The stochastic block-model and its (variational) inference

S. Robin

INRA / AgroParisTech /univ. Paris-Saclay

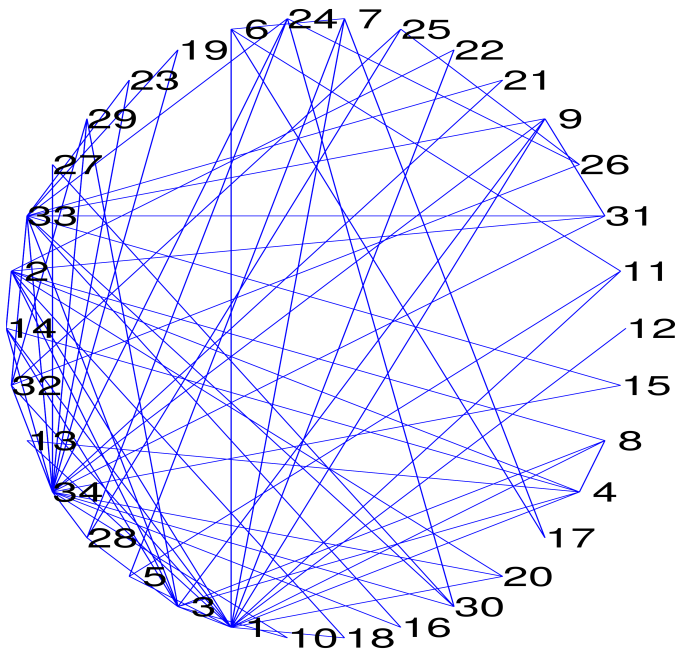
3ème journée du groupe Statistique Mathématique de la SFdS

Graphes aléatoires et Statistique

Paris, IHP, Janvier 2019

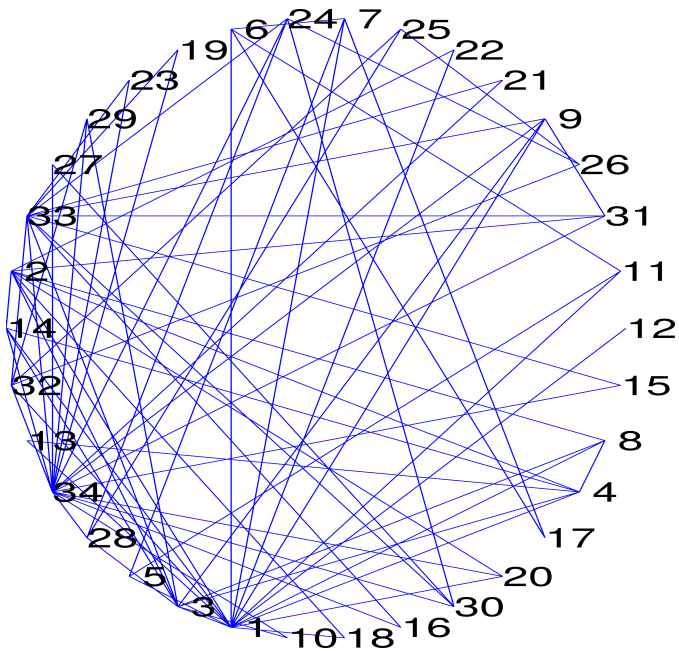
Karate club

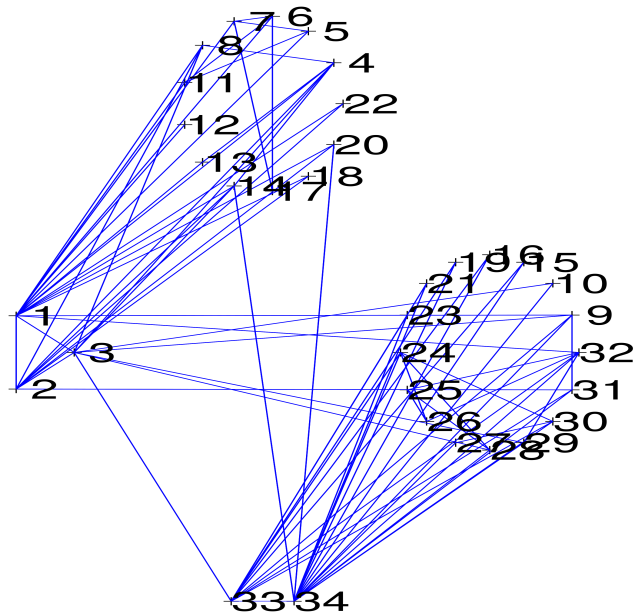
$n = 34$ nodes (individuals), link = mutual friendship



Karate club

$n = 34$ nodes (individuals), link = mutual friendship





Tree ecological network

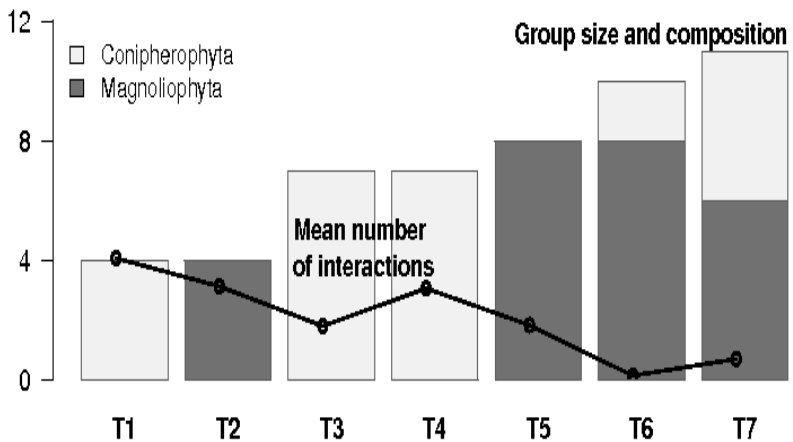
- $n = 51$ tree species,
- Y_{ij} = number of shared parasites,
- x_{ij} = (taxonomic, geographic, genetic distances)

Mariadassou & al., Ann. Applied Stat., 2010

Tree ecological network

- $n = 51$ tree species,
- Y_{ij} = number of shared parasites,
- x_{ij} = (taxonomic, geographic, genetic distances)

Without covariates: $\mathcal{P}(e^{\gamma_{k\ell}})$



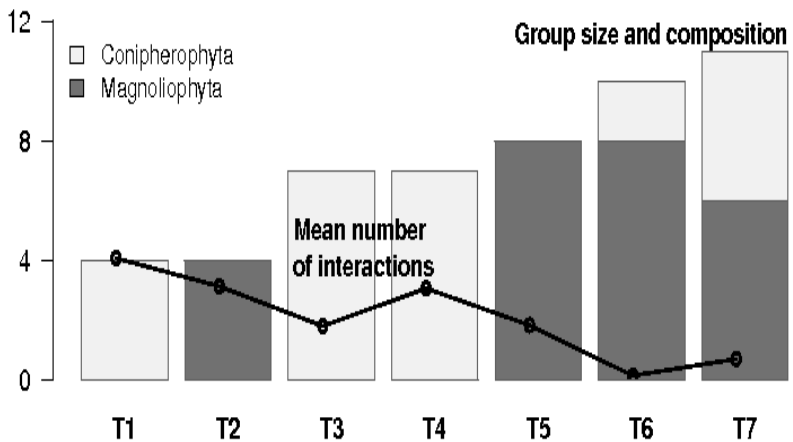
Mariadassou & al., Ann. Applied Stat., 2010

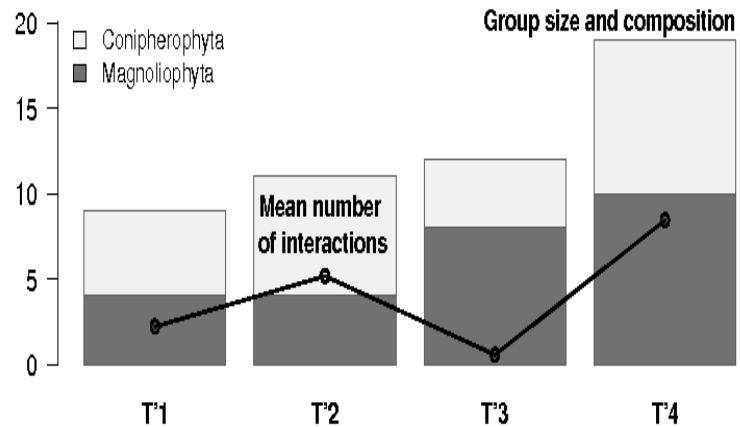
Tree ecological network

- $n = 51$ tree species,
- Y_{ij} = number of shared parasites,
- x_{ij} = (taxonomic, geographic, genetic distances)

Without covariates: $\mathcal{P}(e^{\gamma_{k\ell}})$

With covariates: $\mathcal{P}(e^{\gamma_{k\ell} + x_{ij}^{\mathsf{T}} \beta})$

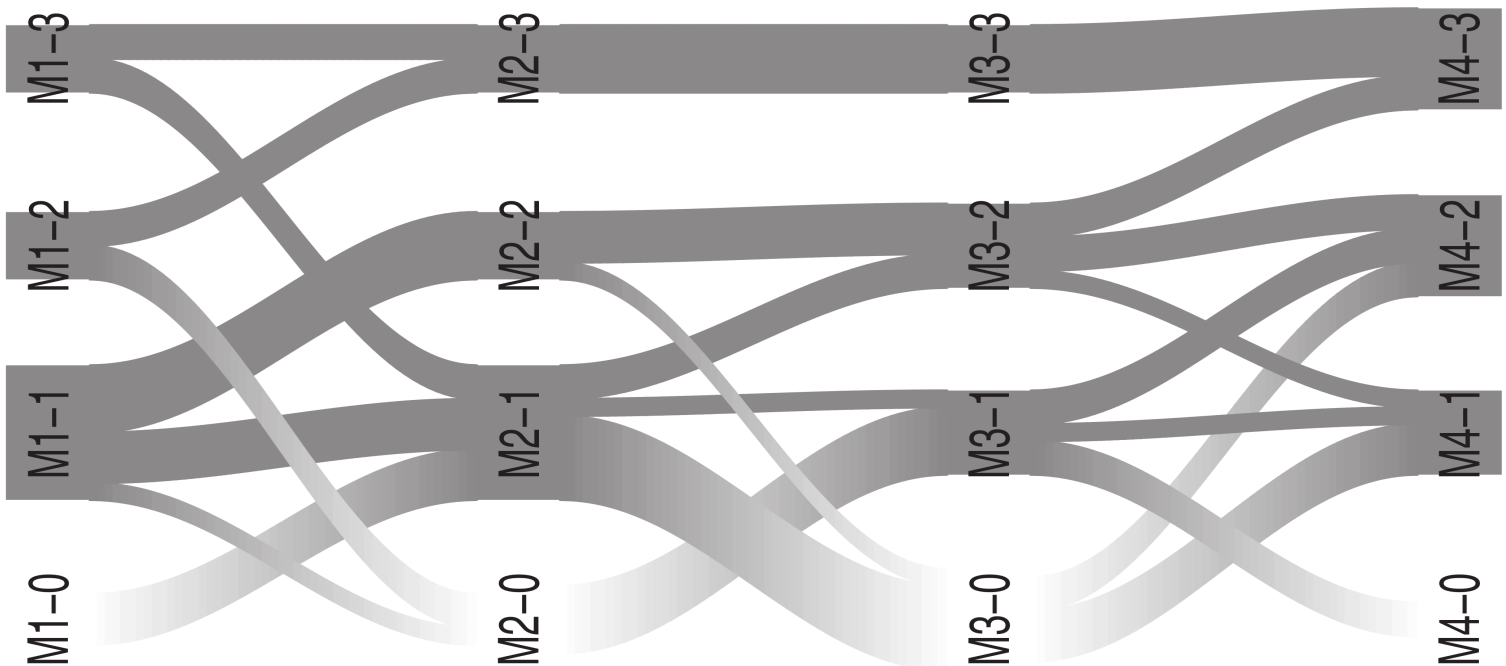




Mariadassou & al., Ann. Applied Stat., 2010

Onager network

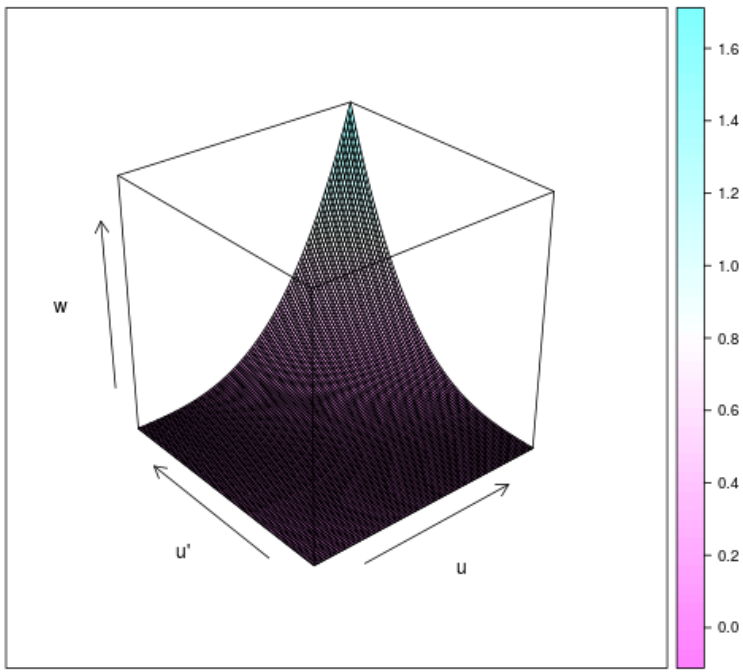
$n = 23$ individuals, $T = 4$ dates, $\widehat{K} = 4$ groups



Matias & Miele, JRSSB, 2017

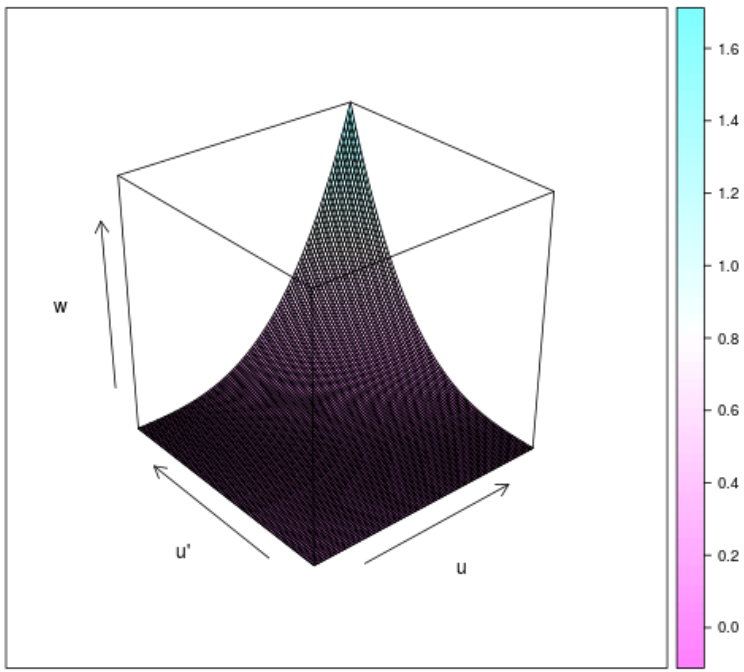
W-graph

A graphon function: $w(u, u')$.

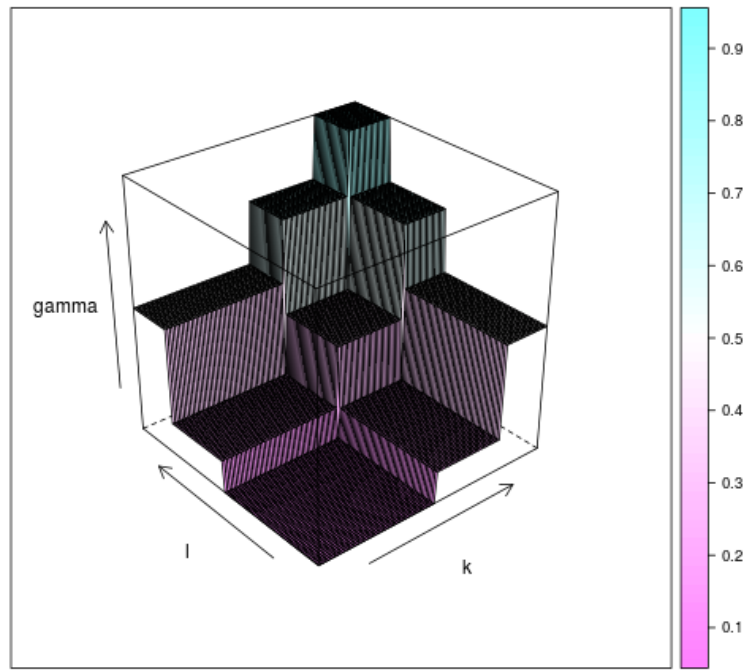


W-graph

A graphon function: $w(u, u')$.



Graphon function of an SBM.



Graphical models (Lauritzen, 96)

Directed graphical models: p faithfull to G (DAG) iff

$$p(U_1, \dots U_m) = \prod_j p(U_j \mid U_{pa_G(j)})$$

where $U_J = \{U_j : j \in J\}$ and $pa_G(j) =$ sets of parents of j in G.

Graphical models (Lauritzen, 96)

Directed graphical models: p faithfull to G (DAG) iff

$$p(U_1, \dots U_m) = \prod_j p(U_j \mid U_{pa_G(j)})$$

where $U_J = \{U_j : j \in J\}$ and $pa_G(j) =$ sets of parents of j in G.

Undirected graphical models: p faithfull to G iff

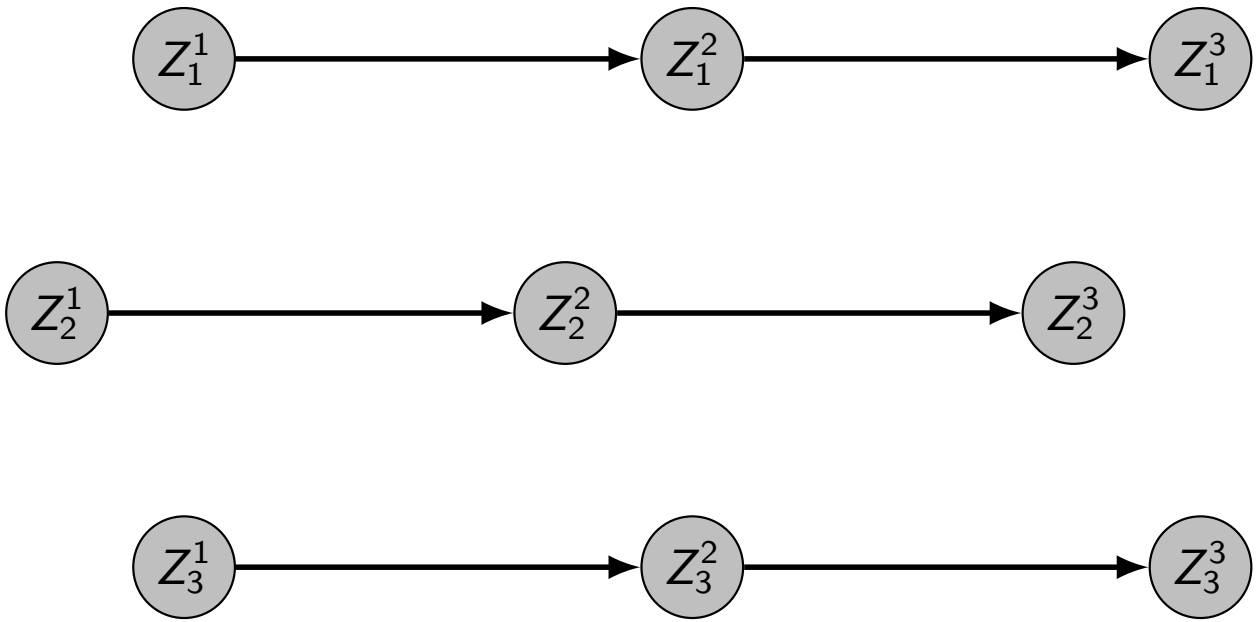
$$p(U_1, \dots U_m) \propto \prod_{C \in \mathcal{C}(G)} \Psi_C(U_C)$$

where $\mathcal{C}(G) =$ set of all maximal cliques of G.

- Directed GM $\rightarrow$ Undirected GM, via moralization.
- Undirected GM: Equivalence between separation and conditional independence.

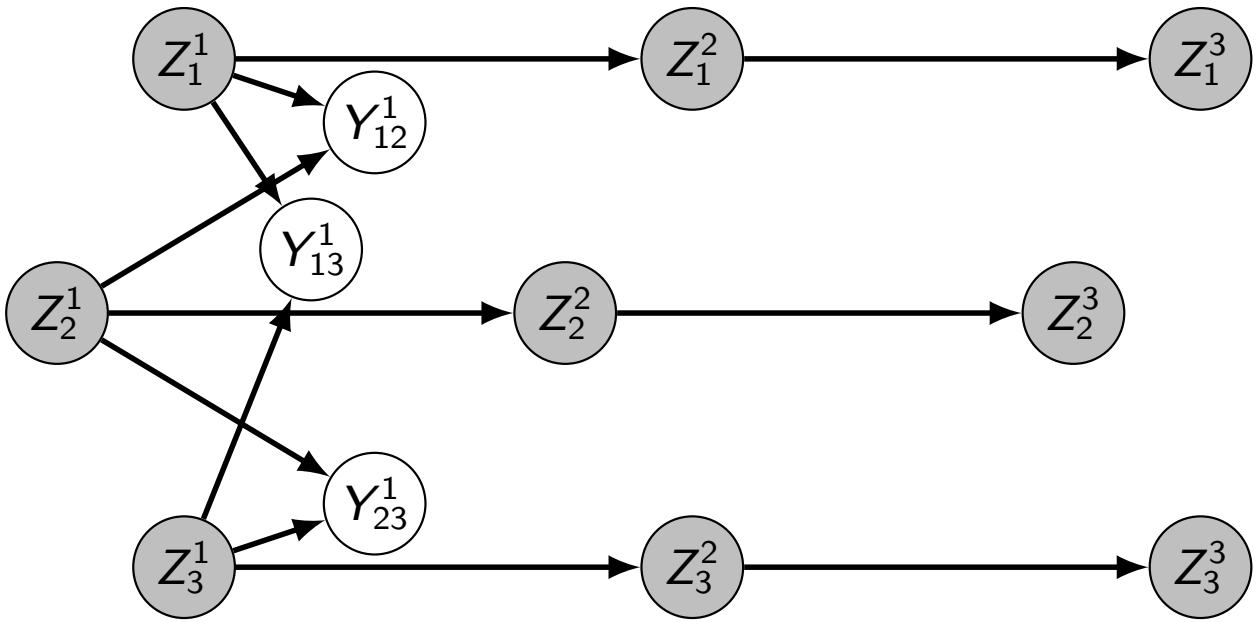
Graphical models for the dynamic SBM

Hidden Markov chains. $n = 3,\ T = 3$



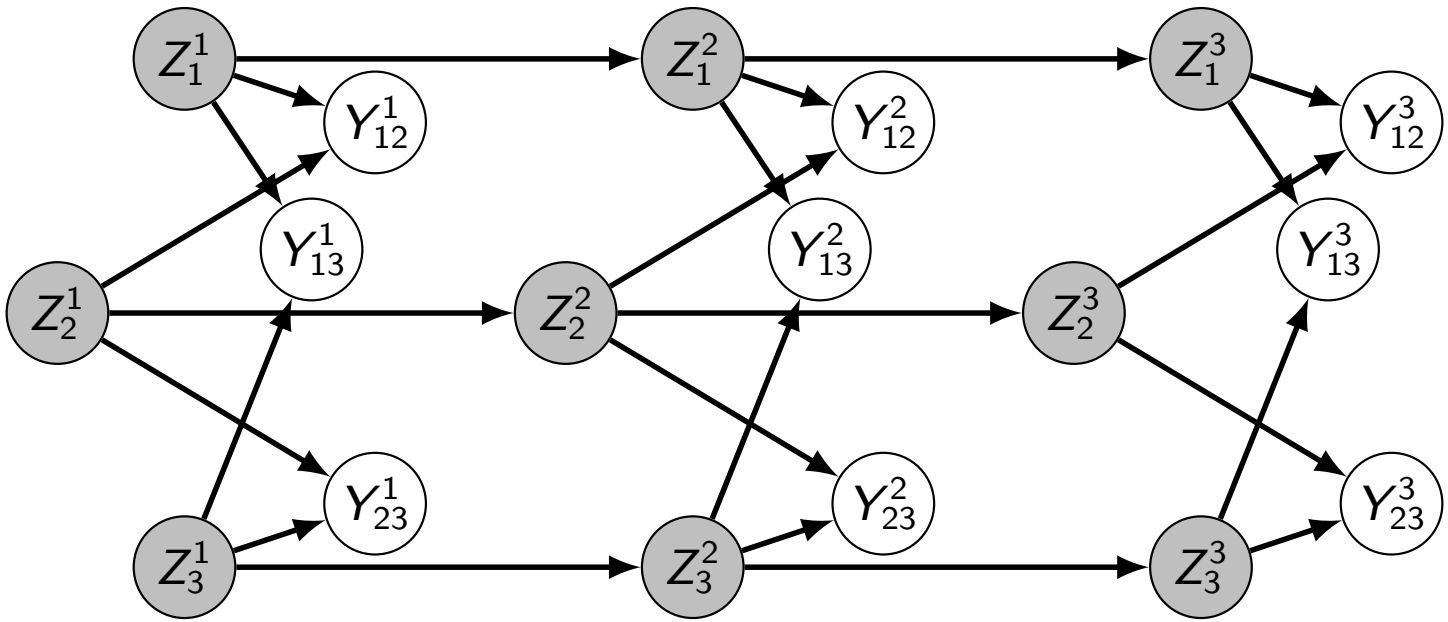
Graphical models for the dynamic SBM

Observed network at $t = 1$.



Graphical models for the dynamic SBM

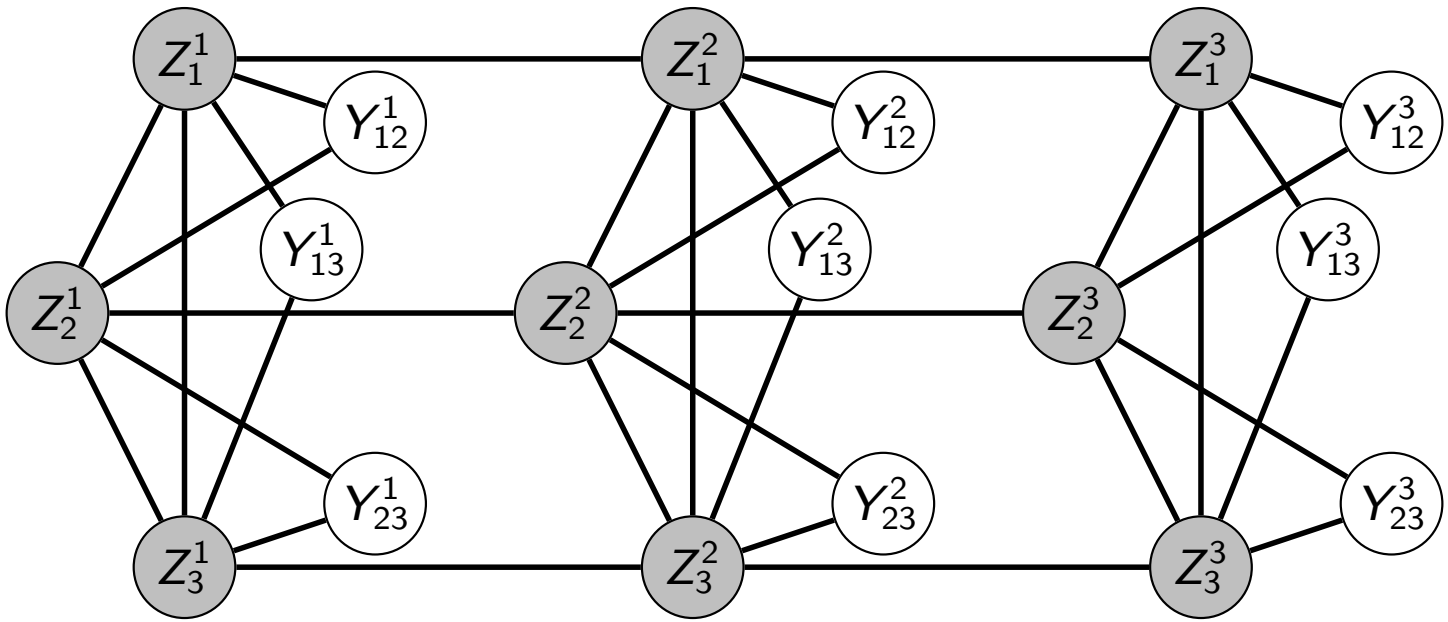
Observed networks at $t = 1, \dots T$.



$\rightarrow$ $(Z^t, Y^t)_t \sim HMM$ with K^n states.

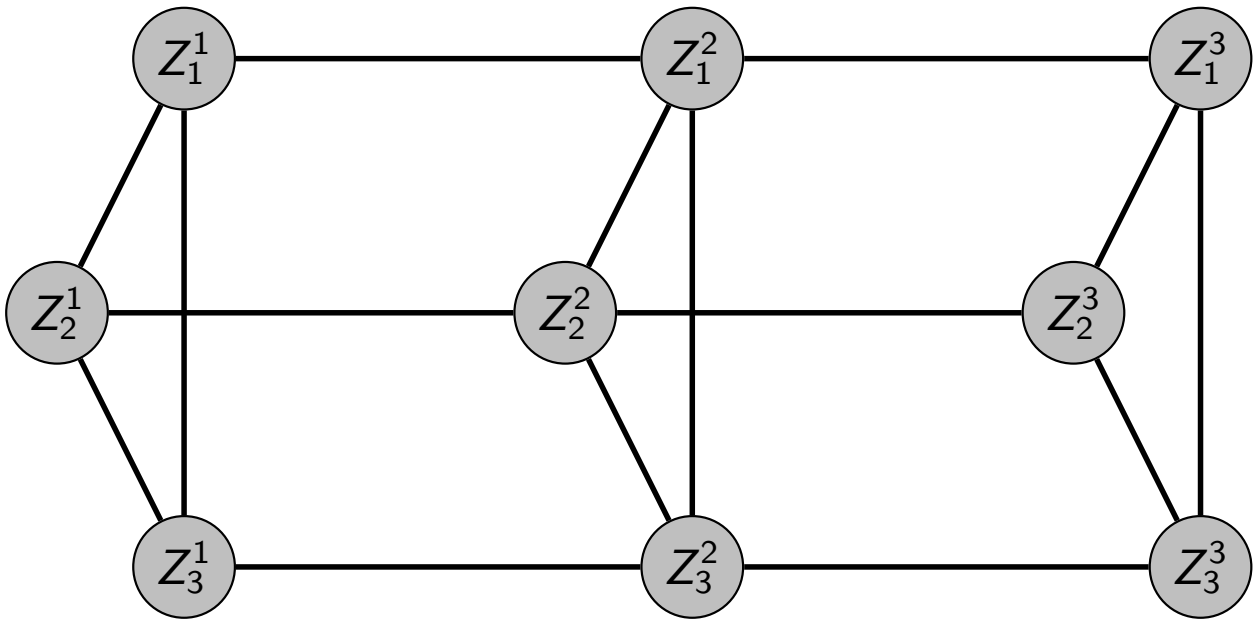
Graphical models for the dynamic SBM

Graph moralization.



Graphical models for the dynamic SBM

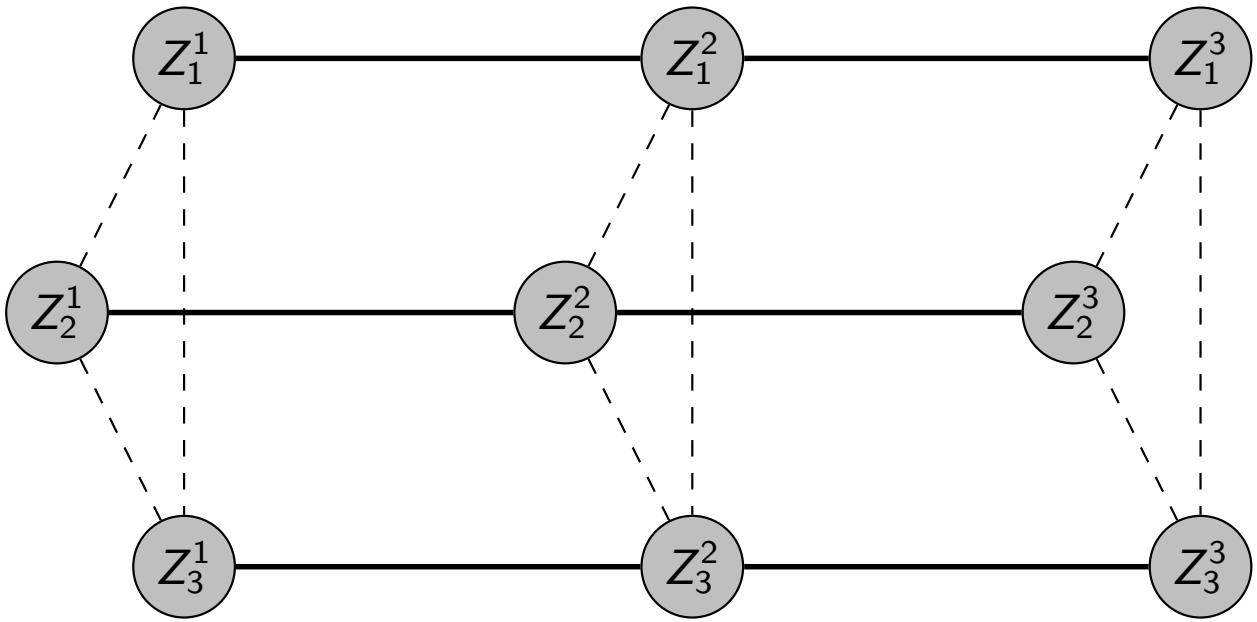
Graphical model of $p(Z \mid Y)$.



$\rightarrow$ (Heterogeneous) Markov chain with K^n states

Graphical models for the dynamic SBM

Variational approximation: $p(Z \mid Y) \simeq \prod_i q_i(Z_i) \neq \prod_{i,t} q_{it}(Z_i^t)$

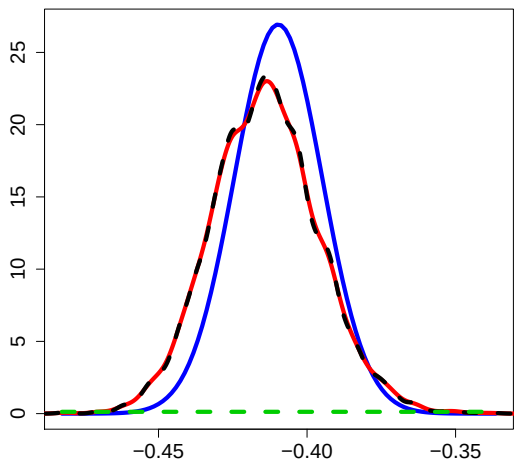


$\rightarrow$ Partial mean-field approximation

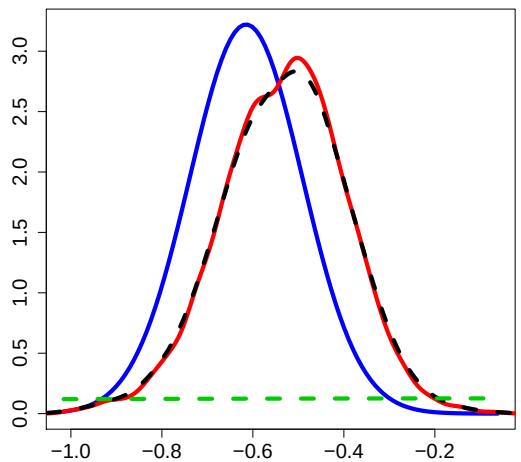
Tree interaction network

$n = 51$ tree species, Y_{ij} = number of shares parasites, Poisson emission, 3 covariates

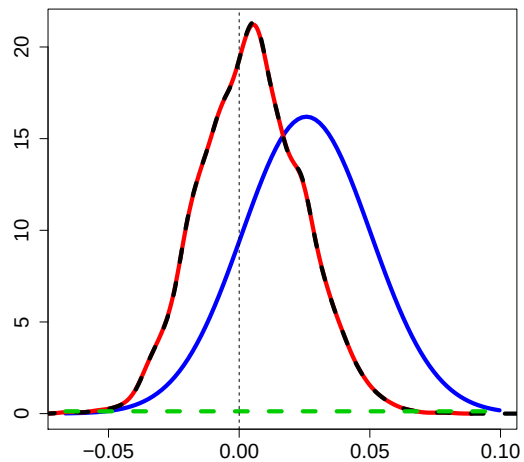
taxonomy



geography

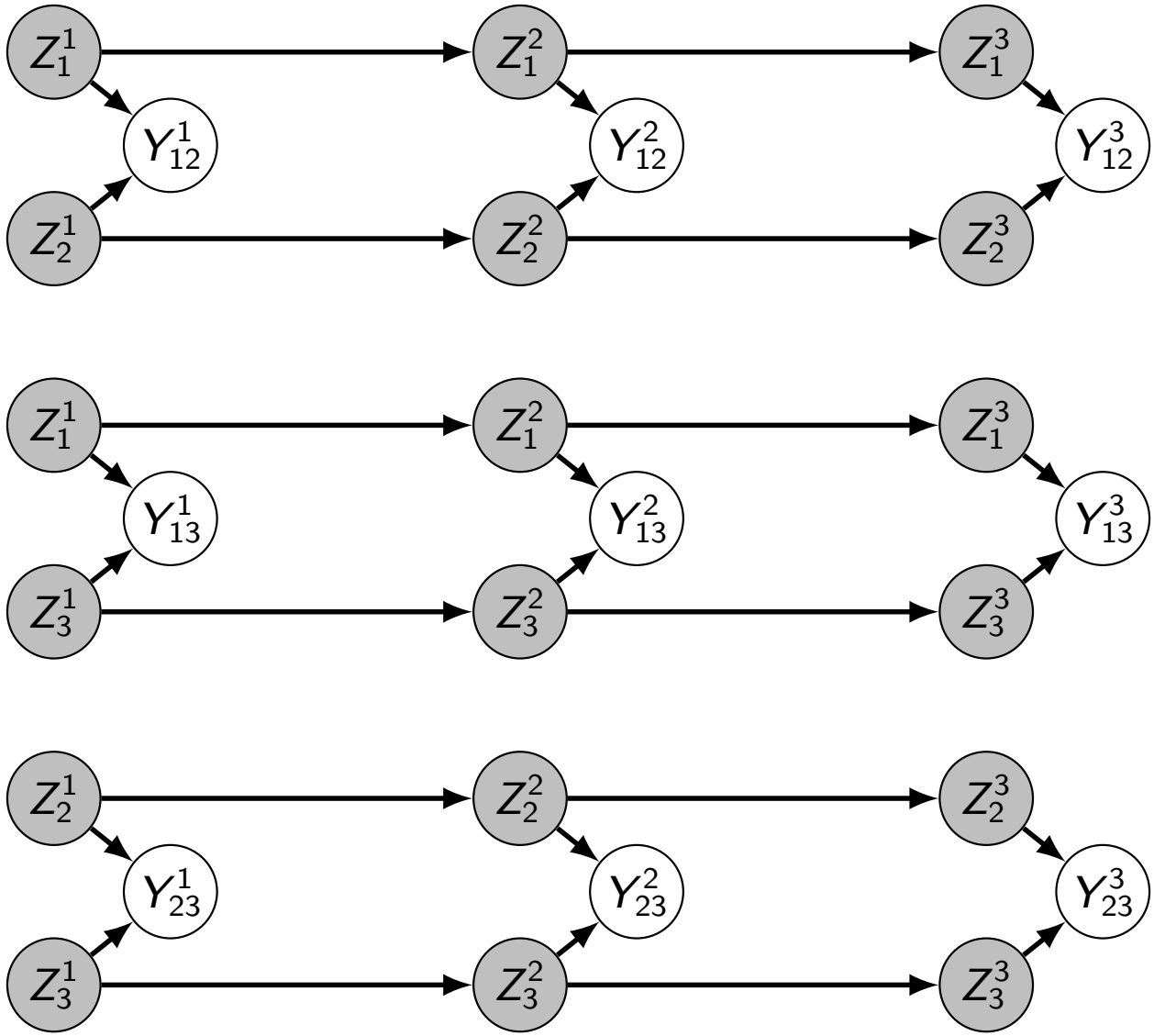


genetics



$p(\beta)$, $\widetilde{p}(\beta \mid \widehat{K})$, $\widehat{p}(\beta \mid Y, \widehat{K})$, $\widehat{p}(\beta \mid Y) = \sum_K \widehat{p}(K \mid Y)\widehat{p}(\beta \mid Y, K)$

Composite likelihood for dynamic SBM



$\rightarrow$ $n(n-1)/2$ HMMs with K^2 states.

References I

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, 2008.

E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 2014.

C. Ambroise and C. Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B*, 74(1):3–35, 2012.

E.A. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pages 3099–3132, 2009.

P. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, pages 1922–1943, 2013.

M. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, university of London, 2003.

J. Beal, M. and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayes. Statist.*, 7:543–52, 2003.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

A. Celisse, J.-J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statis.*, 6:1847–99, 2012.

A. Channarond, J.-J. Daudin, and S. Robin. Classification and estimation in the stochastic block model based on the empirical degrees. *Electron. J. Statis.*, 6:2574–601, 2012.

P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica*, 7(28):33–61, 2008.

References II

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Stat. Comput.*, 18(2):173–83, 2008.

J.-J. Daudin, L. Pierre, and C. Vacher. Model for heterogeneous random networks using continuous latent variables and an application to a tree–fungus network. *Biometrics*, 66(4):1043–1051, 2010.

S. Donnet and S. Robin. Using deterministic approximations to accelerate SMC for posterior sampling. Technical Report 1612.06928, arXiv, 2017.

G. Govaert and M. Nadif. An EM algorithm for the block mixture model. *IEEE Trans. Pattern Anal. Machine Intel.*, 27(4):643–7, 2005.

P.W. Holland and S. Leinhardt. Structural sociometry. *Perspectives on Social Network Research*, pages 63–83, 1979.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–98, 2002.

M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A*, 170(2):301–354, 2007. doi: 10.1111/j.1467-985X.2007.00471.x.

T. Jaakkola. *Advanced mean field methods: theory and practice*, chapter Tutorial on variational approximation methods. MIT Press, 2001.

Y. Jernite, P. Latouche, C. Bouveyron, P. Rivera, L. Jegou, and S Lamassé. The random subgraph model for the analysis of an ecclesiastical network in merovingian gaul. *The Annals of Applied Statistics*, 8(1):377–405, 2014.

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216, 2015.

B. Karrer and M. EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

References III

P. Latouche, E. Birmelé, and C. Ambroise. Overlapping stochastic block models with application to the French political blogosphere. *Ann. Appl. Stat.*, 5(1):309–336, 2011.

P. Latouche, E. Birmelé, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statis. Model.*, 12(1):93–115, 2012.

P. Latouche and S. Robin. Variational bayes model averaging for graphon functions and motif frequencies inference in W-graph models. *Statistics and Computing*, 26(6):1173–1185, 2016.

P. Latouche, S. Robin, and S. Ouadah. Goodness of fit of logistic regression models for random graphs. *Journal of Computational and Graphical Statistics*, 27(1):98–109, 2018.

L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933 – 957, 2006.

Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd, 2005.

M. Mariadassou and C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573, 2015.

C. Matias and V. Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141, 2017.

A.F. McDaid, T.B. Murphy, N. Friel, and N.J. Hurley. Improved bayesian inference for the stochastic block model with application to large networks. *Computational Statistics & Data Analysis*, 60:12–31, 2013.

C. Matias and S. Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proc.*, 47:55–74, 2014.

M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, pages 715–742, 2010.

References IV

K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

E. Szemerédi. Regular partitions of graphs. Technical report, Sanford univ Calif dept of Computer Science, 1975.

T. Tabouy, P. Barbillon, and J. Chiquet. Variational Inference for Stochastic Block Models from Sampled Data. Technical report, arXiv:1707.04141, 2017.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008.