EDGE EXCHANGEABILITY: A NEW FRAMEWORK FOR NETWORK MODELING

HARRY CRANE RUTGERS UNIVERSITY

Joint work with Walter Dempsey, University of Michigan

INTRODUCTION

The most common network models cannot replicate the asymptotic behaviors of sparsity and power law degree distribution.

In many network datasets, the basic units are the edges (e.g., collaborations, interactions, phone calls), suggesting an alternate theory for edge-labeled networks and **edge exchangeable** network models.

We prove that edge exchangeable models can replicate sparsity and power law, and we develop this new framework for network modeling.

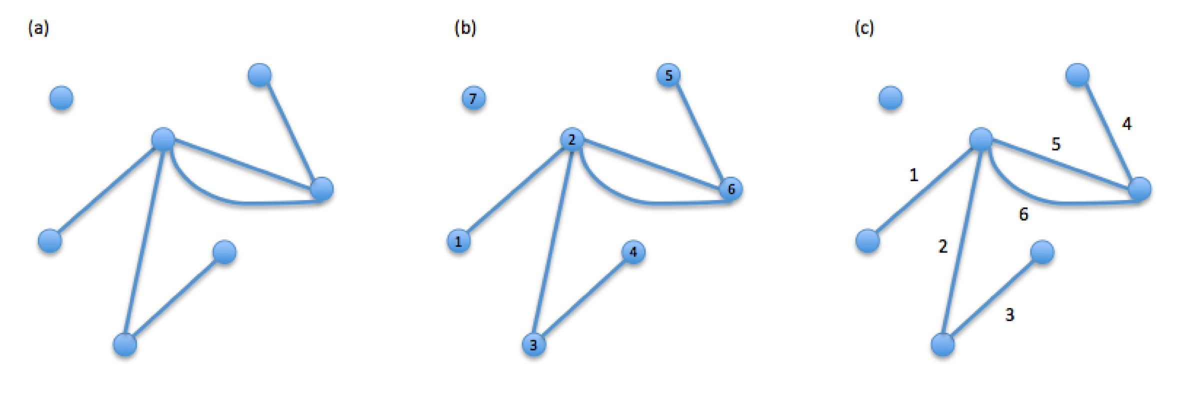

Figure 1: Representations of network data.

EXCHANGEABILITY

Exchangeability: invariance with respect to relabeling.

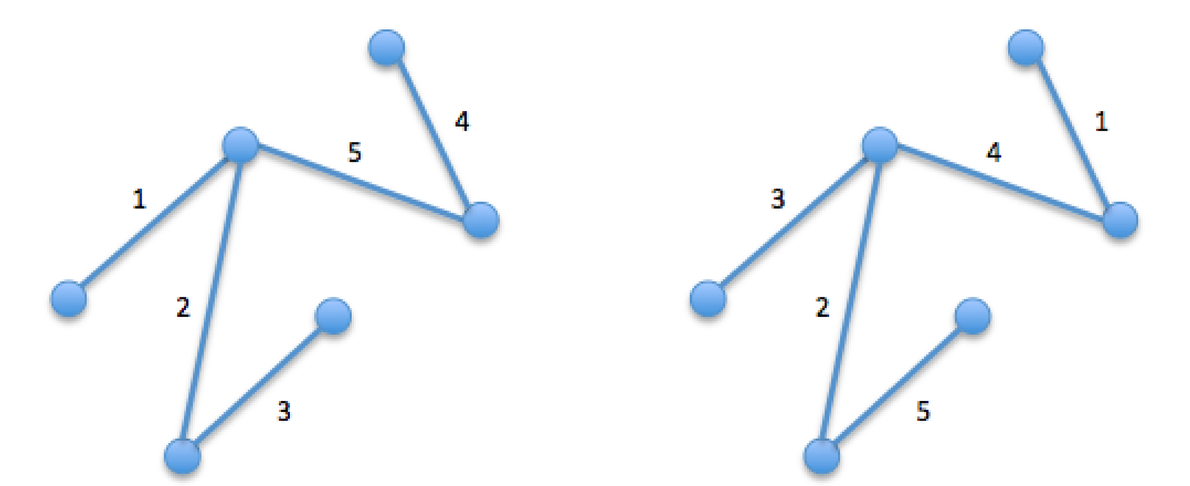

Figure 2: **Edge-exchangeable** models assign equal probability to isomorphic edge-labeled networks.

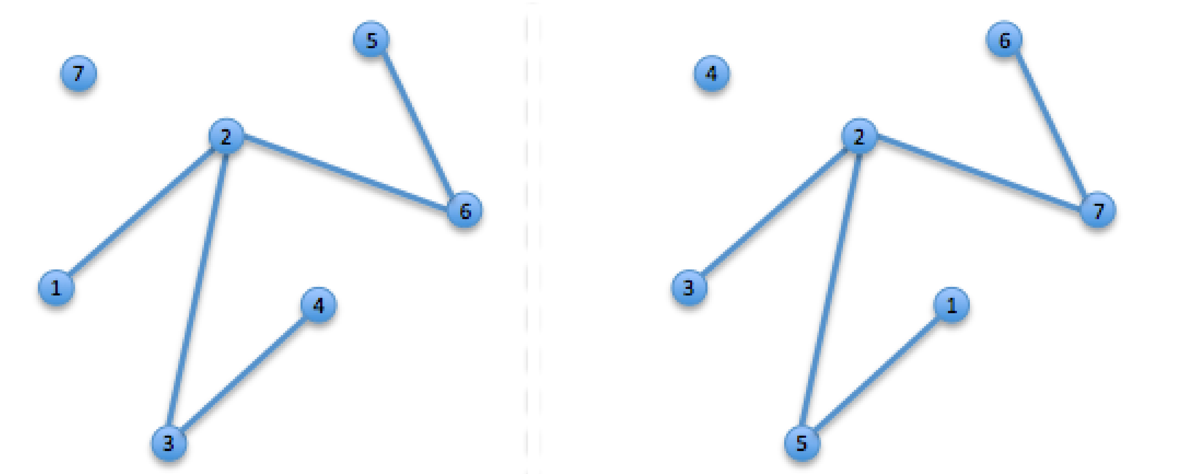

Figure 3: Vertex-exchangeable models assign equal probability to isomorphic vertex-labeled graphs.

NETWORK PROPERTIES

Many modern network datasets (Internet, collaboration networks, Facebook) exhibit

- **sparsity**: have few edges relative to the number of vertices. In particular, a sequence of networks $(G_n)_{n\geq 1}$ is *sparse* if
$$\limsup_{n\to\infty} \frac{\#\text{edges}(G_n)}{\#\text{vertices}(G_n)^2} = 0;$$
- **power law degree distribution**: for large $k \geq 1$ the proportion of vertices of degree k in G_n, written $p_k(G_n)$, satisfies
$$p_k(G_n) \sim k^{-\gamma} \quad \text{as } n\to\infty,$$
for some $\gamma > 1$.

Fact: Vertex-exchangeable models cannot replicate either of these behaviors. (Aldous–Hoover)

EDGE EXCHANGEABILITY

Let ν be a probability distribution on

$$\Delta^{\downarrow} = \{(f_{i,j})_{j\geq i\geq 1} : f_{i,j} \geq 0 \text{ and } \sum_{j\geq i\geq 1}^{\infty} f_{i,j} = 1\}.$$

Generate edges of network by taking $f \sim \nu$ and, given f, letting $X_1, X_2, \ldots$ be conditionally i.i.d.

$$P(X_k = \{i,j\} \mid f) = f_{i,j}, \quad j \geq i \geq 1. \tag{1}$$

For example: $X_1 = \{2,4\}$, $X_2 = \{1,2\}$, $X_3 = \{1,3\}$, $X_4 = \{5,6\}$, $X_5 = \{2,6\}$, $X_6 = \{2,6\}$

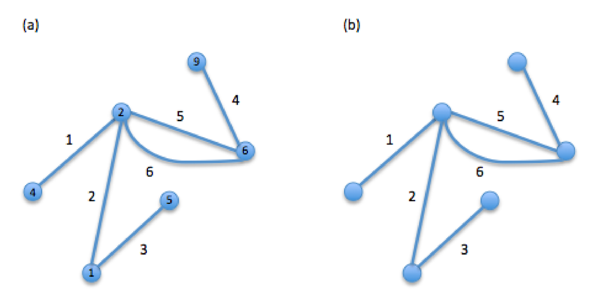

Figure 4: (a) Network with labeled vertices and edges. (b) Edge exchangeable network after removing vertex labels.

Theorem 1: Every edge exchangeable network can be constructed as in (1) for some ν .

HOLLYWOOD MODEL

Let (α, θ) satisfy $0 < \alpha < 1$ and $\theta > -\alpha$. Generate a sequence of edges $X_1, X_2, \ldots$ sequentially by:

$$\mathrm{pr}(X_{n,j} = i \mid H_{n,j}) \propto \propto \begin{cases} D_{n,j}(i) - \alpha, & i = 1, \ldots, V_n(j), \\ \theta + \alpha V_n(j), & i = V_n(j) + 1, \end{cases}$$

where $X_n = (X_{n,1}, X_{n,2})$ are the vertices in the nth edge.

The **Hollywood model** has a closed form expression for random edge-labeled networks of each finite size $n \geq 1$ given by

$$\mathrm{pr}(\mathcal{Y}_n = \mathcal{E}; \alpha, \theta) = \alpha^{v(\mathcal{E})} \frac{(\theta/\alpha)^{\uparrow v(\mathcal{E})}}{\theta^{\uparrow(2n)}} \prod_{k=2}^{\infty} ((1-\alpha)^{\uparrow(k-1)})^{N_k(\mathcal{E})}$$

where $x^{\uparrow j} = x(x+1)\cdots(x+j-1)$ is the ascending factorial function, $v(\mathcal{E})$ is the number of vertices, and $N_k(\mathcal{E})$ is the number of vertices of degree k.

PROPERTIES OF HOLLYWOOD

Theorem 2: The Hollywood model is edge exchangeable for all (α, θ, ν).

Theorem 3: For each $n \geq 1$, let $p_n(k) = N_k(\mathcal{Y}_n)/v(\mathcal{Y}_n)$, $k \geq 1$, be the empirical degree distribution of $\mathcal{Y}_n$. Then, for every $k \geq 1$,

$$p_n(k) \sim \alpha k^{-(\alpha+1)}/\Gamma(1-\alpha) \quad \text{a.s.} \quad \text{as } n \to \infty,$$

where $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ is the gamma function. That is, $(\mathcal{Y}_n)_{n\geq 1}$ has a power law degree distribution with exponent $\gamma = 1 + \alpha \in (1,2)$.

Theorem 4: The expected number of vertices satisfies

$$E(v(\mathcal{Y}_n)) \sim \frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)}(2n)^{\alpha} \quad \text{as } n \to \infty.$$

Furthermore, if $1/2 < \alpha < 1$, then the network is almost surely **sparse**.

Applications: The model also fits well to real network data. See article for more details.

VERTEX COMPONENTS MODEL

Construct $f = (f_{ij})_{i,j\geq 1}$ from random sequence $W = (W_i)_{i\geq 1}$ in infinite simplex

$$\Delta_1 = \{(s_1, s_2, \ldots) : \sum_{i\geq 1} s_i = 1\}$$

by putting

$$f_{ij} = W_i W_j, \quad i,j \geq 1. \tag{2}$$

Stick-breaking: We can generate the sequence $X_1, X_2, \ldots$ at the same time as $W = (W_i)_{i\geq 1}$.

- Let $\{\varphi_i\}_{i\geq 1}$ be a collection of probability densities on $[0,1]$.
- Put $X_{1,1} = 1$ and sample $W_1 \sim \varphi_1$.
- For $n = 1, 2, \ldots$, given $X_1, \ldots, X_n$ and $W_1, \ldots, W_{V_n}$, where V_n is the largest vertex label assigned so far, choose next vertex $X_{n+1,k}$ $(k = 1,2)$ by
$$\mathrm{pr}(X_{n+1,k} = r \mid W_1, \ldots, W_{V_n}) = \begin{cases} W_r, & r = 1, \ldots, V_n, \\ 1 - \sum_{j=1}^{V_n} W_j, & r = V_n + 1. \end{cases}$$
- If $X_{n+1,k} = V_n + 1$, then we choose $W_{V_n+1} \sim \varphi_{n+1}(\cdot/(1 - \sum_{j=1}^{V_n} W_j))$.

Corollary 5: Hollywood model with parameter (α, θ) corresponds to (2) with W from Poisson–Dirichlet(α, θ).

REFERENCES

Main reference:

- **H. Crane** and W. Dempsey. (2016). Edge exchangeable models for network data.

Other references:

- **H. Crane** and W. Dempsey. (2015). A framework for statistical network modeling.
- **H. Crane**. (2016). The ubiquitous Ewens sampling formula (with comments and a rejoinder by the author). *Statistical Science*, **31**(1):1–39.

Email: hcrane@stat.rutgers.edu

Web: www.harrycrane.com