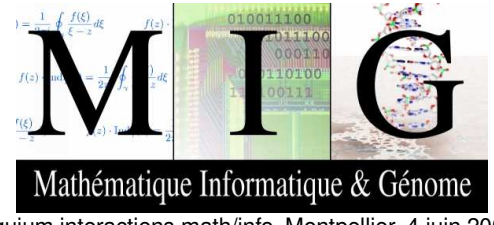
Statistical analysis of biological networks.

Assessing the exceptionality of network motifs

S. Schbath

Jouy-en-Josas/Evry/Paris, France

http://genome.jouy.inra.fr/ssb/

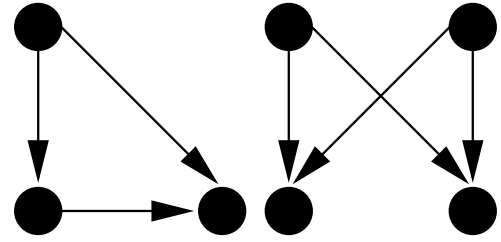


The network revolution

- **Nature of the data**:
  – n individuals (n large),
  – but also n^2 couples.
- **Many scientific fields**:
  sociology, physics, "internet", biology.
- **Biological networks** :
  protein-protein interaction networks, regulatory networks, metabolic networks.
- **Statistical aspects**:
  – network inference,
  – statistical properties of given networks (degrees, diameter, clustering coefficient, modules, motifs etc.),
  – random graph models.

Looking for local structures

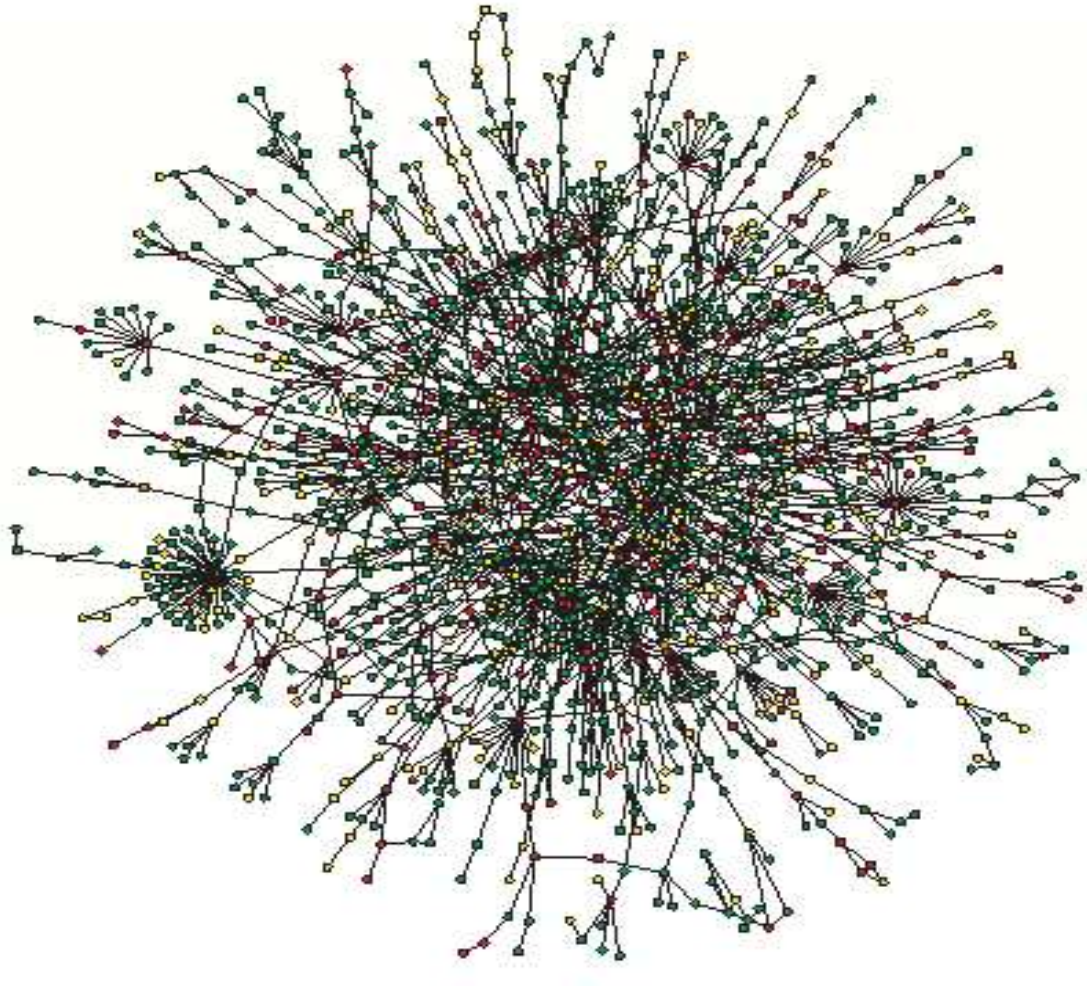
- Breaking-down complex networks into functional modules or **basic building blocks**: [Shen-Orr et al. (02)]
  $\longrightarrow$ patterns of interconnection, **topological motifs**.
- **Focus on exceptional motifs** = motifs appearing more frequently than **expected**. [Milo et al. (02), Shen-Orr et al. (02), Zhang et al. (05)]
- **Interpretation**:
  - they are thought to reflect functional units which combine to regulate the cellular behavior as a whole,
  - mathematical analysis of their dynamics are required. [Mangan and Alon (03), Prill et al. (05)]



- **Transcriptional regulatory networks**: particular regulatory units (e.g. feed-forward loop, bi-fan).

Other kind of motifs

If the nodes are colored (e.g. GO term, EC number etc.):

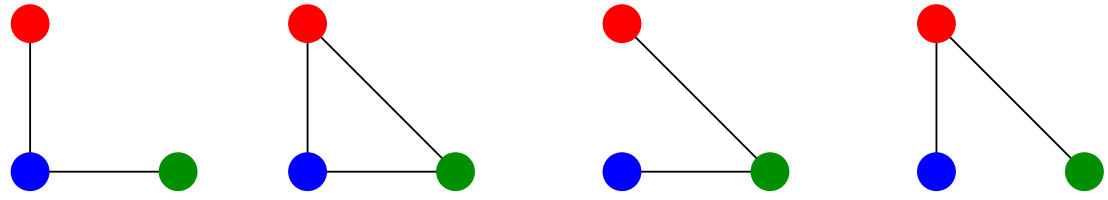


From Barabasi et al.(2004)

Other kind of motifs

If the nodes are colored (e.g. GO term, EC number etc.):

- **Colored motifs** [Lacroix et al. (05, 06)]
  - no topological constraint
  - only a connectedness constraint



- **Topological colored motifs** [Lee et al. (07), Taylor et al. (07)]

How to assess the exceptionality of a motif?

Step 1 **To count** the observed number of occurrences $N_{\text{obs}}(\mathbf{m})$ of a given motif $\mathbf{m}$ (out of my scope)

Its significance is assessed with the p-value $\mathbb{P}\{N(\mathbf{m}) \geq N_{\text{obs}}(\mathbf{m})\}$
(*the probability to get as much occurrences at random*)

Step 2 To choose an appropriate **random graph model**

Step 3 To get the **distribution** of the count $N(\mathbf{m})$ under this model

State of the art (1/2)

Analytical approaches:

- The most popular random graph model is the **Erdös-Rényi model** (nodes are connected independently with proba p)
- Some theoretical works exist on Poisson and Gaussian approximations of topological motif count distribution [Janson et al. (00)]
  BUT
  - only for particular motifs,
  - the Erdös-Rényi model is not a good model for biological networks (e.g. it does not fit the degrees).

State of the art (2/2)

Simulated approaches:

- Random networks are generated by edge swapping, (degrees are preserved)
- Empirical distributions for motif counts are obtained leading either to p-values or to z-scores
  BUT
  - huge number of simulations required to estimate tiny p-values,
  - z-scores are compared to $\mathcal{N}(0,1)$ which is not always appropriate,
  - Edge swapping does not define a probabilistic random graph model.

Our contributions (1/2)

- To propose probabilistic random graph models
  - adapted for biological networks,
  - allowing probabilistic calculations,
  - with efficient estimation procedures.

Daudin, Picard, Robin (08) *A mixture model for random graphs*. Statis. Comput.
Birmelé (07) *A scale-free graph model based on bipartite graphs*. Disc. Appl. Math.

Daudin, Picard, Robin (07) *Mixture model for random graphs: a variational approach*, SSB preprint 4
Zanghi, Ambroise, Miele (07) *Fast online graph clustering via Erdös-Rényi mixture*, SSB preprint 8.
Mariadassou, Robin (07) *Uncovering latent structure in valued graphs: a variational approach*, SSB preprint 10.

Our contributions (2/2)

- To provide general analytical results on motif count distribution:
  - mean and variance of the count in a wide class of random graph models,
  - relevant distribution to approximate the count distribution.

Matias, Schbath, Birmelé, Daudin and Robin (06) *Network motifs: mean and variance for the count*, REVSTAT. 4 31–51.

Picard, Daudin, Schbath and Robin (08) *Assessing the exceptionality of network motifs*, J. Comput. Biol.

Schbath, Lacroix and Sagot *Assessing the exceptionality of coloured motifs in networks*, submitted.

Random graph models

Random graphs

- A random graph G is defined by:
  - a set $\mathcal{V}$ of fixed vertices with $|\mathcal{V}| = n$,
  - a set of random edges $\mathbf{X} = \{X_{ij}, (i,j) \in \mathcal{V}^2\}$ such that

$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected,} \\ 0 & \text{otherwise} \end{cases}$$

  - and a distribution on X_{ij}.
- Examples:
  - the Erdös-Rényi model,
  - the Mixture model for random graph (Mixnet/ERMG),
  - the Expected Degree Distribution model.

Example 1: Erdös-Rényi model

- Edges X_{ij}'s are independent ...
- ... and identically distributed according to $\mathcal{B}(p)$

$$\mathbb{P}(X_{ij} = 1) = p$$

- Degrees are Poisson distributed

$$K_i := \sum_{j \neq i} X_{ij} \sim \mathcal{B}(n-1, p) \approx \mathcal{P}((n-1)p)$$

- It does not fit with biological networks.
  The main reason is: heterogeneity.

Example 2: Mixture model for random graphs

- Vertices are spread into Q groups.
- Conditionally to the group of vertices, edges are independent and

$$X_{ij} \mid \{i \in q, j \in \ell\} \sim \mathcal{B}(\pi_{q,\ell})$$

  $\pi_{q,\ell}$ is the connection probability between groups q and ℓ.
- Degrees are distributed according to a Poisson mixture

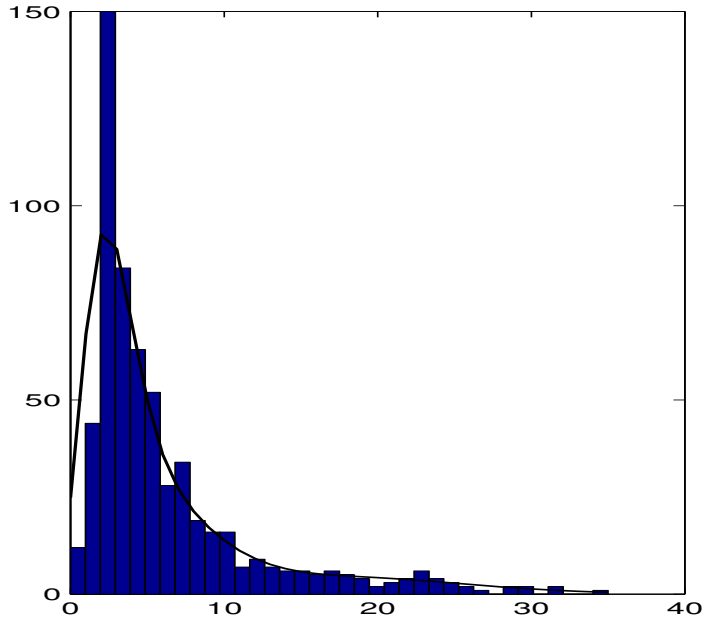
$$K_i \sim \sum_q \alpha_q \mathcal{B}(n-1, \overline{\pi}_q) \text{ with } \overline{\pi}_q = \sum_\ell \alpha_\ell \pi_{q,\ell}$$

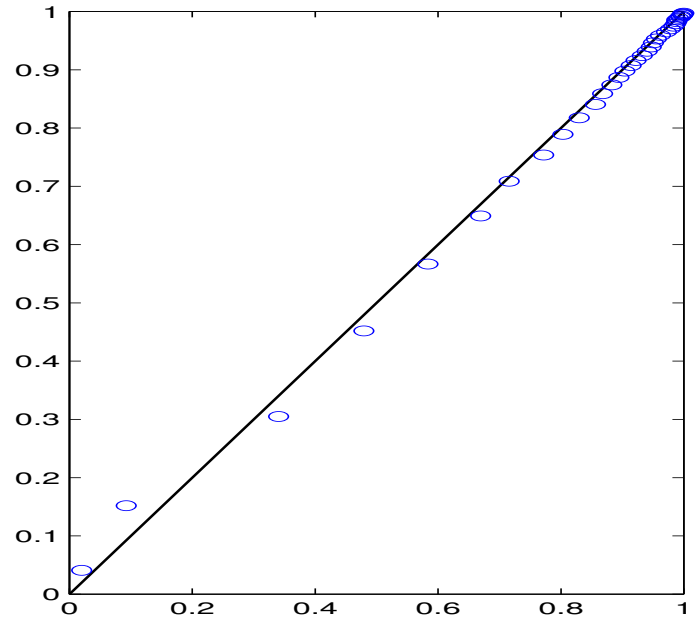
- Better fit to biological networks.

Mixnet fit

- *E. coli* reaction network: 605 vertices, 1782 edges.
  (data curated by V. Lacroix and M.-F. Sagot).
- **Degrees**: Poisson mixture versus empirical distribution

PP-plot

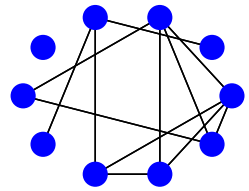
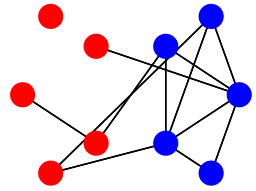
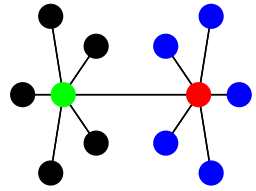
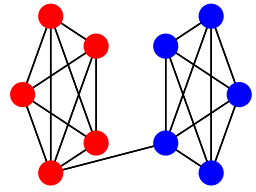




- **Clustering coefficient**:

Empirical	Mixnet ($Q = 21$)	ER ($Q = 1$)
0.626	0.544	0.0098

Mixnet flexibility

Examples	Network	Q	π
Erdös-Rényi		1	p
Independent model (product connectivity)		2	$\begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix}$
Stars		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
Clusters (affiliation network)		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$

Mixnet: estimation procedure

- The aim is to maximize the log-likelihood $\mathcal{L}(\mathbf{X})$,
- . . . but $\mathcal{L}(\mathbf{X})$ is not calculable because of hidden groups ($\mathbf{Z}$, Z_i is the group of node i).
- EM algorithm is classical to fit mixture models,
- . . . but cannot be used because $\mathbb{P}(\mathbf{Z} \mid \mathbf{X})$ is not computable (all vertices are potentially connected, no local dependence)
- Strategy = **variational approach**
  - maximization of $\mathcal{L}(\mathbf{X}) - KL(\mathbb{P}(\mathbf{Z} \mid \mathbf{X}), Q_R(\mathbf{Z}))$ where Q_R is the best approximation of $\mathbb{P}(\mathbf{Z} \mid \mathbf{X})$ within a class of 'nice' distributions. $\Rightarrow$ estimator of $\mathbb{P}(Z_i = q|\mathbf{X})$.
  - iterative algorithm
- Heuristic penalized likelihood criterion inspired from BIC (ICL)

[Daudin, Picard and Robin (07)]

Example 3: Expected Degree Distribution model

- Let f be a given distribution and D_i's i.i.d. $D_i \sim f$.
- Conditionally to the D_i's, edges X_{ij}'s are independent and

$$\mathbb{P}(X_{ij} = 1 \mid D_i, D_j) = \gamma D_i D_j$$

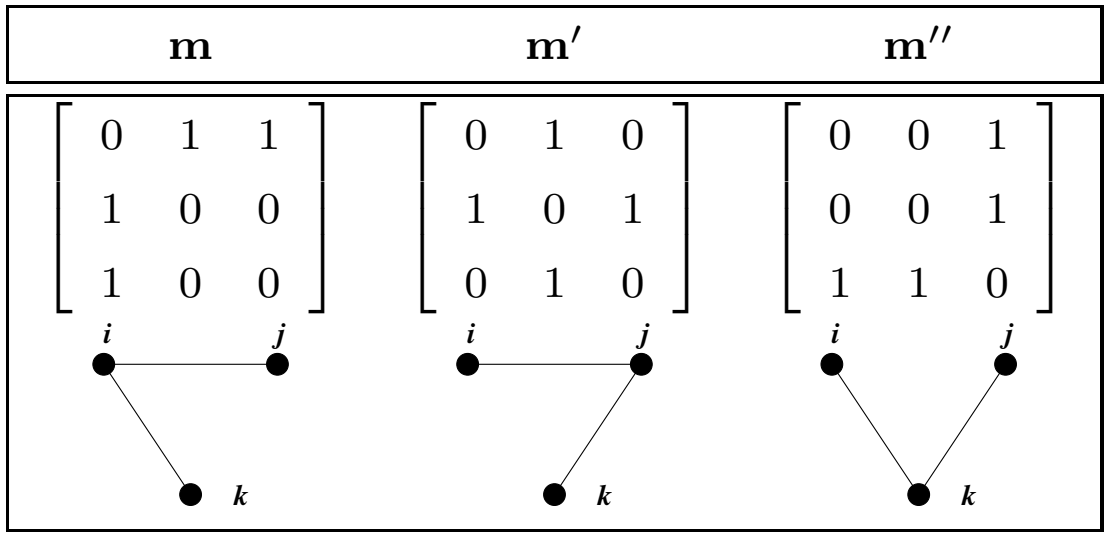
- A suitable value for γ ensures that $\mathbb{E}(K_i|D_i) = D_i$.
- This model generates graphs whose degrees follow a given distribution.

Occurrences of topological motifs

Topological motifs

Let $\mathbf{m}$ be a motif of size k (connected graph with k vertices, $k << n$).

- $\mathbf{m}$ is defined by its adjacency matrix (also denoted by $\mathbf{m}$): $\mathbf{m}_{uv} = 1$ iff nodes $u \leftrightarrow v$ ($\mathbf{m}_{uv} = 0$ otherwise).
- Let $\mathcal{R}(\mathbf{m})$ be the set of non redundant permutations of $\mathbf{m}$ (so-called "versions").
- Ex: 3 versions of the V motif at a **fixed** position (i, j, k).



Occurrences of a motif

- Let $\alpha = (i_1, \ldots, i_k) \in I_k$ be a possible position of $\mathbf{m}$ in G. G_α denotes the subgraph $(V_{i_1}, \ldots, V_{i_k})$.
- Non strict occurrences:

$$\mathbf{m} \text{ occurs at position } \alpha \Leftrightarrow \mathbf{m} \subseteq G_\alpha$$

- Random indicator of occurrence: $Y_\alpha(\mathbf{m})$

$$Y_\alpha(\mathbf{m}) = \mathbf{1}\{\mathbf{m} \text{ occurs at position } \alpha\} = \prod_{1 \leq u,v \leq k} X_{i_u i_v}^{m_{uv}}.$$

- The total count $N(\mathbf{m})$ of motif $\mathbf{m}$ is then:

$$N(\mathbf{m}) = \sum_{\alpha \in I_k} \sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}')$$

- Warning: $N(\mathbf{m}) \neq$ number of induced subgraphs ("$\mathbf{m} = G_\alpha$").

Expected count of a motif

Recall that $N(\mathbf{m}) = \sum_{\alpha \in I_k} \sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}')$

- Under "stationarity" assumption on the random graph model, the distribution of $Y_\alpha(\mathbf{m})$ does not depend on α and let us define

$$\mu(\mathbf{m}) := \mathbb{E} Y_\alpha(\mathbf{m}'), \quad \forall \alpha, \forall \mathbf{m}' \in \mathcal{R}(\mathbf{m}).$$

- The expectation of $N(\mathbf{m})$ is then:

$$\mathbb{E} N(\mathbf{m}) = \binom{n}{k} |\mathcal{R}(\mathbf{m})| \mu(\mathbf{m}).$$

Variance for the count

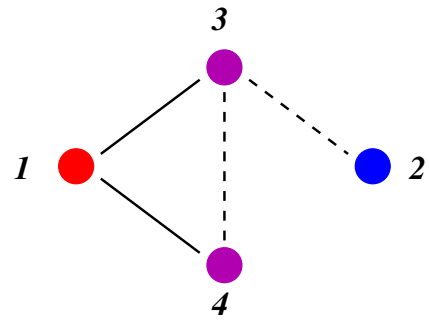
By definition $\mathbb{V}\mathrm{ar}N(\mathbf{m}) = \mathbb{E}N^2(\mathbf{m}) - [\mathbb{E}N(\mathbf{m})]^2$. We then calculate

$$\begin{aligned}
\mathbb{E}N^2(\mathbf{m}) &= \mathbb{E}\left(\sum_{\alpha,\beta\in I_k}\sum_{\mathbf{m}',\mathbf{m}''\in\mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}')Y_\beta(\mathbf{m}'')\right), \\
&= \mathbb{E}\left(\sum_{s=0}^{k}\sum_{|\alpha\cap\beta|=s}\sum_{\mathbf{m}',\mathbf{m}''\in\mathcal{R}(\mathbf{m})} Y_{\alpha\cup\beta}(\mathbf{m}'\underset{s}{\Omega}\mathbf{m}'')\right) \\
&= \sum_{s=0}^{k} C(n,k,s)\sum_{\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''}\mu(\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''),
\end{aligned}$$

where $\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''$ is a **super-motif** composed of the union of two overlapping occurrences of $\mathbf{m}'$ and $\mathbf{m}''$ sharing s common vertices.

Super-motifs

- Example for the V motif:



  - $\mathbf{m}$ occurs at $\alpha = (1, 3, 4)$, $\mathbf{m}'$ occurs at $\beta = (2, 3, 4)$,
  - in this case $\alpha \cap \beta = (3, 4)$ and $s = 2$.

- The adjacency matrix of the super-motifs $\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''$ can be easily derived from the adjacency matrices of $\mathbf{m}'$ and $\mathbf{m}''$ [Picard et al. (07)].

Occurrence probability for mean and variance

The following expressions for the mean and variance of $N(\mathbf{m})$ are valid under any stationary random graph model:

$$\begin{aligned}
\mathbb{E}N(\mathbf{m}) &= \binom{n}{k} |\mathcal{R}(\mathbf{m})| \mu(\mathbf{m}) \\
\mathbb{V}\mathrm{ar}N(\mathbf{m}) &= \sum_{s=0}^{k} C(n,k,s) \sum_{\substack{\mathbf{m}'\Omega\mathbf{m}'' \\ s}} \mu(\mathbf{m}'\underset{s}{\Omega}\mathbf{m}'') - [\mathbb{E}N(\mathbf{m})]^2.
\end{aligned}$$

The final point is now the calculation of the occurrence probability $\mu(\cdot)$ given the adjacency matrix of a motif.

Calculating $\mu(\mathbf{m})$

- The probability of occurrence of a given motif depends on the distribution of the X_{ij}'s.
- Stationary assumption: $\mu(\mathbf{m})$ does not depend on the position of the motif
- In the **Erdös-Rényi model**: $\mu(\mathbf{m}) = \pi^{v(\mathbf{m})}$, with $v(\mathbf{m})$ the number of edges in $\mathbf{m}$.
- In the **Mixnet** model with Q groups with proportion $\alpha_1, \ldots, \alpha_Q$:

$$\mu(\mathbf{m}) = \sum_{c_1=1}^{Q} \ldots \sum_{c_k=1}^{Q} \alpha_{c_1} \ldots \alpha_{c_k} \prod_{1 \leq u < v \leq k} \pi_{c_u, c_v}^{m_{uv}}.$$

- In the **EDD** model ($D \sim f$): $\mu(\mathbf{m}) = \gamma^{m_{++}/2} \prod_{u=1}^{k} \mathbb{E} D^{m_{u+}}$

Motif count distribution

- Exact distribution unknown.
- Several approximations exist (or are used) in the literature:
  - Gaussian distribution
  - Poisson distribution

Motif count distribution

- Exact distribution unknown.
- Several approximations exist (or are used) in the literature:
  - Gaussian distribution
    BUT not adapted for rare motifs
  - Poisson distribution
    BUT mean$\neq$ variance

Motif count distribution

- Exact distribution unknown.
- Several approximations exist (or are used) in the literature:
  - Gaussian distribution
    BUT not adapted for rare motifs
  - Poisson distribution
    BUT mean$\neq$ variance
  - $\Rightarrow$ Compound Poisson distribution?

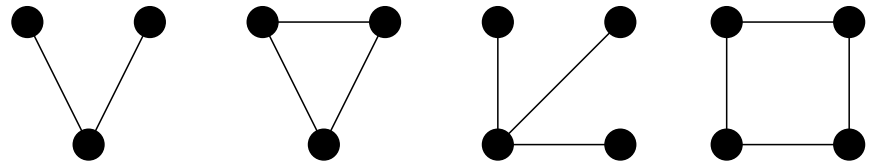
Compound Poisson distribution

- Distribution of $\sum_{i=1}^{Z} T_i$ when $Z \sim \mathcal{P}(\lambda)$ and T_i's iid.
- Particularly adapted for the count of clumping events: Z is the number of clumps and T_i is the size of the i-th clump.

- All network motifs are overlapping: they occur in clumps.
- We proposed to use a **Geometric-Poisson**(λ, a) distribution, i.e. when $T_i \approx \mathcal{G}(1-a)$
  - analogy with sequence motifs [Schbath (95)],
  - (λ, a) can be calculated according to $\mathbb{E}N(\mathbf{m})$ and $\mathbb{V}\text{ar}N(\mathbf{m})$.

Simulation study

- Aim: to compare the Gaussian, Poisson and Geometric-Poisson approximations for the motif count distribution.
- Random graph model: Mixnet with 2 groups.
- Simulation design:
  - the number of vertices $n = 20, 200$
  - the mean connectivity $\bar{\pi} = 1/n, 2/n$
  - the within/between group connectivity $\gamma = 0.1, 0.5, 0.9$
  - the proportion of the groups $\alpha = 0.1, 0.9$
- 4 motifs:

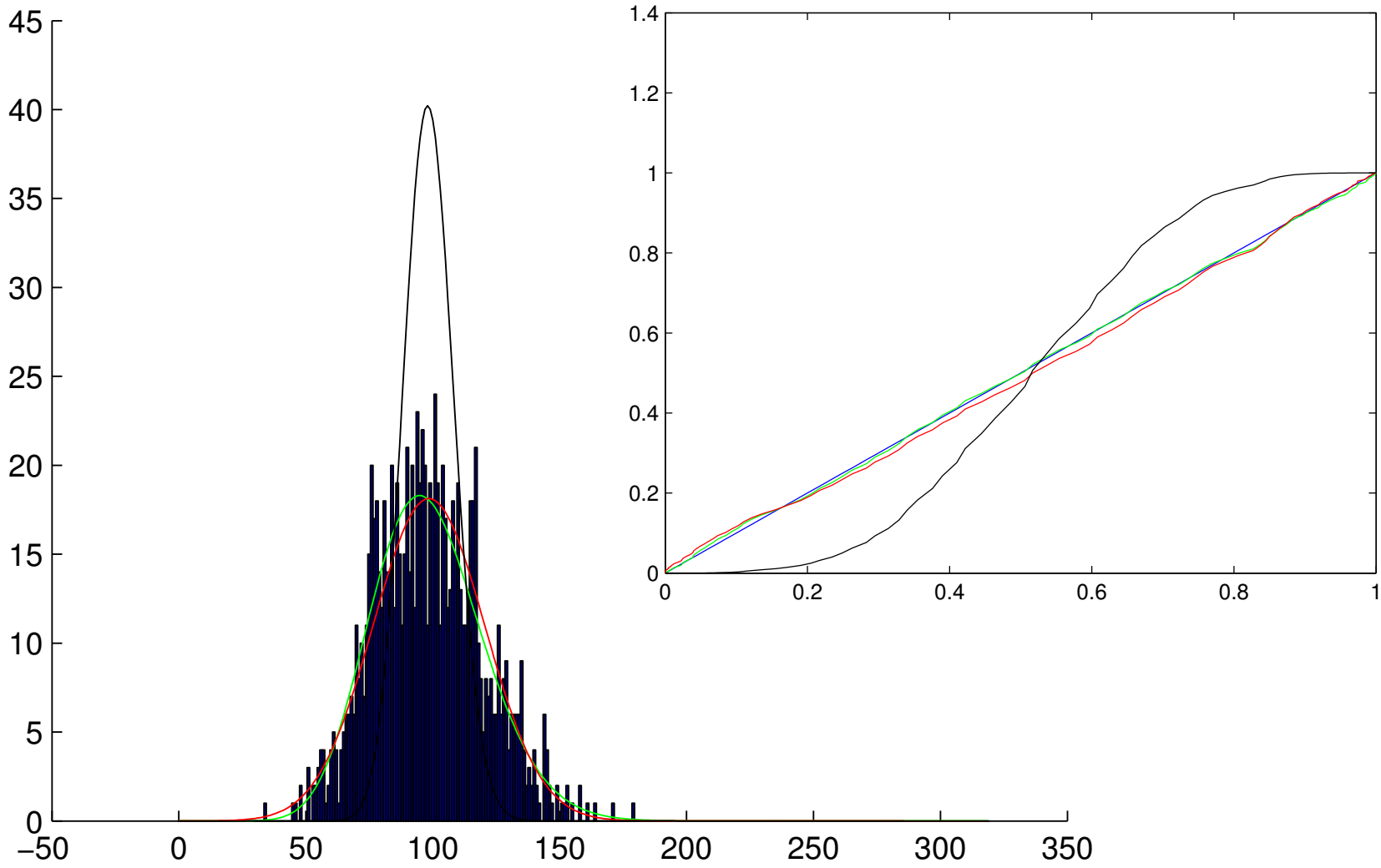


V triangle 4-star square

- We cover a large range for $\mathbb{E}N(\mathbf{m})$ (from 0.07 to 1075.5).

Expectedly frequent motif count distribution

Gaussian (–), Poisson (–) and Geometric-Poisson (–)



Approximation for expectedly frequent motif

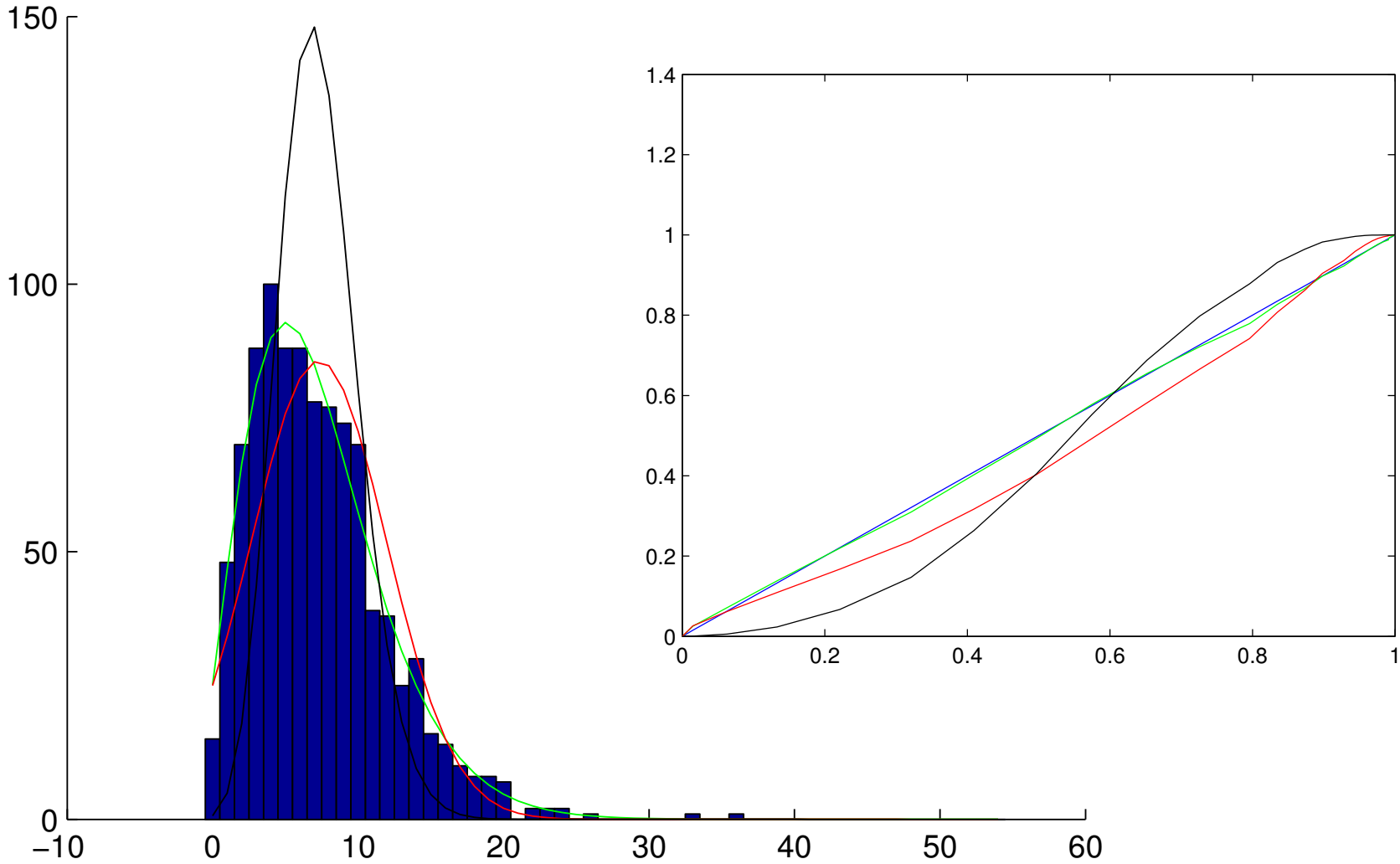
simulation parameters				motif V							
n	$\overline{\pi}$	α (%)	γ (%)	$\mathbb{E}$	$\mathbb{V}$	λ	$\frac{1}{1-a}$	D_G	D_{GP}	$\hat{F}_G$	$\hat{F}_{GP}$
200	0.5	10	10	159.5	2034.0	23.1	6.66	20.4	19.7	2.5	**1.6**
200	0.5	10	90	104.9	590.5	31.6	3.33	15.2	14	1.9	**1.2**
200	0.5	50	10	98.5	484.0	33.3	2.27	13.1	12.6	1.1	**0.7**
200	0.5	50	50	98.5	484.0	33.2	2.27	14.3	13.2	1.6	**1.1**
200	0.5	50	90	98.5	488.4	33.1	2.27	14.5	14.8	2.5	**0.9**

Criteria to assess the goodness-of-fit:

- D_G (resp. D_{GP}): **total variation distance** between the empirical dist. and the Gaussian (resp. Geometric-Poisson) dist.
- $\hat{F}_G$ (resp. $\hat{F}_{GP}$): **empirical proba. of exceeding the 0.99 quantile** of the Gaussian (resp. Geometric-Poisson) dist.

Expectedly rare motif count distribution

Gaussian (–), Poisson (–) and Geometric-Poisson (–)



Approximation for expectedly rare motif

simulation parameters				motif □							
n	$\overline{\pi}$	α (%)	γ (%)	$\mathbb{E}$	$\mathbb{V}$	λ	$\frac{1}{1-a}$	D_G	D_{GP}	$\hat{F}_G$	$\hat{F}_{GP}$
200	1	10	10	7.31	21.72	3.68	2	11.8	5.4	3.2	**0.9**
200	1	10	90	2.57	3.42	2.21	1.16	9.3	2.7	3.6	**0.5**
200	1	50	10	2.74	3.69	2.33	1.17	12.3	3.6	4.7	**1.2**
200	1	50	50	1.94	2.40	1.74	1.11	11.3	2.0	3.2	**1.6**
200	1	50	90	2.74	3.72	2.32	1.17	10.8	4.5	3.7	**0.7**

Criteria to assess the goodness-of-fit:

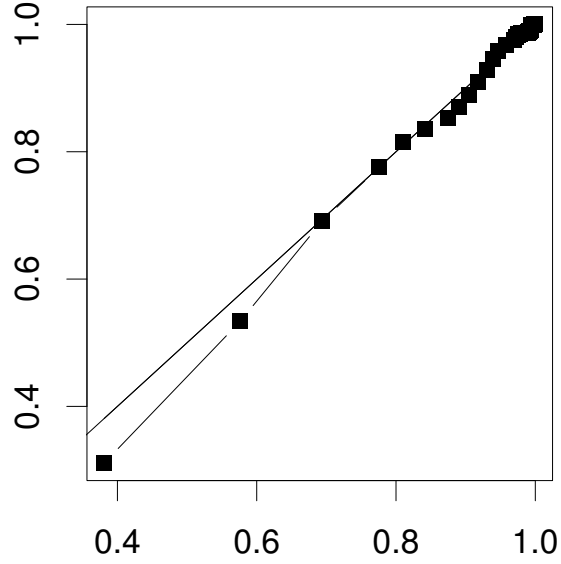
- D_G (resp. D_{GP}): total variation distance between the empirical dist. and the Gaussian (resp. Geometric-Poisson) dist.
- $\hat{F}_G$ (resp. $\hat{F}_{GP}$): empirical proba. of exceeding the 0.99 quantile of the Gaussian (resp. Geometric-Poisson) dist.

Conclusions for the simulation study

- Validation of our analytical expressions for $\mathbb{E}N$ and $\mathbb{V}\mathrm{ar}N$.
- The Poisson approximation is not satisfactory.
- The Geometric-Poisson approximation outperforms the Gaussian approximation for both criteria in all cases, especially for “rare” motifs.
- The 0.99 quantile is underestimated by the Gaussian approximation: $\longrightarrow$ false positive results.
- The total variation distance is high for both approximations in some cases, especially for frequent and highly self overlapping motifs.
- The clumps size distribution is probably not geometric . . .

Application to the *H. pylori* PPI network

- Protein-protein interaction network: 706 proteins (nodes) and 1420 interactions (edges).
- Mixnet was fitted to the network and 4 groups of connectivity were selected using a model selection criterion.
- Goodness-of-fit for the degree distribution (PP-plot):

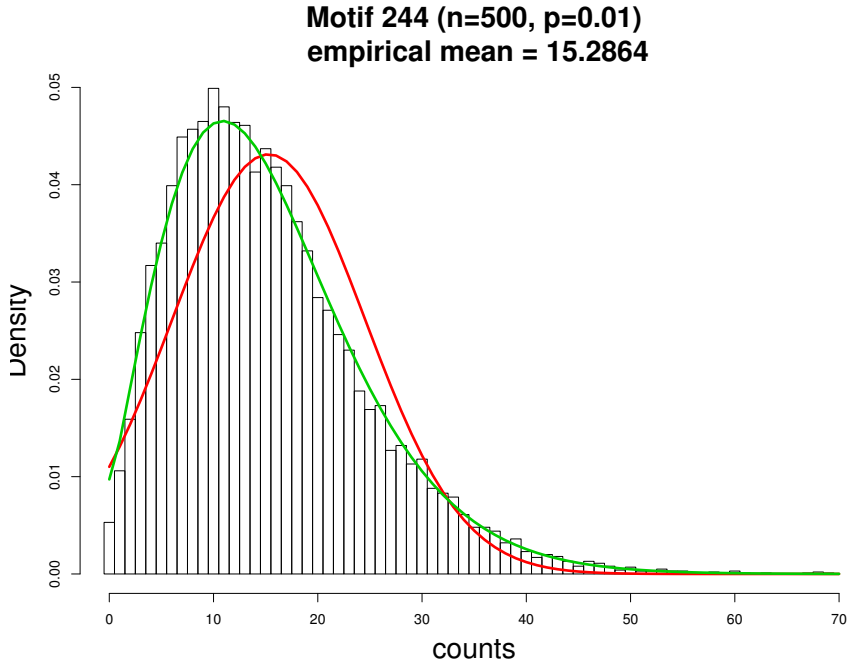


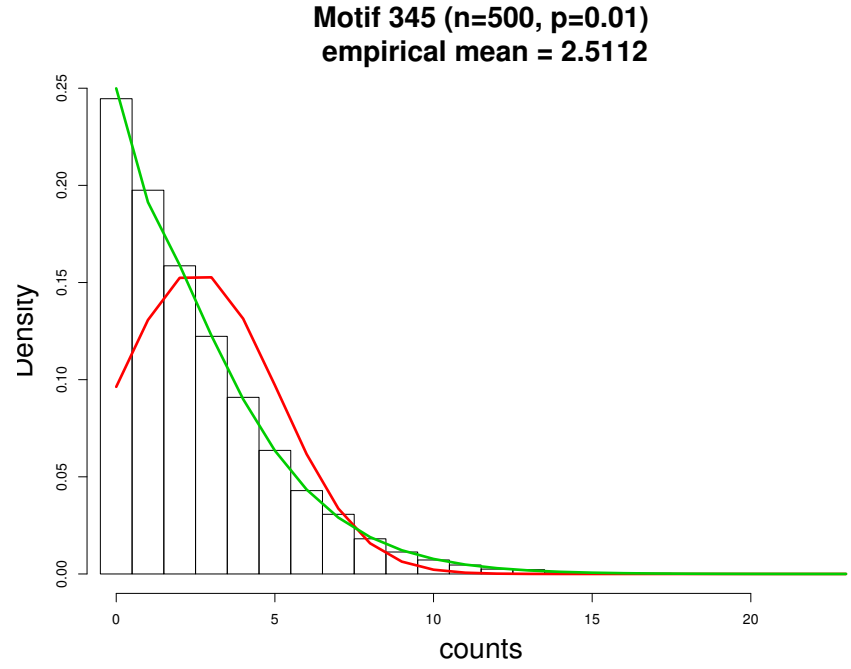
Exceptional motifs of size 3 and 4

Motif	N_{obs}	$\mathbb{E}_{\text{mixnet}} N$	$\sigma_{\text{mixnet}}(N)$	$\mathbb{P}(\mathcal{GP} \leq N_{\text{obs}})$	$\mathbb{P}(\mathcal{GP} \geq N_{\text{obs}})$
	14113	13602	2659		$4.06\,10^{-1}$
	75	66.9	20.4		$3.31\,10^{-1}$
	98697	94578	27039		$4.12\,10^{-1}$
	112490	93741	27257		$2.34\,10^{-1}$
	1058	516.6	208.7		$1.33\,10^{-2}$
	3535	2897	1120		$2.63\,10^{-1}$
	79	34.8	20.0		$3.11\,10^{-2}$
	0	0.17	0.45	$8.5\ 10^{-1}$	

What about colored motifs?

- Ongoing work : colored Erdös-Rényi model.
- Analytical mean and variance for motifs of size up to 5 (difficulty in the variance: connectivity term).
- Simulations: Geometric-Poisson distribution performs better than the Gaussian one.





Conclusions & future directions

- We have proposed a flexible mixture model to fit biological networks.
- We proposed statistical methods to assess the exceptionality of network motifs *without simulations*.
- The Geometric-Poisson approximation for the count distribution works well (better than the Gaussian) on simulated data.
- For topological motifs:
  - The method to calculate the moments of the count is general and can be applied to any random graph model with stationary distribution.
  - Results can be easily generalized to directed motifs.

Directions:

- Distribution of the clump size.
- For colored motifs:
  - General formula for the variance.
  - Generalization to Mixnet.

Acknowledgments

AgroParisTech

Jean-Jacques Daudin

Michel Koskas

Stéphane Robin

Stat&Génome, Evry

Christophe Ambroise

Etienne Birmelé

Catherine Matias

LBBE, Lyon

Vincent Miele

Franck Picard

Marie-France Sagot

Vincent Lacroix

Mixnet: model selection procedure

- Heuristic penalized likelihood criterion inspired from BIC (ICL)
- The completed log-likelihood $\mathcal{L}(\mathbf{X}, \mathbf{Z})$ is the sum

$$\sum_i \sum_q \mathbb{1}\{Z_i = q\} \log \alpha_q \qquad + \qquad \sum_{i,j>i} \sum_{q,\ell} \mathbb{1}\{Z_i = q\}\mathbb{1}\{Z_j = \ell\} \log b(X_{ij}; \pi_{q,\ell})$$

$(Q-1)$ independent proportions α_q's and n terms

$Q(Q+1)/2$ probabilities $\pi_{q,\ell}$'s and $n(n-1)/2$ terms

- The heuristic criterion is then:

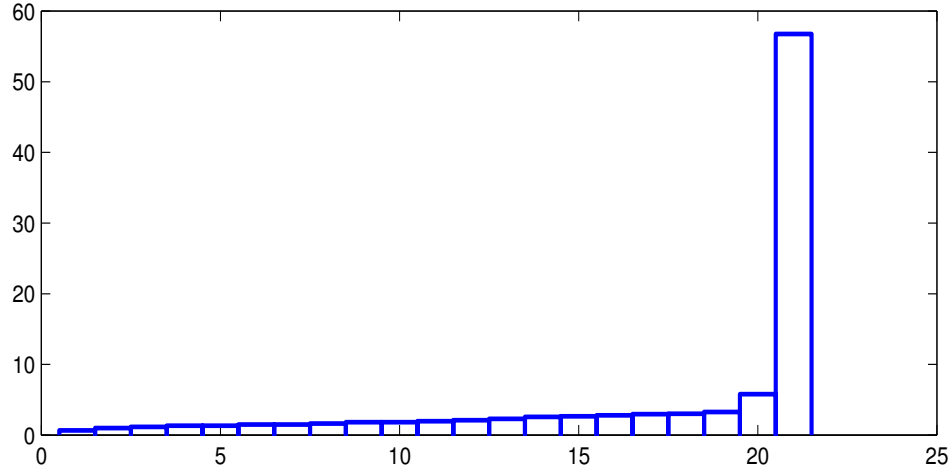
$$-2\widehat{\mathcal{L}}(\mathbf{X}, \mathbf{Z}) + (Q-1)\log n + \frac{Q(Q+1)}{2} \log\left[\frac{n(n-1)}{2}\right].$$

Illustration of Mixnet

The Mixnet has been adjusted to

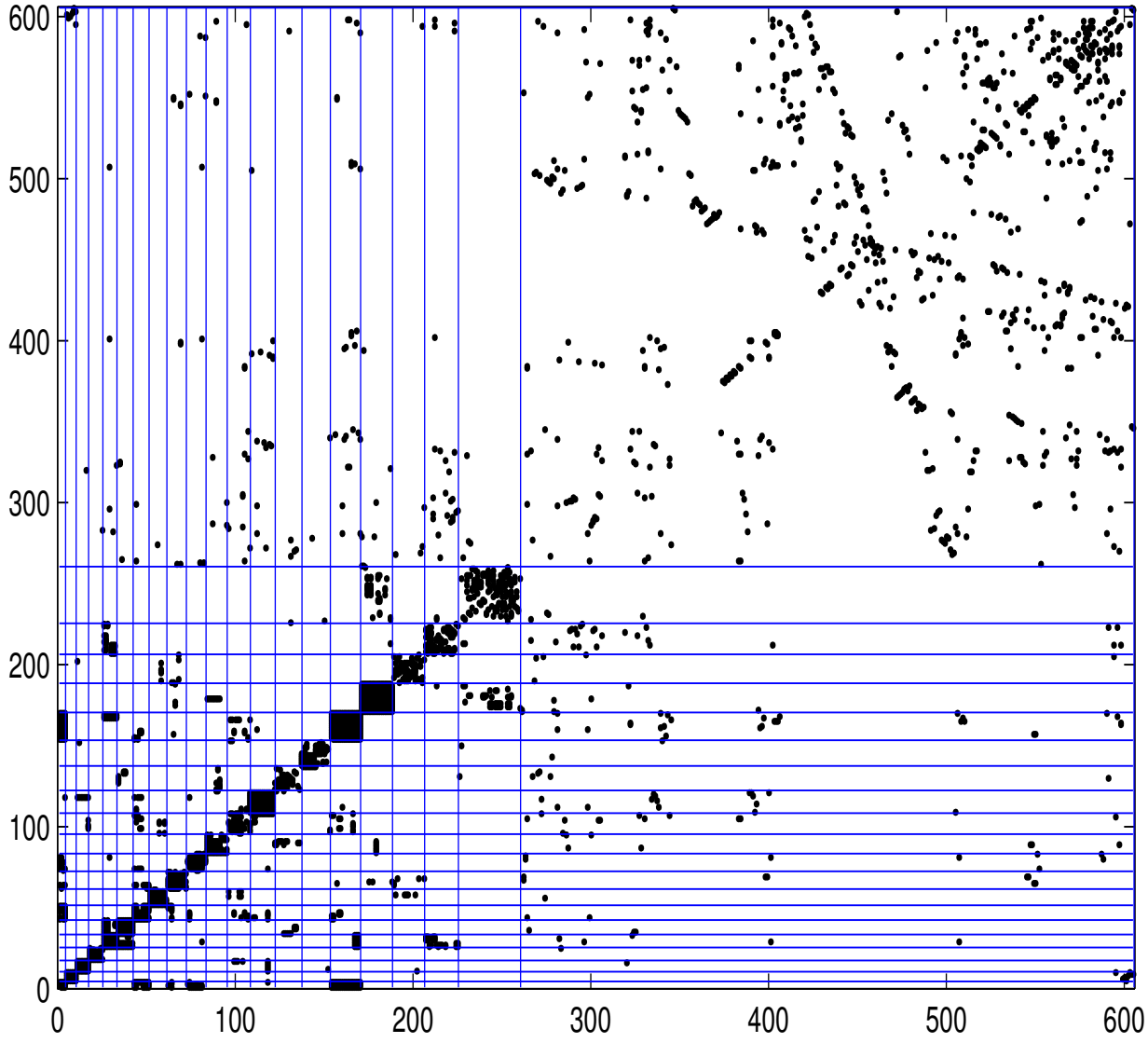
- *E. coli* reaction network: 605 vertices, 1782 edges. (data curated by V. Lacroix and M.-F. Sagot).
- $Q = 21$ groups selected

Group proportions $\widehat{\alpha}_q$ (%).



- Many small groups correspond to cliques or pseudo-cliques.

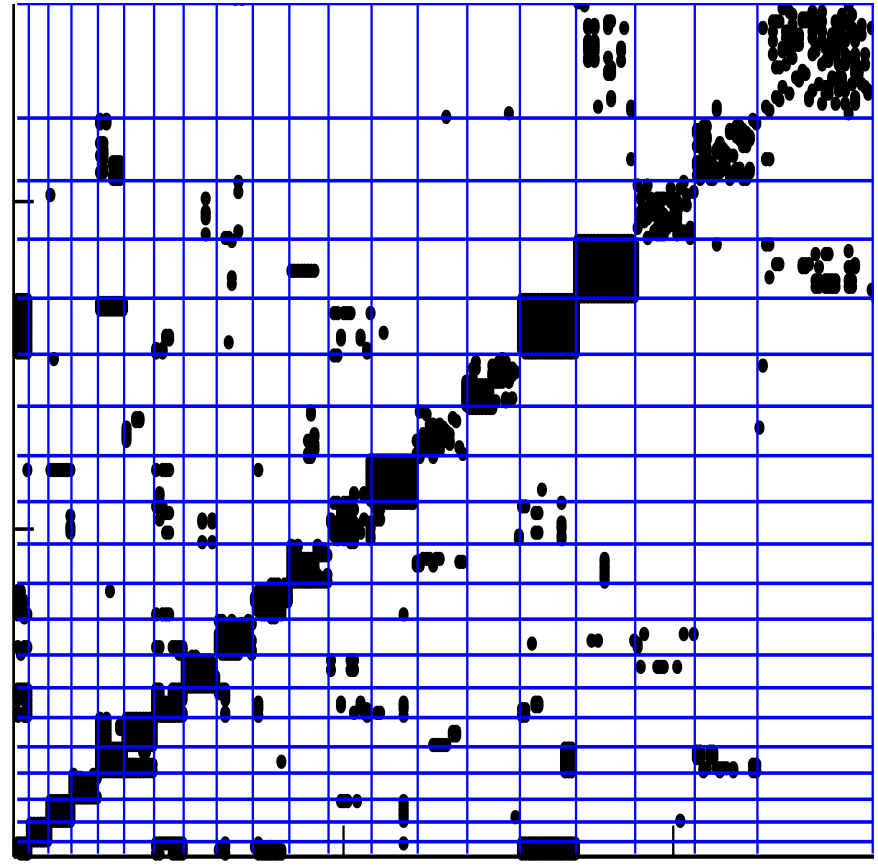
Dot plot representation of the network

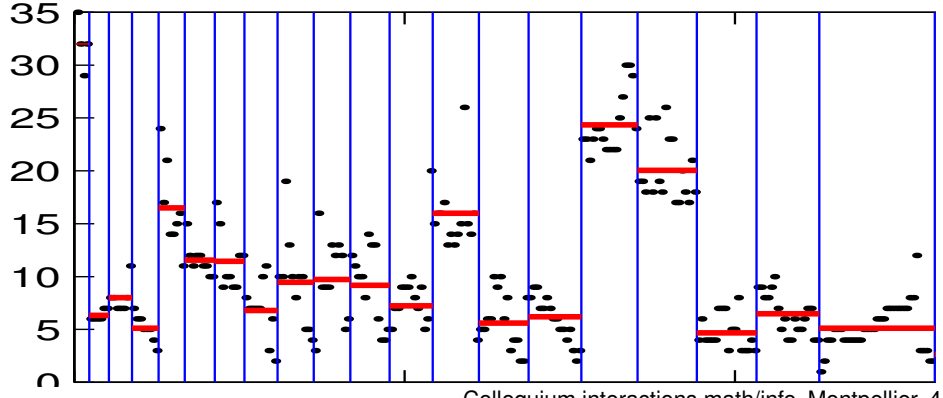


Dot plot representation (zoom)

Sub-matrix of π:

q, ℓ	1	10	16
1	**1.0**		
10	*.43*	.67	
16	**1.0**	*0*	**1.0**





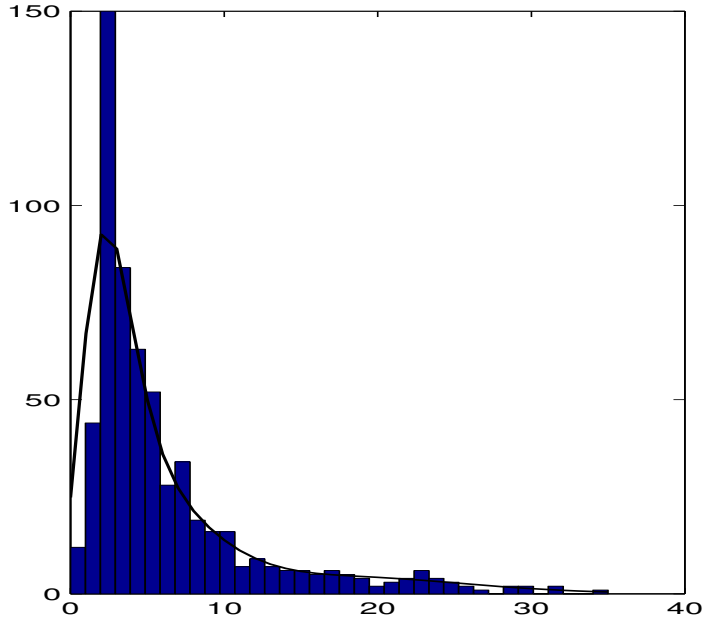
Vertex degrees K_i's.

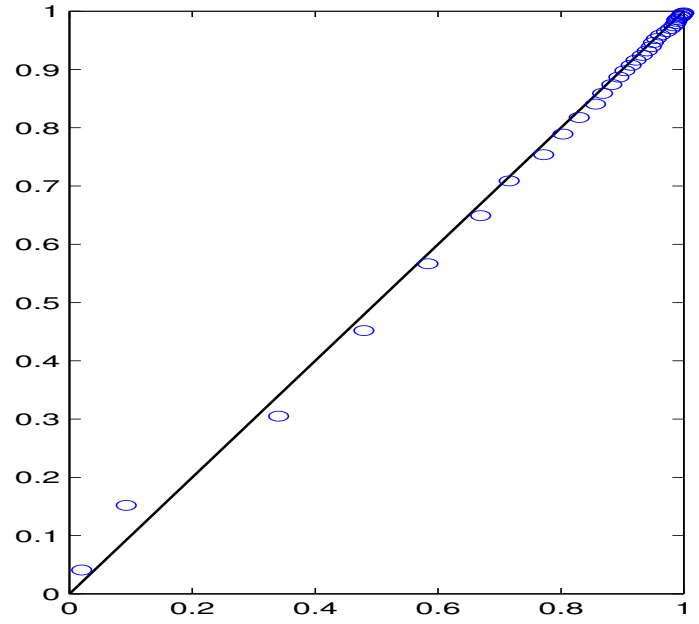
Mean degree in the last group: $\overline{K}_{21} = 2.6$

Model fit

- **Degrees**: Poisson mixture versus empirical distribution

PP-plot





- **Clustering coefficient**:

Empirical	Mixnet ($Q = 21$)	ER ($Q = 1$)
0.626	0.544	0.0098