

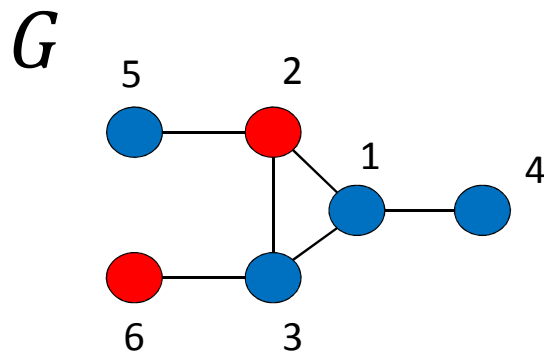
# An analytical model for the statistical significance of colored topology motifs

- Introduction
- Related works
- Definitions
  - Colored Graphs
  - Induced and non induced colored motifs
  - Random Models
    - The Expected Degree Distribution (EED) random Model
    - A toy example of Mean and Variance computation
    - Mean and Variance Computation
- Significance computation
  - Simulation based Pvalue assessment
  - Towards an analytical approach to compute the significance of a non-induced motif: The Geometric-Poisson
  - The Induced case
    - The Kocay Lemma: The induced occurrences as a linear combination of non induced occurrences
    - An algorithm for the computation of coefficients
    - Computing the variance of a mixture
- Experimental analysis

# Colored graph

A colored graph  $G(V,E,C,c)$  is a graph where:

- $V$  is the set of nodes;
- $E$  is the set of edges;
- $C$  is a set of labels or colors;
- $c$  is a function that assigns a color to each node in  $V$ .



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

# Induced and non-induced Subgraph isomorphism

## Definition (uncolored graphs)

Let  $H = (V_H, E_H)$  and  $G = (V, E)$  be graphs. A subgraph isomorphism from  $H$  to  $G$  is a function  $f: V_H \rightarrow V$  such that if  $(u, v) \in E_H$  then  $(f(u), f(v)) \in E$ .  $f$  is an induced subgraph isomorphism if in addition if  $(u, v) \notin E_H$  then  $(f(u), f(v)) \notin E$ .

## Definition (colored graphs)

Let  $H = (V_H, E_H, C_H, c_H)$  and  $G = (V, E, C, c)$  be colored graphs. A subgraph isomorphism from  $H$  to  $G$  is a function  $f: V_H \rightarrow V$  such that if  $(u, v) \in E_H$  then  $(f(u), f(v)) \in E$  and  $c_H(u) = c(f(u))$  and  $c_H(v) = c(f(v))$ .  $f$  is an induced subgraph isomorphism if in addition if  $(u, v) \notin E_H$  then  $(f(u), f(v)) \notin E$ .

# Motifs

- Given a graph  $G$ , a labeled topology that occurs frequently in  $G$  is also called *motif*.
- Formally we have the following Definition

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

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

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  $m_C$ .

# 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;
- Examples of reference models are:
  - The Erdos-Renyi model (ER model);
  - The Fixed degree distribution model (FDD model);
  - The Expected degree distribution model (EDD model);
  - The Erdos-Renyi mixture for graphs model (ERMG model).

# Random models

- ER model:
  - Probability of connecting two nodes does not depend on the nodes;
  - No fit to the connectivity heterogeneity in real networks.
- FDD model:
  - Nodes keep the same degree (just swapping edges repeatedly);
  - Used to compute simulation p-values (Mfinder, etc...);
- EDD model:
  - Generate graphs where node degrees follows a given distribution;
- ERMG model:
  - Based on mixture distributions used to model heterogenous connectivity.

## Description of EDD model (Chung and Lu, 2002)

- Generate graphs where node degrees follow a given distribution;
- Let  $P_{out}(d)$  and  $P_{in}(d)$  the probabilities that a node has out-degree  $d$  and in-degree  $d$ , respectively;
- $P_{out}(d)$  and  $P_{in}(d)$  are computed from the degree distributions in the input network;
- Let  $D_{out}(i)$  and  $D_{in}(i)$  the expected out-degree and the in-degree of node  $i$ , respectively;
- $D_{out}(i)$  and  $D_{in}(i)$  are sampled according to  $P_{out}(d)$  and  $P_{in}(d)$ .

# Description of EDD model (Chung and Lu, 2002)

- Given two nodes  $i$  and  $j$ , with  $i \neq j$  an edge between them exists with probability:

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

where  $N$  is the number of nodes in the network.

- To ensure that  $P(i, j) \leq 1$  we must assume:

$$D_{out}(i) \times D_{in}(j) \leq \sum_{k=1}^N D_{out}(k)$$

otherwise we put  $P(i, j) = 1$ .



# Occurrence probability of a colored motif in the EDD model

- Let  $m_C$  a colored motif of  $k$  nodes, with colors in the multiset  $C_m$ ;
- Probability to assign colors in  $C_m$  to the  $k$  nodes of  $m$  follows a multinomial distribution (Schbath et al., 2009):

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

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

# Occurrence probability of a colored motif in the EDD model – undirected graph

- Probability of observing  $m$  as non-induced motif of the EDD graph, regardless its colors (Picard et al., 2008):

$$\mu(m) = \lambda^{m_{++}/2} \prod_{u=1}^k \mathbb{E}(P_{out}^{m_{u+}})$$

where:

- $\lambda = 1/[(N - 1)\mathbb{E}(D_{out})]$  and  $\mathbb{E}(D_{out})$  is the average out-degree in the EDD graph;
- $m_{++}$  is the total number of outgoing edges in  $m$ ;
- $m_{u+}$  is the number of outgoing edges from node  $u$  in  $m$ ;
- $\mathbb{E}(P_{out}^{m_{u+}})$  is the  $m_{u+}$ -th moment of distribution  $P_{out}$ :

$$\mathbb{E}(P_{out}^i) = \sum_{d=1}^{maxOutDeg} P_{out}(d) \times d^i$$

# Occurrence probability of a colored motif in the EDD model – directed graph

- Probability of observing  $m$  as non-induced motif of the EDD graph, regardless its colors:

$$\mu(m) = \lambda^{m_{++}} \prod_{u=1}^k \mathbb{E}(P_{out}^{m_{u+}}) \mathbb{E}(P_{in}^{m_{u-}})$$

where:

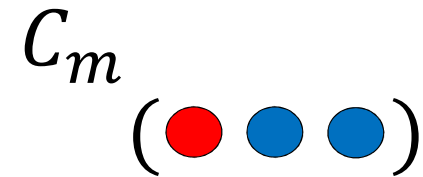
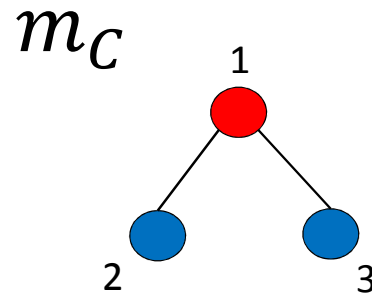
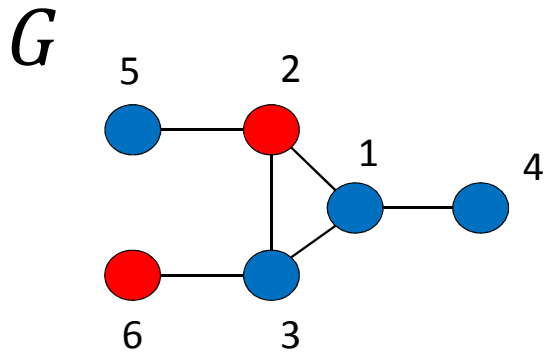
- $\lambda = 1/[(N - 1)\mathbb{E}(D_{out})]$  and  $\mathbb{E}(D_{out})$  is the average out-degree in the EDD graph;
- $m_{++}$  is the total number of outgoing edges in  $m$ ;
- $m_{u+}$  is the number of outgoing edges from node  $u$  in  $m$ ;
- $m_{u-}$  is the number of ingoing edges from node  $u$  in  $m$ ;
- $\mathbb{E}(P_{out}^{m_{u+}})$  and  $\mathbb{E}(P_{in}^{m_{u-}})$  are the  $m_{u+}$ -th and the  $m_{u-}$ -th moment of distributions  $P_{out}$  and  $P_{in}$ , respectively;

# Occurrence probability of a colored motif in the EDD model

- Probability of observing the colored motif  $m_C$ :

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

# Toy example



$$\gamma(C) = \frac{3!}{2!1!} \times \frac{4}{6} \times \frac{4}{6} \times \frac{2}{6} = 0.445$$

$$\mu(m) = \lambda^2 \prod_{u=1}^3 \mathbb{E}(P_{out}^{m_{u+}}) = \lambda^2 \times \mathbb{E}(P_{out}^2) \times \mathbb{E}(P_{out}) \times \mathbb{E}(P_{out}) = \left(\frac{1}{10}\right)^2 \times 2 \times 1.333 \times 1.333 = 0.0355$$

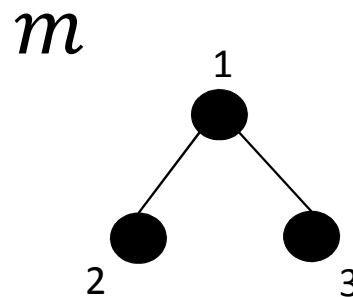
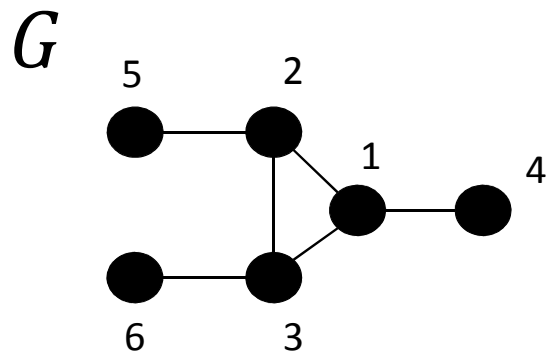
$$\sigma(m_C) = \gamma(C) \times \mu(m) = 0.016$$

# Mean and variance of a colored motif non-induced count

- 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;
- Three important assumptions:
  - The occurrence probability of a given motif does not depend on the occurrence position (exchangeability assumption of the underlying random model);
  - Disjoint occurrences are independent one another;
  - Colors are independent from topologies;
- ER, EDD and ERMG are exchangeable models;

# Non-redundant permutations

- An uncolored motif  $m$  of  $k$  nodes can occur in different positions within a graph  $G$ ;
- We can represent the location of  $m$  in  $G$  through a  $k$ -uple of indexes,  $\alpha = (i_1, i_2, \dots, i_k)$  representing the positions of its nodes;
- Example (non-induced occurrences):




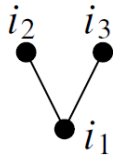
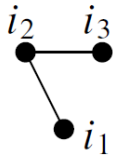
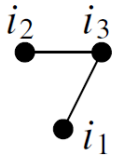
$m$  occurs in  $G$  at positions:

- (2,5,1)
- (2,5,3)
- (3,6,2)
- (3,6,1)
- (1,3,4)
- (1,2,4)
- (1,2,3)
- (2,1,3)
- (3,1,2)

# Non-redundant permutations

- A motif  $m$  in a position  $\alpha$  can occur in different «configurations», where each configuration correspond to a permutation of indexes in  $\alpha$ ;
- Some permutations of indexes result in the same motif;
- We need to consider only the set  $R(m)$  of non-redundant permutations of motif  $m$ ;
- Example:

TABLE 1. NONREDUNDANT PERMUTATIONS OF THE  
 MOTIF AT POSITION  $\alpha = (i_1, i_2, i_3)$

$\mathbf{m}$	$\mathbf{m}'$	$\mathbf{m}''$
$\begin{bmatrix} 0 & 1 & 1 \\ \cdot & 0 & 0 \\ \cdot & \cdot & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ \cdot & 0 & 1 \\ \cdot & \cdot & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 1 \\ \cdot & 0 & 1 \\ \cdot & \cdot & 0 \end{bmatrix}$
		



# Non-redundant permutations

- Let  $\rho(m) = |R(m)|$ ;
- $\rho(m) = k! / |\text{aut}(m)|$ , where  $\text{aut}(m)$  is the set of automorphisms of  $m$ ;
- For a path of 3 nodes,  $\rho(m) = 3!/2 = 3$ ;
- Method to compute  $R(m)$ :
  - 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.
- Complexity:  $O(k!^2)$ .

# Mean of a colored motif count

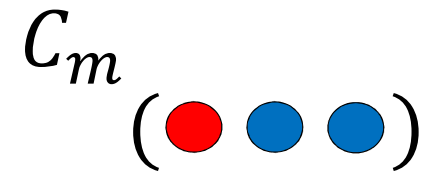
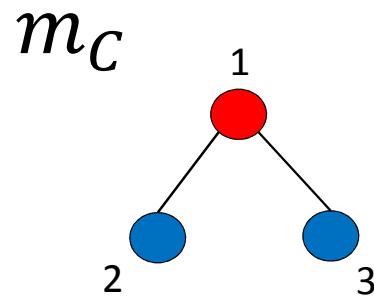
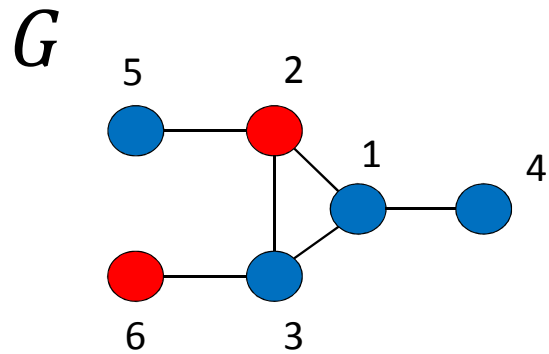
- For the exchangeability assumption, each permutation of  $m$  has the same probability of occurrence;
- Mean of the count of a colored motif  $m_c$  in a graph  $G$  with  $N$  nodes:

$$\mathbb{E}N(m_c) = \binom{N}{k} \times \rho(m) \times \sigma(m_c)$$

where:

- $\binom{N}{k}$  is the number of all possible locations of  $m$  in  $G$ ;
- $\rho(m)$  is the number of non-redundant permutations of  $m$ ;
- $\sigma(m_c)$  is the occurrence probability of  $m_c$  according to an exchangeable random model

# Toy example



$$\mathbb{E}N(m_C) = \binom{N}{k} \times \rho(m) \times \sigma(m_C) = \binom{6}{3} \times 3 \times 0.016 = 0.96$$

# Overlapping occurrences and super-motifs

- For the computation of the variance of the count of a colored motif, we need to consider overlapping occurrences;
- Two occurrences of a motif overlap if they share at least one node;
- We introduce the concept of super-motif, which is a motif formed by two overlapping occurrences (non-redundant permutations) of a given motif;
- Given two NRPs of a motif  $m$ ,  $m'$  and  $m''$ , 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.
- The notion of super-motif is transitive.

# Computation of a super-motif with overlap $s$

- Break the adjacency matrices of  $m'$  and  $m''$  such that:

$$\mathbf{m}' = \left( \begin{array}{c|c} \mathbf{m}'_{11} & \mathbf{m}'_{12} \\ \hline \mathbf{m}'_{21} & \mathbf{m}'_{22} \end{array} \right) \quad \mathbf{m}'' = \left( \begin{array}{c|c} \mathbf{m}''_{11} & \mathbf{m}''_{12} \\ \hline \mathbf{m}''_{21} & \mathbf{m}''_{22} \end{array} \right)$$

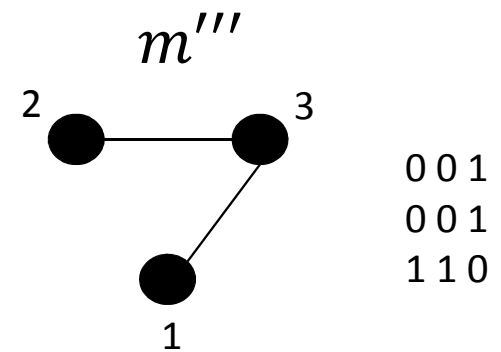
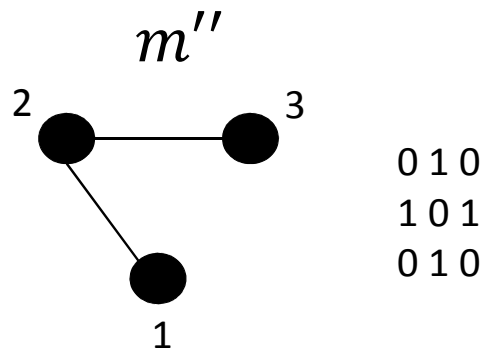
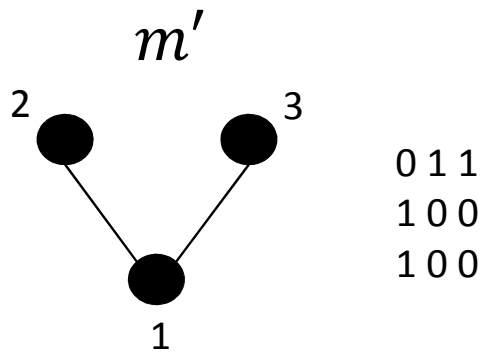
$(k-s) \times (k-s)$      $(k-s) \times s$                        $s \times s$      $s \times (k-s)$   
 $s \times (k-s)$                        $s \times s$                        $(k-s) \times s$      $(k-s) \times (k-s)$

- Adjacency matrix of super-motif (max function = logical OR):

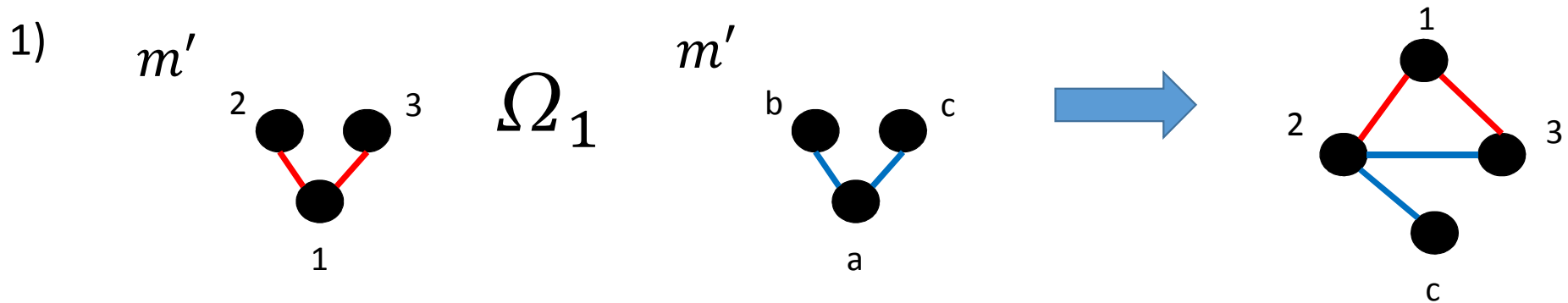
$$\mathbf{m}' \underset{s}{\Omega} \mathbf{m}'' = \left( \begin{array}{c|c|c} \mathbf{m}'_{11} & \mathbf{m}'_{12} & \mathbf{0} \\ \hline \mathbf{m}'_{21} & \max(\mathbf{m}'_{22}, \mathbf{m}''_{11}) & \mathbf{m}''_{12} \\ \hline \mathbf{0} & \mathbf{m}''_{21} & \mathbf{m}''_{22} \end{array} \right)$$

# Example: the 9 super-motifs of a path of 3 nodes with overlap $s=2$

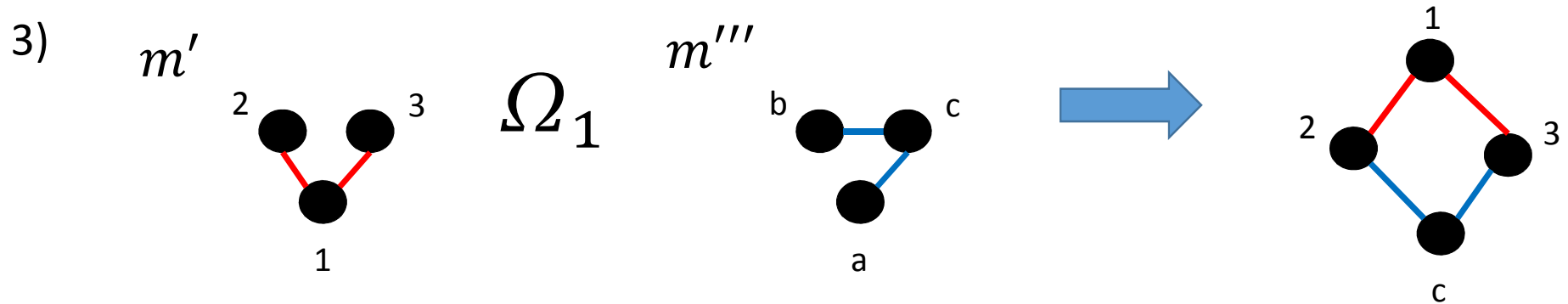
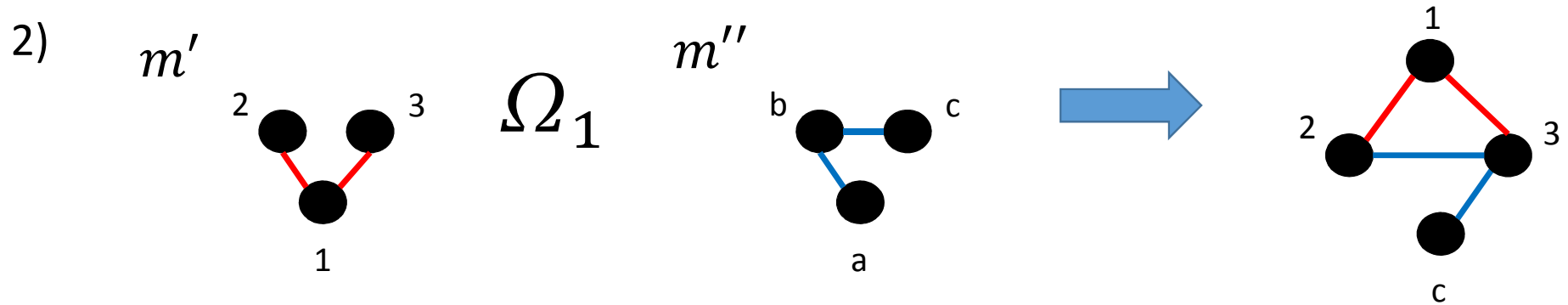
- Non-redundant permutations of a path with 3 nodes:



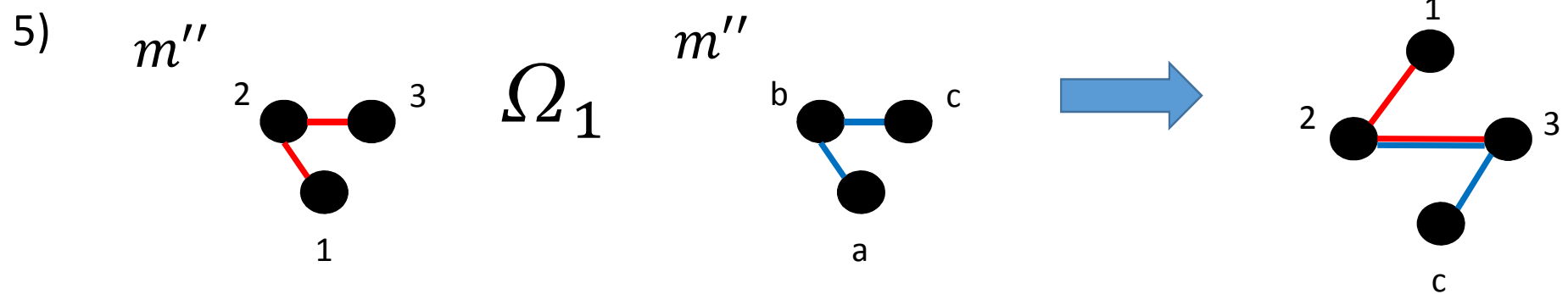
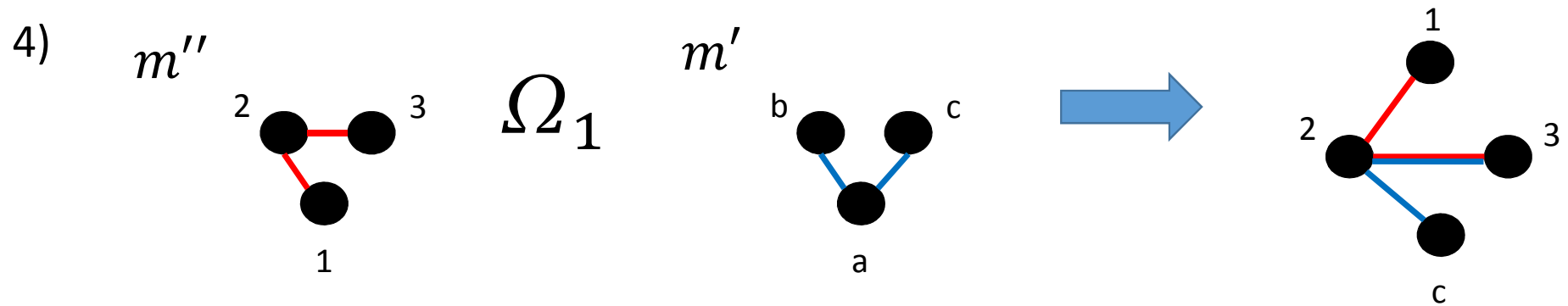
- All super-motifs with overlap  $s=2$  obtained from a path with 3 nodes:



Example: the 9 super-motifs of a path of 3 nodes with  $s=1$

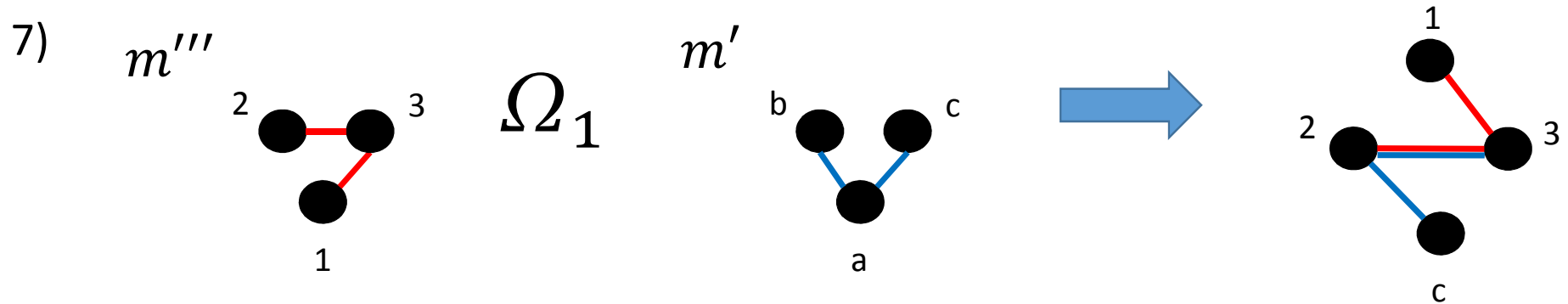
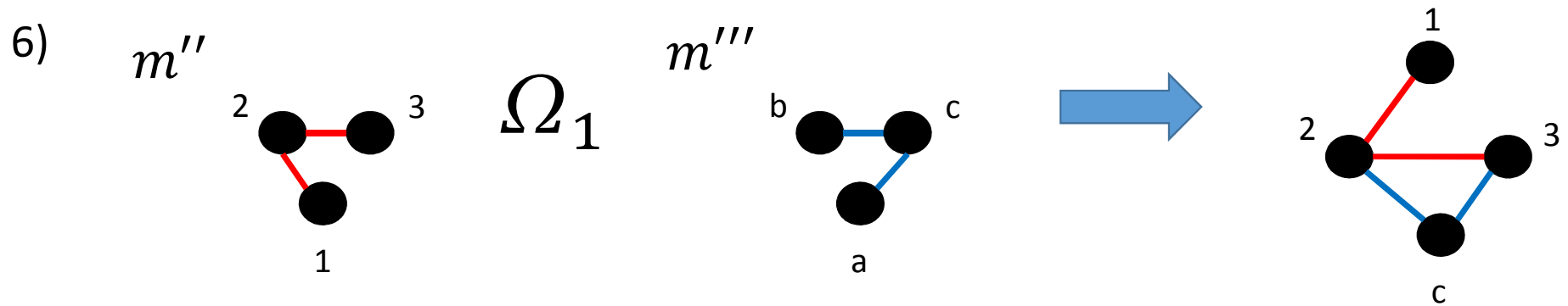


Example: the 9 super-motifs of a path of 3 nodes with  $s=1$

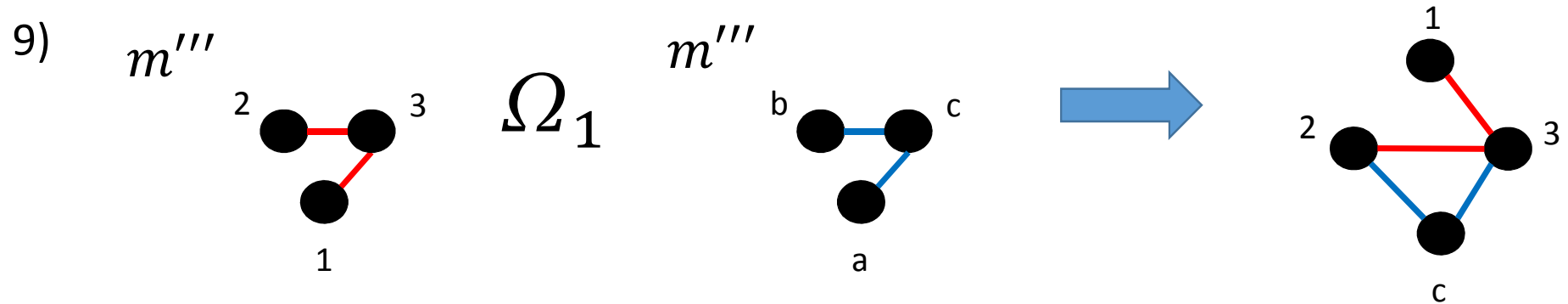
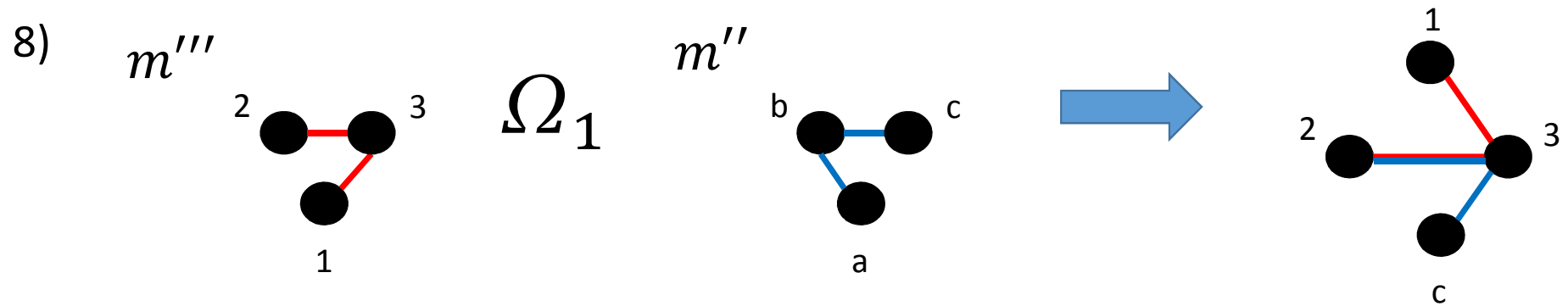




Example: the 9 super-motifs of a path of 3 nodes with  $s=1$



Example: the 9 super-motifs of a path of 3 nodes with  $s=1$



# Super-motifs and overlapping colors

- Super-motifs inherit the colors of ancestor motifs;
- Due to node overlappings, one or more colors can overlap;
- Colors can overlap in many ways, so the same super-motif can be colored with different super-sets of colors;
- As for the super-motifs, we can define an overlapping operations for two multi-sets of colors  $C_1$  and  $C_2$  with overlap  $s$ :  $C_1 \Pi_s C_2$ ;
- $C_1 \Pi_s C_2 = \{C_1 \cap_s C_2\}$ , i.e. the set of all possible intersections of  $C_1$  and  $C_2$  with  $s$  elements;
- $C \Pi_s C$  is equivalent to the set of all subsets of  $C$  with  $s$  elements.

# Probability of observing colored super-motifs

- The probability of observing a super-motif (regardless of its colors) can be easily computed as  $\mu(m' \Omega_s m'')$ ;
- Let  $C$  a multiset of colors of  $m'$  and  $m''$  and  $C^*$  the multiset of colors of the  $s$  overlapping nodes of  $m'$  and  $m''$ ;
- Let  $C^- = C \setminus C^*$ , where  $\setminus$  is the set difference operator;
- Probability of observing the super-multiset of colors  $C \Pi_s C$  in the graph (Schbath et al., 2009):

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

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

# Probability of observing colored super-motifs

- Probability of observing a colored super-motif generated from colored motifs  $m'_C$  and  $m''_C$  with overlap  $s$ :

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

# Expectation of the squared count

- The calculation of variance is based on the expectation of the squared count of a colored motif (i.e. the moment 2);
- The expectation is given by the contribution of two terms, one is related to pairs of disjoint occurrences and one is related to pairs of overlapping occurrences (with different degrees of overlap);
- In both cases we must consider:
  - All possible locations of the two occurrences of a motif  $m$  in the graph;
  - All possible non-redundant permutations of  $m$ ;

# Expectation of the squared count

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

- $k$  is the number of nodes of motif  $m$ ;
- $N$  is the number of nodes of the input network;
- $\binom{N}{N-2k, k, k}$  is the number of all possible combinations of locations of two non-redundant permutations of  $m$  with no overlap;
- $\binom{N}{k-s, s, k-s, n-2k+s}$  is the number of all possible combinations of locations of two non-redundant permutations of  $m$  with overlap  $s$ .

# Variance of the count

- Variance of the count of a colored motif  $m_C$ :

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



# P-value Computation

- To decide whether a motif  $m$  is overrepresented in a given network, one needs to calculate the probability

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

- Where
  - $N_{\text{obs}}(m)$  is the observed number of occurrences of  $m$
  - $N(m)$  is a Random Variable representing the number of occurrences of the motif under the chosen reference model.

# Simulation based approach

- A common approach on the approximation of  $P[N(m) \geq N_{\text{obs}}(m)]$  relies on the usage of permutation test.
- $L$  random graphs are generated and the counts of the motifs are computed. The probability is then approximated as:

$$P[N(m) \geq N_{\text{obs}}(m)] \approx \#(N_i(m) \geq N_{\text{obs}}(m))/L$$

# Towards an analytical p-value computation: *Overlapping motifs*

- Network motifs come in clusters (*clumps*) since they can overlap;
- Clusters result in numerous occurrences of a motif with a reduced number of vertices;
- Given a network we can observe a certain number of clusters according to the overlapping of the motifs;
- This number can be modeled as a random variable  $X_1$ .

# Towards an analytical p-value computation: *Overlapping motifs within a cluster*

- Suppose we have a set of clusters according to the intersection of pairs of motifs;
- We can build a random variable  $X_2$  in which we sample several times a cluster until we observe the size of the cluster we are looking for.

# Towards an analytical p-value computation: *The Geometric-Poisson approximation*

- The Geometric-Poisson distribution is suitable to describe how the count of events occurring in clumps (in our case motifs) may vary;
- We assume that:
  - $X_1$  (modeling the number of clumps) has a Poisson distribution;
  - $X_2$  (modeling the probability of observing a certain cluster/clump size) has a Geometric distribution;
  - The clump sizes are independent each other with a common distribution.

# Polya-Aeppli distribution

- The Polya-Aeppli (denoted by PA) distribution (or geometric Poisson) is obtained when the clump size has a *geometric distribution*  $G(1 - a)$ , therefore the mean size of a clump is  $1/(1 - a)$ .
- In this case,  $X \sim PA(\lambda, a)$  where  $X$  is a random variable representing the the number of observed events:

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

# Polya-Aeppli distribution

- The Polya-Aeppli distribution can be used as an approximation of the distribution of the count of  $N(m)$ .
- The first two moments of  $PA(\lambda, a)$  are defined as  $\frac{\lambda}{(1-a)}$  and  $\frac{\lambda(1+a)}{(1-a)^2}$
- By using the mean and variance of  $N(m_C)$  we can obtain  $\lambda$  and  $a$ :

$$a = \frac{[\text{Var}N(m_C) - \mathbb{E}N(m_C)]}{[\text{Var}N(m_C) + \mathbb{E}N(m_C)]}$$

$$\lambda = (1 - a) \times N(m_C)$$

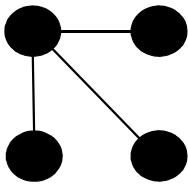
# Extension of the model to the induced case

- The equations described for computing the mean and the variance of a motif count refer to the non-induced case;
- We want to extend the model to compute the mean and the variance of the number of induced occurrences of a motif;
- We need a theoretical results to relate the number of non-induced occurrences to the number of induced occurrences (Kocay Lemma);



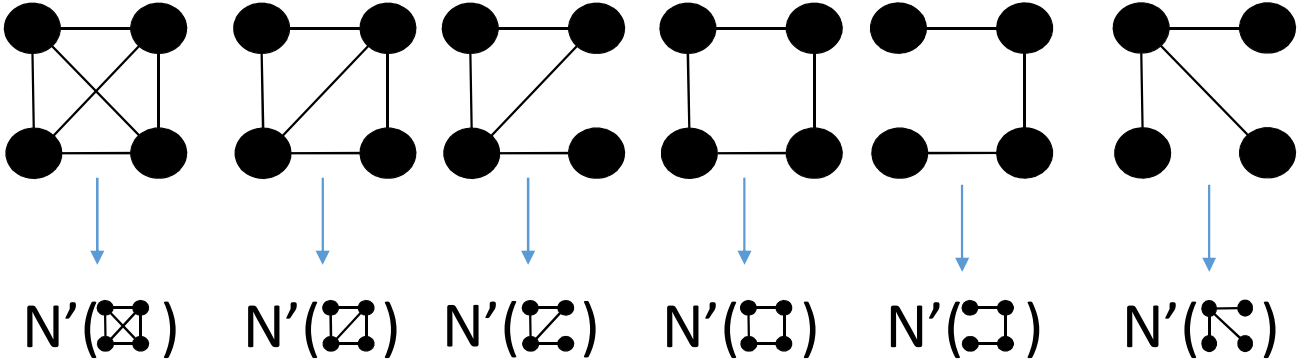
# Kocay Lemma (Kocay, 1981)

- Suppose we want to count the number of non-induced occurrences  $N$  of a certain subgraph with  $k$  nodes;
- By applying the Kocay lemma we can express this number as a linear combination of the number of induced occurrences of all 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.

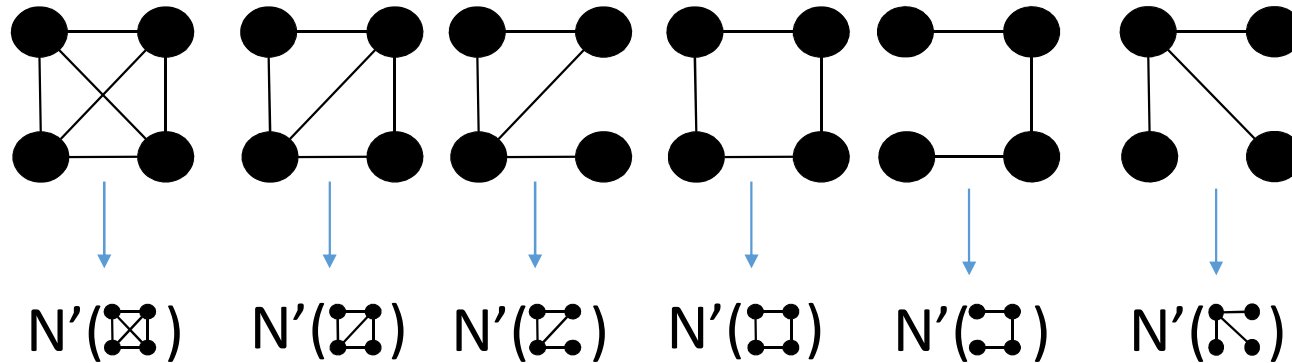
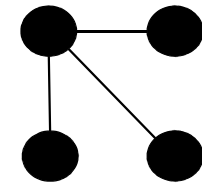


# Example

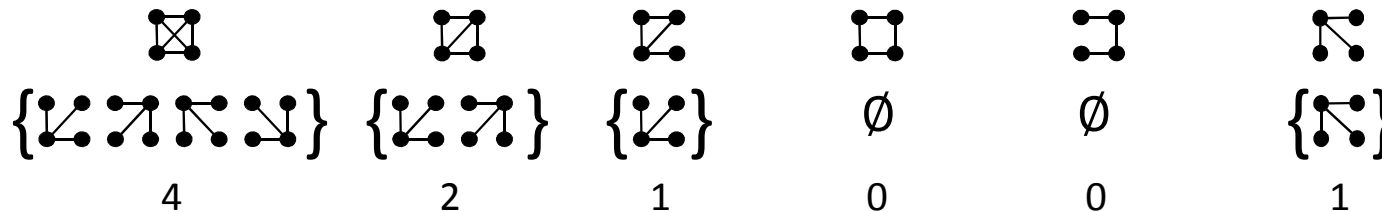
- Suppose we have a star topology with 4 nodes and we wish to find its non-induced occurrences within a target graph  $G$ .
- Suppose we know the number of occurrences of all induced topologies with 4 nodes (we use  $N'$  to denote the number of induced occurrences);

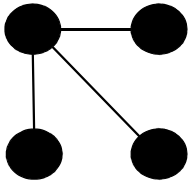


# Example



- The coefficients of the linear combination can be determined by counting the occurrences of the topology within each topology of 4 nodes:





# Example

- Then we obtain the following linear combination:


$$4N'(\text{graph with all edges}) + 2N'(\text{graph with top and bottom edges}) + N'(\text{graph with top and right edges}) + 0 \times N'(\text{graph with top and left edges}) + 0 \times N'(\text{graph with top and bottom edges}) + N'(\text{graph with top and right edges}) = N(\text{graph with all edges})$$

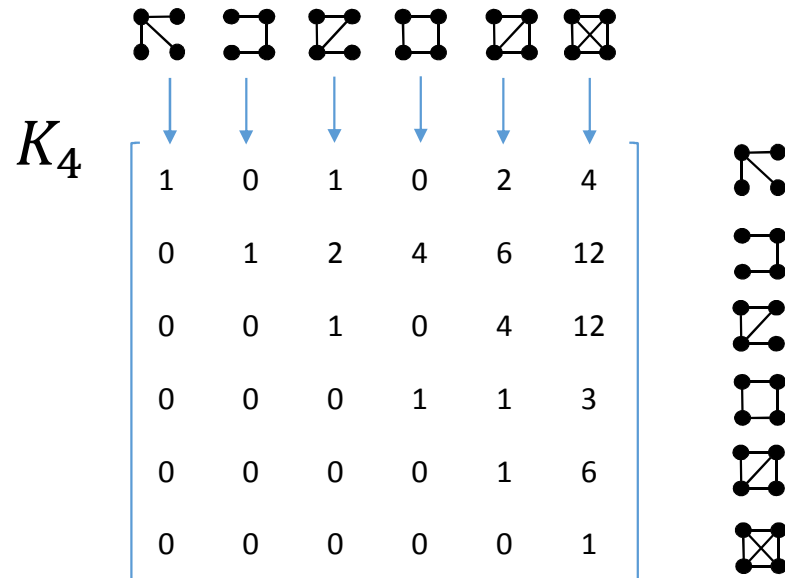
# Example

- The coefficient for all topologies can be represented using a matrix notation;
- We denote with  $K_k$  the Kocay matrix for topologies of size k;
- Each row refers to a specific topology m. We denote with  $K_k(m)$  the corresponding row.

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

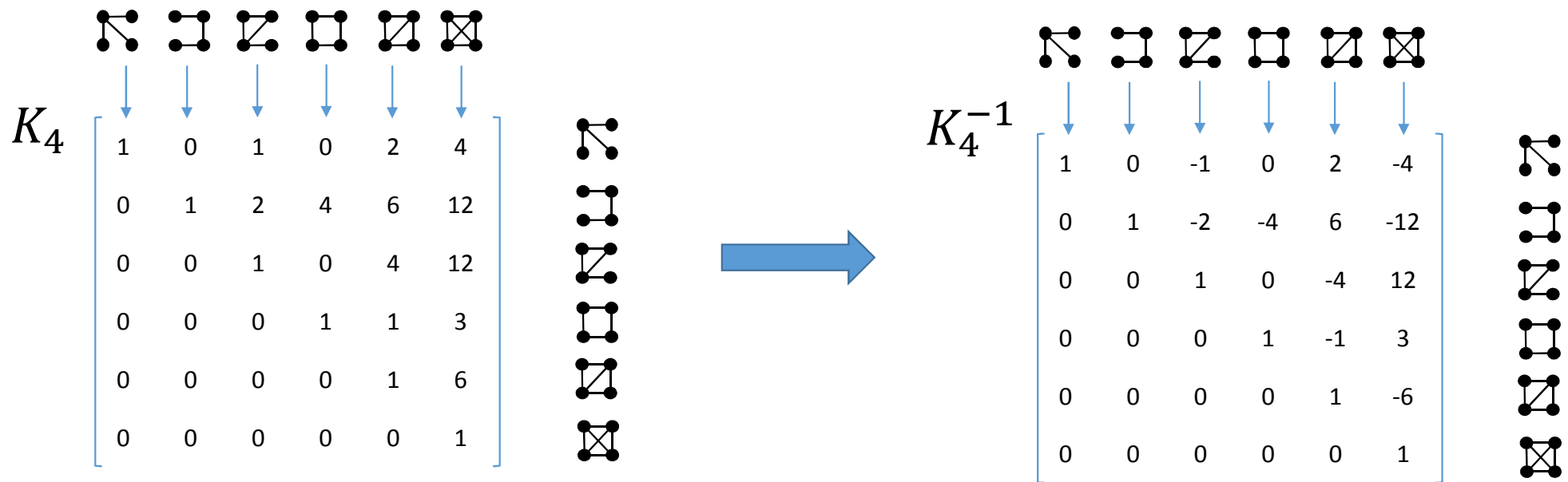
# Computing the coefficients for all the topologies

- The coefficients in the column of the matrix can be obtained by running a motif finding algorithm computing the non induced occurrences of size  $k$  of each topology of size  $k$ .
- For example if the target topology is  we obtain as number of occurrences the last column of the matrix.



# Inverse of Kocay matrix

- 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.



# Mean of the induced count of a colored motif

- The mean of the induced count of a colored motif  $m_C$  with  $k$  nodes is the mean of a linear combination of non-induced count, i.e. a linear combination of the means of non-induced counts of all topologies of size  $k$ ;
- The coefficients of the linear combination are the elements of a row of the inverse Kocay matrix;

$$\mathbb{E}N'(m_C) = \mathbb{E}[K_k^{-1}(m)N(m_C)] = K_k^{-1}(m) \times \mathbb{E}N(m_C)$$



# Variance of the induced count of a colored motif

- Likewise, the variance of the induced count of a colored motif  $m_C$  with  $s$  nodes is the variance of a linear combination of non-induced count of all topologies of size  $s$ ;
- The variance of a linear combination of  $N$  random variables implies the computation of the covariance:

$$\mathbb{V}\left(\sum_{i=1}^N a_i X_i\right) = \sum_{i=1}^N a_i^2 \mathbb{V}(X_i) + 2 \sum_{1 \leq i < j \leq N} a_i a_j \text{Cov}(X_i, X_j)$$

$$\text{Cov}(X_i, X_j) = \mathbb{E}X_i X_j - \mathbb{E}X_i \mathbb{E}X_j$$

# Variance of the induced count of a colored motif

- Variance of the induced count of a colored motif  $m_C$  with  $k$  nodes:

$$\begin{aligned} & \mathbb{V}N'(m_C) \\ &= \sum_{\bar{m} \in M_k} [K_k^{-1}(m, \bar{m})]^2 \mathbb{V}N(\bar{m}_C) + \sum_{\bar{m}, \bar{\bar{m}} \in M_k : \bar{m} \neq \bar{\bar{m}}} K_k^{-1}(m, \bar{m}) K_k^{-1}(m, \bar{\bar{m}}) \text{Cov}(\bar{m}_C, \bar{\bar{m}}_C) \end{aligned}$$

where:

- $K_k^{-1}(m, \bar{m})$  is the number of induced occurrences of  $m$  in  $\bar{m}$ ;
- $M_k$  is the set of all topologies with  $s$  nodes.

# Covariance of the non-induced counts of two motifs

- Given two colored motifs  $\bar{m}_C$  and  $\bar{\bar{m}}_C$  defined on the same multiset of node colors  $C$ , the covariance of the non-induced counts of  $\bar{m}_C$  and  $\bar{\bar{m}}_C$  is defined as:

$$\text{Cov}(N(\bar{m}_C), N(\bar{\bar{m}}_C)) = \mathbb{E}N(\bar{m}_C)N(\bar{\bar{m}}_C) - \mathbb{E}N(\bar{m}_C)\mathbb{E}N(\bar{\bar{m}}_C)$$

where:

$$\begin{aligned} & \mathbb{E}N(\bar{m}_C)N(\bar{\bar{m}}_C) \\ &= \binom{N}{N-2k, k, k} \sum_{m' \in R(\bar{m}), m'' \in R(\bar{\bar{m}})} \sigma(m'_C)\sigma(m''_C) + \sum_{s=1}^k \binom{N}{k-s, s, k-s, N-2k+s} \sum_{m' \in R(\bar{m}), m'' \in R(\bar{\bar{m}})} \sigma(m'_C, m''_C, s) \end{aligned}$$