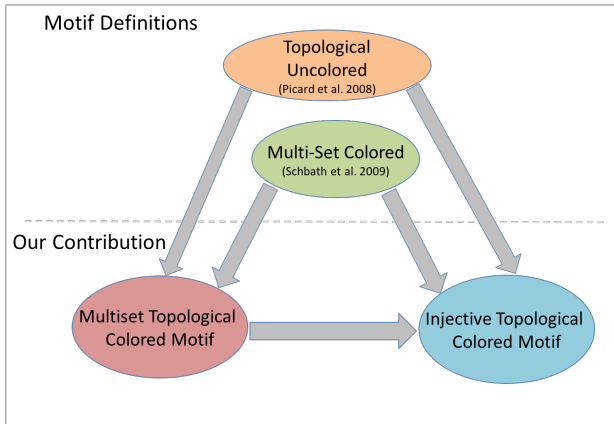
What We Bring to the Table

- There is work that finds analytical models for particular topologies and work that looks for particular combinations of colors regardless of topology. We want to combine the two.
- For the work concerning colors, we want to treat the case where degree and color are related (e.g. carbon atoms have a valence of four, whereas hydrogen only one) and where not.
- We want to treat directed/undirected graphs, induced/non-induced motifs, various mappings of color to nodes.
- More realistic random model than Erdős-Rényi: Expected Degree Distribution (EDD model).

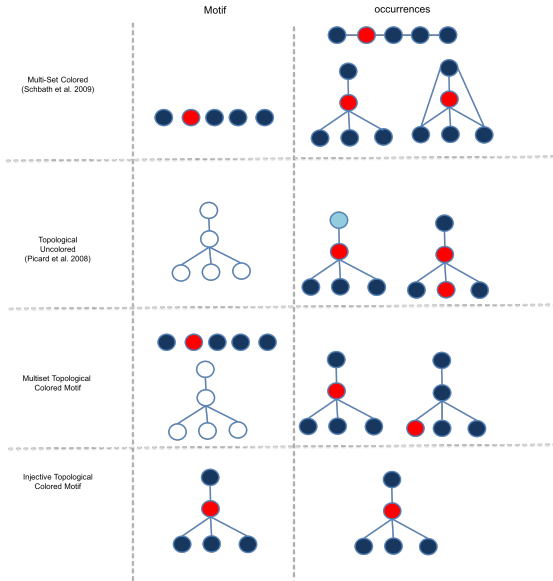
Colored Motif Models

We propose two new definitions of colored motif:

- *Multiset Topological colored motifs*
 - A subgraph of k nodes with a given topology having nodes belonging to a multiset of colors M , e.g. a star configuration of one rock musician and four fans. This is close to the Schbath definition but we include topology.
- *Injective topological colored motif*
 - A topology and a specific color assignment to each node in the topology. In this case a motif is a subgraph of k nodes having colors assigned to each node of a given topology, e.g. a star configuration of one rock musician and four fans, with a rock musician at the center.



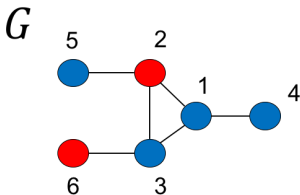
Motif hierarchy. Four different definitions of motifs. When definition A points to definition B, the set of motifs responding to a query according to A is a superset of those responding to a query according to B. All the definitions apply to both directed and undirected graphs.



Example of Motif occurrences within the motif hierarchy.

Colored graph

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 $c : V \rightarrow C$ is a function that assigns a color to each node in V . G is undirected means that if $\forall (u, v) \in E$, then $(v, u) \in E$, i.e. all neighbor relationships go both ways.



$$C = \{\text{red, blue}\}$$

Multiset Topological Colored Motif [formalization]

Let $G(V, E, C, c)$ be a colored 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 colored subgraph (induced or non-induced) of G having C_m as the multiset of node colors of the nodes V_m .

Let $N_{obs}(m)$ be the number of topologically isomorphic non-redundant occurrences of m in G having the same multiset of colors 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 .

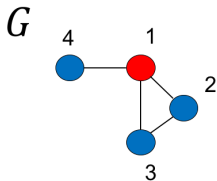
Injective Topological Colored Motif [formalization]

Let $G(V, E, C, c)$ be a colored 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 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 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 .

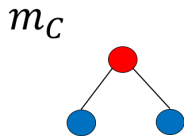
Two kind of motifs: Induced and Non-Induced (e.g. for a path)



Induced occurrences:

4-1-2

4-1-3



Non-Induced occurrences:

4-1-2

4-1-3

2-1-3

Expected Degree Distribution (EDD) Random Model

EDD on graphs where node colors are independent of node degrees

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

Road Map [lots of moving parts]

- **Find the probability of the occurrence of a non-induced colored motif at a given position in the graph.**
- Find expected number of instances of that motif in the whole graph. Also variance.
- Pólya-Aeppli model to calculate p-value.
- Kocay mapping in order to handle induced patterns.
- Experiments evaluating accuracy and performance.

Throughout: generalize to different motif definitions, to directed graphs and to the case where colors influence degrees..

Probability of an edge in the Expected Degree Distribution Model

- An edge between two nodes i and j (with $i \neq j$) is generated with probability:

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

where $\gamma = 1 / [(N - 1) \times \mathbb{E}[Deg]]$ and $D(i)$ is the degree of node i within the input graph. Thus, γ is approximately the number of edges. When $i = j$ we have $P(i, j) = 0$.

- We won't use the min function later, because ultimately we want to know how many instances of the motif there are so even if this probability becomes greater than one, that will not matter in the global estimate.
- We have $N - 1$ rather than N because there are no self-loops.

Finding Probability of an Occurrence at a Position in a Graph: example

- Let's say we have a graph with degrees 1, 2, ..., maxdeg
- Suppose you want to compute the probability of observing a topology of three nodes in a graph that resembles the input graph, for example a path X-X-X.
- We have the following list of degree assignments:
 - 1,1,1
 - 1,2,1
 - ...
 - 1,2,3
 - 1,2,4
 - 1,2,5
 - ...
 - maxdeg, maxdeg, maxdeg

Occurrence Probability Example Continued

- Generalizing for any degree assignments i, j, k we have terms to sum up of the form
 $P(i) \times P(j) \times P(k) \times \gamma \times i \times j \times \gamma \times j \times k$ Note that the middle term appears twice.
- Consider now $E(deg) \times E(deg^2) \times E(deg)$

$$\left[\sum_i^{maxdegree} P(i) \times i \right] \times \left[\sum_j^{maxdegree} P(j) \times j \times j \right] \times \left[\sum_k^{maxdegree} P(k) \times k \right]$$

- The cross terms describe a sum corresponding to each degree combination.
- Note that certain impossible cases (e.g. degrees of 1, 1, 1 for a path of length two) are still here so this is an approximation [Chung and Lu, 2002].

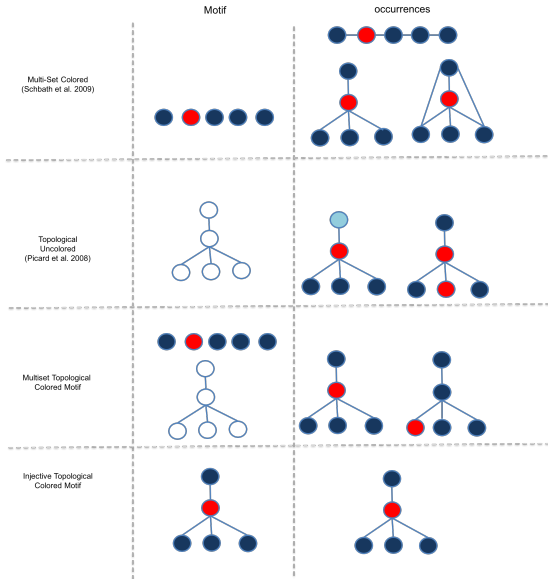
Occurrence Probability of Non-induced subgraph at a Position: general case

Occurrence Probability for General Potential Motif

Generalizing from our example on three nodes, we obtain a product of moments of the Deg distribution:

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

m_{++} is twice the total number of edges in a candidate motif m , m_{u+} is the valence of node u in m and $\mathbb{E}[Deg^{m_{u+}}]$ is the m_{u+} -th moment of distribution Deg . k is the number of nodes in m .



Have to determine how to distribute the colors in the bottom two cases.

Probability of observing motif colors: Multiset motif model

- In the multiset motif model (e.g. we just care that there is one rock musician and four fans) the color assignment to nodes is independent of their degrees.
- The assignment of colors in C to the k nodes of m_C follows a multinomial distribution:

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

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

Probability of observing motif colors: Injective motifs

For an injective topological colored motif (e.g. where we care about where the rock musician is), the probability of observing the color assignment to nodes of m_C is:

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

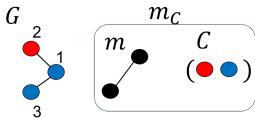
where $P(c_u)$ is the probability of observing the color c_u of motif node u in the graph.

Colored Motif Probability: Multiset and Injective




The probability of observing the colored motif m_C is probability of finding the topology times the probability of finding those colors on that topology:

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

Concrete Example to Show Calculation [self-study]



(a)

i) 	$I_2 = \{(1,2), (2,1), (1,1), (2,2)\}$ $\mu(m) = \sum_{(d_i, d_j) \in I_2} P(\text{Deg} = d_i) P(\text{Deg} = d_j) \gamma d_i d_j$ $= \gamma^{m+1/2} \prod_{u=1}^2 \mathbb{E}(\text{Deg}^{m_u}) = \frac{2}{3}$
ii) 	$v(C) = \frac{2!}{1!1!} \times \frac{1}{3} \times \frac{2}{3} = \frac{4}{9}$
iii) 	$\sigma(m_C) = v(C_m) \times \mu(m) = \frac{4}{9} \times \frac{2}{3} = \frac{8}{27}$

(b)

Explanation [self-study]

Occurrence probability for a non-induced multiset topological colored motif under the EDD random model with color-degree independence on an undirected graph.

(a) Input graph $G(V, E)$, input motif m_C . Since we have only two different degrees within G , the Deg distribution assumes values

within the set $1, 2$, $P(Deg = 1) = \frac{2}{3}$, $P(Deg = 2) = \frac{1}{3}$,

$\mathbb{E}[Deg] = P(Deg = 1) \times 1 + P(Deg = 2) \times 2 = \frac{4}{3}$,

$\gamma = \frac{1}{(|V| - 1) \times \mathbb{E}[Deg]} = \frac{3}{8}$.

Explanation part 2 [self-study]

Probability of the motif.

- i) Probability of the topology computation: m_u indicates the degree of node u within the motif. In this case m_u is 1 for both nodes of the motif. m_{++} is twice the number of edges of the motifs, in this case 2.
- ii) Probability of the multiset of colors comes from the multinomial distribution.
- iii) The probability of the motif is obtained as the product of the probabilities (i) and (ii).

Directed graphs – finding non-induced colored subgraphs

On directed graphs we have the following equation:

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

where $\gamma = 1 / [(N - 1) \times \mathbb{E}[Deg_{out}]]$ (equivalently, 1/the number of edges)

The probability of observing an uncolored topology of k nodes is:

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

In this case, m_{++} is the number of edges

EDD on graphs where node colors influence node degrees

- Given the Deg distribution of degrees, we can define a number of EDD conditional 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 .
- $P(Deg = x|c)$ is the probability of sampling a node in G with a degree x given the color c .

Occurrence probability a colored motif topology (non-induced)

The motif m_C with k nodes, given a color assignment C to its nodes has the following probability:

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

where $Deg|c_u$ is the degree distribution for nodes of color c_u in the input network.

In the case of directed graphs we have:

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

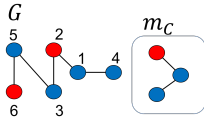
Probability of Injective motifs

The probability of observing the injective colored motif m_C is:

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

where $\nu(C)$ is the product of nodes color probabilities.

Example for Intuition [self-study]



(a)

$$I_3 = \{(1,1,1), (1,1,2), (1,2,1), (2,1,1), (1,2,2), (2,1,2), (2,2,1), (2,2,2)\}$$

$$\begin{aligned} \sigma(m_C) &= \mu(m_C|C) \times \nu(C) = \\ & \left[\sum_{(i,j,k) \in I_3} P(d_i|R)P(d_j|B)P(d_k|B) \gamma d_i \gamma d_j \gamma d_k \right] \times \\ & \prod_{c_u \in \{R,B,B\}} P(c_u) = \\ & \gamma^2 \prod_{u=1}^3 \mathbb{E}(\text{Deg}^{m_u+} | c_u) \prod_{u=1}^3 P(c_u) = 0.018 \end{aligned}$$

(b)

Example When Colors Influence Degrees [self-study]

Occurrence probability for a non-induced injective topological colored motif under the EDD random model with color-degree dependence on an undirected graph. (a) Input graph $G(V, E)$, input motif m_C . We have two different degrees within G , the Deg distribution assumes values within the set $1, 2$,

$P(Deg = 1) = \frac{1}{3}$, $P(Deg = 2) = \frac{2}{3}$, $\mathbb{E}[Deg] = \frac{5}{3}$, $\gamma = \frac{3}{25}$. The probability of the two colors are $P(R) = \frac{1}{3}$, $P(B) = \frac{2}{3}$. We have to define the two color conditioned degree distributions: the $Deg|R$ distribution assumes values within the set $1, 2$, $P(Deg = 1|R) = \frac{1}{2}$, $P(Deg = 2|R) = \frac{1}{2}$, $\mathbb{E}[Deg|R] = \frac{3}{2}$, the $Deg|B$ distribution assumes values within the set $1, 2$,

$P(Deg = 1|B) = \frac{1}{4}$, $P(Deg = 2|B) = \frac{3}{4}$, $\mathbb{E}[Deg|B] = \frac{7}{4}$,
 $\mathbb{E}[Deg^2|B] = \frac{13}{4}$.

Example When Colors Influence Degrees – part 2 [self-study]

(b) Probability of the motif. Generate the set I_3 containing all degree triples with colors 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 color times the probabilities of the color.

Probability Multiset colored motifs

- The color assignments to the nodes within the motif influence the computation of the conditional moments.
- Let $\hat{C} = \{C_1, C_2, \dots, C_l\}$ be the set of all possible color assignments to nodes of m_C obtained by sampling without replacement from C .
- The probability of observing the multiset motif:

$$\sigma(m_C) = \sum_{C_i \in \hat{C}} \mu(m_C | C_i) \times \nu(C_i)$$

where $\nu(C)$ is the product of the nodes color probabilities.

Expectation and Variance of Non-Induced Motifs in Whole Graph

- We first describe a procedure to compute exact mean and variance of the number of non-induced occurrences of a colored motif under any random graph model.
- According to the exchangeability assumption, the occurrence probability of a given motif does not depend on the occurrence position and disjoint occurrences are independent of one another.

Road Map

- Find the probability of the occurrence of a non-induced colored motif at a given position in the graph.
- **Find expected number of instances of that motif in the whole graph. Also variance.**
- Pólya-Aeppli model to calculate p-value.
- Kocay mapping in order to handle induced patterns.
- Experiments evaluating accuracy and performance.

Throughout: generalize to different motif definitions, to directed graphs and to the case where colors influence degrees..

Expectation and Variance of Non-Induced Motifs in Whole graph

- A motif m_C of k nodes can occur in different positions within a graph G . The number of such positions (combination of nodes) is $\binom{N}{k}$. We represent each such position as a tuple of node identifiers in increasing order by node identity:
 $\alpha = (i_1, i_2, \dots, i_k)$ where $i_1 < i_2 < \dots < i_k$.
- We introduce a random variable $Y_\alpha(m_C)$ which equals one if the topology m_C occurs at position α and 0 otherwise.

Expectation and Variance of Non-Induced Motifs

- A motif m_C in a position α can occur in different *configurations*. (Two configurations on the same nodes are distinct if they are not isomorphic.)
- Some permutations of the indexes yield the same motif, we need to consider only the set of its Non-Redundant Permutations (NRP) denoted with $R(m_C)$.
- $\rho(m_C) = |R(m_C)|$ is the number of Non-Redundant Permutations of m_C .
- Computation of $R(m_C)$:
 - Generate all possible $k!$ simultaneous permutations of the rows and columns of the adjacency matrix of m .
 - For each permutation, build the corresponding adjacency matrix and check the latter for redundancy.

Expectation and Variance of Non-Induced Motifs

- We have the following random variable representing the number of instances of the motif:

$$N(m_C) = \sum_{\alpha} \sum_{m'_C \in R(m_C)} Y(m'_C)$$

- This is the sum over all positions α of the number of non-redundant motifs at that position.

Expectation of $N(m_C)$

- The expectation of the count of a colored motif m_C with structure m and multiset of colors C in a graph G with N nodes is (number of combinations times multiplicity for each combination times occurrence probability of a single assignment of topology and colors)

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

Variance of $N(m_C)$

- The variance of the random variable is

$$\mathbb{V}[N(m_C)] = \mathbb{E}[N^2(m_C)] - \mathbb{E}[N(m_C)]^2$$

- The expectation of $N^2(m_C)$ is computed considering that $N^2(m_C)$ can be expressed as:

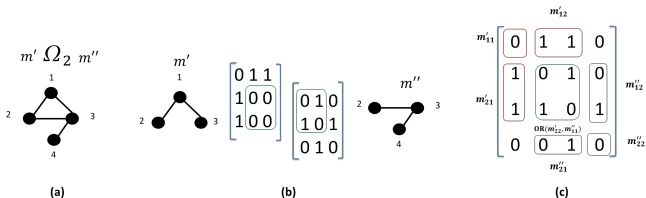
$$\begin{aligned} N^2(m_C) &= \left(\sum_{\alpha} \sum_{m'_C \in R(m_C)} Y_{\alpha}(m'_C) \right)^2 \\ &= \sum_{\alpha} \sum_{m'_C \in R(m_C)} \sum_{\alpha'} \sum_{m''_C \in R(m_C)} Y_{\alpha}(m'_C) \cdot Y_{\alpha'}(m''_C) \end{aligned}$$

- 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.

Super motifs

- A 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.

Example



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 motifs of 3 nodes sharing two nodes. (b) Two non-redundant permutations of a path with 3 nodes with the corresponding adjacency matrices (notice that in this example we have two other possible paths 1, 3, 4 and 1, 2, 3). The overlapping regions are represented (highlighted in green) 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.

Probability of super-motif

The probability of observing the multiset of colors $C_1 \Pi_s C_2$ in the motif is:

$$\nu(C_1 \Pi_s C_2) = \sum_{C^* \subset C: |C^*|=s} \frac{\nu(C^*) [\nu(C \setminus C^*)]^2}{s(C^*)}$$

where $s(C^*)$ is the multiplicity of subset C^* in C , the colon means "such that" and the backslash is set minus, and C is the resulting multiset of $C_1 \Pi_s C_2$.

The probability of observing a colored super-motif generated from colored motifs is the following (μ is the occurrence probability of the topology at a given position).

$$\sigma(m'_{C_1}, m''_{C_2}, s) = \mu(m'_{C_1} \Omega_s m''_{C_2}) \times \nu(C_1 \Pi_s C_2)$$

Expectation of $N^2(m_C)$

The expectation of the squared count has two components: the occurrences that don't overlap (with their possible multiplicities) and the occurrences that do overlap.

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

Injective Topological Colored Motif

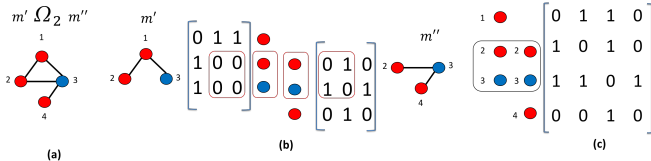
To apply to this case...

- Change the computation of the motif occurrence probability (already done).
- Extend the original definition of non-redundant permutations of a topology.
- Introduce the concept of non-redundant colored permutations of an injective colored motif.

Expectation and Variance of $N(m_C)$ of Injective motifs

- 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 multiset topological colored motifs providing the proper probability of the motif.
- The variance uses the concept of supermotif which has to be defined carefully, taking into account the node colors.

Example



Colored 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 colored 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 colors of the nodes be compatible. The overlapping involves two nodes, the colors 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

Multiset Topological Colored Motifs with Color-Degree Dependence in Input Network

We need to define a new random variable $N^*(m_C)$ representing the number of occurrences of motif m_C . This variable will be a linear combination of random variables $N(m_C)$ coming from the injective case. The number of random variables in the linear combination is determined by the number of possible non-redundant color assignments to the motif nodes according to the multiset of colors C_m . We have the following random variable:

$$N^*(m_C) = \sum_{C_i \in C_m} N(m_{C_i}).$$

Multiset Topological Colored Motifs with Color-Degree Dependence in Input Network

To compute the expectation of $N^*(m_C)$ we can observe that it is a linear operator. Therefore, we have:

$$\begin{aligned}\mathbb{E}[N^*(m_C)] &= \sum_{C_i \in C_m} \mathbb{E}[N(m_{C_i})] = \sum_{C_i \in C_m} \binom{N}{k} \pi(m_{C_i}) \sigma(m_{C_i}) = \\ &= \binom{N}{k} \sum_{C_i \in C_m} \pi(m_{C_i}) \sigma(m_{C_i})\end{aligned}$$

As regards the computation of the variance, we experimentally observed that the $N(m_{C_i})$ random variables are empirically uncorrelated, consequently for the variance we used the following equation:

$$\mathbb{V}[N^*(m_C)] = \sum_{C_i \in C_m} \mathbb{V}[N(m_{C_i})].$$

where $\mathbb{V}[N(m_{C_i})]$ is the variance of the injective colored motif m_{C_i} .

Road Map

- Find the probability of the occurrence of a non-induced colored motif at a given position in the graph.
- Find expected number of instances of that motif in the whole graph. Also variance.
- **Pólya-Aeppli model to calculate p-value.**
- Kocay mapping in order to handle induced patterns.
- Experiments evaluating accuracy and performance.

Throughout: generalize to different motif definitions, to directed graphs and to the case where colors influence degrees..

Assessing the motif 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_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 Pólya-Aeppli distribution

We model $N(m_C)$ as the Pólya-Aeppli distribution.

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):

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

Pólya-Aeppli distribution parameters

The mean and the variance of $PA(\lambda, \alpha)$ are defined as:

$$\mathbb{E}[X] = \frac{\lambda}{1 - \alpha}$$

$$\mathbb{V}[X] = \frac{\lambda(1 + \alpha)}{(1 - \alpha)^2}$$

By making use of the mean and variance obtained using the exchangeable random graph model we can deduce the parameters of the distribution as:

$$\alpha = \frac{\mathbb{V}[N(m_C)] - \mathbb{E}[N(m_C)]}{\mathbb{V}[N(m_C)] + \mathbb{E}[N(m_C)]}$$

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

Road Map

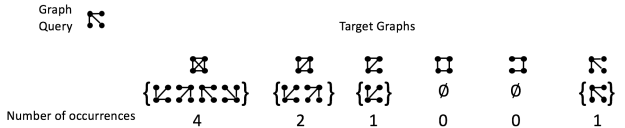
- Find the probability of the occurrence of a non-induced colored motif at a given position in the graph.
- Find expected number of instances of that motif in the whole graph. Also variance.
- Pólya-Aeppli model to calculate p-value.
- **Kocay mapping in order to handle induced patterns.**
- Experiments evaluating accuracy and performance.

Throughout: generalize to different motif definitions, to directed graphs and to the case where colors influence degrees..

The Kocay lemma

- Suppose we want to count the number of non-induced occurrences N_{obs} of a certain subgraph with k nodes.
- The Kocay lemma shows how to express this as a linear combination of the number of induced occurrences of all the possible topologies with k nodes.
- Therefore to construct such a relation we have to find the coefficients of the linear combination.
- Later, we will invert this process to find the mean of the induced motifs from non-induced motifs.

Example



Linear combination of N as function of N'

$$4N'(\text{Star}_4) + 2N'(\text{Star}_3) + N'(\text{Star}_2) + 0 \times N'(\text{Two Nodes}) + 0 \times N'(\text{Two Nodes}) + N'(\text{Star}_4) = N(\text{Star}_4)$$

The Kocay coefficients shows how to express the number of non-induced occurrences, denoted by N , as a linear combination of induced occurrences, denoted by N' . We show an example using a star topology of four nodes. The coefficients of the linear combination can be determined by counting the occurrences of the star topology within each topology of (in this case) four nodes.

The Kocay matrix for a topology of 4 nodes

$$K_4 \begin{bmatrix} 1 & 0 & 1 & 0 & 2 & 4 \\ 0 & 1 & 2 & 4 & 6 & 12 \\ 0 & 0 & 1 & 0 & 4 & 12 \\ 0 & 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} N'(\text{graph 1}) \\ N'(\text{graph 2}) \\ N'(\text{graph 3}) \\ N'(\text{graph 4}) \\ N'(\text{graph 5}) \\ N'(\text{graph 6}) \end{bmatrix} = \begin{bmatrix} N(\text{graph 1}) \\ N(\text{graph 2}) \\ N(\text{graph 3}) \\ N(\text{graph 4}) \\ N(\text{graph 5}) \\ N(\text{graph 6}) \end{bmatrix}$$

The Kocay coefficients in matrix form. The N' terms correspond to the number of induced subgraphs of a particular form of some parent graph. The N terms correspond to the number of non-induced subgraphs of some form.

- 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.
- 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.
- 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-induced motifs of size k . Let M^k be the set of all possible topologies with k nodes. We have:

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

where t_C is a colored motif with topology t and multiset of colors C .

Mean and Variance of $N'(m_C)$ Induced Multiset motifs occurrences

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

The variance of $N'(m_C)$ requires the computation of the covariance.

$$\begin{aligned} \mathbb{V}[N'(m_C)] &= \sum_{t \in M^k} [K_k^{-1}(m, t)]^2 \mathbb{V}[N(t_C)] + \\ &\sum_{t', t'' \in M^k \mid t' \neq t''} K_k^{-1}(m, t') K_k^{-1}(m, t'') \text{Cov}\left(N(t'_C), N(t''_C)\right) \end{aligned}$$

Covariance Calculation

We have that

$$\text{Cov} \left(N(t'_C), N(t''_C) \right) = \mathbb{E}[N(t'_C)N(t''_C)] - \mathbb{E}[N(t'_C)]\mathbb{E}[N(t''_C)] \text{ where:}$$

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

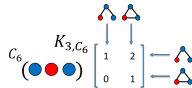
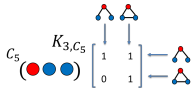
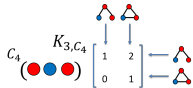
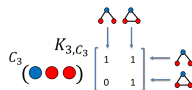
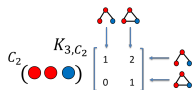
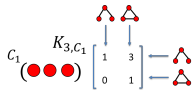
$$\sum_{s=1}^k \binom{N}{k-s, s, k-s, N-2k+s} \sum_{m' \in R(t'), m'' \in R(t'')} \sigma(m'_C, m''_C, s)$$

Injective Topological Colored Motif

- As for multiset motifs, we 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 denoted as $K_{k,C}$.

Example

Induced graphs on top and non-induced to the right.



Road Map

- Find the probability of the occurrence of a non-induced colored motif at a given position in the graph.
- Find expected number of instances of that motif in the whole graph. Also variance.
- Pólya-Aeppli model to calculate p-value.
- Kocay mapping in order to handle induced patterns.
- **Experiments evaluating accuracy and performance.**

Throughout: generalize to different motif definitions, to directed graphs and to the case where colors influence degrees..

Experimental Analysis

- We analyzed the accuracy of the analytical model in the identification of statistically significant graph 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 the permutation-test (i.e. the simulation-based approach).
- In several cases, the Pólya-Aeppli (PA) distribution provides a better fit of the empirical distribution of motif counts in a sample of EDD graphs than the Gaussian distribution.
- The analytical model usually vastly outperforms the simulation-based method in terms of running time. The speed-up is greatest for non-induced motifs.

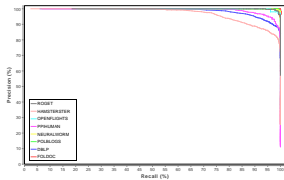
Dataset

The dataset of colored graphs used for testing the analytical method consists of eight real graphs and two artificial graphs.

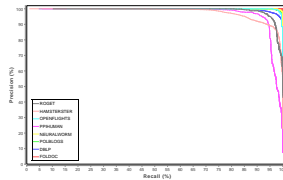
Graph	Orientation	Nodes	Edges	Node colors
ROGET	Undirected	1,010	3,648	6
HAMSTERSTER	Undirected	2,426	16,631	16
OPENFLIGHTS	Undirected	2,939	15,677	5
PPIHUMAN	Undirected	9,506	37,054	11
NEURALWORM	Directed	279	2,990	3
POLBLOGS	Directed	1,224	19,022	2
DBLP	Directed	12,591	49,744	8
FOLDOC	Directed	13,356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

Accuracy of the model

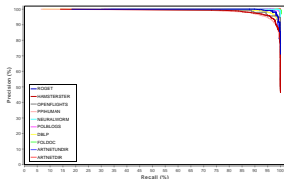
Non-induced occurrences



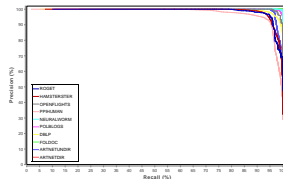
(a) multiset motifs -
color-degree independence



(b) injective motifs -
color-degree independence



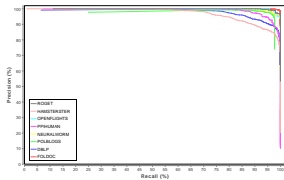
(c) multiset motifs -
color-degree dependence



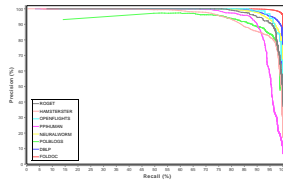
(d) injective motifs -
color-degree dependence

Accuracy of the model

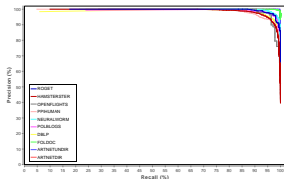
Induced occurrences



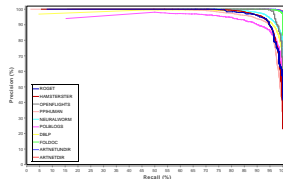
(e) multiset motifs -
color-degree independence



(f) injective motifs -
color-degree independence



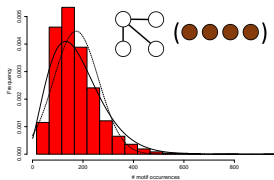
(g) multiset motifs -
color-degree dependence



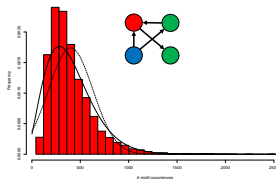
(h) injective motifs -
color-degree dependence

Comparison between Pólya-Aeppli and Gaussian distribution

Non-induced occurrences



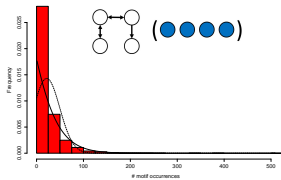
(i) multiset motif, ROGET graph with color-degree independence



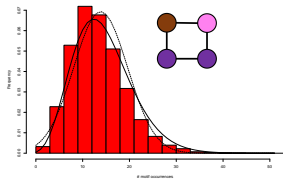
(j) injective motif, NEURALWORM graph with color-degree independence

Comparison between Pólya-Aeppli and Gaussian distribution

Induced occurrences



(k) multiset motif,
NEURALWORM graph with
color-degree independence



(l) injective motif, ROGET
graph with color-degree
independence

Running times

Non-induced occurrences

Running times (secs) of analytical model-based algorithm vs simulation-based algorithm for the computation of non-induced colored motifs (D=color-degree dependence, I=color-degree independence).

Graph	k	Simulation-based algorithm				Analytical model-based algorithm			
		I-Multiset	I-Injective	D-Multiset	D-Injective	I-Multiset	I-Injective	D-Multiset	D-Injective
ROGET	3	27.92	26.91	27.22	26.78	0.03	0.03	0.03	0.03
	4	175.52	206.40	206.73	195.36	0.28	0.64	0.76	0.64
HAMSTERSTER	3	283.18	256.22	257.05	255.94	0.35	0.31	0.43	0.35
	4	11550.39	10081.02	9824.35	9647.71	19.08	20.83	30.97	20.87
OPENFLIGHTS	3	380.10	342.15	339.97	339.95	0.48	0.37	0.51	0.37
	4	16345.06	13084.28	12922.77	12736.22	27.24	26.43	43.00	26.36
PIIHUMAN	3	3413.27	3362.08	3360.45	3364.48	2.16	2.55	2.75	2.23
	4	25736.71	21034.03	20846.42	20590.24	34.66	33.38	46.69	33.55
NEURALWORM	3	23.19	26.25	25.48	25.10	0.05	0.05	0.08	0.06
	4	782.68	3032.65	2365.53	2353.22	4.83	9.83	12.18	9.87
POLBLOGS	3	374.59	325.48	325.71	327.74	0.74	0.70	1.12	0.71
	4	34755.42	30363.26	30465.09	30263.51	131.38	130.06	234.78	131.38
DBLP	3	7639.44	7565.49	7559.17	7558.59	4.29	4.37	4.41	4.09
	4	60142.39	52601.67	53544.69	53486.89	111.39	101.11	140.00	101.31
FOLDOC	3	9352.80	9239.93	9269.15	9243.77	5.76	6.50	7.38	6.36
	4	95923.70	130550.67	132947.10	129241.64	300.57	820.60	902.32	824.58
ARTNETUNDIR	3	-	-	109.67	93.57	-	-	0.14	0.11
	4	-	-	191.78	186.16	-	-	0.57	0.47
ARTNETDIR	3	-	-	155.13	154.29	-	-	0.13	0.09
	4	-	-	1085.88	1078.60	-	-	3.95	3.83
Average performance ratio of analytical vs simulation						724x	673x	588x	725x

Running times

Induced occurrences

Graph	k	Simulation-based algorithm				Analytical model-based algorithm			
		I-Multiset	I-Injective	D-Multiset	D-Injective	I-Multiset	I-Injective	D-Multiset	D-Injective
ROGET	3	27.90	26.29	26.52	26.11	0.02	0.03	0.03	0.03
	4	175.62	145.61	155.15	144.87	0.19	3.83	3.91	3.86
HAMSTERSTER	3	283.50	250.70	252.23	251.30	0.30	0.30	0.41	0.34
	4	11427.55	8787.20	8842.34	8735.41	14.21	51.17	56.63	51.31
OPENFLIGHTS	3	379.11	341.44	340.70	339.43	0.36	0.34	0.45	0.33
	4	16381.74	13011.54	12899.35	12751.47	18.41	17.79	24.62	17.87
PIIHUMAN	3	3422.87	3366.29	3362.88	3363.07	2.58	2.55	2.78	2.55
	4	25662.37	20029.44	20225.28	20459.98	31.66	65.08	75.88	65.24
NEURALWORM	3	23.04	20.01	20.41	20.15	0.03	0.07	0.08	0.07
	4	779.08	648.55	682.28	649.20	14.69	506.15	509.97	512.08
POLBLOGS	3	375.09	324.76	323.70	326.32	0.50	0.48	0.64	0.47
	4	34911.58	29315.97	29126.04	29130.96	56.39	167.71	183.10	169.42
DBLP	3	7639.00	7555.49	7560.25	7562.57	4.00	4.30	4.60	4.27
	4	59834.26	51616.89	53181.85	52238.83	118.61	4172.64	4243.94	4212.15
FOLDOC	3	9340.76	9166.24	9174.23	9183.45	5.21	9.84	9.96	9.27
	4	96164.98	82440.02	87260.57	83560.30	144.02	154798.84	155903.10	155974.89
ARTNETUNDIR	3	-	-	92.26	91.86	-	-	0.08	0.08
	4	-	-	172.69	167.53	-	-	3.52	3.50
ARTNETDIR	3	-	-	150.89	150.17	-	-	0.52	0.50
	4	-	-	395.96	386.33	-	-	7019.75	7017.91
Average performance ratio of analytical vs simulation						1034x	525x	472x	531x

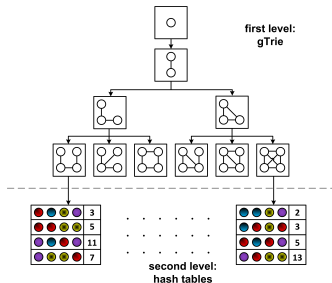
The gLabTrie – only if time

- The analytical model enables the evaluation of the significance of a motif based on the number of its occurrences in the input graph.
- Because of the combinatorial explosion of the number of colored subgraphs, employing time- and memory-efficient algorithms is compulsory for this task.
- The algorithm we describe is based on an extension of G-Trie that handles colored graphs.
- Our structure, called gLabTrie, handles directed and undirected vertex-colored graphs and multiset and injective induced and non-induced topological colored motifs.

The glabtrie algorithm, data structures

We use a two-level data structure called gLabTrie for organizing colored subgraphs.

- The first level organizes uncolored subgraphs in a gTrie, a tree where every node represents a subgraph and the children of a node represent supergraphs of the parent's subgraph.
- Every leaf node of the gTrie is associated to a second-level structure, a hash table that stores all color assignments and associates them with counters.



Example of a gLabTrie. The data structure stores the counters of all colored subgraphs with 4 vertices.

The counting procedure

The Census Algorithm

The algorithm takes as input a graph and an empty gLabTrie and returns a filled gLabTrie.

- It starts by reordering the network nodes according to a predefined color order. This step is necessary for correctly grouping automorphic colored subgraphs.
- The remainder of the algorithm enumerates all subgraphs one by one and increases the counter values of the corresponding entries in the gLabTrie.
- Enumeration is based on the recursive procedure *Match* that matches paths of the gTrie with all possible subgraphs of the input graph.

What the gLabTrie achieves

- The gLabTrie counting algorithm applies to undirected graphs and injective topological colored induced motifs. Directed graphs are managed similarly.
- For multiset topological colored motifs, all subgraphs that share the same topology and the same ordered multiset of colors are grouped together and their counts are summed up.
- Non-induced subgraphs are handled by post-processing the results: counters of non-induced motifs are computed by applying the Kocay matrix to the vector of counters of induced motifs.

Summary: Efficient Way to Find the p -value of Motifs

Applies to....

- Directed and undirected graphs.
- Induced and non-induced subgraphs.
- Degree may or may not depend on color (for almost everything of interest, e.g. chemical, degree depends on color).
- Two kinds of associations of color to nodes – e.g. 5 nodes in star configuration with 4 fans and one rock star vs. 5 nodes in star configuration with 4 fans and one rock star in center.

Summary: Analytical Techniques

- Occurrence probability of a motif at a particular position approximated as a product of moments.
- Number of occurrences based on counting auto-morphisms and overlaps.
- Pólya-Aeppli model to calculate p-value.
- Kocay mapping from induced to non-induced forms in order to evaluate p-value of induced subgraphs.
- gLabTrie to find motifs in a given graph efficiently.

Summary: Experimental Results

- Accuracy is uniformly high compared to simulation.
- Performance is hundred of times better than simulation in all cases but for induced graphs in the injective case. Reason: the variance calculation is expensive

Summary: Take-Away Message

- Finding occurrences of a colored subgraph is fast for small sizes.
- Finding p-value is slow if done naively.
- Our system and software can help you do this analytically and fast.

Some Future Work

- Extend this work to large subgraphs.
- Extend to labeled edges.
- Recursively find motifs.

References



P. Erdos and A. Renyi, "On random graphs", *Publicationes Mathematicae* 6, pp. 290-297, 1959.



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.



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.



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.



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.



W. Kocay, "An extension of Kelly's lemma to spanning subgraphs", *Congr. Num.* 31, pp. 109-120, 1981.



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.



M. Mongiovì, 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.