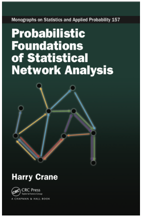
Probabilistic Foundations of Statistical Network Analysis
Chapter 2: Binary relational data

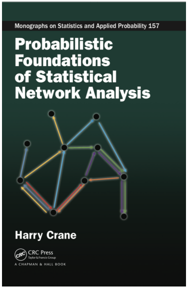
Harry Crane

Based on Chapter 2 of *Probabilistic Foundations of Statistical Network Analysis*


Monographs on Statistics and Applied Probability 157
Probabilistic Foundations of Statistical Network Analysis
Harry Crane
CRC Press
A CHAPMAN & HALL BOOK

Book website: `http://www.harrycrane.com/networks.html`

Table of Contents


Monographs on Statistics and Applied Probability 157
Probabilistic Foundations of Statistical Network Analysis
Harry Crane
CRC Press
A CHAPMAN & HALL BOOK

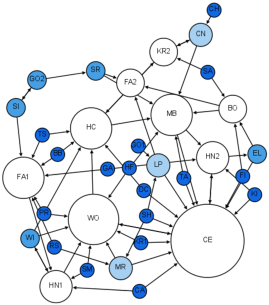
Chapter 1	Orientation
2	**Binary relational data**
3	Network sampling
4	Generative models
5	Statistical modeling paradigm
6	Vertex exchangeability
7	Getting beyond graphons
8	Relative exchangeability
9	Edge exchangeability
10	Relational exchangeability
11	Dynamic network models

Outline

1. Scenario: International Relations Data
2. Dyad independence model
3. Exponential Random Graph Model (ERGM)
4. Scenario: Friendships in a high school
5. Network inference under sampling

Basic setup

- Many networks represent relational information among a fixed collection of individuals:
  - Friendships among co-workers
  - International relations among countries
  - Connectivity among neurons
- Vertices are fixed and known prior to observing the relations (edges) among them.
- Typically represented as a **graph** $G = (V, E)$ with vertex set V and edge set $E \subseteq V \times V$.



- Sociogram from Moreno (1930s).

Scenario 1: International Relations Data

- Let $[n] = \{1, \ldots, n\}$ index a set of countries (e.g., USA, England, China, Russia, etc.).
- $\mathbf{Y} = (Y_{ij})_{1 \leq i,j \leq n}$ be the binary relational data with $Y_{ij} = 1$ if i imports goods from j and $Y_{ij} = 0$ otherwise.

	USA	Russia	China	England	$\cdots$
USA	–	0	1	1	$\cdots$
Russia		–	1	0	$\cdots$
China			–	0	$\cdots$
England				–	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

- Assume that $\mathbf{Y}$ is observed without any further information about the countries, such as GDP, geographical location, etc.
- **Goal**: describe any interesting patterns among the trade relationships among these countries.

Summarizing network structure

- Scenario 1: Data for fixed collection of countries (no sampling).
- Sociometric studies: number of vertices small/moderate, but network still too complex to visualize.
- Model serves as tool for summarizing network structure. (Exploratory Data Analysis).

Properties of good model:

- Easily interpretable parameters.
- Computationally feasible.
- No need for sophisticated generative models or sampling constraints.

Common approach:

- Compute summary statistics of interest.
- Analyze how network structure depends on these statistics.
- For example:
  - reciprocity: both *i* and *j* import from one another
  - differential attractiveness: popularity compared to other vertices
  - transitivity: if *i* imports from *j* and *j* imports from *k*, how likely that *i* imports from *k*?

Dyad independence model

- **Dyad**: $D_{ij} = (Y_{ij}, Y_{ji})$ (relationship for pair i and j)
- Define a probability distribution p_{ij} for each dyad D_{ij}, $1 \leq i < j \leq n$:

$$p_{ij}(z, z') := \Pr(D_{ij} = (z, z')), \quad z, z' \in \{0, 1\}. \tag{1}$$

- p_1 **model**: Given $\mathbf{p} = (p_{ij})_{1 \leq i < j \leq n}$ and the assumption that dyads $(D_{ij})_{1 \leq i < j \leq n}$ are independent according to (1), $\mathbf{Y} = (Y_{ij})_{1 \leq i,j \leq n}$ has distribution

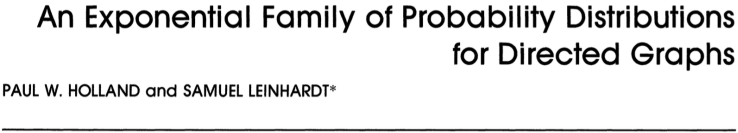
$$\begin{aligned} \Pr(\mathbf{Y} = \mathbf{y}; \mathbf{p}) &= \prod_{1 \leq i < j \leq n} p_{ij}(y_{ij}, y_{ji}) & (2) \\ &\propto \exp\left\{ \sum_{1 \leq i < j \leq n} \rho_{ij} y_{ij} y_{ji} + \sum_{1 \leq i \neq j \leq n} \theta_{ij} y_{ij} \right\} & (3) \end{aligned}$$

for each $\mathbf{y} = (y_{ij})_{1 \leq i,j \leq n} \in \{0, 1\}^{n \times n}$, where

$$\rho_{ij} = \log\left(\frac{p_{ij}(0,0) p_{ij}(1,1)}{p_{ij}(0,1) p_{ij}(1,0)}\right) \quad \text{and}$$

$$\theta_{ij} = \log(p_{ij}(1,0)/p_{ij}(0,0)).$$

p_1 model (Holland and Leinhardt)



An Exponential Family of Probability Distributions for Directed Graphs

PAUL W. HOLLAND and SAMUEL LEINHARDT*

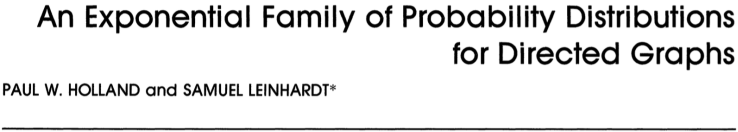
$$\Pr(\mathbf{Y} = \mathbf{y}; \mathbf{p}) \quad \propto \quad \exp\left\{ \sum_{1 \leq i < j \leq n} \rho_{ij} y_{ij} y_{ji} + \sum_{1 \leq i \neq j \leq n} \theta_{ij} y_{ij} \right\}$$

for

$$\begin{aligned} \rho_{ij} &= \rho \quad \text{and} \\ \theta_{ij} &= \theta + \alpha_i + \beta_j. \end{aligned}$$

- ρ indicates the relative probability that two generic vertices *reciprocate* their relation to one another;
- α_i and β_i capture the *differential attactiveness* of each vertex i, which indicate how strongly (relative to other vertices) i is to have outgoing links (α_i) and incoming links (β_i).

p_1 model (Holland and Leinhardt)



An Exponential Family of Probability Distributions for Directed Graphs

PAUL W. HOLLAND and SAMUEL LEINHARDT*

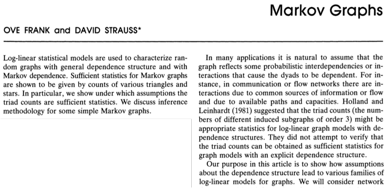
Benefits:

- Interpretable parameters
- Computable in closed form
- Consistent with respect to selection sampling (more later)

Drawbacks:

- Address only specific attributes (reciprocity, differential attractiveness)
- Not flexible enough for most applications of interest

Exponential random graph model (ERGM)



Markov Graphs

OVE FRANK and DAVID STRAUSS*

Log-linear statistical models are used to characterize random graphs with general dependence structure and with Markov dependence. Sufficient statistics for Markov graphs are shown to be given by counts of various triangles and stars. In particular, we show under which assumptions the triad counts are sufficient statistics. We discuss inference methodology for some simple Markov graphs.

In many applications it is natural to assume that the graph reflects some probabilistic interdependencies or interactions that cause the dyads to be dependent. For instance, in communication or flow networks there are interactions due to common sources of information or flow and due to available paths and capacities. Holland and Leinhardt (1981) suggested that the triad counts (the numbers of different induced subgraphs of order 3) might be appropriate statistics for log-linear graph models with dependence structures. They did not attempt to verify that the triad counts can be obtained as sufficient statistics for graph models with an explicit dependence structure.

Our purpose in this article is to show how assumptions about the dependence structure lead to various families of log-linear models for graphs. We will consider network

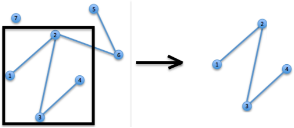
- Real-valued parameters $\theta_1, \dots, \theta_k \in \mathbb{R}$.
- Sufficient statistics $T_1, \dots, T_k : \{0,1\}^{n \times n} \to \mathbb{R}$.
- **Definition**: The *exponential random graph model* (ERGM) with (natural) parameter $\theta = (\theta_1, \dots, \theta_k)$ and (canonical) sufficient statistic $T = (T_1, \dots, T_k)$ assigns probability

$$\Pr(\mathbf{Y} = \mathbf{y}; \theta, T) = \frac{\exp\{\sum_{i=1}^k \theta_i T_i(\mathbf{y})\}}{\sum_{\mathbf{y}^* \in \{0,1\}^{n \times n}} \exp\{\sum_{i=1}^k \theta_i T_i(\mathbf{y}^*)\}} \tag{4}$$

  to each $\mathbf{y} \in \{0,1\}^{n \times n}$.
- p_1 model and Erdős–Rényi model have form of (4).
- Much more general than p_1 model, but difficult to compute normalizing constant and lacks consistency under subsampling.

Scenario 2: Friendships in a high school

- High school with N students.
- Sample $n < N$ students and observe the friendships among them.



- Unlike previous (IR) scenario, the observed relationships here are only a sample of the total population of friendships of interest.
- Using the observation $\mathbf{Y}_n$ to infer patterns in the population $\mathbf{Y}_N$ requires an assumption about how the sampled students are related to the population of all students.
- Inference about $\mathbf{Y}_N$ based on $\mathbf{Y}_n$ entails an assumption that $\mathbf{Y}_n$ is somehow representative of the population $\mathbf{Y}_N$, raising the question:

> *In what way is the observed data $\mathbf{Y}_n$ representative of the relationships $\mathbf{Y}_N$ for the whole population?*

Network inference under sampling

- Arises in high school friendship scenario, not International Trade scenario.
- Consider how the observed friendships vary if obtained under the following different scenarios:
  1. n students are sampled uniformly among all freshman, i.e., first year students, in the school;
  2. n students are sampled uniformly among all senior, i.e., final year students, in the school;
  3. n students are sampled uniformly among all students in the school; and
  4. all students who write for the school newspaper, of which there are n in total, are sampled.
- Scenarios 1-3: sampling mechanism is the same but population is different.
- Scenario 4: population is same as in 3, but sampling mechanism differs — sampled students are known to already have similar interests, i.e., writing for the newspaper, and therefore more likely than randomly selected students to be friends.
- Also notice: number of observed individuals in Scenario 4 is determined by number of students who write for the newspaper — not specified *a priori* by data analyst as in scenarios 1–3.

Effects of observation/sampling mechanism often overlooked in network modeling.

Moving forward

- Sampling considerations not exclusive to network modeling — all well-specified statistical models must account for observation mechanism.
- In many classical settings the observation mechanism is obvious and, therefore, overlooked.
  - e.g., i.i.d. assumption establishes implicit relationship between observed data and rest of population — all observations independent and from same distribution.
  - Even in this case, assumption must be scrutinized with respect to circumstances of the given problem.
- Departures from i.i.d. have led to new frameworks, e.g., time series, hidden Markov models, etc.

- Some recent progress on sampling in network modeling, but most of the focus has been on *selection sampling*.
- Selection sampling unrealistic for most practical applications.

References to p_1 model and ERGM:

- *Frank and Strauss. Markov Graphs.*
- *Holland and Leinhardt. An exponential family of probability distributions for directed graphs.*
- *Wasserman and Pattison. Logit models and logistic regression for social networks.*