UNIVERSITY OF OXFORD

DEPARTMENT OF STATISTICS

Generating and assessing synthetic networks

Gesine Reinert

Department of Statistics
University of Oxford

Erlangen AI Hub Conference 2025
June 9, 2025

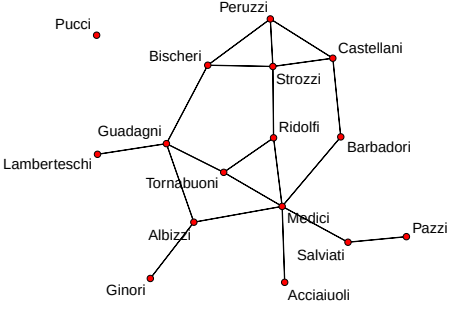
1. Motivation: Synthetic networks
2. Theoretical guarantees: Stein's method and Stein discrepancies
3. Stein's method to characterise exponential random graph models
4. Stein's method for assessing goodness of fit to an ERGM
5. Assessing the quality of graph generators
6. Stein's method for generating fidelitous and diverse networks
7. Some discussion and future directions

Based on joint work with Wenkai Xu (Warwick)

Outline

Padgett's Florentine marriage network

Cosimo de Medici



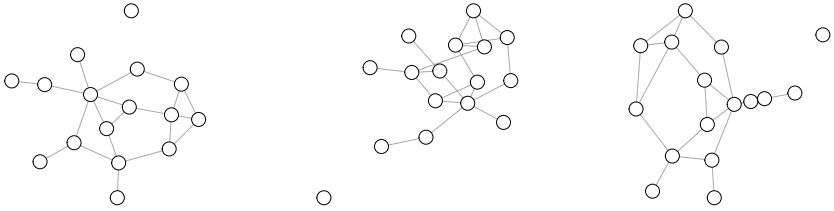
Niccolo Strozzi

Padgett and Ansell (1993), Wasserman and Faust (1994)

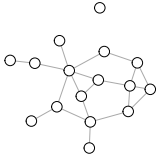
Synthetic networks

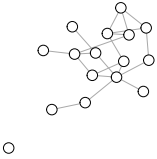
- Synthetitic data are increasingly used in computational statistics and machine learning, for example for
  - privacy;
  - data augmentation;
  - method development.
- One way forward: Set up a parametric model; estimate the parameters; simulate from the model using the estimated parameters.
- Another way forward: Draw samples from the data; simulate from the data using these samples.

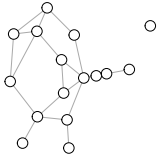
Synthetic data are easy (?)

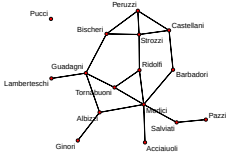


Synthetic data are easy (?)









Desirable features of synthetic data generators

- faithful to the distribution of topological features of interest in the data;
- different enough from the original data to provide variability;
- theoretical guarantees to mitigate risk and assess resilience.

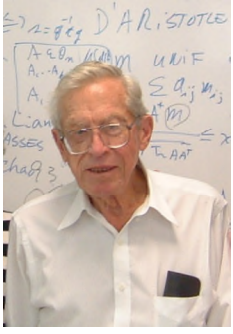
Outline

Stein's method

Starting point
assess distance between distributions

Typical situation
One distribution is relatively simple, the other distribution is more complicated, often based on n random elements.

Aim
explicit bounds, which usually depend on n.



Stein's method in a nutshell

For p a target distribution find a Stein operator $\mathcal{A}_p$ and a Stein class $\mathcal{F}(\mathcal{A}_p)$: if $X \sim p$

$$\mathbb{E}\mathcal{A}_p f(X) = 0 \text{ for all } f \in \mathcal{F}(\mathcal{A}_p) \qquad \text{(Stein characterisation)}$$

For $h \in \mathcal{H}$ a large function class find $f = f_h \in \mathcal{F}(\mathcal{A}_p)$ solving

$$h(x) - \mathbb{E}h(X) = \mathcal{A}_p f(x) \qquad \text{(Stein equation)}.$$

Then for any random element W, $h \in \mathcal{H}$

$$\mathbb{E}h(W) - \mathbb{E}h(X) = \mathbb{E}\mathcal{A}_p f(W).$$

Stein discrepancies

If $X \approx W$ in distribution and if $\mathcal{A}_p$ is a Stein operator for X then, intuitively $\mathbb{E}[\mathcal{A}_p g(W)] \approx 0$ for all sufficiently regular functions g.

Gorham and Mackey (2017); Chwialkowski et al. 2016, Liu et al. 2017
Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) associated with kernel k, inner product $\langle \cdot, \cdot \rangle$ and unit ball $B_1(\mathcal{H})$.

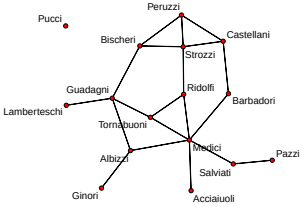
Let $Y \sim q$. The kernel Stein discrepancy (KSD) between p and q is

$$\mathrm{KSD}(p, q; k) = \sup_{f \in B_1(\mathcal{H})} |\mathbb{E}[\mathcal{A}_p f(Y)]|.$$

Outline

Exponential random graph models

Exponential random graph models are often used for social networks.



Graphs $\mathcal{G}_n^{\text{lab}}$: simple, undirected, n labelled vertices, described by

$$x = (x_{i,j}) \in \{0,1\}^{\binom{n}{2}}; \quad 1 \leq i < j \leq n;$$

$x_{i,j} = 1$ if there is an edge between vertices i and j.

Fix $t_1, \ldots, t_k$ which are scaled counts of subgraphs of $\mathcal{G}_n^{\text{lab}}$; $t_1(x)$ is the number of edges in the graph x.

The random graph $X \in \mathcal{G}_n^{\text{lab}}$ follows the exponential random graph model (ERGM) with parameters $\beta = (\beta_1, \ldots, \beta_k) \in \mathbb{R}^k$ if for $x \in \mathcal{G}_n^{\text{lab}}$,

$$\mathbb{P}(X = x) = \frac{1}{\kappa_n(\beta)} \exp\left(\sum_{\ell=1}^{k} \beta_\ell t_\ell(x)\right),$$

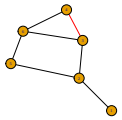
where $\kappa_n(\beta)$ is a normalizing constant. The normalizing constant is usually intractable.

If $k = 1$: ER graph $\mathcal{G}(n, p)$ with independent edges having probability

$$p = e^{2\beta_1}/(1 + e^{2\beta_1}).$$

A Stein operator

Glauber dynamics: Each pair of vertices s has an independent exponentially distributed clock.



When it rings for s, we resample the edge indicator at vertex pair s according to the conditional probability of an edge at s, given the rest of the network.

Notation:

$$x^{(s,1)} \text{ is } x \text{ with } s = 1;$$

$$x^{(s,0)} \text{ is } x \text{ with } s = 0;$$

$$x^{-(s)} \text{ is } x \text{ without } x_s.$$

For a function $h : \{0,1\}^N \to \mathbb{R}$ let $\Delta_s h(x) = h(x^{(s,1)}) - h(x^{(s,0)})$.

The transition probability

For a vertex pair s, the conditional probability of an edge at s in the ERGM given the rest of the network is

$$q_\beta(x^{(s,1)}|x^{-(s)}) = \frac{\exp\left\{\sum_{\ell=1}^k \beta_\ell t_\ell(x^{(s,1)})\right\}}{\exp\left\{\sum_{\ell=1}^k \beta_\ell t_\ell(x^{(s,1)})\right\} + \exp\left\{\sum_{\ell=1}^k \beta_\ell t_\ell(x^{(s,0)})\right\}}$$

which simplies to

$$q(x^{(s,1)}|\Delta_s t_\ell(x), \ell = 1, \ldots, k).$$

Only the changes in counts relative to s need to be computed.
For a $\mathcal{G}(n,p)$ random graph we have $q(x^{(s,1)}|x^{-(s)}) = p$.

Stein operator

The generator $\mathcal{A}_\beta$ of this Markov process on $\mathcal{G}_n$ is

$$\begin{aligned}\mathcal{A}_\beta f(x) &= \frac{1}{N}\sum_{s\in[N]}\Big[q_\beta(x^{(s,1)}|x^{-(s)})(f(x^{(s,1)})-f(x)) \\ &\qquad +(1-q_\beta(x^{(s,1)}|x^{-(s)}))(f(x^{(s,0)})-f(x))\Big] \\ &= \frac{1}{N}\sum_{s\in[N]}\Big[q_\beta(x^{(s,1)}|x^{-(s)})\Delta_s f(x)+\big(f(x^{(s,0)})-f(x)\big)\Big].\end{aligned}$$

This is a Stein operator as for X following the corresponding ERGM model,

$$\mathbb{E}\mathcal{A}_\beta f(X)=0.$$

Chatterjee and Diaconis (2013) showed that in some regime, ERGMs with t_ℓ counting subgraphs H_ℓ with E$_e$*ll* edges are asymptotically close to $\mathcal{G}(n,p)$. Let e_ℓ be the number of edges of H_ℓ. Key functions on $[0,1]$:

$$\Phi(a) = \sum_{\ell=1}^{k} \beta_\ell e_\ell a^{e_\ell - 1}; \quad |\Phi(a)| := \sum_{\ell=1}^{k} |\beta_\ell|\, e_\ell\, a^{e_\ell - 1}; \quad \phi(a) = \frac{e^{2\Phi(a)}}{e^{2\Phi(a)} + 1}.$$

Assumption A: $\frac{1}{2}|\Phi'|(1) < 1$ and $a^* \in [0,1]$ satisfies $a^* = \phi(a^*)$

The so-called *subcritical regime* regime relates to the large deviation behaviour of ER random graphs. For Ising models: high temperature regime.

In *R. and Ross (2019)* we bound the distance by comparing their Stein operators.

Outline

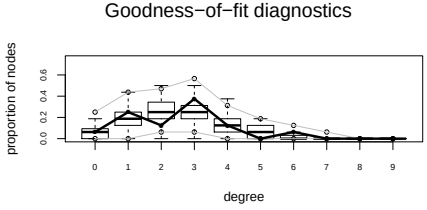
Goodness of fit to an ERGM

The ERGM likelihood has intractable normalizing constant.

Only one network is observed.

We can simulate from a given ERGM.

For assessing goodness of fit the standard approach (*Hunter et al. 2008*) are Monte Carlo tests based on a collection of edge based statistics.



Lospinoso and Snijders 2019 use the Mahalanobis distance between these statistics.

The graph kernel Stein statistic

(*R. and Xu 2021*) Based on our ERGM(β) Stein operator

$$\mathcal{A}_\beta f(x) = \frac{1}{N} \sum_{s \in [N]} \left[q_\beta(x^{(s,1)} | x^{-(s)}) \Delta_s f(x) + f(x^{(s,0)}) - f(x) \right]$$

we define the graph kernel Stein statistic for a network x as

$$\text{gKSS}(\beta, x) = \sup_{f \in B_1(\mathcal{H})} |\mathbb{E} \mathcal{A}_\beta f(x)|.$$

Here $\mathcal{H}$ is a RKHS. We can calculate this supremum explicitly using the RKHS properties.

Theoretical guarantees

Under Assumption A, ERGM(β) $\approx \mathcal{G}(n, a^*)$.

Then gKSS(β, x) $\approx$ gKSS for $\mathcal{G}(n, a^*)$.

In $\mathcal{G}(n, a^*)$ the edge indicators are independent. Under additional assumptions, for $\mathcal{G}(n, a^*)$,

$$\sqrt{n}\text{gKSS}^2 \text{ is approximately normal}$$

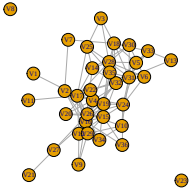
with mean and variance functions of the kernel k.

Stein's method gives explicit bounds on these approximations.

We simulate networks from the null distribution and assess how much the observed statistic differs from the simulated statistics.

Examples

Lazega's lawyer network (*Lazega, 2001*) consists of a network between 36 laywers; Lazega suggests an ER model



A Glasgow friendship network of 50 secondary school students (*Steglich et al., 2006*)

We fit

- An ER $\mathcal{G}(n, p)$ with p the maximum likelihood estimate
- an E2ST model with edges, 2-stars and triangles,
- An ER $\mathcal{G}(n, a^*)$ with a^* calculated from the E2ST fit.

	n	ER (mle)	E2ST	ER(a^*)
Lawyer	36	0.280	0.012	0.152
Teenager	50	0.016	0.060	0.336
Florentine	16	0.52	0.16	0.43

Red: rejected at 5 % level; 100 samples to simulate the null distribution

ER may be rejected whereas ER(a^*) may be accepted.

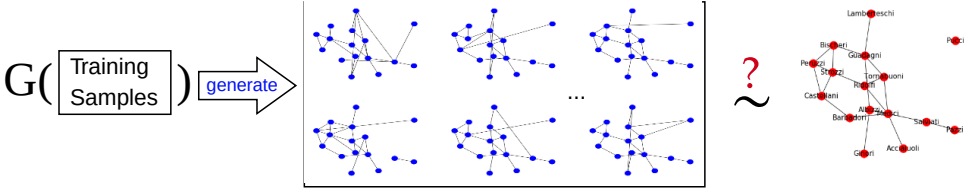
Florentine marriage network: The MLE is $1/6 = 0.1667$ while $a^* = 0.1737$ when estimated from the E2ST model.

Outline

Assessing the quality of graph generators

Suppose a synthetic network generator generates samples which are meant to come from the same distribution of that of an observed network, without knowing that distribution.

How can we assess the quality of the implicit graph generator?



Idea

Use the synthetic networks to estimate the marginal transition probabilities of a Glauber-type Markov chain starting from the initial network, using summary statistics t, which do not have to be sufficient statistics.

We then construct a Stein operator based on the estimated marginal transition probabilities.

From the Stein operator we can construct a kernelised goodness of fit test which we call AgraSSt.

With this test statistic we can test whether the observed network comes from the distribution which underlies the synthetic data generator.

Assessing synthetic graph generators

- GraphRNN (*You et al. 2018*) is an architecture to generate graphs from learning two recurrent neural networks (RNNs), one a vertex-level RNN and the other an edge-level RNN.
- NetGAN (*Bojchevski et al. 2018*) utilises an adversarial approach by training an interplay between a generator and a discriminator neural network on graph data.
- CELL (*Rendsburg et al. 2020*) improves on NetGAN idea by solving a low-rank approximation problem based on a cross-entropy objective.

Statistics for model assessment

- Deg (*Ouadah et al. 2020*) is a degree-based statistics for goodness-of-fit testing of exchangeable random graphs based on the estimated variance of the degree distribution.
- TV_deg denotes the total variation distance between degree distributions.
- MDdeg is the Mahalanobis distance between degree distributions.
- AgraSSt with $t(x^{(s)})$ the edge density of $x^{(s)}$.

Example: Florentine marriage network

	AgraSSt	Deg	MDdeg	TV_deg	density
GraphRNN	0.01	0.11	0.26	0.03	0.188
NetGAN	0.16	0.18	0.09	0.06	0.176
CELL	0.23	0.36	0.69	0.18	0.165

p-values for models trained from the Florentine marriage network; 100 samples to simulate the null distribution; rejected null at significance level $\alpha = 0.05$ is marked red.

Florentine marriage network: edge density 0.167

Outline

Idea

A distribution can be characterised by a Stein operator.

Why not use the Stein operator to generate synthetic networks?

The Stein operator can often be interpreted as transition operator of a Markov process with target distribution as stationary distribution.

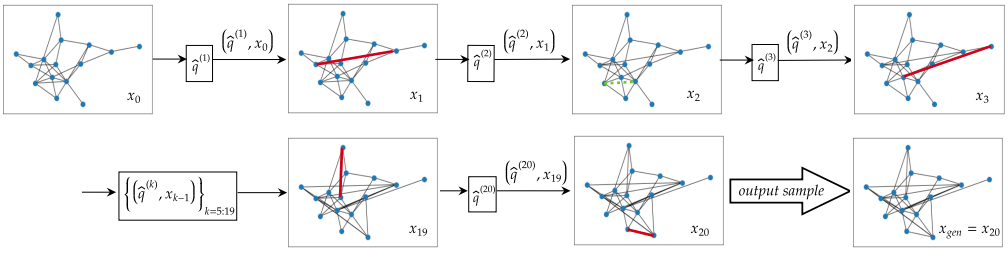
Make an initial estimate of the Stein operator from the data.

Use the Markov process to generate a new sample.

Re-estimate the Stein operator from the new sample.

Iterate.

The SteinGen method



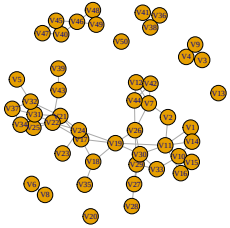
$\hat{q}^{(k)}$ is the (re-)estimated conditional probability based on the current graph x_{k-1}; Glauber dynamics using $\hat{q}^{(k)}$ on the current x_{k-1} generates the next graph sample x_k.

We compare against ...

- samples generated using an ERGM with parameters estimated by MPLE, a maximum pseudo-likelihood estimator from *Schmid and Desmarais (2017)*;
- samples generated using an ERGM with parameters estimated by CD, an estimator based on a local approximation of the gradient of the log likelihood near the observed data, by *Asuncion et al. 2010*;
- CELL, *Rendsburg et al. 2020*;
- Stein_nr, a SteinGen version which estimates the target only once and then proceeds via Gibbs sampling.

Real-world data:Teenager friendship network

We use the Glasgow teenager friendship network of 50 secondary school students (*Steglich et al. 2006*).



Three evaluations:

1. Proportion of rejections of the AgraSSt goodness-of-fit test;
2. Hamming distance to the original network;
3. standard summary statistics.

	Density	2Stars	Triangles	AgraSSt	Hamming
MPLE	0.0421 (2.42e-2)	329 (80.4)	75.52 (43.4)	0.68	0.106 (2.22e-2)
CD	0.2900 (1.10e-2)	4537 (538)	4146 (668)	0.92	0.211 (1.03e-2)
CELL	0.0450 (3.46e-4)	220 (14.1)	22.50 (7.73)	0.12	0.0423 (3.32e-3)
NetGAN	0.1120 (1.38e-6)	227 (13.3)	9.28 (2.53)	0.34	0.0820 (5.07e-3)
SteinGen_nr	0.0516 (1.02e-3)	362 (14.9)	88.90 (24.8)	0.06	0.0912 (9.95e-3)
SteinGen	0.0445 (9.49e-4)	364 (84.1)	85.75 (10.7)	0.08	0.107 (1.32e-2)
Teenager	0.0458	368	86.00	pval=0.64	

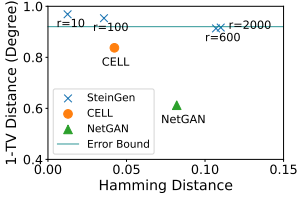
Fidelity: how well can we do?

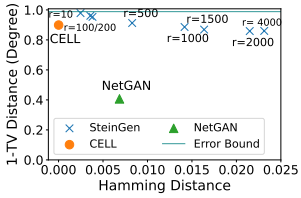
$$d_{TV}(G^{(1)}, G^{(2)}) = \frac{1}{2}\sum_{k=0}^{n-1}\left|\frac{1}{n}\sum_{v=1}^{n}1(\deg^{(1)}(v)=k) - \frac{1}{n}\sum_{v=1}^{n}1(\deg^{(2)}(v)=k)\right|$$

For r generated networks $G^{(i)}$, $i = 1, \ldots, r$ and $G^{(0)}$ the observed network, for $\mathcal{G}(n,p)$ we can show that approximately,

$$\frac{1}{r}\sum_{i=1}^{r}\mathbb{E}d_{TV}(G^{(0)}, G^{(i)}) \in \left[4p(1-p)\sqrt{\frac{1}{n\pi}}, \sqrt{\frac{1}{n\pi}}\right].$$

Comparing realisations to the input network





Diversity (Hamming distance) against fidelity (1 minus total variation distance of the degree distribution, averaged over the simulations)
Ideal: top right corner
Left:Teenager network (50 nodes); right: Yeast protein interactions (2617 nodes)

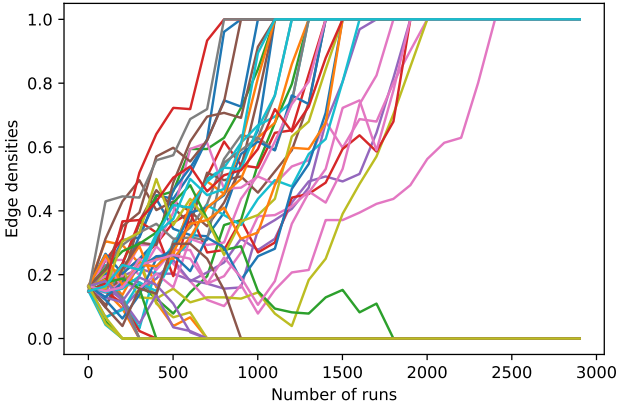
More on SteinGen behaviour

The SteinGen method with re-estimation gives a Markov chain.

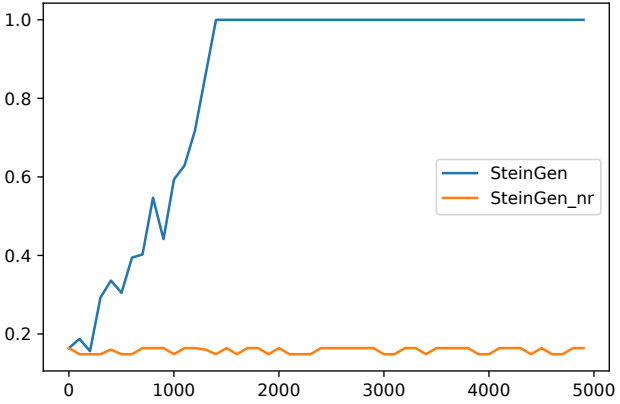
Its absorbing states are the full and the empty graph.

Eventually it will converge to one of these states, and SteinGen fails.

Stein_nr (Gibbs sampling after initial transition probability estimation) does not have this issue.



Florentine family network, Bernoulli model; $n = 16, N = 120$



Florentine family network, Bernoulli model; $n = 16, N = 120$

Outline

Beyond networks

AgraSST can be extended to d-dimensional continuous distributions (*Xu and R. 2022*). We call it NP-KSD for *non-parametric kernel Stein discrepancy*.

Suppose we have a collection of conditional probabilities for the d one-dimensional marginals (which we could estimate via relative frequencies).

Each conditional distribution gives rise to a conditional Stein operator.

Summing over the conditional Stein operators gives a Stein operator for the distribution.

With a Stein operator, we can define a kernelised Stein statistic for testing.

Future: use this idea also for synthetic data generation.

Using Stein's method...

We used Stein's method to

- characterise a network distribution;
- devise a kernelised goodness of fit test;
- assess the quality of synthetic network generators;
- generate synthetic data;
- ... and give theoretical guarantees (not shown).

Work in progress:

- Other models for random networks (*Fatima + R. 2025*);
- network time series.

Future work:

- Include node and edge attributes;
- hypergraph time series (allowing for hyperedges having more than 2 nodes);
- scaling properties;
- general point clouds.

Fatima, A. & R., G. (2025). A kernelised Stein discrepancy for assessing the fit of inhomogeneous random graph models. arXiv:2505.21580.
R., G. & Xu, W. (2024). SteinGen: Generating Fidelitous and Diverse Graph Samples. arXiv:2403.1857.
Xu, W. & R., G. (2022) A Kernelised Stein Statistic for Assessing Implicit Generative Models. NeurIPS 35: 7277-7289.
Xu, W. and R., G. (2022) AgraSSt: Approximate Graph Stein Statistics for Interpretable Assessment of Implicit Graph Generators. NeurIPS 35: 24268-24279; also arXiv:2203.03673.
Xu, W. and R., G. (2021) A Stein Goodness-of-test for Exponential Random Graph Models. In: International Conference on Artificial Intelligence and Statistics, pp. 415-423.
R., G. and Ross, N. (2019) Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. The Annals of Applied Probability 29.5: 3201-3229.