Comparing ERGM to LOLOG

Comparing the Real-World Performance of Exponential-family Random Graph Models and Latent Order Logistic Models for Social Network Analysis

Duncan A. Clark and Mark S. Handcock

Department of Statistics
University of California - Los Angeles

UCLA

Paper available at
`https://doi.org/10.1111/rssa.12788`

INSNA Sunbelt, July 14th, 2022

Overview

- Exponential-family Random Graph Models (ERGM) are able to represent complex network generation processes.
- A new family, Latent Order Logistic (LOLOG) Models, developed by Ian E. Fellows, have similar properties
- LOLOG posit the existence of a latent discrete temporal dimension along which the network edges form.
- How do we compare the two families?
  - Both are fully general and have intuitive parameters.
- The fundamental question here is:
  - How well do typical ERGM/LOLOG family members fit the sorts of data that social networkers model?
- Also of interest are computational complexity, degeneracy, stability, diagnostics, etc

Network Modeling Refresher:

Let Y be a random graph whose realization is $y \in \mathscr{Y} = \{a \in \{0,1\}^{n \times n} \quad | \quad \forall i, j \quad a_{i,i} = 0\}$.

The number of nodes n, and matrix of nodal covariates X are fixed and known.

Social network models in our context are probability mass functions (PMFs) over $\mathscr{Y}$ typically parameterized in an socially interpretable way.

As the social network generation process is typically complex, modeling is intrinsically challenging.

LOLOG and ERGM are alternative specifications of the distribution of Y.

Exponential-family Random Graph Model (ERGM)

An ERGM for the network can be expressed as

$$p_E(y|\theta) = \frac{\exp(\theta \cdot g(y))}{c(\theta)} \qquad y \in \mathscr{Y} \tag{1}$$

where $g(y)$ is a $d-$vector valued function defining a set of sufficient statistics

$\theta \in \mathbb{R}^d$ is a vector of parameters

$c(\theta)$ the normalizing constant.

Each ERGM family member is defined by the choice of sufficient statistics.

ERGM Challenges

ERGMs are easy to specify, but challenging to specify correctly and often to estimate:

- $c(\theta)$ is a sum over all possible graphs (undirected case : $\mathcal{O}(2^{n^2})$) so evaluating likelihood is impossible
- State of the art estimation is MCMC MLE algorithm.
- Sampling from ERGMs requires full MCMC burn-in, which can be computationally expensive
- ERGMs are prone to degeneracy, even with many standard statistics. Other statistics have been developed.
- Tapering, also developed by Ian E. Fellows, is also very helpful though under-utilized.

ERGM and Change Statistics

Toggling tie variables $Y_{i,j}$ gives the so-called change statistics. Letting y_{ij}^{+} be the graph y with the tie variable (i, j) toggled on and $y_{i,j}^{-}$, with the tie variable toggled off we define:

$$c_{i,j} = g(y_{i,j}^{+}) - g(y_{i,j}^{-}) \tag{2}$$

These are used in the MCMC sampling. As well as giving the following logistic regression style interpretation where $y_{i,j}^{c}$ is the observed graph y excluding the tie variable $Y_{i,j}$.

$$\log\left(\frac{P(Y_{i,j}=1|y_{i,j}^{c})}{P(Y_{i,j}=0|y_{i,j}^{c})}\right) = \theta \cdot \left(g(y_{i,j}^{+}) - g(y_{i,j}^{-})\right) \tag{3}$$

LOLOG Specification

LOLOG posit the existence of a latent discrete temporal dimension along which the network edges form.

The social forces that result in the edges are modeled sequentially.

A LOLOG model is specified by two components. First is *the probability of a graph given a specified order of edge formation, s*:

$$p(y|s,\theta) = \prod_{t=1}^{|y|} \frac{1}{Z_t(s)} \exp\left(\theta \cdot C_{s,t}\right) \tag{4}$$

where $s \in \mathscr{S}_{|y|}$ is the set of possible edge formation orders, and LOLOG change statistics are defined as

$$C_{s,t} = g(y_t, s_{\leq t}) - g(y_{t-1}, s_{\leq t-1}) \tag{5}$$

where $s_{\leq t}$ denotes the first t elements of $s \in \mathscr{S}_{|y|}$.

LOLOG Specification

The $Z_t(s)$ sequentially specify the normalizing constants.

Let y_t^+ be the graph y_{t-1} with the edge s_t added, then

$$Z_t(s) = \exp\left(g(y_t^+, s_{\leq t}) - g(y_{t-1}, s_{\leq t-1})\right) + 1 \tag{6}$$

LOLOG Specification

The second component is *a model for the edge order permutations,* $p(s)$. The LOLOG distribution for Y is then:

$$
\begin{aligned}
p_L(y|\theta) &= \sum_s p(y|s,\theta)p(s) \\
&= \sum_s \left(p(s) \prod_{t=1}^{|y|} \frac{1}{Z_t(s)} \exp\left(\theta \cdot C_{s,t}\right) \right)
\end{aligned} \tag{7}
$$

Comparing ERGM to LOLOG

- How to they compare on the population of networks that social network researchers analyze?
- In this study, we consider the population of networks from *Social Networks*, the premier INSNA journal (peer reviewed)
- Requested data from all articles using ERGM, and compared to fits with LOLOG models.
- After various exclusions, 35 networks from 14 peer review papers.

Ensemble Description I

Table: Properties of each network contained in the ensemble. The ensemble includes directed and undirected networks from various applications ranging in size from 16 nodes to 1681 nodes

Description	Nodes	Edges	Directed	Nodal Covariates
Add Health	1681	1236	Undirected	4
School Friends	Various	Varies	Directed	3
Kapferer's Tailors	39	267	Undirected	0
Florentine Families	16	15	Undirected	2
German Schoolboys	53	53	Directed	4
Employee Voice	27	104	Directed	3
Employee Voice	24	53	Directed	3
Employee Voice	30	126	Directed	3
Employee Voice	31	139	Directed	3
Employee Voice	37	149	Directed	3

Ensemble Description II

Employee Voice	39	155	Directed	3
Office Layout	67	211	Directed	0
Office Layout	69	203	Directed	0
Office Layout	109	458	Directed	0
Office Layout	119	872	Directed	0
Disaster Response	20	148	Directed	7
Company Boards	808	1997	Undirected	0
Company Boards	808	1740	Undirected	0
Company Boards	808	1682	Undirected	0
Company Boards	808	1622	Undirected	0
Swiss Decisions	24	282	Directed	8
Swiss Decisions	23	294	Directed	8
Swiss Decisions	20	169	Directed	8
Swiss Decisions	25	224	Directed	8
Swiss Decisions	24	248	Directed	8
Swiss Decisions	20	227	Directed	8
Swiss Decisions	22	256	Directed	8

Ensemble Description III

Swiss Decisions	19	138	Directed	8
Swiss Decisions	26	280	Directed	8
Swiss Decisions	26	316	Directed	8
University Emails	1133	10903	Undirected	0
School Friends	22	177	Directed	1
School Friends	24	161	Directed	1
School Friends	22	103	Directed	1
Online Links	158	1444	Directed	3
Online Links	150	1382	Undirected	3

A rubic for comparison of models

1. Are we able to recreate the published ERGM qualitatively?
   We asked this to screen out network data where our usage differs qualitatively from the original, for whatever reason. This is to help ensure we were using the data correctly, so that our comparison is valid.
2. Do the recreations of the published ERGM fit the network well?
   This is to assess the validity of the published ERGM results, and to assess if ERGM is a good model for the published case study.
3. Are we able to fit the LOLOG with the published ERGM terms?
   This is to assess the LOLOG on terms likely favorable to the ERGM. Typically, published ERGM will have undergone model selection criteria to choose terms that had good fit compared to other possible ERGM. This criteria assesses the flexibility of the LOLOG model class.

A rubic for comparison of models

1. Does the LOLOG model with the published ERGM terms fit well?
2. Are we able to fit the LOLOG model with ERGM Markov terms (that are often degenerate in ERGM)?
   Markov terms, such as k-stars and triangles, often lead to near-degenerate models
3. Is a better fit achieved with LOLOG than the published ERGM?
4. Do the published ERGM and best-fitting LOLOG models have consistent interpretations?
5. Which model do we believe to be more useful?

Detailed Example: Sailer's Office Layouts

- Four networks of daily social interactions between workers within four different office spaces
- An ERGM based analysis was originally carried out in Sailer and McCulloch (2012).
- The networks are directed and have 69, 63, 109 and 120 nodes.
- Have covariates for usefulness, team membership and floor in building.

ERGM fits

Table: Office layout ERGM fits as per the published results. In all cases the selected measure of distance is negative and significant suggesting that close office workers, are more likely to interact, even after allowing for team, floor, usefulness as well as social structure in the form of reciprocity and transitivity.

	University 2005	University 2008	Research Institute	Publisher
Edges	-3.4 (0.37)***	-4.41 (0.2)***	-4.1 (0.12)***	-5.07 (0.15)***
Reciprocity	0.38 (0.45)	0.62 (0.31)***	2.39 (0.2)***	-1.26 (0.19)***
GWESP(0.5)	1.36 (0.14)***	1.24 (0.11)***	0.92 (0.07)***	2.09 (0.09)***
Usefulness	0.7 (0.15)***	0.54 (0.11)***	0.81 (0.04)***	1.31 (0.05)***
Team Match	0.78 (0.18)***	0.56 (0.1)***	NA	NA
Floor Match	0.15 (0.26)	0.58 (0.14)***	NA	NA
Metric Distance	-0.04 (0.01)***	-0.01 (0)***	-0.01 (0)***	NA
Topo Distance	NA	NA	NA	-0.06 (0)***

*** p-value < 0.001 , ** p-value < 0.01, * p-value < 0.05

LOLOG model fits

Table: Office layout LOLOG fit with the same terms as the published ERGM. Model fits show broad qualitative agreement with the published results using the ERGM in Table 3

	University 2005	University 2008	Research Institute	Publisher
Edges	-1.69 (0.38)***	-3.67 (0.36)***	-3.18 (0.13)***	-1.63 (0.09)***
Reciprocity	1.99 (0.34)***	1.96 (0.31)***	3.9 (0.25)***	0.64 (0.2)***
GWESP(0.5)	0.55 (0.12)***	0.87 (0.13)***	0.73 (0.09)***	-0.22 (0.06)***
Usefulness	1.02 (0.15)***	0.81 (0.14)***	1.21 (0.05)***	1.89 (0.06)***
Team Match	1.29 (0.19)***	0.72 (0.19)***	NA	NA
Floor Match	-0.28 (0.3)	1.08 (0.29)***	NA	NA
Metric Distance	-0.07 (0.01)***	-0.02 (0.01)***	-0.02 (0)***	NA
Topo Distance	NA	NA	NA	-0.1 (0)***

*** p-value < 0.001 , ** p-value < 0.01, * p-value < 0.05

Overall Results

We fitted many hundreds of models, very broad summary comments are as follows:

- In many cases we were not able to recreate the published ERGM, while it was possible the GOF on important properties was often poor.
- We were able to use the same terms from the ERGM in a LOLOG model for around 50% of networks.
- When fitting LOLOG models with ERGM terms the LOLOG usually did not fit well
- We could usually fit LOLOG models with terms that were degenerate under ERGM. This achieved better fit.
- Full results are contained in Duncan Clark's thesis (Clark 2022).

Discussion

- LOLOG can be fit to most members of an ensemble of network data sets published with ERGM fits.
- There is likely a strong selection bias towards networks that are well suited to ERGM.
- The ERGM and LOLOG qualitative interpretations were typically consistent.
- LOLOG models are at least the equal of the ERGM, in terms of GOF and interpretability.
- There as strong evidence that the LOLOG model is useful for modeling real social network data

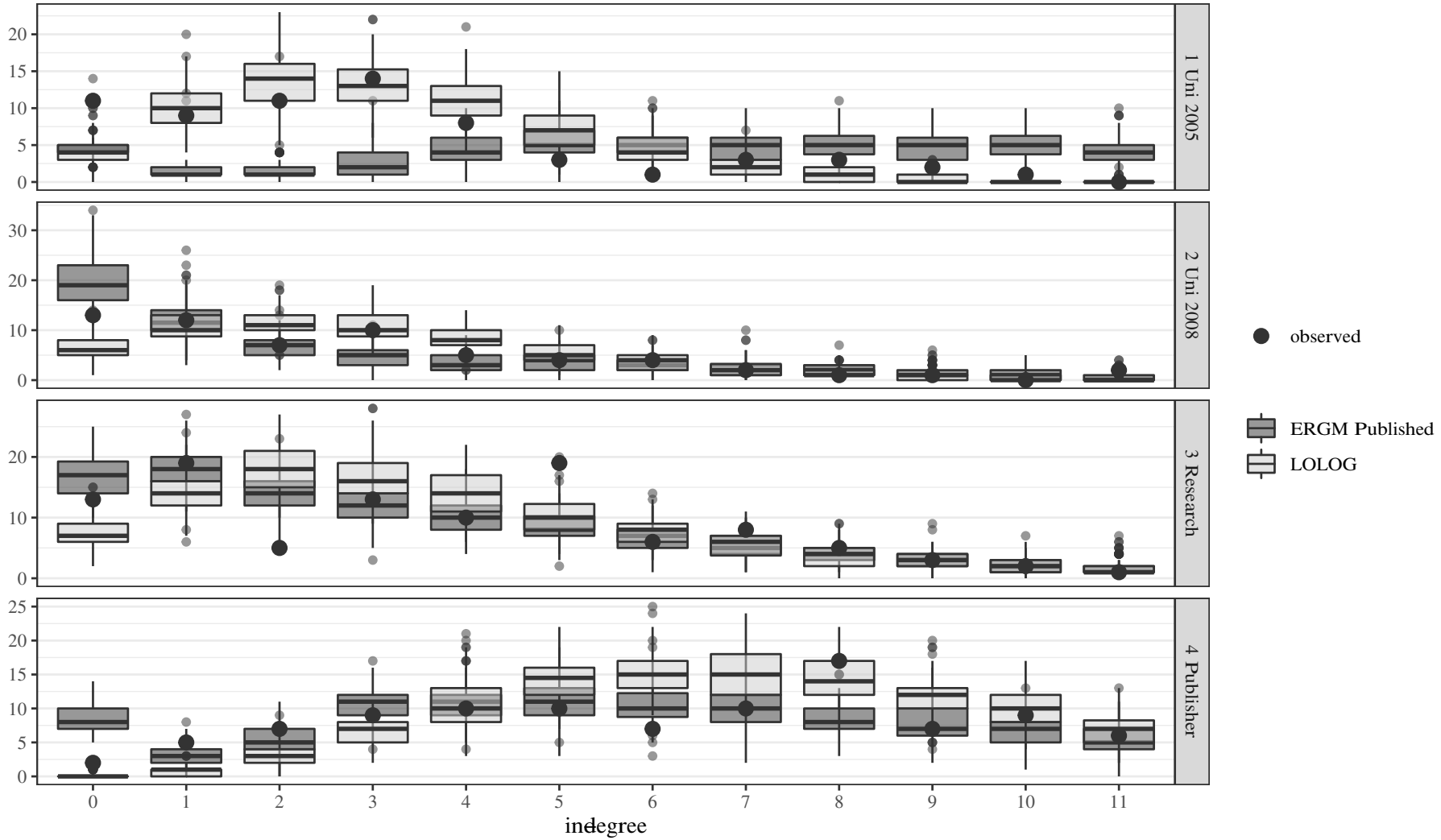
References I


Fellows, I. E. (2018, April).
A new generative statistical model for graphs: The latent order logistic (lolog) model.


Sailer, K. and I. McCulloch (2012).
Social networks and spatial configuration—how office layouts drive social interaction.
Social Networks 34, 47–58.

GOF comparisons

Figure: In-degree goodness of fit comparison plot for Office layout networks.



LOLOG improved fit

Table: Office layout LOLOG fit with GWESP and 2- and 3- in- and out-stars.

	University 2005	University 2008	Research Institute	Publisher
Edges	-3.2 (0.67)***	-5.04 (0.59)***	-4.04 (0.22)***	-4.87 (1.19)***
Reciprocity	2.03 (0.77)***	1.11 (0.45)***	4.7 (0.52)***	3.16 (1.27)***
GWESP(0.5)	0.33 (0.2)	0.49 (0.16)***	0.77 (0.11)***	0.01 (0.26)
Out-2-Star	1.39 (0.26)***	0.65 (0.16)***	0.41 (0.07)***	0.69 (0.15)***
Out-3-Star	-0.28 (0.07)***	-0.07 (0.03)***	-0.04 (0.01)***	-0.02 (0)***
In-2-Star	0.26 (0.22)	0.25 (0.15)	0.21 (0.12)	0.73 (0.54)
In-3-Star	-0.04 (0.05)	-0.03 (0.02)	-0.09 (0.03)***	-0.18 (0.1)
Usefulness	1.07 (0.2)***	0.75 (0.16)***	1.28 (0.07)***	2.98 (0.61)***
Team Match	1.93 (0.31)***	1.14 (0.25)***	NA	NA
Floor Match	-0.24 (0.47)	1.35 (0.43)***	NA	NA
Metric Distance	-0.09 (0.01)***	-0.02 (0.01)***	-0.02 (0)***	NA
Topo Distance	NA	NA	NA	-0.24 (0.06)***

*** p-value < 0.001 , ** p-value < 0.01, * p-value < 0.05

Improved fit GOF

Figure: In-degree goodness of fit comparison plot for Office layout networks.