

Some statistical approaches in random graph modeling

Catherine MATIAS

CNRS, Laboratoire Statistique & Génome, Évry, FRANCE

<http://stat.genopole.cnrs.fr/~cmatias>



Outline

Molecular interactions networks

Some statistical networks models

- Exponential random graphs

- (Overlapping) Stochastic block models

- Latent space models

Analyzing networks: (probabilistic) node clustering

Outline

Molecular interactions networks

Some statistical networks models

Exponential random graphs

(Overlapping) Stochastic block models

Latent space models

Analyzing networks: (probabilistic) node clustering

What kind of networks? (1/3)

Protein interactions networks (PIN)

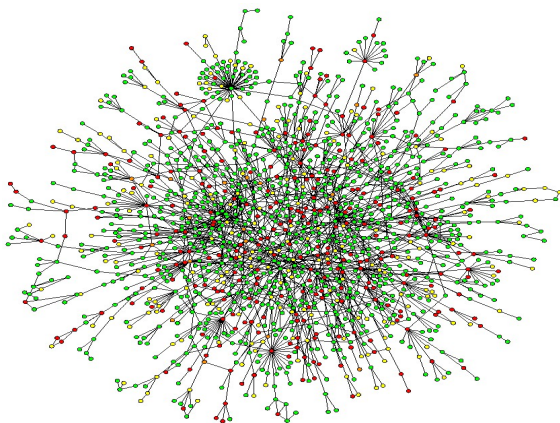


Figure: Yeast Protein Interaction Network. Source:
<http://www.bordalierinstitute.com/images/yeastProteinInteractionNetwork.jpg>

What kind of networks? (1/3)

Protein interactions networks (PIN)

- ▶ Describe possible **physical interactions** between proteins (formation of protein complex, phosphorylation cascade).
- ▶ Public databases store interactions known from the literature.
- ▶ Many interactions are based on yeast two-hybrid experiments, inducing **many false positive**.

What kind of networks? (2/3)

Metabolic networks

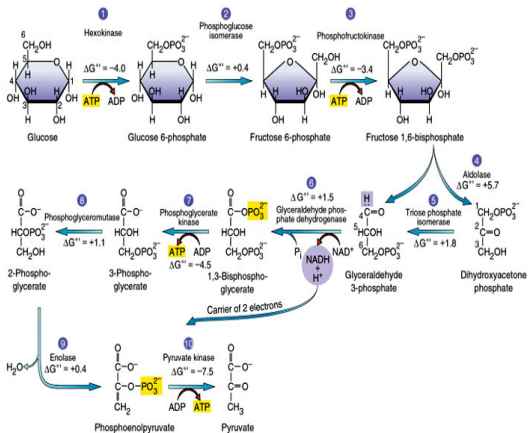


Figure: Glycolysis pathway.

What kind of networks? (2/3)

Metabolic networks

- ▶ Describe **chemical reactions** between metabolites (small molecules) transforming a substrate to a product.
- ▶ Most reactions need to be catalyzed by enzymes and are considered to be reversible.
- ▶ The metabolic networks are mostly inferred using comparative genomics techniques, inducing **many false negatives**.
- ▶ Modeled using oriented hypergraphs.

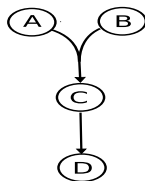


Figure: Oriented hypergraph modeling a metabolic network. Source: V. Lacroix.

What kind of networks? (3/3)

Gene regulatory networks

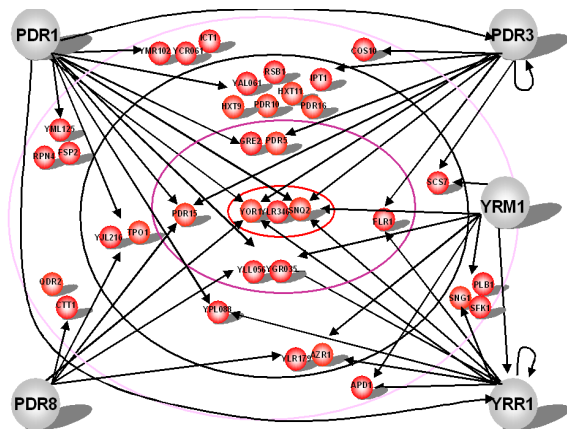


Figure: PDR network in *S. cerevisiae*. Source: Lab. Génomique de la levure, ENS.

What kind of networks? (3/3)

Gene regulatory networks

- ▶ Describe regulations (inhibitions or activations) of gene expressions, by other genes.
- ▶ Oriented graph, with positive or negative label.
- ▶ Either static or dynamic (in time).
- ▶ Mostly **statistically inferred** from transcription data sets.

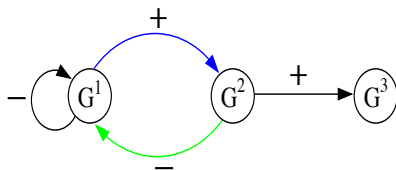


Figure: Example of a regulatory motif. Source: S. Lèbre.

Challenges

Main issues

- ▶ Analyzing large data sets (thousands of nodes and edges), with many **noise**.
- ▶ Identifying **structures** (motifs, groups, etc).
- ▶ Observed networks are **sampled** from existing ones.

Outline

Molecular interactions networks

Some statistical networks models

- Exponential random graphs

- (Overlapping) Stochastic block models

- Latent space models

Analyzing networks: (probabilistic) node clustering

Some models

Some famous models

- ▶ Erdős Rényi random graph
- ▶ Degree distribution (power law, fixed degree sequence, etc)
- ▶ Preferential attachment (dynamic model)
- ▶ ...

Here, we are going to focus on (static) 'statistical' models,

- ▶ Exponential random graph model (ERGM) [Frank & Strauss 86].
- ▶ Stochastic block model or MixNet [Frank & Harary 82, Holland *et al.* 83, Snijders & Nowicki 97, Daudin *et al.* 08].
- ▶ Overlapping stochastic block models (OSBM) [Latouche *et al.* 11a] or mixed membership SBM [Airoldi *et al.* 08].
- ▶ Latent space models [Hoff *et al.* 02, Handcock *et al.* 07].

We refer to [Goldenberg *et al.* 10] for a recent overview.

Some models

Some famous models

- ▶ Erdős Rényi random graph
- ▶ Degree distribution (power law, fixed degree sequence, etc)
- ▶ Preferential attachment (dynamic model)
- ▶ ...

Here, we are going to focus on (static) 'statistical' models,

- ▶ Exponential random graph model (ERGM) [Frank & Strauss 86].
- ▶ Stochastic block model or MixNet [Frank & Harary 82, Holland *et al.* 83, Snijders & Nowicki 97, Daudin *et al.* 08].
- ▶ Overlapping stochastic block models (OSBM) [Latouche *et al.* 11a] or mixed membership SBM [Airoldi *et al.* 08].
- ▶ Latent space models [Hoff *et al.* 02, Handcock *et al.* 07].

We refer to [Goldenberg *et al.* 10] for a recent overview.

Some models

Some famous models

- ▶ Erdős Rényi random graph
- ▶ Degree distribution (power law, fixed degree sequence, etc)
- ▶ Preferential attachment (dynamic model)
- ▶ ...

Here, we are going to focus on (static) 'statistical' models,

- ▶ Exponential random graph model (ERGM) [Frank & Strauss 86].
- ▶ Stochastic block model or MixNet [Frank & Harary 82, Holland *et al.* 83, Snijders & Nowicki 97, Daudin *et al.* 08].
- ▶ Overlapping stochastic block models (OSBM) [Latouche *et al.* 11a] or mixed membership SBM [Airoldi *et al.* 08].
- ▶ Latent space models [Hoff *et al.* 02, Handcock *et al.* 07].

We refer to [Goldenberg *et al.* 10] for a recent overview.

Exponential random graphs (1/2)

Notations

- ▶ $X = (X_{ij})_{1 \leq i, j \leq n}$ the (binary) adjacency matrix of the graph
- ▶ $S(X)$ a known vector of graph statistics on X
- ▶ θ a vector of parameters

$$\mathbb{P}_{\theta}(X = x) = \frac{1}{c(\theta)} \exp(\theta^{\top} S(x)), \quad c(\theta) = \sum_{\text{graphs } y} \exp(\theta^{\top} S(y))$$

Examples

- ▶ $S(X)$ may contain the number of edges, triangles, k -stars, ...
- ▶ S may also contain **covariates**.

Remarks

- ▶ $S(X)$ becomes a vector of **sufficient statistics**
- ▶ $c(\theta)$ is not computable

Exponential random graphs (1/2)

Notations

- ▶ $X = (X_{ij})_{1 \leq i, j \leq n}$ the (binary) adjacency matrix of the graph
- ▶ $S(X)$ a known vector of graph statistics on X
- ▶ θ a vector of parameters

$$\mathbb{P}_{\theta}(X = x) = \frac{1}{c(\theta)} \exp(\theta^{\top} S(x)), \quad c(\theta) = \sum_{\text{graphs } y} \exp(\theta^{\top} S(y))$$

Examples

- ▶ $S(X)$ may contain the number of edges, triangles, k -stars, ...
- ▶ S may also contain **covariates**.

Remarks

- ▶ $S(X)$ becomes a vector of **sufficient statistics**
- ▶ $c(\theta)$ is not computable

Exponential random graphs (1/2)

Notations

- ▶ $X = (X_{ij})_{1 \leq i, j \leq n}$ the (binary) adjacency matrix of the graph
- ▶ $S(X)$ a known vector of graph statistics on X
- ▶ θ a vector of parameters

$$\mathbb{P}_{\theta}(X = x) = \frac{1}{c(\theta)} \exp(\theta^{\top} S(x)), \quad c(\theta) = \sum_{\text{graphs } y} \exp(\theta^{\top} S(y))$$

Examples

- ▶ $S(X)$ may contain the number of edges, triangles, k -stars, ...
- ▶ S may also contain **covariates**.

Remarks

- ▶ $S(X)$ becomes a vector of **sufficient statistics**
- ▶ $c(\theta)$ is not computable

Exponential random graphs (1/2)

Notations

- ▶ $X = (X_{ij})_{1 \leq i, j \leq n}$ the (binary) adjacency matrix of the graph
- ▶ $S(X)$ a known vector of graph statistics on X
- ▶ θ a vector of parameters

$$\mathbb{P}_{\theta}(X = x) = \frac{1}{c(\theta)} \exp(\theta^{\top} S(x)), \quad c(\theta) = \sum_{\text{graphs } y} \exp(\theta^{\top} S(y))$$

Examples

- ▶ $S(X)$ may contain the number of edges, triangles, k -stars, ...
- ▶ S may also contain **covariates**.

Remarks

- ▶ $S(X)$ becomes a vector of **sufficient statistics**
- ▶ $c(\theta)$ is not computable

Exponential random graphs (2/2)

Issues on parameter estimation

- ▶ Maximum likelihood estimation is difficult
- ▶ Maximum pseudo-likelihood estimators may be used [Frank & Strauss 86]. **Quality of approximation ?**
- ▶ MCMC approaches [Hunter *et al.* 11]: may be slow to converge
- ▶ Very different values of θ can give rise to essentially the same distribution
- ▶ [Chatterjee & Diaconis 11] established a 'degeneracy' of these models, which are 'ill-posed'

Other issues

- ▶ What about clustering the nodes ?

Exponential random graphs (2/2)

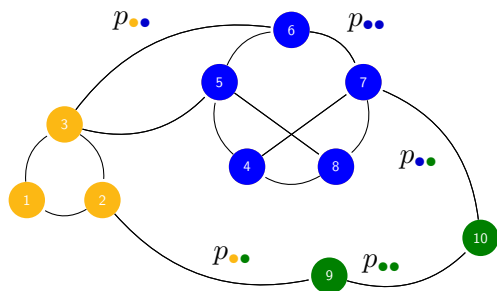
Issues on parameter estimation

- ▶ Maximum likelihood estimation is difficult
- ▶ Maximum pseudo-likelihood estimators may be used [Frank & Strauss 86]. **Quality of approximation ?**
- ▶ MCMC approaches [Hunter *et al.* 11]: may be slow to converge
- ▶ Very different values of θ can give rise to essentially the same distribution
- ▶ [Chatterjee & Diaconis 11] established a 'degeneracy' of these models, which are 'ill-posed'

Other issues

- ▶ What about clustering the nodes ?

Stochastic block model (binary graphs)

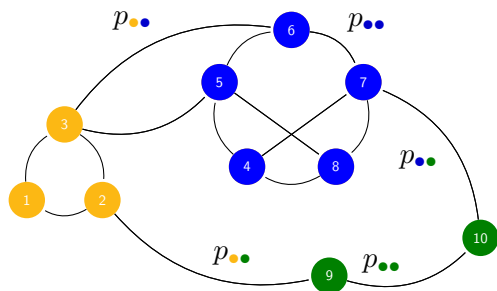


$$n = 10, Z_{5\bullet} = 1 \\ X_{12} = 1, X_{15} = 0$$

Binary case

- ▶ Q groups (=colors $\bullet\bullet\bullet$).
- ▶ $\{Z_i\}_{1 \leq i \leq n}$ i.i.d. vectors $Z_i = (Z_{i1}, \dots, Z_{iQ}) \sim \mathcal{M}(1, \pi)$, where $\pi = (\pi_1, \dots, \pi_Q)$ group proportions. Z_i is not observed,
- ▶ Observations: edges indicator X_{ij} , $1 \leq i < j \leq n$,
- ▶ Conditional on the $\{Z_i\}$'s, the random variables X_{ij} are independent $\mathcal{B}(p_{Z_i Z_j})$.

Stochastic block model (binary graphs)



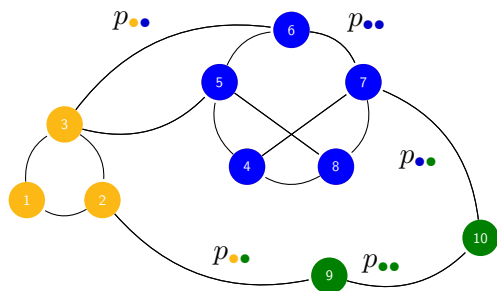
$$n = 10, Z_{5\bullet} = 1$$

$$X_{12} = 1, X_{15} = 0$$

Binary case

- ▶ Q groups (=colors $\bullet\bullet\bullet$).
- ▶ $\{Z_i\}_{1 \leq i \leq n}$ i.i.d. vectors $Z_i = (Z_{i1}, \dots, Z_{iQ}) \sim \mathcal{M}(1, \pi)$, where $\pi = (\pi_1, \dots, \pi_Q)$ group proportions. Z_i is not observed,
- ▶ Observations: edges indicator X_{ij} , $1 \leq i < j \leq n$,
- ▶ Conditional on the $\{Z_i\}$'s, the random variables X_{ij} are independent $\mathcal{B}(p_{Z_i Z_j})$.

Stochastic block model (binary graphs)

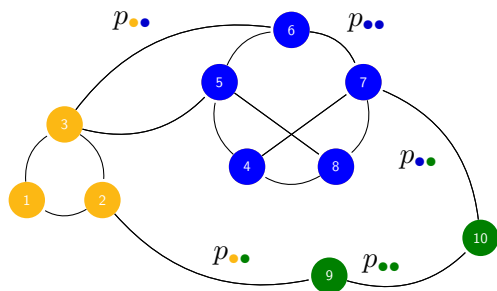


$$n = 10, Z_{5\bullet} = 1 \\ X_{12} = 1, X_{15} = 0$$

Binary case

- ▶ Q groups (=colors $\bullet\bullet\bullet$).
- ▶ $\{Z_i\}_{1 \leq i \leq n}$ i.i.d. vectors $Z_i = (Z_{i1}, \dots, Z_{iQ}) \sim \mathcal{M}(1, \pi)$, where $\pi = (\pi_1, \dots, \pi_Q)$ group proportions. Z_i is not observed,
- ▶ Observations: edges indicator X_{ij} , $1 \leq i < j \leq n$,
- ▶ Conditional on the $\{Z_i\}$'s, the random variables X_{ij} are independent $\mathcal{B}(p_{Z_i Z_j})$.

Stochastic block model (binary graphs)

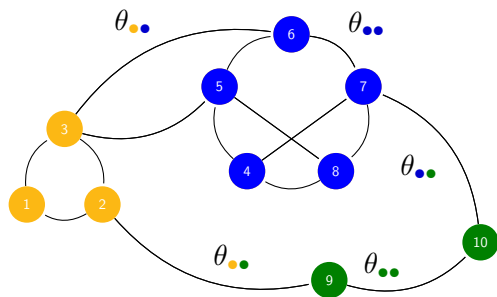


$$n = 10, Z_{5\bullet} = 1$$
$$X_{12} = 1, X_{15} = 0$$

Binary case

- ▶ Q groups (=colors $\bullet\bullet\bullet$).
- ▶ $\{Z_i\}_{1 \leq i \leq n}$ i.i.d. vectors $Z_i = (Z_{i1}, \dots, Z_{iQ}) \sim \mathcal{M}(1, \pi)$, where $\pi = (\pi_1, \dots, \pi_Q)$ group proportions. Z_i is not observed,
- ▶ Observations: edges indicator X_{ij} , $1 \leq i < j \leq n$,
- ▶ Conditional on the $\{Z_i\}$'s, the random variables X_{ij} are independent $\mathcal{B}(p_{Z_i Z_j})$.

Stochastic block model (weighted graphs)



$$n = 10, Z_{5\bullet} = 1$$
$$X_{12} \in \mathbb{R}, X_{15} = 0$$

Weighted case

- ▶ Observations: weights X_{ij} , where $X_{ij} = 0$ or $X_{ij} \in \mathbb{R}^s \setminus \{0\}$,
- ▶ Conditional on the $\{Z_i\}$'s, the random variables X_{ij} are independent with distribution

$$\mu_{Z_i Z_j}(\cdot) = p_{Z_i Z_j} f(\cdot, \theta_{Z_i Z_j}) + (1 - p_{Z_i Z_j}) \delta_0(\cdot)$$

(Assumption: f has continuous cdf at zero).

SBM properties

Results

- ▶ **Identifiability** of parameters [Allman *et al.* 09, Allman *et al.* 11].
- ▶ **Parameter estimation / node clustering** procedures:
exact EM approach is not possible

SBM properties

Results

- ▶ **Identifiability** of parameters [Allman *et al.* 09, Allman *et al.* 11].
- ▶ **Parameter estimation / node clustering** procedures:
variational EM [Daudin *et al.* 08, Picard *et al.* 09], variational Bayes [Latouche *et al.* 11b], online variational EM [Zanghi *et al.* 08], other methods [Ambroise & Matias 10] . . .

SBM properties

Results

- ▶ **Identifiability** of parameters [Allman *et al.* 09, Allman *et al.* 11].
- ▶ **Parameter estimation / node clustering** procedures:
variational EM [Daudin *et al.* 08, Picard *et al.* 09], variational Bayes [Latouche *et al.* 11b], online variational EM [Zanghi *et al.* 08], other methods [Ambroise & Matias 10] . . .
- ▶ **Model selection criteria** [Daudin *et al.* 08, Latouche *et al.* 11b]

SBM properties

Results

- ▶ **Identifiability** of parameters [Allman *et al.* 09, Allman *et al.* 11].
- ▶ **Parameter estimation / node clustering** procedures:
variational EM [Daudin *et al.* 08, Picard *et al.* 09], variational Bayes [Latouche *et al.* 11b], online variational EM [Zanghi *et al.* 08], other methods [Ambroise & Matias 10] . . .
- ▶ **Model selection criteria** [Daudin *et al.* 08, Latouche *et al.* 11b]

Remaining challenges

- ▶ Behavior of the nodes posterior dist. / Quality of variational approx. ?
- ▶ Consistency of the MLE ?
(ongoing works of Céliste, Daudin & Pierre and Mariadassou & Matias)

Overlapping SBM / Mixed membership SBM

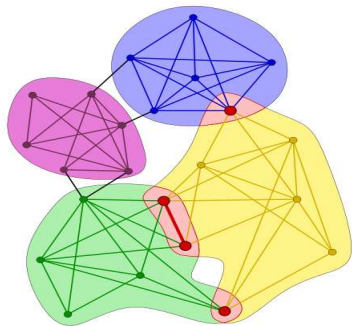


Figure: Overlapping mixture model. Source: Palla *et al.*, Nature, 2005.

Nodes may belong to many classes
[Latouche *et al.* 11a, Airoldi *et al.* 08].

OSBM [Latouche et al. 11a]

Model

- ▶ $Z_i = (Z_{i1}, \dots, Z_{iQ}) \sim \prod_{q=1}^Q \mathcal{B}(\pi_q)$
- ▶ $X_{ij} | Z_i, Z_j \sim \mathcal{B}(g(p_{Z_i Z_j}))$ where $g(x) = (1 + e^{-x})^{-1}$ (logistic function) and

$$p_{Z_i Z_j} = Z_i^\top W Z_j + Z_i^\top U + V^\top Z_j + \omega$$

W is a $Q \times Q$ real matrix while U and V are Q -dimensional real vectors and ω real number.

Results [Latouche et al. 11a]

- ▶ Parameter's identifiability
- ▶ Variational Bayes approach + variational logistic Bayes
- ▶ Model selection criterion

Issues

- ▶ Quality of (double) variational approximation ?

OSBM [Latouche et al. 11a]

Model

- ▶ $Z_i = (Z_{i1}, \dots, Z_{iQ}) \sim \prod_{q=1}^Q \mathcal{B}(\pi_q)$
- ▶ $X_{ij} | Z_i, Z_j \sim \mathcal{B}(g(p_{Z_i Z_j}))$ where $g(x) = (1 + e^{-x})^{-1}$ (logistic function) and

$$p_{Z_i Z_j} = Z_i^\top W Z_j + Z_i^\top U + V^\top Z_j + \omega$$

W is a $Q \times Q$ real matrix while U and V are Q -dimensional real vectors and ω real number.

Results [Latouche et al. 11a]

- ▶ Parameter's identifiability
- ▶ Variational Bayes approach + variational logistic Bayes
- ▶ Model selection criterion

Issues

- ▶ Quality of (double) variational approximation ?

OSBM [Latouche et al. 11a]

Model

- ▶ $Z_i = (Z_{i1}, \dots, Z_{iQ}) \sim \prod_{q=1}^Q \mathcal{B}(\pi_q)$
- ▶ $X_{ij} | Z_i, Z_j \sim \mathcal{B}(g(p_{Z_i Z_j}))$ where $g(x) = (1 + e^{-x})^{-1}$ (logistic function) and

$$p_{Z_i Z_j} = Z_i^\top W Z_j + Z_i^\top U + V^\top Z_j + \omega$$

W is a $Q \times Q$ real matrix while U and V are Q -dimensional real vectors and ω real number.

Results [Latouche et al. 11a]

- ▶ Parameter's identifiability
- ▶ Variational Bayes approach + variational logistic Bayes
- ▶ Model selection criterion

Issues

- ▶ Quality of (double) variational approximation ?

Latent space models [Handcock *et al.* 07]

Model

- ▶ Z_i i.i.d. vectors in a *latent space* \mathbb{R}^d .
- ▶ Conditional on $\{Z_i\}$, the $\{X_{ij}\}$ are independent Bernoulli r.v.
 $\log\text{-odds}(X_{ij} = 1 | Z_i, Z_j, U_{ij}, \theta) = \theta_0 + \theta_1^T U_{ij} - \|Z_i - Z_j\|$,
where $\log\text{-odds}(A) = \log \mathbb{P}(A) / (1 - \mathbb{P}(A))$; $\{U_{ij}\}$ set of covariate vectors and θ parameters vector.
- ▶ This may be extended to weighted networks

Results [Handcock *et al.* 07]

- ▶ Two-stage maximum likelihood or MCMC procedures are used to infer the model's parameters
- ▶ Assuming Z_i sampled from mixture of multivariate normal, one may obtain a clustering of the nodes.

Issues

- ▶ No model selection procedure to infer the 'effective' dimension d of latent space and the number of groups

Latent space models [Handcock *et al.* 07]

Model

- ▶ Z_i i.i.d. vectors in a *latent space* \mathbb{R}^d .
- ▶ Conditional on $\{Z_i\}$, the $\{X_{ij}\}$ are independent Bernoulli r.v.
 $\log\text{-odds}(X_{ij} = 1 | Z_i, Z_j, U_{ij}, \theta) = \theta_0 + \theta_1^T U_{ij} - \|Z_i - Z_j\|$,
where $\log\text{-odds}(A) = \log \mathbb{P}(A) / (1 - \mathbb{P}(A))$; $\{U_{ij}\}$ set of covariate vectors and θ parameters vector.
- ▶ This may be extended to weighted networks

Results [Handcock *et al.* 07]

- ▶ Two-stage maximum likelihood or MCMC procedures are used to infer the model's parameters
- ▶ Assuming Z_i sampled from mixture of multivariate normal, one may obtain a clustering of the nodes.

Issues

- ▶ No model selection procedure to infer the 'effective' dimension d of latent space and the number of groups

Latent space models [Handcock *et al.* 07]

Model

- ▶ Z_i i.i.d. vectors in a *latent space* \mathbb{R}^d .
- ▶ Conditional on $\{Z_i\}$, the $\{X_{ij}\}$ are independent Bernoulli r.v.
 $\log\text{-odds}(X_{ij} = 1 | Z_i, Z_j, U_{ij}, \theta) = \theta_0 + \theta_1^T U_{ij} - \|Z_i - Z_j\|$,
where $\log\text{-odds}(A) = \log \mathbb{P}(A) / (1 - \mathbb{P}(A))$; $\{U_{ij}\}$ set of covariate vectors and θ parameters vector.
- ▶ This may be extended to weighted networks

Results [Handcock *et al.* 07]

- ▶ Two-stage maximum likelihood or MCMC procedures are used to infer the model's parameters
- ▶ Assuming Z_i sampled from mixture of multivariate normal, one may obtain a clustering of the nodes.

Issues

- ▶ No model selection procedure to infer the 'effective' dimension d of latent space and the number of groups

Outline

Molecular interactions networks

Some statistical networks models

- Exponential random graphs

- (Overlapping) Stochastic block models

- Latent space models

Analyzing networks: (probabilistic) node clustering

Clustering the nodes of a network

Probabilistic approach

- ▶ Using either mixture or overlapping mixture models, one may recover nodes groups.
- ▶ These groups reflect a common 'connectivity behaviour'.

Non probabilistic approach = community detection

- ▶ Many clustering methods try to group the nodes that belong to the same **clique**.
- ▶ Here the nodes in the same groups tend to be connected with each other.

Clustering the nodes of a network

Probabilistic approach

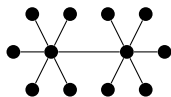
- ▶ Using either mixture or overlapping mixture models, one may recover nodes groups.
- ▶ These groups reflect a common 'connectivity behaviour'.

Non probabilistic approach = community detection

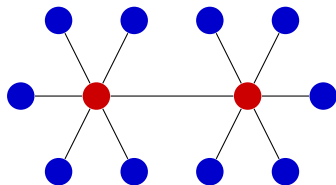
- ▶ Many clustering methods try to group the nodes that belong to the same **clique**.
- ▶ Here the nodes in the same groups tend to be connected with each other.

Major difference between probabilistic/non probabilistic approach

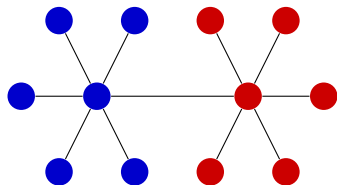
Observation of



may lead to either



MixNet model



Clustering based on cliques

Estimating and Understanding Exponential Random Graph Models.

arXiv:1102.2650, 2011.



[Daudin *et al.* 08] J-J. Daudin, F. Picard and S. Robin.

A mixture model for random graphs.

Statist. Comput., 18(2):173-183, 2008.



[Frank & Harary 82] O. Frank and F. Harary.

Cluster inference by using transitivity indices in empirical graphs.

J. Amer. Statist. Assoc., 77(380):835-840, 1982.



[Frank & Strauss 86] O. Frank and D. Strauss.

Markov graphs.

J. Amer. Statist. Assoc., 81(395):832-842, 1986.



[Goldenberg *et al.* 10] A. Goldenberg, A.X. Zheng, S.E.

Fienberg and E.M. Airoldi.

A Survey of Statistical Network Models.

Found. Trends Mach. Learn., 2(2):129-233, 2010.

Overlapping Stochastic Block Models With Application to the French Political Blogosphere.

Annals of Applied Statistics, 5(1):309-336, 2011.



[Latouche *et al.* 11b] P. Latouche, E. Birmelé and C. Ambroise.

Variational Bayesian Inference and Complexity Control for Stochastic Block Models.

Statistical Modelling, (arXiv:0912.2873), to appear.



[Picard *et al.* 09] F. Picard, V. Miele, J-J. Daudin, L. Cottret and S. Robin.

Deciphering the connectivity structure of biological networks using MixNet.

BMC Bioinformatics, 10:1-11, 2009.



[Snijders & Nowicki 97] T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure.

J. Classification 14(1):75-100, 1997.



[Zanghi *et al.* 08] H. Zanghi, C. Ambroise and V. Miele.

Fast Online Graph Clustering via Erdős Rényi Mixture.
Pattern Recognition, 41(12):3592-3599, 2008.