Statistical Network Analysis with Bergm

Lampros Bouranis

Marie Sklodowska-Curie Research Fellow
Department of Statistics
Athens University of Economics and Business, Greece

bouranis@aueb.gr

October 2021

Development team

Publication details

- Journal of Statistical Software (Accepted).
- arXiv e-print: *https://arxiv.org/abs/2104.02444*.

Cornell University

arXiv.org > stat > arXiv:2104.02444

Statistics > Computation

[Submitted on 6 Apr 2021]

Statistical Network Analysis with Bergm

Alberto Caimo, Lampros Bouranis, Robert Krause, Nial Friel

Recent advances in computational methods for intractable models have made network data increasingly amenable to statistical analysis. Exponer
graph models (ERGMs) emerged as one of the main families of models capable of capturing the complex dependence structure of network data ir
of applied contexts. The Bergm package for R has become a popular package to carry out Bayesian parameter inference, missing data imputatior
selection and goodness-of-fit diagnostics for ERGMs. Over the last few years, the package has been considerably improved in terms of efficiency

Outline

Getting Bergm

Bergm: Bayesian Exponential Random Graph Models

Bayesian analysis for exponential random graph models using advanced computational algorithms. More information can be found at: <https://acaimo.github.io/Bergm/>.

Version:	5.0.3
Depends:	ergm, R ($\geq$ 2.10)
Imports:	network, coda, MCMCpack, Matrix, mvtnorm, matrixcalc, statnet.common
Published:	2021-06-15
Author:	Alberto Caimo [aut, cre], Lampros Bouranis [aut], Robert Krause [aut] Nial Friel [ctb]
Maintainer:	Alberto Caimo <acaimo.stats at gmail.com>

- **Bergm** version 5 of the package available on CRAN.
- Considerable improvement in terms of usability for practitioners and performance since its early versions (Caimo and Friel, 2014).
- Type the following commands to obtain **Bergm** from CRAN and load it in R:

```
R> install.packages("Bergm")
R> library("Bergm")
```

Motivation

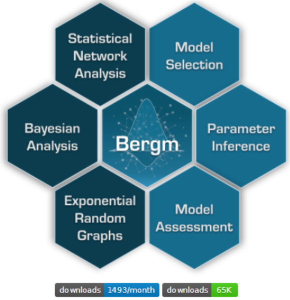
- Statistical models with intractable likelihood functions abound:
  - Image analysis.
  - Spatial statistics.
  - Statistical network analysis.
- We focus on ERGMs - widely used in statistical network analysis.
- Bayesian inference for ERGMs is challenging because of the intractability of both the likelihood and the marginal likelihood.
- Advanced computational methods developed in the last decade have made it computationally feasible to model increasingly large network data using ERGMs on several thousands of nodes.
- The development of user-friendly software has always represented an essential aspect of the research activity in this area.

Motivation

Bergm aims to provide a set of tools for both developers and end-users.

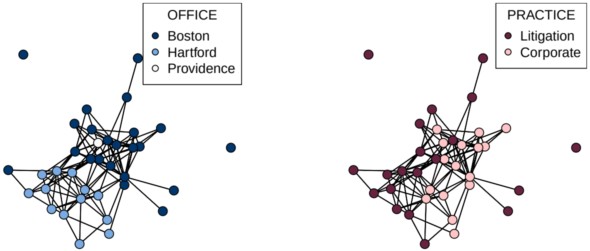
Several competing functions (based on different statistical approaches) are included, for carrying out:

- Bayesian parameter inference.
- Missing data imputation.
- Model selection.
- Goodness-of-fit diagnostics.



- Can be computationally intensive, but is easy to use and represents an attractive way of analysing networks.
- Several applications of **Bergm**: neuroscience (Sinke et al., 2016), organisation science (Caimo and Lomi, 2014; Tasselli and Caimo, 2019) and political science (Henning et al., 2019).

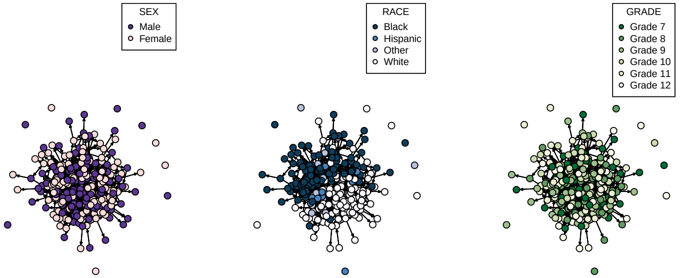
Network data - Lazega's law firm



- Binary undirected collaborative relations between 36 partners in a Northeastern US corporate law firm (Lazega, 2001).
- Member attributes available: seniority, formal status, office in which they work, gender, law school attended, etc.

```
R> data(lazega)
```

Network data - Faux Dixon High School



- Simulation of a binary directed in-school friendship network (Resnick et al., 1997). See *?faux.dixon.high* for the ERGM that was fit to the original data, generating the network dataset.
- The network comprises 248 nodes representing students. Information on the following nodal attribute variables is available: sex, race, grade.

```
R> data(faux.dixon.high)
R> dixon ← faux.dixon.high
```

The exponential random graph model (ERGM)

Purpose of ERGMs

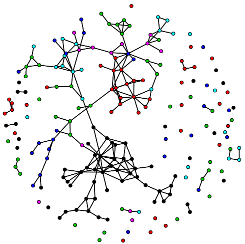
1. Capture the complex dependence structure of an observed network.
2. Identify the relational effects that are supposed to describe the link creation process.

$$f(y \mid \theta) = \frac{q(y \mid \theta)}{z(\theta)} = \frac{\exp\left\{\theta^\top s(y)\right\}}{z(\theta)}$$

- Random adjacency matrix $y \in \mathcal{Y}$ with n nodes: $y_{ij} = \{0;1\}$.
- $s(y) \in \mathbb{R}^d_+$, $\theta \in \Theta \subseteq \mathbb{R}^d$.
- $z(\theta)$ is a normalising constant,

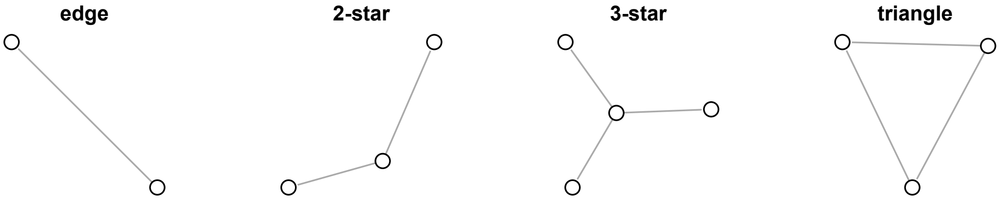
$$z(\theta) = \sum_{\textit{all possible graphs}} \exp\left\{\theta^\top s(y)\right\}$$

- $2^{\binom{n}{2}}$ possible undirected graphs of n nodes.



Model specification: network statistics

- Probability of observing a given graph depends on certain ”local” graph configurations.



- Inclusion of covariate information x is allowed:

$$s(y,x) = \sum_{i \neq j} y_{ij} \times \begin{cases} x_i + x_j & : \text{"main effect"}, \\ \mathbb{1}_{\{x_i = x_j\}} & : \text{"homophily effect"}. \end{cases}$$

Bayesian exponential random graph models (BERGMs)

- Allows for prior (expert-judgement) information, eg:
  - Info from previous studies about network effects (e.g. estimates of ERGM parameters).
  - Network sparsity.
- Probabilistic treatment of uncertainty.
- Combination of prior and current information - included in the likelihood.
- Bayesian approach to ERGMs is difficult because of likelihood intractability!

Bayesian exponential random graph models (BERGMs)

- Challenging to sample from the doubly-intractable posterior distribution

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)p(\theta)}{\pi(y)} = \frac{f(y \mid \theta)p(\theta)}{\int_\Theta f(y \mid \theta)p(\theta)\, d\theta}.$$

- Naive MH algorithm proposes the move $\theta \to \theta'$ with probability

$$\begin{aligned} A(\theta, \theta') &= \min\left\{1, \frac{f(y \mid \theta')}{f(y \mid \theta)} \frac{p(\theta')}{p(\theta)} \frac{h(\theta \mid \theta')}{h(\theta' \mid \theta)}\right\} \\ &= \min\left\{1, \frac{q(y \mid \theta')}{q(y \mid \theta)} \frac{p(\theta')}{p(\theta)} \frac{h(\theta \mid \theta')}{h(\theta' \mid \theta)} \times \frac{z(\theta)}{z(\theta')}\right\}. \end{aligned}$$

- Estimating the model evidence $\pi(y)$ is also a challenge...

Exchange algorithm (Murray et al., 2006)

Samples from the augmented distribution

$$\pi(\theta', y', \theta \mid y) \propto f(y \mid \theta) p(\theta) h(\theta' \mid \theta) f(y' \mid \theta'),$$

whose marginal distribution for θ is the posterior of interest.

1. Gibbs update of (θ', y'):
   (i) Draw $\theta' \sim h(\cdot \mid \theta)$.
   (ii) Draw $y' \sim f(\cdot \mid \theta')$.
2. Propose move from θ to θ' with probability:

$$\min\left\{1, \frac{q(y \mid \theta')}{q(y \mid \theta)} \frac{p(\theta')}{p(\theta)} \frac{h(\theta \mid \theta')}{h(\theta' \mid \theta)} \frac{q(y' \mid \theta)}{q(y' \mid \theta')} \times \frac{z(\theta) z(\theta')}{z(\theta') z(\theta)}\right\}.$$

Approximate exchange algorithm (Caimo and Friel, 2011)

- Crucially, the Exchange requires a draw from $f(y' \mid \theta')$ at each iteration. Perfect sampling is an obvious approach, if this is possible.
- Pragmatic solution: take a realisation from a long MCMC run (M transitions) with stationary distribution $f(y' \mid \theta')$ as an approximate draw:
  1. Gibbs update of (θ', y'):
     (i) Draw $\theta' \sim h(\cdot \mid \theta)$.
     (ii) Draw $y' \sim R^M(\cdot \mid \theta')$ via MCMC ["tie-no-tie" (TNT) sampler].
  2. Propose move from θ to θ' with probability:

$$\min\left\{1, \frac{q(y \mid \theta')}{q(y \mid \theta)} \frac{p(\theta')}{p(\theta)} \frac{h(\theta \mid \theta')}{h(\theta' \mid \theta)} \frac{q(y' \mid \theta)}{q(y' \mid \theta')} \times \frac{z(\theta)z(\theta')}{z(\theta')z(\theta)}\right\}.$$

Approximate exchange algorithm (Caimo and Friel, 2011)

- Intuitively we expect the number of auxiliary iterations, M, to be proportional to the # of dyads of the graph, n^2.
- Everitt (2012) showed that, under regularity conditions, the invariant distribution of approximate exchange converges to the true target as # of auxiliary iterations, M, increases.
- Conservative approach: choose a large M...
- Bottleneck: exponentially long mixing time for auxiliary draw from the likelihood - A **computationally intensive procedure** for larger graphs!

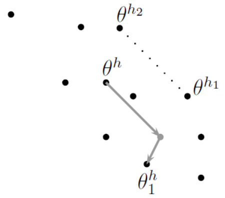
Parallel adaptive direction sampler

- **Computational challenge:** Often, strong correlation between model parameters & high posterior density region can be thin $\rightarrow$ slow mixing (Caimo and Friel, 2011).
- **Improving chain mixing and convergence:** a parallel adaptive direction sampler (ADS) for the Gibbs update of θ' (Gilks et al., 1994):
  - i^{th} iteration: a collection of H different chains interacting with one another.
  - State space $\{\theta_1, \dots, \theta_H\}$ with target distribution

$$\pi(\theta_1 \mid y) \otimes \cdots \otimes \pi(\theta_H \mid y).$$

- **Parallel ADS move for each chain $h = 1, \dots, H$ [$\theta_h^i \rightarrow \theta_h^{i+1} = \theta_h'$]:**

1. Select at random h_1 and h_2 without replacement from $\{1, \dots, H\} \setminus h$.
2. Sample $\varepsilon \sim \mathcal{N}_d(0, \Sigma)$.
3. Propose $\theta_h' = \theta_h^i + \gamma\left(\theta_{h_1}^i - \theta_{h_2}^i\right) + \varepsilon$.



Implementation

- Network statistics:
  - Density effect captured by the number of edges (*edges*).
  - Homophily effect between lawyers working in the same office (*nodematch("Office")*) and in the same practice area (*nodematch("Practice")*).
  - Transitivity effect captured by the geometrically weighted edgewise shared partners statistic (GWESP) (Snijders et al., 2006).

```
# Define the model:
R> m1 ← lazega ~ edges + nodematch("Office") +
+         nodematch("Practice") + gwesp(0.5, fixed = TRUE)

# Prior assumptions (multivariate Normal distribution):
R> M.prior ← c(-4, 0.5, 0.5, 1)
R> S.prior ← diag(4, 4)

# Implement the parallel ADS procedure:
R> p.m1 ← bergm(m1, nchains = 8, aux.iters = 2500,
+                prior.mean = M.prior, prior.sigma = S.prior,
+                burn.in = 500, main.iters = 3000, gamma = 0.6)
```

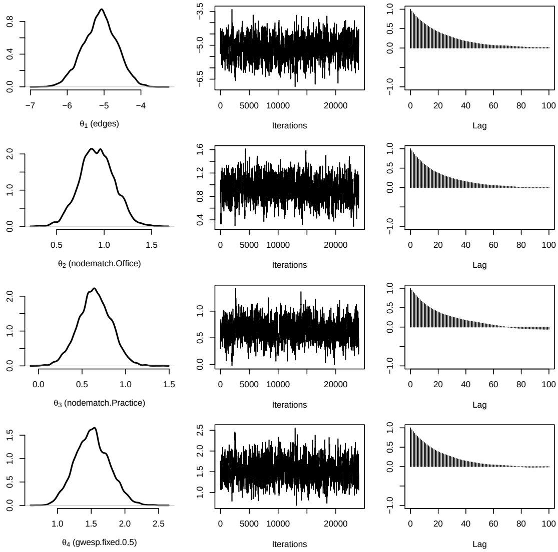
Posterior outputs

- **CPU time**: $\approx$ 2.6 mins.

```
R> summary(p.m1)

 Posterior Density Estimate for Model: y ~ edges +
    nodematch("Office") + nodematch("Practice") +
    gwesp(0.5, fixed = TRUE)

                              Mean   SD Naive SE Time-series SE
theta1 (edges)              -5.110 0.450 0.003 0.0195
theta2 (nodematch.Office)    0.925 0.181 0.001 0.0072
theta3 (nodematch.Practice)  0.645 0.186 0.001 0.0076
theta4 (gwesp.fixed.0.5)     1.517 0.251 0.003 0.0105

                              2.5%    25%    50%    75%  97.5%
theta1 (edges)              -6.022 -5.417 -5.095 -4.799 -4.251
theta2 (nodematch.Office)    0.577  0.801  0.922  1.046  1.278
theta3 (nodematch.Practice)  0.276  0.521  0.644  0.770  1.006
theta4 (gwesp.fixed.0.5)     1.041  1.345  1.511  1.689  2.027

Acceptance rate: 0.2
```

Posterior outputs

```
R> plot(p.m1, lag = 100)
```



Missing data augmentation (Krause et al., 2020)

- **Missing network data types:**
  1. Actor non-response: all outgoing ties of an actor/node are missing. 'Non-response' implies that data is collected via self-reports of network actors and stems from classical survey research.
  2. Tie non-response: some, but not all ties of an actor/node are missing. Data collection methods, like link tracing or snowball sampling, might lead more often to tie non-response.
- **Effect on descriptive network statistics depends on:**
  - The amount of missing data
  - The network structure
  - The descriptive statistic in question
  - How the missing data are treated

Missing data augmentation (Krause et al., 2020)

- **Treatment – Multiple data imputation**:
  - Assumption 1: missing data is ignorable, i.e the probability for data to be missing is independent from the missing values themselves and only dependent on the observed data ["missing at random"].
  - Assumption 2: limited to the setting where the set of unobserved tie variables is known and fixed.
  - Assumption 3: all covariates/nodal attributes are known and fixed.
  - Missing network data imputed using draws from the posterior distribution of the tie variable that is generated to obtain parameter estimates.
  - Allows to retain the augmented networks, thus achieving proper multiple imputations.
  - Shown to provide reliable estimates of $\pi(\theta \mid y)$ (Koskinen et al., 2010), and low biases in descriptive statistics, even with high missing data rates.

Missing data augmentation (Krause et al., 2020)

Convention: u represents the observed part of the data; v represents the unobserved part of the data. The network can be re-assembled as $y = (u, v)$.

Algorithm 1: Bergm with Multiple data imputation

Initialise $s(y^*)$ with $s(u)$; Initialise θ.
for $k = 1, \ldots, K$ **do**
  Generate θ' using the ADS proposal procedure.
  Draw $y' \sim R^M(\cdot \mid \theta')$ via MCMC ["tie-no-tie" (TNT) sampler].
  Update $\theta \to \theta'$ with the log of the probability:

$$\log \alpha = \min \left(0, [\theta - \theta']^\top [s(y') - s(y^*)] + \log \left[\frac{p(\theta')}{p(\theta)} \right] \right).$$

  if θ' accepted **then**
    Draw $v^* \sim f(\cdot \mid \theta', u)$ and generate a new $y^* = (u, v^*)$.
  end if
end for

Implementation

- *nImp*: retain a specified number of imputed networks y^* from the estimation procedure, after the burnin phase.

```
# Lazega's data is fully observed; randomly set all
# outgoing ties of 4 nodes (11%) to missing:
R> set.seed(1)
R> missV ← sample(1:36, 4)
R> lazega[missV, ] ← lazega[, missV] ← NA

# Implement the parallel ADS procedure:
R> p.m1.M ← bergmM(m1, nchains = 8, aux.iters = 3000,
+                  prior.mean = M.prior, prior.sigma = S.prior,
+                  burn.in = 200, main.iters = 3000,
+                  gamma = 0.6, nImp = 10)

# Obtain the imputed networks:
R> p.m1.M$impNets
```

Posterior summaries

- **CPU time**: $\approx$ 10.4 mins.

```
R> summary(p.m1.M)

 Posterior Density Estimate for Model: y ~ edges +
    nodematch("Office") + nodematch("Practice") +
    gwesp(0.5, fixed = TRUE)

                              Mean    SD Naive SE Time-series SE
theta1 (edges)              -4.782 0.438 0.002 0.019
theta2 (nodematch.Office)    0.860 0.188 0.001 0.008
theta3 (nodematch.Practice)  0.567 0.195 0.001 0.009
theta4 (gwesp.fixed.0.5)     1.369 0.248 0.002 0.011

                              2.5%    25%    50%    75%  97.5%
theta1 (edges)              -5.722 -5.060 -4.767 -4.489 -3.953
theta2 (nodematch.Office)    0.485  0.736  0.864  0.979  1.235
theta3 (nodematch.Practice)  0.176  0.441  0.569  0.698  0.955
theta4 (gwesp.fixed.0.5)     0.918  1.199  1.355  1.524  1.885

Acceptance rate: 0.19
```

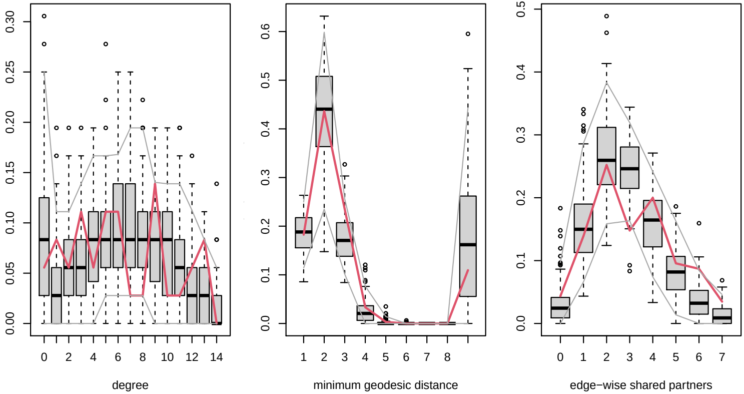
Goodness-of-fit diagnostics

- Bayesian goodness-of-fit (GOF) procedure – evaluate the model GOF in terms of posterior predictive assessment (Caimo and Friel, 2011).
- Observed network is compared with a randomly simulated network sample (sample size determined by *sample.size*) drawn from the estimated posterior distribution using *aux.iters* iterations for the network simulation step.
- High-level characteristics not modeled explicitly:
  - Degree distributions (for degrees > 3);
  - Minimum geodesic distance (the proportion of pairs of nodes whose shortest connected path is of length $l = 1, 2, \ldots$);
  - # of edge-wise shared partners (# of edges in the network that share l neighbours in common ($l = 1, 2, \ldots$).

```
R> set.seed(1)
R> bgof(p.m1, aux.iters = 5000, sample.size = 100,
+        n.deg = 15, n.dist = 9, n.esp = 8)
```

Goodness-of-fit diagnostics

- Red lines represent the observed network GOF statistic values; boxplots represent the GOF statistics of the simulated networks.
- The structure of the observed graph can be considered as a possible realisation of the posterior density.



Bayesian model selection

Model selection - ERGM context

Translates into the choice of which subset of network statistics should be included into the model (Caimo and Friel, 2013).

- Model set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \ldots\}$ for data $y \in \mathcal{Y}$.
- Within each model:

$$\pi(\theta_m \mid y, \mathcal{M}_m) = \frac{f(y \mid \theta_m, \mathcal{M}_m) p(\theta_m \mid \mathcal{M}_m)}{\pi(y \mid \mathcal{M}_m)}.$$

- Constant of proportionality (”Marginal likelihood”/”Evidence”):

$$\pi(y \mid \mathcal{M}_m) = \int_{\Theta_m} f(y \mid \theta_m, \mathcal{M}_m) p(\theta_m \mid \mathcal{M}_m) \, \mathrm{d}\theta_m.$$

Bayesian model selection

- Summarise pairwise comparison of models $\mathcal{M}_m, \mathcal{M}_{m'}$ by:

$$\frac{\pi(\mathcal{M}_m \mid y)}{\pi(\mathcal{M}_{m'} \mid y)} = \frac{\pi(y \mid \mathcal{M}_m)}{\pi(y \mid \mathcal{M}_{m'})} \times \frac{p(\mathcal{M}_m)}{p(\mathcal{M}_{m'})}$$

$$\text{posterior odds} = \text{BF}_{m,m'} \times \text{prior odds}$$

- The **Bergm** package assumes a multivariate Normal prior $\mathcal{N}_{d_m}(\mu_m, \Sigma_m)$ for θ_m, that leads to a marginal likelihood which is finite.

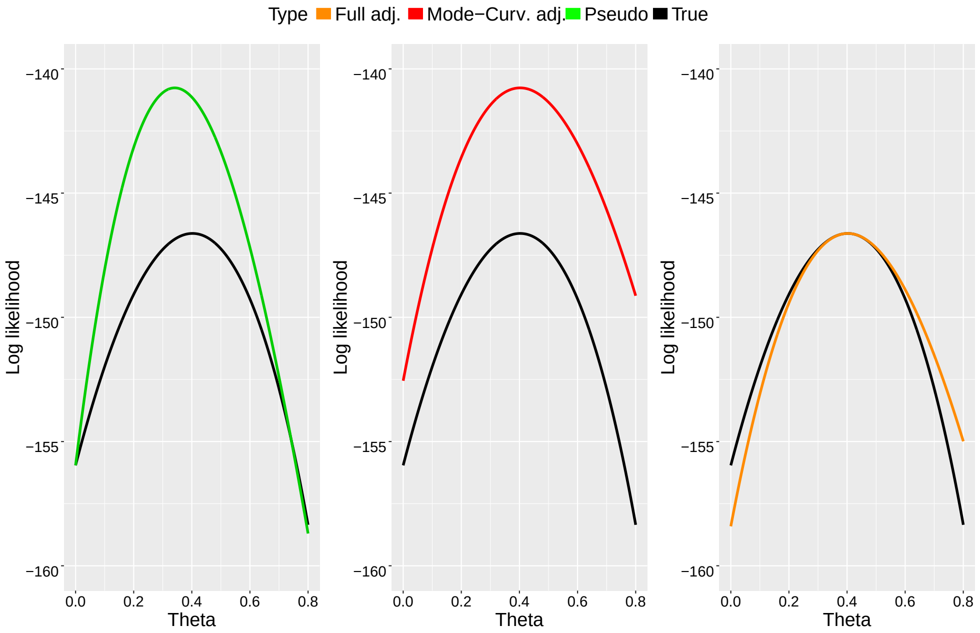
Pseudo-likelihood adjustment for large networks

$$f_{\text{PL}}(y \mid \theta) = \prod_{i \neq j \,\cup\, i<j} p(y_{ij} \mid y_{-ij}, \theta) = \prod_{i \neq j \,\cup\, i<j} \frac{p(y_{ij} = 1 \mid y_{-ij}, \theta)^{y_{ij}}}{\{1 - p(y_{ij} = 1 \mid y_{-ij}, \theta)\}^{y_{ij}-1}},$$

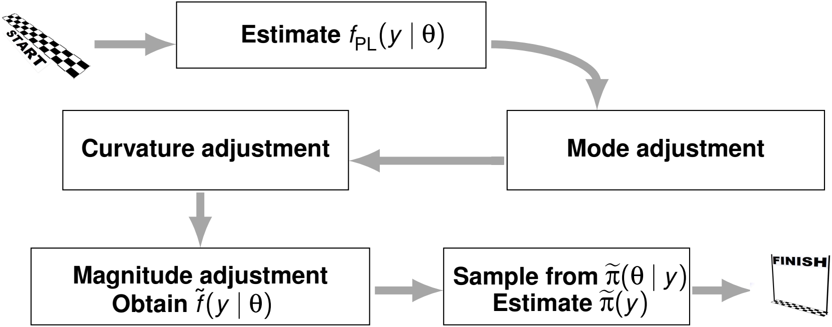
where y_{-ij} denotes $\text{y} \setminus \{\text{y}_{ij}\}$.

- Each factor in the product is a Bernoulli random variable (Strauss and Ikeda, 1990).
- Estimation is equivalent to logistic regression - fast computation.
- Assumes the collection $y_{ij} \mid y_{-ij}$ are mutually independent - problematic inference.
- Pseudo-posterior: $\pi_{\text{PL}}(\theta \mid y) \propto f_{\text{PL}}(y \mid \theta) p(\theta)$.

Pseudo-likelihood adjustment for large networks



Pseudo-likelihood adjustment for large networks



Pseudo-likelihood adjustment for large networks

- Adjustments lead to the model-specific fully adjusted pseudo-likelihood (Bouranis et al., 2018)

$$\tilde{f}_m(y \mid \theta_m) = C_m \cdot f_{PL,m}(y \mid \hat{\theta}_{MPLE,m} + Q_m(\theta_m - \hat{\theta}_{MLE,m})).$$ Details

- Approximate the true posterior distribution by

$$\widetilde{\pi}(\theta \mid y, \mathcal{M}_m) = \frac{\tilde{f}(y \mid \theta_m, \mathcal{M}_m) p(\theta_m \mid \mathcal{M}_m)}{\widetilde{\pi}(y \mid \mathcal{M}_m)} = \frac{\tilde{f}(y \mid \theta_m, \mathcal{M}_m) p(\theta_m \mid \mathcal{M}_m)}{\int_{\Theta_m} \tilde{f}(y \mid \theta_m, \mathcal{M}_m) p(\theta_m \mid \mathcal{M}_m)\, \mathrm{d}\theta_m}.$$

- Approximate $BF_{mm'}$ by

$$\widetilde{BF}_{mm'} = \frac{\widetilde{\pi}(y \mid \mathcal{M}_m)}{\widetilde{\pi}(y \mid \mathcal{M}_{m'})} = \frac{\int_{\Theta_m} \tilde{f}(y \mid \theta_m, \mathcal{M}_m) p(\theta_m \mid \mathcal{M}_m)\, \mathrm{d}\theta_m}{\int_{\Theta_{m'}} \tilde{f}(y \mid \theta_{m'}, \mathcal{M}_{m'}) p(\theta_{m'} \mid \mathcal{M}_{m'})\, \mathrm{d}\theta_{m'}}.$$

Evidence estimation techniques

- Evidence estimation: well studied problem in the last 30 years (Ardia et al., 2012).
- Many techniques to estimate intractable multi-dimensional integrals, eg.:
  - Chib and Jeliazkov's *one-block Metropolis-Hastings* method (2001).
  - Power posteriors (Friel and Pettitt, 2008).
- Most require a tractable likelihood.
- With our adjustments we can borrow methods from the Bayesian toolbox.
- In **Bergm**, the *evidence()* function estimates $\widetilde{\pi}(y \mid \mathcal{M}_m)$.

Application

- Model $\mathcal{M}_1$ – almost identical to the model used to generate the simulated data (see *?faux.dixon.high*):
  - density (*edges*), mutuality (*mutual*) and transitivity (*gwesp*) effects;
  - homophily effects for *race*, *sex* and *grade*;
  - # of nodes of in-degree 0 and 1 & # of nodes of out-degree 0 and 1.

```
R> m1 ← dixon ~ edges + mutual + absdiff("grade") +
+     nodefactor("race") +
+     nodefactor("grade") +
+     nodefactor("sex") +
+     nodematch("race", diff = TRUE, levels = c("B","O","W")) +
+     nodematch("grade", diff = TRUE) +
+     nodematch("sex", diff = FALSE) +
+     idegree(0:1) + odegree(0:1) +
+     gwesp(0.1,fixed = TRUE)

R> M.prior1 ← c(-5, rep(0, 26))
R> S.prior1 ← diag(5, 27)
```

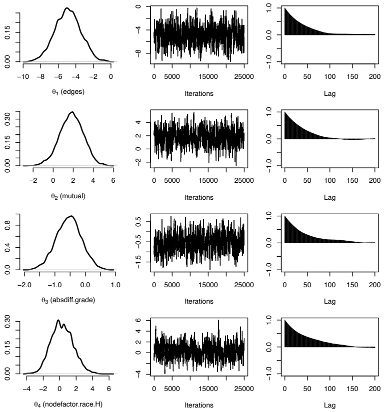
Application

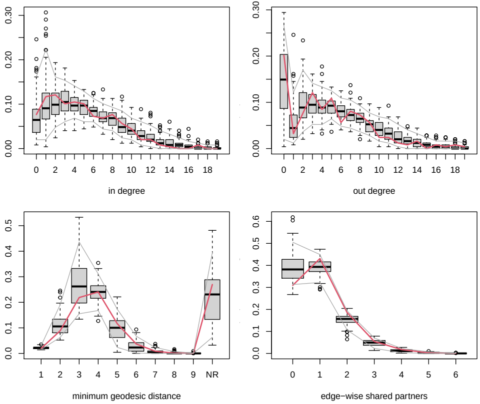
- *evidence()* function includes an additional feature that allows for estimation of $\log \widetilde{\pi}(y \mid \mathcal{M}_m)$.

```
R> cj1 ← evidence(formula = m1, seed = 1,
+     prior.mean = M.prior1, prior.sigma = S.prior1,
+     burn.in = 5000, main.iters = 30000,
+     aux.iters = 2500, n.aux.draws = 50, aux.thin = 50,
+     ladder = 200, V.proposal = 0.5, estimate = "CD",
+     evidence.method = "CJ", num.samples = 25000)

R> cj1$log.evidence

R> summary(cj1)
```

Application

```
R> plot(cj1)

R> bgof(cj1, sample.size = 100, aux.iters = 5000,
+        n.ideg = 20, n.odeg = 20, n.dist = 10, n.esp = 7)
```





Application

- Model $\mathcal{M}_2$ – assumes that the terms for mutuality and transitivity are removed from $\mathcal{M}_1$.

```
R> m2 <- dixon ~ edges + absdiff("grade") +
+     nodefactor("race") +
+     nodefactor("grade") +
+     nodefactor("sex") +
+     nodematch("race", diff = TRUE, levels = c("B","O","W")) +
+     nodematch("grade", diff = TRUE) +
+     nodematch("sex", diff = FALSE) +
+     idegree(0:1) + odegree(0:1)

R> M.prior2 <- c(-5, rep(0, 24))
R> S.prior2 <- diag(5, 25)

R> cj2 <- evidence(formula = m2, seed = 1,
+     prior.mean = M.prior2, prior.sigma = S.prior2,
+     burn.in = 5000, main.iters = 30000,
+     aux.iters = 2500, n.aux.draws = 50, aux.thin = 50,
+     ladder = 200, V.proposal = 0.5, estimate = "CD",
+     evidence.method = "CJ", num.samples = 25000)
```

Application

- The transitivity and mutuality effect are important connectivity features of the observed network.
- The homophily effect of race, sex and grade can help explain the complexity of the observed network data.

Model	Log evidence estimate	CPU (mins)	BF_{12}
$\mathcal{M}_1$	$-38,064.65$	2.93	3.68×10^{58}
$\mathcal{M}_2$	$-38,199.50$	1.13	

Concluding remarks

- Aim of this tutorial: a useful introduction to the main capabilities of the package & some of the algorithms and methods behind it.
- **Bergm** aims to help researchers and practitioners in two ways:
  - A simple, efficient and complete range of tools for conducting Bayesian inference for ERGMs.
  - A platform that can be easily customised, extended and adapted to address different requirements.
- **Bergm** is under continual maintenance and periodic significant upgrading.
- Future developments:
  - Treatment of missing nodal attributes like age & gender.
  - Uncertainty quantification of the Monte Carlo estimates of the evidence.
  - Extensions to weighted networks (Caimo and Gollini, 2020).
  - Extentions to multiplex networks (Krause and Caimo, 2019).

References

Ardia, D., Baştürk, N., Hoogerheide, L., and van Dijk, H. (2012). A comparative study of monte carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis*, 56:3398–3414.

Bouranis, L., Friel, N., and Maire, F. (2018). Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2018.1448832.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55.

Caimo, A. and Friel, N. (2013). Bayesian model selection for exponential random graph models. *Social Networks*, 35(1):11–24.

Caimo, A. and Friel, N. (2014). Bergm: Bayesian exponential random graphs in R. *Journal of Statistical Software*, 61(2):1–25.

Caimo, A. and Gollini, I. (2020). A multilayer exponential random graph modelling approach for weighted networks. *Computational Statistics & Data Analysis*, 142:106825.

Caimo, A. and Lomi, A. (2014). Knowledge sharing in organizations: A Bayesian analysis of the role of reciprocity and formal structure. *Journal of Management*, 41:665–691.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96:270–281.

Everitt, R. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *Journal of Computational and Graphical Statistics*, 24(4):940–960.

Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24:709–723.

References (cont.)

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society, Series B*, 70(3):589–607.

Gilks, W. R., Roberts, G. O., and George, E. I. (1994). Adaptive direction sampling. *Statistician*, 43(1):179–189.

Henning, C., Aßmann, C., Hedtrich, J., Ehrenfels, J., and Krampe, E. (2019). What drives participatory policy processes: Grassroot activities, scientific knowledge or donor money? a comparative policy network approach. *Social Networks*, 58:78–104.

Koskinen, J. H., Robins, G. L., and Pattison, P. E. (2010). Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation. *Statistical Methodology*, 7(3):366–384.

Krause, R. and Caimo, A. (2019). Multiple imputation for Bayesian exponential random multi-graph models. *International Workshop on Complex Networks*, pages 63–72.

Krause, R., Huisman, M., Steglich, C., and Snijders, T. (2020). Missing data in cross-sectional networks – an extensive comparison of missing data treatment methods. *Social Networks*, 62:99–112.

Lazega, E. (2001). *The Collegial Phenomenon: The Social Mechanisms of Cooperation among Peers in a Corporate Law Partnership*. Oxford University Press.

Murray, I., Ghahramani, Z., and MacKay, C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366.

Oates, C., Papamarkou, T., and Girolami, M. (2016). The controlled thermodynamic integral for Bayesian model evidence evaluation. *Journal of the American Statistical Association*, 111(514):634–645.

Resnick, M., Bearman, P., Blum, R., Bauman, K., Harris, K., Jones, J., Tabor, J., Beuhring, T., Sieving, R., Shew, M., Ireland, M., Bearinger, L., and Udry, J. (1997). Protecting adolescents from harm. findings from the national longitudinal study on adolescent health. *Journal of the American Medical Association*, 278(10):823–32.

Ribatet, M., Cooley, D., and Davison, A. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Stat. Sin.*, 22:813–845.

References (cont.)

Sinke, M. R., Dijkhuizen, R. M., Caimo, A., Stam, C. J., and Otte, W. M. (2016). Bayesian exponential random graph modeling of whole-brain structural networks across lifespan. *Neuroimage*, 135:79–91.

Snijders, T., Pattison, P., Robins, G., and Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36:99–153.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *J. Am. Stat. Assoc.*, 85:204–212.

Tasselli, S. and Caimo, A. (2019). Does it take three to dance the tango? Organizational design, triadic structures and boundary spanning across subunits. *Social Networks*, 59:10–22.

Chib and Jeliazkov's method (2001)

- Following from Bayes formula:

$$\widetilde{\pi}(y) = \frac{\tilde{f}(y \mid \theta)p(\theta)}{\widetilde{\pi}(\theta \mid y)}.$$

- Estimate $\log \widetilde{\pi}(y)$ as

$$\log \widetilde{\pi}(y) = \log \tilde{f}(y \mid \theta^*) + \log p(\theta^*) - \log \hat{\widetilde{\pi}}(\theta^* \mid y).$$

- *One block Metropolis-Hastings* approach:

$$\hat{\widetilde{\pi}}(\theta^* \mid y) = \frac{M^{-1} \sum_{m=1}^{M} \widetilde{\alpha}(\theta^{(m)}, \theta^*) h(\theta^{(m)}, \theta^*)}{L^{-1} \sum_{l=1}^{L} \widetilde{\alpha}(\theta^*, \theta^{(l)})}.$$

- $h(\theta, \theta')$ is the candidate generating density.
- $\{\theta^{(m)}\}$ draws from $\widetilde{\pi}(\theta \mid y)$; $\{\theta^{(l)}\}$ draws from $h(\theta^*, \theta)$.

Power posteriors (Friel and Pettitt, 2008)

- Define the power posterior at inverse temperature t by

$$\widetilde{\pi}_t(\theta \mid y) \propto \tilde{f}(y \mid \theta)^t p(\theta), \; t \in [0,1]$$
$$z(y \mid t) = \int_\theta \tilde{f}(y \mid \theta)^t p(\theta) \, d\theta.$$

- Key identity: $\frac{d}{dt} \log z(y \mid t) = \mathbb{E}_{\theta|y,t} \log \tilde{f}(y \mid \theta)$.
- Consequently:

$$\log \widetilde{\pi}(y) = \log \left\{ \frac{z(y \mid t=1)}{z(y \mid t=0)} \right\} = \int_0^1 \mathbb{E}_{\theta|y,t} \log \tilde{f}(y \mid \theta) \, dt.$$

- Discretise the inverse temperatures $0 = t_0 < t_1 < \ldots < t_m = 1$: Friel and Pettitt (2008) recommend $t_j = (j/m)^5, j = 0, \ldots m$.

Power posteriors - Reducing the error (Friel et al., 2014)

- Two sources of error:
  - Sampling error (expensive): estimate $\mathbb{E}_{\theta|y,t_j} \log \tilde{f}(y \mid \theta), \ \forall t_j$.
  - Discretisation error (cheap): approximate integral with numerical integration (trapezium rule):

$$\log \widetilde{\pi}(y) = \sum_{j=1}^{m} (t_j - t_{j-1}) \left[\frac{\mathbb{E}_{\theta|y,t_{j-1}} \log \tilde{f}(y \mid \theta) + \mathbb{E}_{\theta|y,t_j} \log \tilde{f}(y \mid \theta)}{2} \right].$$

- Reducing discretisation error for the trapezium rule:

$$\log \widetilde{\pi}(y) = \sum_{j=1}^{m} (t_j - t_{j-1}) \left[\frac{\mathbb{E}_{\theta|y,t_{j-1}} \log \tilde{f}(y \mid \theta) + \mathbb{E}_{\theta|y,t_j} \log \tilde{f}(y \mid \theta)}{2} \right] - \frac{(t_j - t_{j-1})^2}{12} \left[\mathbb{V}_{\theta|y,t_{j-1}} \log \tilde{f}(y \mid \theta) - \mathbb{V}_{\theta|y,t_j} \log \tilde{f}(y \mid \theta) \right].$$

Power posteriors - Control variates (Oates et al., 2016)

- Aim: improve efficiency (variance reduction) of evidence estimate by:

$$\begin{aligned} k(\theta) &= \log \tilde{f}(y \mid \theta), \\ \mathbb{E}[\tilde{k}(\theta)] &= \mathbb{E}[k(\theta) + h(\theta \mid y, t)], \\ \mathbb{V}[\tilde{k}(\theta)] &< \mathbb{V}[k(\theta)]. \end{aligned}$$

- ZV control variates:

$$\begin{aligned} u(\theta \mid y, t) &= \nabla_\theta \log \widetilde{\pi}_t(\theta \mid y), \\ h(\theta \mid y, t) &= \Delta_\theta[P(\theta \mid \phi(y, t))] + \nabla_\theta[P(\theta \mid \phi(y, t))] \cdot u(\theta \mid y, t). \end{aligned}$$

- Low order polynomials P (degree 2 here) with coefs $\phi(y, t)$.
- Controlled thermodynamic integral:

$$\log \widetilde{\pi}(y) = \int_0^1 \mathbb{E}_{\theta \mid y, t} \left[\log \tilde{f}(y \mid \theta) + h(\theta \mid y, t) \right] \, dt.$$

Pseudo-likelihood adjustment for large networks

- Invertible and differentiable mapping

$$g : \begin{cases} \Theta \to \Theta \\ \theta \mapsto \hat{\theta}_{MPLE} + Q(\theta - \hat{\theta}_{MLE}) \end{cases}$$

General

- Estimate the maxima of the likelihood and the pseudolikelihood:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \log f(y \mid \theta), \quad \hat{\theta}_{MPLE} = \arg\max_{\theta} \log f_{\mathrm{PL}}(y \mid \theta).$$

- Mode and curvature adjusted pseudolikelihood:

$$f_{\mathrm{PL}}(y \mid g(\theta)) = f_{\mathrm{PL}}(y \mid \hat{\theta}_{MPLE} + Q(\theta - \hat{\theta}_{MLE})), \quad Q \in \mathcal{M}(\mathbb{R}^d).$$

- Fully adjusted pseudolikelihood:

$$\tilde{f}(y \mid \theta) = C \cdot f_{\mathrm{PL}}(y \mid g(\theta)).$$

Pseudo-likelihood adjustment for large networks

- $\hat{\theta}_{MLE}$: estimate with MC-MLE (Geyer and Thompson, 1992).
  - Available in R's **ergm** package.
  - Maximise the log-likelihood ratio

$$\ell(\theta) - \ell(\theta_0) = (\theta - \theta_0)^\top s(y) - \log\left\{\frac{z(\theta)}{z(\theta_0)}\right\}.$$

  - Useful property:

$$\frac{z(\theta)}{z(\theta_0)} = \sum_{y \in \mathcal{Y}} \frac{q(y \mid \theta)}{q(y \mid \theta_0)} \frac{q(y \mid \theta_0)}{z(\theta_0)} = \mathbb{E}_{y|\theta_0}\left[\frac{q(y \mid \theta)}{q(y \mid \theta_0)}\right].$$

  - Unbiased IS estimate: simulate draws $y'_1, \ldots, y'_K \sim f(\cdot \mid \theta_0)$ and set

$$\widehat{\frac{z(\theta)}{z(\theta_0)}} = \frac{1}{K}\sum_{k=1}^{K} \frac{q(y'_k \mid \theta)}{q(y'_k \mid \theta_0)}.$$

- $\hat{\theta}_{MPLE}$: standard optimisation methods (BFGS).

Pseudo-likelihood adjustment for large networks

Mode adjustment

$\nabla_\theta \log f(y|\theta)|_{\hat\theta_{MLE}} = Q \nabla_\theta \log f_{\text{PL}}(y|\theta)|_{\hat\theta_{MPLE}}$

Curvature adjustment

$\nabla^2_\theta \log f(y|\theta)|_{\hat\theta_{MLE}} = Q^T \nabla^2_\theta \log f_{\text{PL}}(y|\theta)|_{\hat\theta_{MPLE}} Q$

- Estimate the first two moments on the basis of
  - Gradient: $\nabla_\theta \log f(y \mid \theta) = s(y) - \mathbb{E}_{y|\theta}[s(y)]$,
  - Hessian: $\nabla^2_\theta \log f(y \mid \theta) = -\mathbb{V}_{y|\theta}[s(y)]$,

  using Monte Carlo draws from $f(y \mid \theta)$.
- Moments of $\log f_{\text{PL}}(y \mid \theta)$ in closed form.
- Cholesky decompositions (Ribatet et al., 2012):

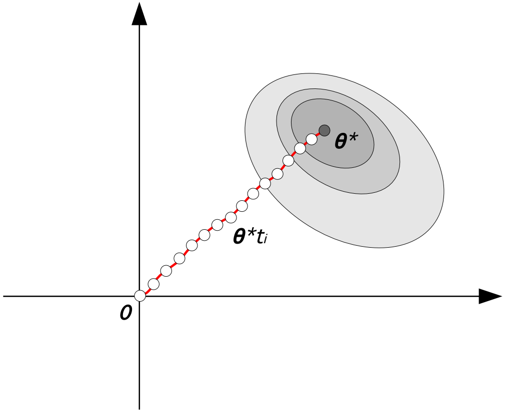
$$\left.\begin{aligned} -\nabla^2_\theta \log f(y \mid \theta)|_{\hat\theta_{MLE}} &= N^T N \\ -\nabla^2_\theta \log f_{\text{PL}}(y \mid \theta)|_{\hat\theta_{MPLE}} &= M^T M \end{aligned}\right\} \Rightarrow Q = M^{-1} N.$$

Pseudo-likelihood adjustment for large networks

Magnitude adjustment

$$\tilde{f}(y \mid \hat{\theta}_{MLE}) = f(y \mid \hat{\theta}_{MLE}) \Leftrightarrow C = \frac{q(y \mid \hat{\theta}_{MLE}) \cdot z^{-1}(\hat{\theta}_{MLE})}{f_{\text{PL}}(y \mid g(\hat{\theta}_{MLE}))}$$

- Aux. variable $t \in [0,1]$: $0 = t_0 < t_1 < \ldots < t_L = 1.$



Pseudo-likelihood adjustment for large networks

- Consider:

$$\frac{z(\hat{\theta}_{MLE})}{z(0)} = \frac{z(t_L \hat{\theta}_{MLE})}{z(t_0 \hat{\theta}_{MLE})} = \prod_{j=0}^{L-1} \frac{z(t_{j+1}\hat{\theta}_{MLE})}{z(t_j \hat{\theta}_{MLE})}.$$

- Note that:

$$\frac{z(t_{j+1}\hat{\theta}_{MLE})}{z(t_j \hat{\theta}_{MLE})} = \mathbb{E}_{y_j \mid t_j \hat{\theta}_{MLE}} \left[\frac{q(y_j \mid t_{j+1}\hat{\theta}_{MLE})}{q(y_j \mid t_j \hat{\theta}_{MLE}))} \right].$$

- Unbiased IS estimate: simulate draws $y'_{j1}, \ldots, y'_{jK} \sim f(\cdot \mid t_j \hat{\theta}_{MLE})$ and set

$$\widehat{\frac{z(t_{j+1}\hat{\theta}_{MLE})}{z(t_j \hat{\theta}_{MLE})}} = \frac{1}{K} \sum_{k=1}^{K} \frac{q(y'_{jk} \mid t_{j+1}\hat{\theta}_{MLE})}{q(y'_{jk} \mid t_j \hat{\theta}_{MLE})}.$$

- Easy to find $z(0) = \begin{cases} 2^{\binom{n}{2}}, & n : \text{\# of nodes (undirected graphs)} \\ 2^N, & N : \text{size of lattice.} \end{cases}$