Modelling a network of point processes: an approach based on exchangeability

Patrick Rubin-Delanchy

University of Oxford & Heilbronn Institute for Mathematical Research

with contributions from

Niall Adams (Imperial College London), Nick Heard (Imperial College London) and Dan Lawson (University of Bristol)

Statistical Aspects of Cyber-Security, Royal Statistical Society
10th March 2016

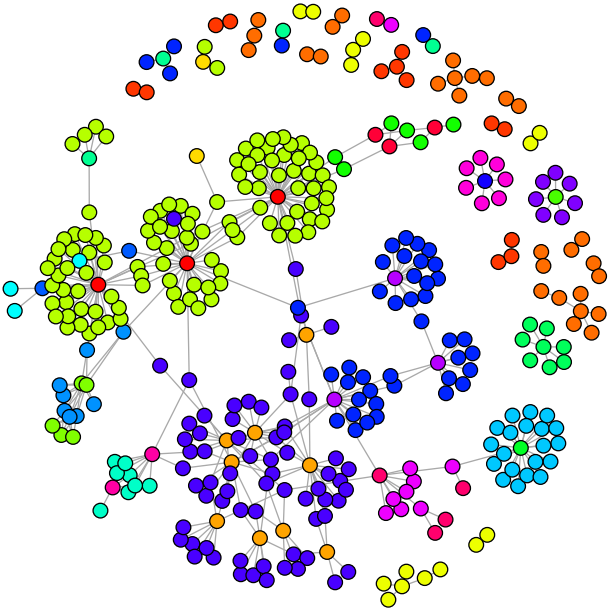


Figure : NetFlow (1 min), Los Alamos National Laboratory (LANL) network (Kent, 2016)

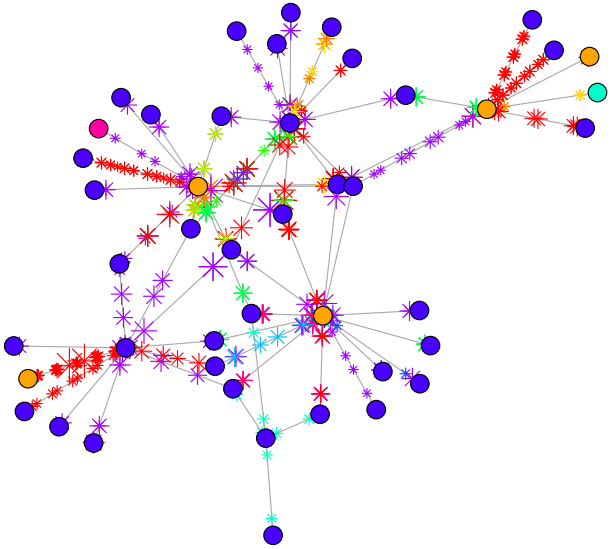


Figure : NetFlow (5 min), LANL network (Kent, 2016)

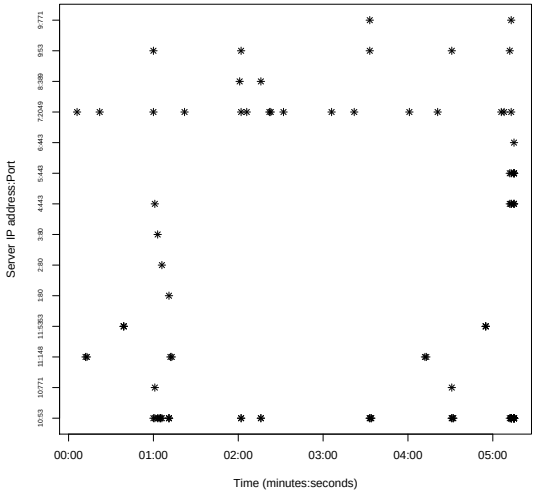
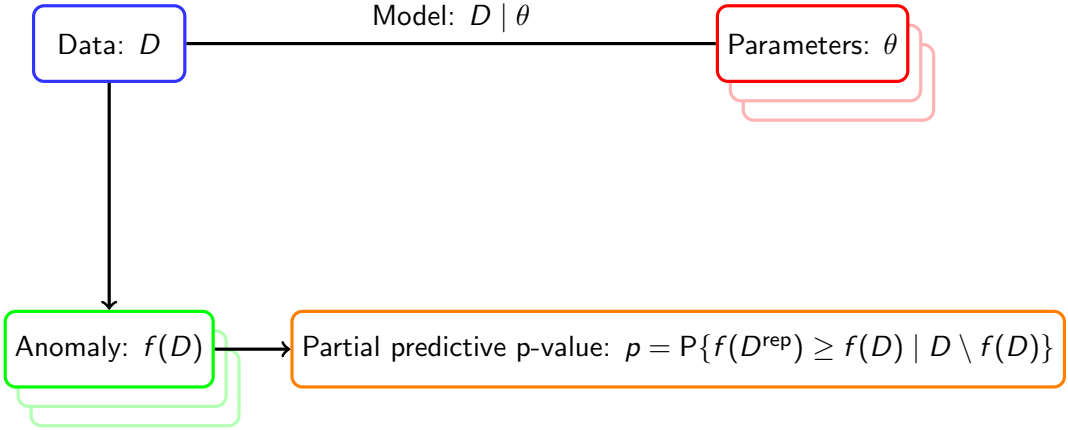


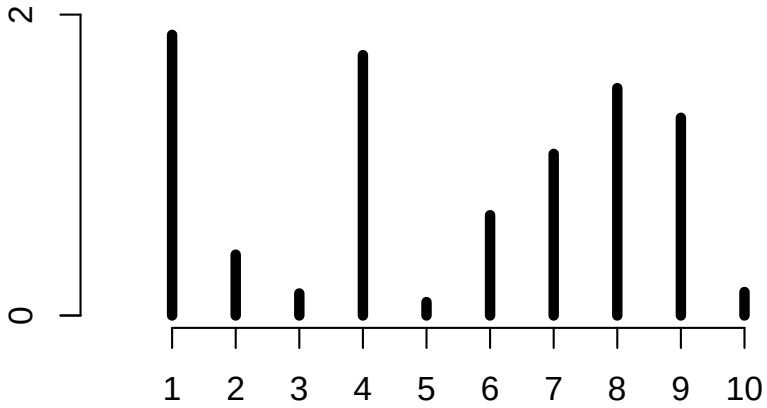
Figure : NetFlow (5 min), generated by a computer on the Imperial College London network

Bayesian anomaly detection



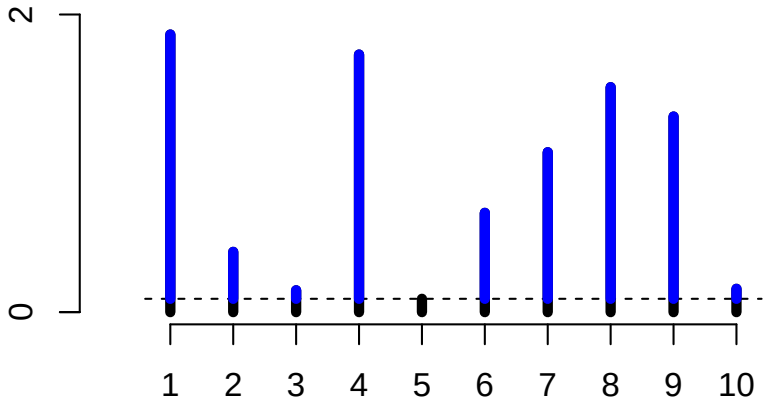
See Bayarri and Berger (2000).

Bayesian anomaly detection



- $X_1, \ldots, X_{10}$ independent exponential RVs with unknown rate λ
- $f(D) = \min(X_i)$ (testing for outlying small value)
- Compute $p = \mathrm{P}\{f(D^{\mathrm{rep}}) \geq f(D) \mid D \setminus f(D)\}$
- Partial posterior on λ uses $\tilde{X}_1, \ldots, \tilde{X}_9$

Bayesian anomaly detection



- $X_1, \ldots, X_{10}$ independent exponential RVs with unknown rate λ
- $f(D) = \min(X_i)$ (testing for outlying small value)
- Compute $p = \mathrm{P}\{f(D^{\mathrm{rep}}) \geq f(D) \mid D \setminus f(D)\}$
- Partial posterior on λ uses $\tilde{X}_1, \ldots, \tilde{X}_9$

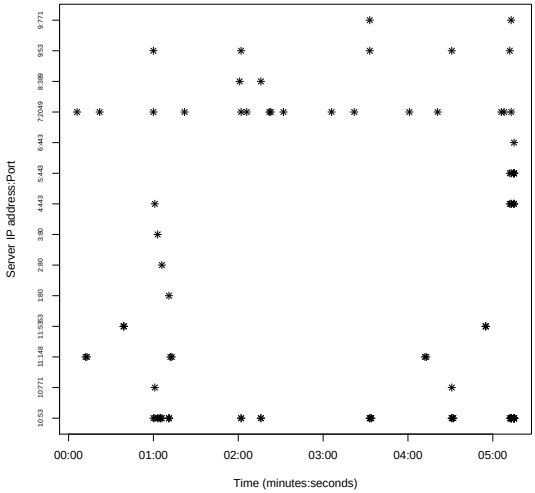


Figure : NetFlow (5 min), generated by a computer on the Imperial College London network

Sessionization

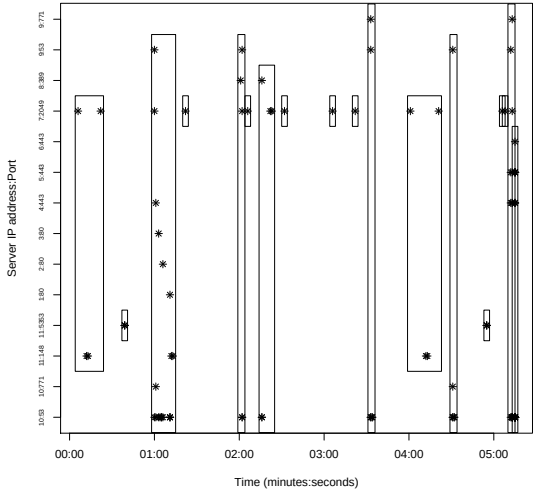
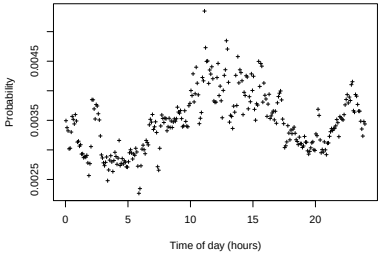
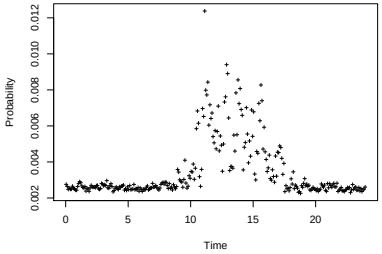
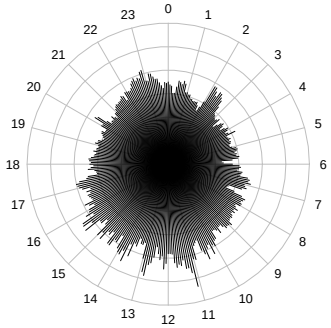


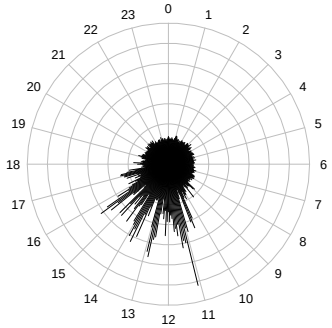
Figure : Sessionization of Netflow — Bayesian temporal clustering

Filtering


Probability
0.0025
0.0035
0.0045
0
5
10
15
20
Time of day (hours)


Probability
0.002
0.004
0.006
0.008
0.010
0.012
0
5
10
15
20
Time


0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23


0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Point process modelling

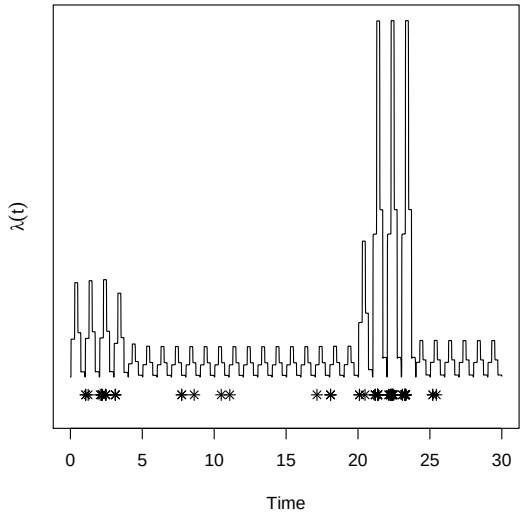


Figure : Non-parametric Bayesian intensity estimation — Enron email data

De Finetti's representation theorem

Let $X_1, X_2, \ldots$ be an infinite exchangeable sequence of Bernoulli variables.

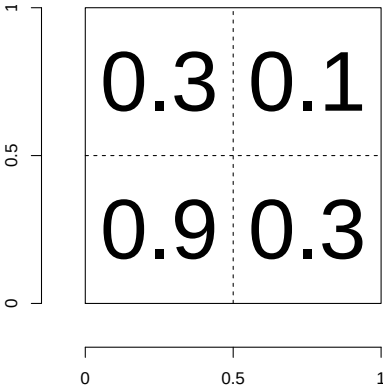
Then X_i are independent conditional on a common success probability, p, where p is a random variable on $[0, 1]$.

Aldous-Hoover theorem

Let G be an infinite undirected exchangeable graph.

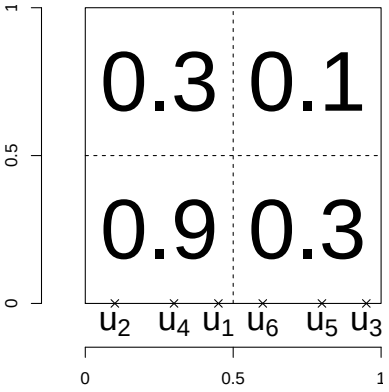
Then each edge $i \leftrightarrow j$ is a conditionally independent Bernoulli variable with success probability $f(u_i, u_j)$ where

- $u_1, u_2, \ldots$ are independent uniform random variables on $[0, 1]$.
- f is a *random* symmetric function from $[0, 1]^2 \to [0, 1]$ (called a graphon).



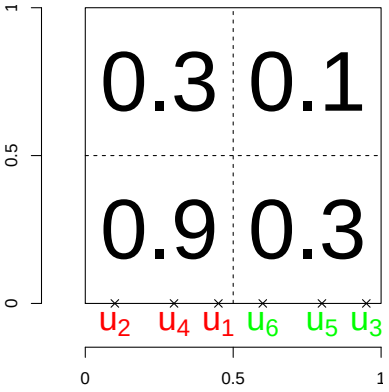
Generative model:

- Generate $f \sim \mu$, for some probability measure μ.
- Generate $u_1, \ldots, u_N \overset{\text{i.i.d}}{\sim} \text{Uniform}[0,1]$
- Let each edge $i \leftrightarrow j$ occur with probability $f(u_i, u_j)$.



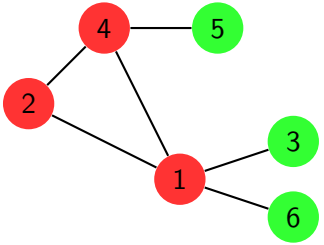
Generative model:

- Generate $f \sim \mu$, for some probability measure μ.
- Generate $u_1, \ldots, u_N \overset{\text{i.i.d}}{\sim} \text{Uniform}[0,1]$
- Let each edge $i \leftrightarrow j$ occur with probability $f(u_i, u_j)$.



Generative model:

- Generate $f \sim \mu$, for some probability measure μ.
- Generate $u_1, \ldots, u_N \overset{\text{i.i.d}}{\sim} \text{Uniform}[0,1]$
- Let each edge $i \leftrightarrow j$ occur with probability $f(u_i, u_j)$.



Block approximation

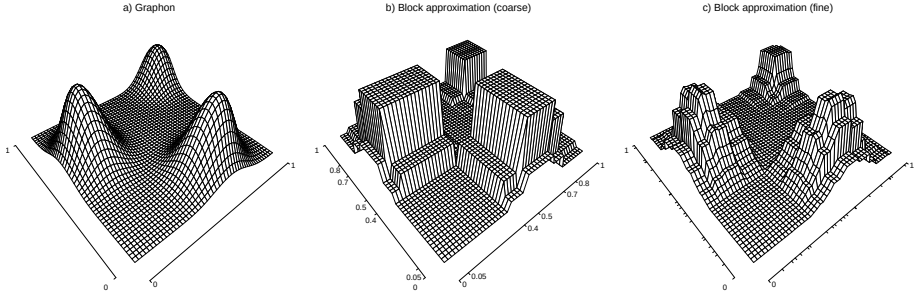


Figure : A graphon and its block approximations

Stochastic block model

- Partitioning of nodes into communities represented a vector of labels ℓ
- Marginal likelihood of the graph

$$\mathrm{P}(G \mid \ell) = \prod_{k \leq l} \int_0^1 p^{n_{kl}} (1-p)^{o_{kl} - n_{kl}} \mathrm{d}F(p),$$

where

1. n_{kl} number of edges between communities k and l
2. o_{kl} number of possible edges
3. F prior distribution function on edge probability (assuming an IID prior)

Example (Stochastic block model: conjugate prior)

$$\mathrm{P}(G \mid \ell) = \prod_{k,l} \frac{B\left(n_{kl} + \alpha_{kl}, o_{kl} - n_{kl} + \beta_{kl}\right)}{B(\alpha_{kl}, \beta_{kl})}$$

Stochastic block model

- Partitioning of nodes into communities represented a vector of labels ℓ
- Marginal likelihood of the graph

$$\mathrm{P}(G \mid \ell) = \prod_{k \leq l} \int_0^1 p^{n_{kl}} (1-p)^{o_{kl} - n_{kl}} \mathrm{d}F(p),$$

where

1. n_{kl} number of edges between communities k and l
2. o_{kl} number of possible edges
3. F prior distribution function on edge probability (assuming an IID prior)

Example (Stochastic block model: conjugate prior)

$$\mathrm{P}(G \mid \ell) = \prod_{k,l} \frac{B\left(n_{kl} + \alpha_{kl}, o_{kl} - n_{kl} + \beta_{kl}\right)}{B(\alpha_{kl}, \beta_{kl})}$$

Stochastic block model

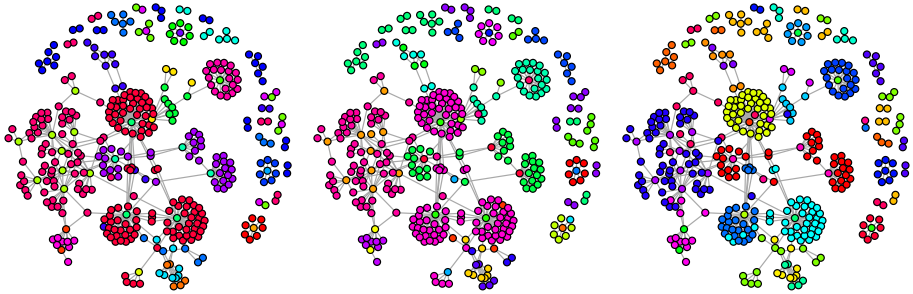
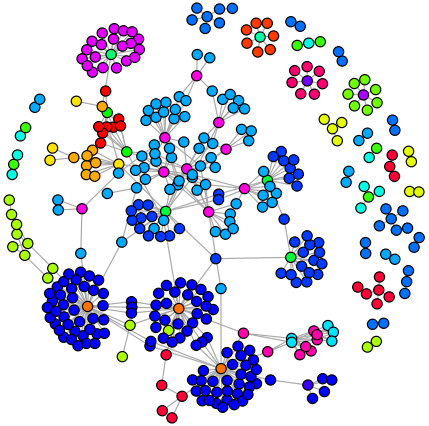


Figure : Three posterior samples from fitted stochastic block model, LANL network

Port prediction



	N1	445	N46	N17	80	389	137	88	443	N20
	52	0	0	0	21	0	0	0	0	0
	0	20	0	14	0	21	0	8	0	0
	0	0	0	2	0	1	0	0	0	7
	0	0	25	0	0	0	0	0	0	0
	0	18	0	0	0	0	1	0	0	0
	0	1	2	0	0	0	0	0	0	0
	0	0	0	0	0	0	14	0	0	0
	0	0	0	1	0	0	0	0	0	0
	0	4	2	0	1	0	0	0	0	0
	0	3	0	1	0	0	0	0	1	0

Figure : Port prediction contingency table. Pearson's Chi-square $p \approx 0$

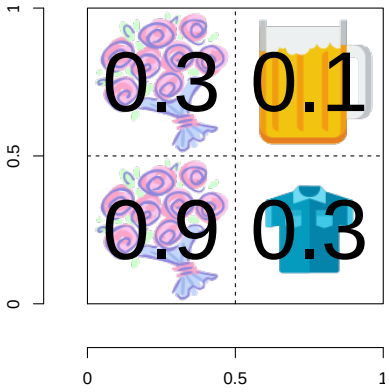
Aldous-Hoover theorem revisited

Let $(X_{ij}), 1 \leq i, j < \infty$ denote an infinite row- and column-exchangeable array.

Then there exists a function g such that $X_{ij} = g(\alpha, \xi_i, \eta_j, \lambda_{ij}), 1 \leq i, j < \infty$, where $\alpha, \xi_i, \eta_j, \lambda_{ij}$ are independent uniform random variables.

Source	Dest.	Port
A	B	53
C	A	80
A	D	53
A	E	80
D	A	80
C	E	80
C	F	53
F	B	25
⋮	⋮	⋮

Sender	Receiver	Gift
Anna	Daniel	
Mary	Anna	
Anna	Bob	
Anna	Julie	
Bob	Anna	
Mary	Julie	
Mary	George	
George	Daniel	
⋮	⋮	⋮



Sender	Receiver	Gift
Anna	Daniel	
Mary	Anna	
Anna	Bob	
Anna	Julie	
Bob	Anna	
Mary	Julie	
Mary	George	
George	Daniel	
⋮	⋮	⋮

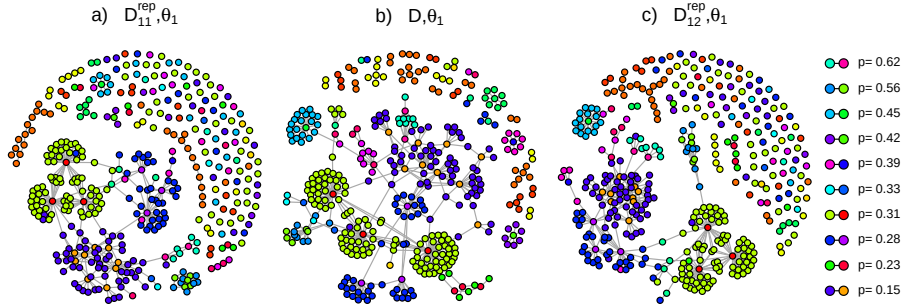
A generic model for NetFlow

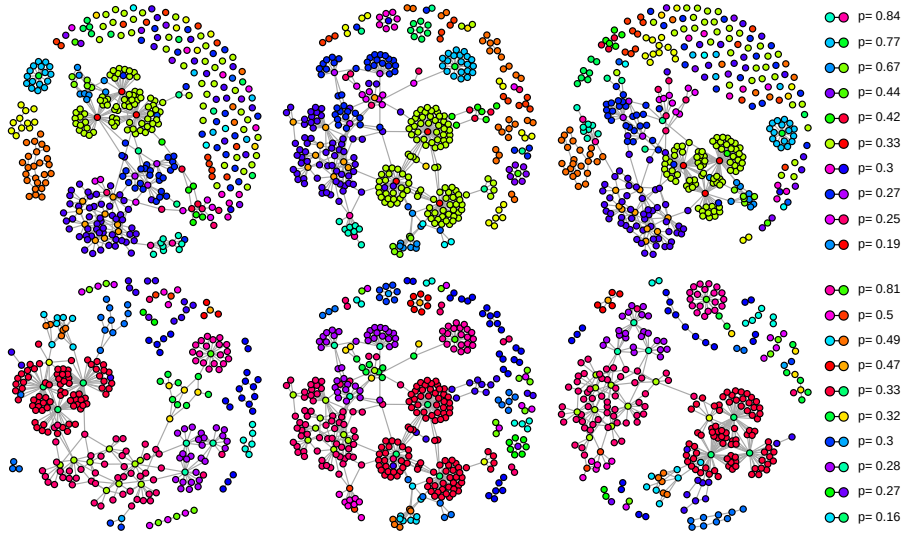
- Partitioning of nodes into communities represented a vector of labels ℓ
- Marginal likelihood of the network

$$p(D \mid \ell) = \prod_{k,l} \int p(\theta_{kl}) \prod_i p(D_{kl}^{(i)} \mid \theta_{kl}) \mathrm{d}\theta_{kl}$$

where

1. k and l are community indices
2. $D_{kl}^{(i)}$ are the edge processes with source in community k and destination in community l (including processes with no events)


a) D_{11}^{rep},θ_1
b) D,θ_1
c) D_{12}^{rep},θ_1
p= 0.62
p= 0.56
p= 0.45
p= 0.42
p= 0.39
p= 0.33
p= 0.31
p= 0.28
p= 0.23
p= 0.15


a) D_{11}^{rep},θ_1
b) D,θ_1
c) D_{12}^{rep},θ_1
p= 0.84
p= 0.77
p= 0.67
p= 0.44
p= 0.42
p= 0.33
p= 0.3
p= 0.27
p= 0.25
p= 0.19
p= 0.81
p= 0.5
p= 0.49
p= 0.47
p= 0.33
p= 0.32
p= 0.3
p= 0.28
p= 0.27
p= 0.16
d) D_{21}^{rep},θ_2
(− singletons)
e) D,θ_2
f) D_{22}^{rep},θ_2
(− singletons)

Posterior predictive p-value

The posterior predictive p-value is (Meng, 1994; Gelman et al., 1996, Eq. 2.8, Eq. 7)

$$P = \mathrm{P}\{f(D^{\mathrm{rep}}, \theta) \geq f(D, \theta) \mid D\}, \tag{1}$$

where θ represents the model parameters, D is the observed dataset, D^{rep} is a hypothetical replicated dataset generated from the model with parameters θ, and $\mathrm{P}(\cdot \mid D)$ is the joint posterior distribution of $(\theta, D^{\mathrm{rep}})$ given D.

In words: if a new dataset were generated from the same model and parameters, what is the probability that the new discrepancy would be as large?

Discussion: Guttman (1967), Box (1980), Rubin (1984), Bayarri and Berger (2000), Hjort et al. (2006)

Posterior predictive p-values in practice: Huelsenbeck et al. (2001), Sinharay and Stern (2003), Thornton and Andolfatto (2006), Steinbakk and Storvik (2009)

Theory of posterior predictive p-values: Rubin-Delanchy and Lawson (2015)

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598.

Bayarri, M. and Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142.

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 83–100.

Heard, N. A., Lawson, D. J., and Rubin-Delanchy, P. (2014). Filtering automated polling traffic in computer network flow data. In *Proceedings of the IEEE Joint Intelligence and Security Informatics Conference (JISIC)*.

Hjort, N. L., Dahl, F. A., and Steinbakk, G. H. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, 101(475):1157–1174.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314.

Kent, A. D. (2016). Cybersecurity data sources for dynamic network research. In *Dynamic Networks and Cybersecurity*. Imperial College Press, London. To appear.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 22(3):1142–1160.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Rubin-Delanchy, P. and Heard, N. A. (2014). A test for dependence between two point processes on the real line. *arXiv preprint arXiv:1408.3845*.

Rubin-Delanchy, P. and Lawson, D. J. (2015). Posterior predictive p-values and the convex order. *arXiv preprint arXiv:1412.3442*.

Sinharay, S. and Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111(1):209–221.

Steinbakk, G. H. and Storvik, G. O. (2009). Posterior predictive p-values in Bayesian hierarchical models. *Scandinavian Journal of Statistics*, 36(2):320–336.

Thornton, K. and Andolfatto, P. (2006). Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a netherlands population of drosophila melanogaster. *Genetics*, 172(3):1607–1619.