Statistical network modeling: challenges and perspectives

Harry Crane

Department of Statistics
Rutgers

August 1, 2017

Statistical network modeling: challenges and perspectives

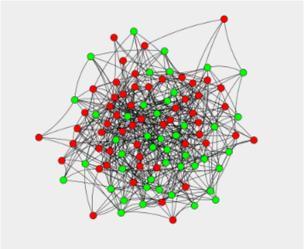
Harry Crane

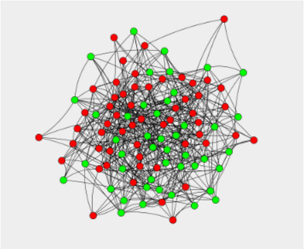
Department of Statistics
Rutgers

August 1, 2017

Subject of upcoming book on network modeling (Summer 2018)
Based on joint work with Walter Dempsey (Harvard)
$\rightarrow$ Will speak Thursday, August 4 at 11:35 AM in Room CC-309.

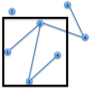
Modeling network data: What is it? Where does it come from?



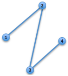


Assumed Setting:

Population



Observed network (sample)



Guiding Question:

How to model network data in the presence of sampling?

Key Points and Further Reading

Important directions of future statistics research:

1. *Network sampling*
2. *Network dynamics*

* mostly ignored in the current statistics literature

References:

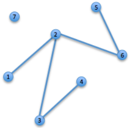
- H. Crane and W. Dempsey. (2015). A framework for statistical network modeling. arXiv:1509.08185.
- H. Crane and W. Dempsey. (2016). Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*.
- H. Crane. Principles of Statistical Network Modeling (tentative title). Forthcoming, 2018.
- H. Crane. (2017). Exchangeable graph-valued Feller processes. *Probability Theory and Related Fields*.
- H. Crane. (2017). Combinatorial Lévy processes. *Annals of Applied Probability*.

Available at **www.harrycrane.com**

Conventional Definition:

A (parameterized) statistical model is a family of probability distributions $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, each defined on the sample space.

- Population or Sample model? And what's the connection?

	Population	Observed network (sample)
		???
Model	$\{P_\theta : \theta \in \Theta\}$	???

- **Problem**: How to draw sound inferences about population model based on sampled network?
- Need to model data in a manner consistent with
  - (i) population model and
  - (ii) sampling mechanism.

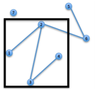
Revised Definition:

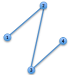
A (parameterized) statistical model is

(i) *a family of probability distributions $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, each defined on the sample space, along with*

(ii) *a description of the sampling scheme $(\mathbf{S}_{n,N})_{1 \leq n \leq N}$ used in observing a sample of each size n.*

Population	Observed network (sample)
$\mathbf{Y}_N$	$\mathbf{S}_{n,N}(\mathbf{Y}_N)$





Model $\{P_\theta : \theta \in \Theta\}$???

- Sampling scheme $\mathbf{S}_{m,n}$ necessary to establish relationship between observation and population.

Sampling mechanism often (almost always) left out of model specification.

Preview

Will discuss 3 network modeling scenarios:

1. Scenario 1: Vertex sampling
   - ERGMs, graphon models, Aldous–Hoover theorem, vertex exchangeability
2. Scenario 2: Edge sampling
   - edge exchangeability, Hollywood model
3. Scenario 3: Time-varying networks
   - Graph-valued Markov chains
   - Modeling network dynamics under sampling

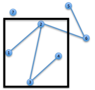
Scenario 1: Inference from sampled networks

Only n of all N Facebook account users are sampled and their relationships to one another are recorded as an array $\mathbf{Y}_n = (Y_{ij})_{1 \leq i,j \leq n}$ with

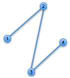
$$Y_{ij} = \begin{cases} 1, & i \text{ and } j \text{ Facebook friends}, \\ 0, & \text{otherwise}. \end{cases}$$

We are interested in using information $\mathbf{Y}_n$ to infer structure of entire Facebook network.

Population
$\mathbf{Y}_N$



Sample
$\mathbf{Y}_n$



How should $\mathbf{Y}_n$ be modeled if we are interested in inferring the structure of relationships among all students in the population $\mathbf{Y}_N$?

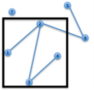
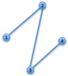
$\rightarrow$ Depends on how $\mathbf{Y}_n$ was sampled from $\mathbf{Y}_N$.

Given any sufficient statistics $(T_1, \dots, T_k)$ and parameters $(\theta_1, \dots, \theta_k)$, assign probability

$$\Pr(\mathbf{Y} = \mathbf{y}; \theta, T) \propto \exp\left\{\sum_{i=1}^{k} \theta_i T_i(\mathbf{y})\right\}, \quad \mathbf{y} = (y_{ij})_{1 \le i,j \le N} \in \{0,1\}^{N \times N}.$$

Holland and Leinhardt (1981), Frank and Strauss (1986), Wasserman and Pattison (1996), Wasserman and Faust (1994).

- **Typical approach**: Estimate θ by fitting ERGM (θ) to $\mathbf{Y}_n$, obtain $\hat{\theta}_n$ and use as estimate for θ in population.

	Population $\mathbf{Y}_N$	**Sample** $\mathbf{Y}_n$
		
Model	ERGM (θ)	???
Parameter	θ	θ

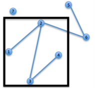
Consistency under subsampling for ERGM

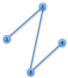
Theorem (Shalizi–Rinaldo)

Model for $\mathbf{S}_{n,N}(\mathbf{Y}_n)$ *is ERGM* (θ) *if and only if sufficient statistics* T *have* separable increments*.*

- separable increments: in-degree, out-degree, reciprocity
- non-separable increments: transitivity (triangles), most higher order statistics

Unclear how to proceed with inference for θ *based on sample.*





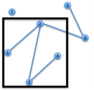
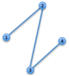
	Population $\mathbf{Y}_N$	**Sample** $\mathbf{Y}_n$
Model	ERGM (θ)	???
Parameter	θ	θ
Estimate	???	$\hat{\theta}_n$

Originally appeared as ϕ-processes in Diaconis–Freedman (1980), also Aldous (1983).

- Graphon $\phi : [0,1] \times [0,1] \to [0,1]$
- Associate i.i.d. sequence $U_1, U_2, \ldots$ of Uniform$[0,1]$ random variables to vertices $1, 2, \ldots$.
- Construct $\mathbf{Y}_N = (Y_{ij})_{1 \le i,j \le N}$ by

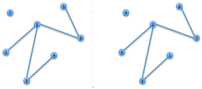
$$\Pr(Y_{ij} = 1 \mid U_1, U_2, \ldots) = \phi(U_i, U_j) \tag{1}$$

conditionally independently for all $1 \le i,j \le N$.

	Population $\mathbf{Y}_N$	**Sample** $\mathbf{Y}_n$
		
Model	graphon (ϕ)	graphon (ϕ)
Parameter	ϕ	ϕ
Estimate	$\hat{\phi}_n$	$\hat{\phi}_n$

Implications of Graphon models

Vertex exchangeable: Assign same probability to



Theorem (Aldous–Hoover)

Every vertex exchangeable model for countable population is a mixture of graphon models.

- Vertex-exchangeability $\Rightarrow$ sampled vertices representative of the population of all vertices $\Rightarrow$ not realistic in most applications.
- Cannot explain sparse behavior.

Graphon models achieve consistency under subsampling, but vertex exchangeability implies unrealistic sampling scheme and network properties.

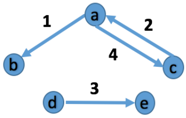
Scenario 2: Phone calls from a database

Entries are sampled uniformly at random from a large database of phone calls (or emails). Each observation (C_i, R_i) contains identity of the caller C_i and receiver R_i on the ith sampled call.

Interested in inferring the structure of connections among users in the database.

Caller	**Receiver**	**Time of Call**	...
555-7892 (a)	555-1243 (b)	15:34	...
550-9999 (c)	555-7892 (a)	15:38	...
555-1200 (d)	445-1234 (e)	16:01	...
555-7892 (c)	550-9999 (a)	15:38	...
⋮	⋮	⋮	⋱

Call sequence $X_1 = (a, b)$, $X_2 = (c, a)$, $X_3 = (d, e)$, $X_4 = (a, c)$ induces network:



Interaction Networks

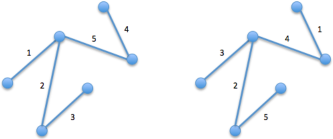
Dataset	vertices	**edges**
Actor collaborations	actors	**movies**
Enron email corpus	employees	**emails**
Karate club dataset	club members	**social interactions**
Wikipedia voting	Wikipedia admin.	**votes**
US Airport	airports	**flights**
Scientific collaborations	scientists	**articles**
UC Irvine online community	members	**online messages**
Political blogs	Websites	**hyperlinks**

- These datasets are driven by interactions
- Edges are the units — not represented as a (vertex-labeled) graph

Edge exchangeable models

- Phone calls are sampled uniformly from the database $\Rightarrow$ exchangeable sequence of pairs $(C_1, R_1), (C_2, R_2), \ldots$.

- Assign same probability to



Edge exchangeability $\Longleftrightarrow$ Size-biased vertex sampling

Hollywood model:

- Easy for estimation, prediction, and testing questions.
- Sparse with probability 1 for $1/2 < \alpha < 1$.
- Power law with exponent $\alpha + 1$ for $0 < \alpha < 1$.

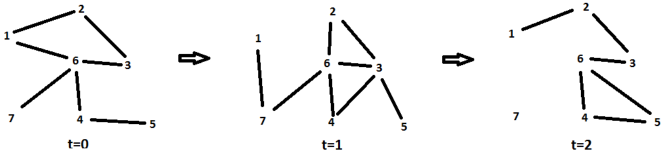
H. Crane and W. Dempsey. (2016). Edge exchangeable models for interaction networks. Journal of the American Statistical Association, *to appear.*

Scenario 3: Dynamics in social networks

For each day $t = 0, 1, 2, \ldots$ an array $\mathbf{Y}(t) = (Y_{ij}(t))_{1 \leq i,j \leq N}$ records interactions among a sample of Twitter users with

$$Y_{ij}(t) = \begin{cases} 1, & i \text{ retweeted or liked a tweet by } j, \\ 0, & \text{otherwise.} \end{cases}$$

Interested in modeling how past interaction behavior may be indicative of future behavior for the population of all Twitter users.



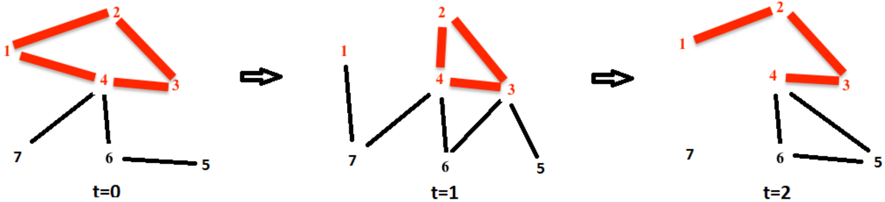
What kinds of Markov models can be fit to this data?

Scenario 3: Dynamics in social networks

For each day $t = 0, 1, 2, \ldots$ an array $\mathbf{Y}(t) = (Y_{ij}(t))_{1 \leq i,j \leq N}$ records interactions among a sample of Twitter users with

$$Y_{ij}(t) = \begin{cases} 1, & i \text{ retweeted or liked a tweet by } j, \\ 0, & \text{otherwise.} \end{cases}$$

Interested in modeling how past interaction behavior may be indicative of future behavior for the population of all Twitter users.



What kinds of Markov models can be fit to this data?

Illustration: TERGM

- θ: parameter vector
- $\eta(\theta)$: natural parameter
- $g(G, G')$: sufficient statistic for transition from G to G'

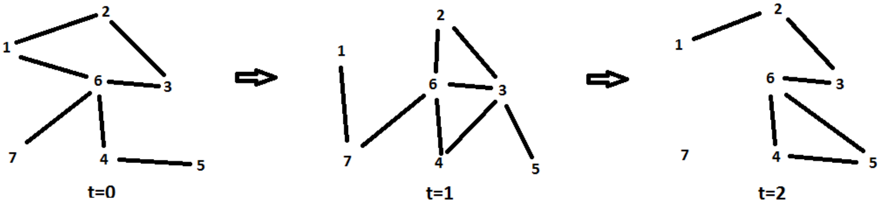
$$\mathbb{P}\{\Gamma_{t+1} = G' \mid \Gamma_t = G; \theta\} \propto \exp\{\eta(\theta) \cdot g(G, G')\}.$$

Observations:

- Typically choose g so that $g(G, G') = g(G^{\sigma}, G'^{\sigma})$ for arbitrary permutations $\sigma : [n] \to [n]$.
- Not projective in general (Shalizi & Rinaldo, 2013)
- Robins and Pattison (2001), Hanneke, Fu and Xing (2010), Kravitsky & Handcock (2012)

What do exchangeability and projective Markov properties imply about transition behavior?

Time-varying network models

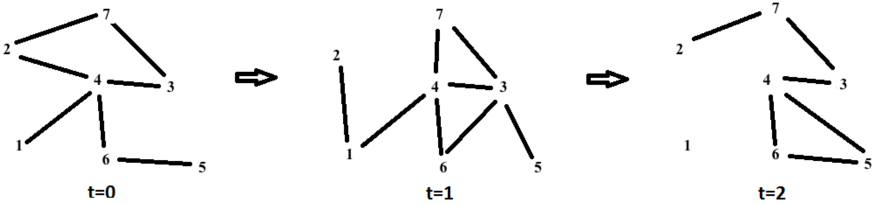


Basic assumptions:

$(\mathbf{Y}(t))_{t\geq 0}$ *is a Markov process satisfying*

- exchangeability*: for all* $\sigma : \mathbb{N} \to \mathbb{N}$, $(\mathbf{Y}^{\sigma}(t))_{t\geq 0}$ *has same finite-dimensional distributions as* $\mathbf{Y}$*.*
- projective Markov property*:* $(\mathbf{Y}(t)|_{[n]})_{t\geq 0}$ *is a Markov chain for every* $n = 1, 2, \ldots$*.*

Time-varying network models

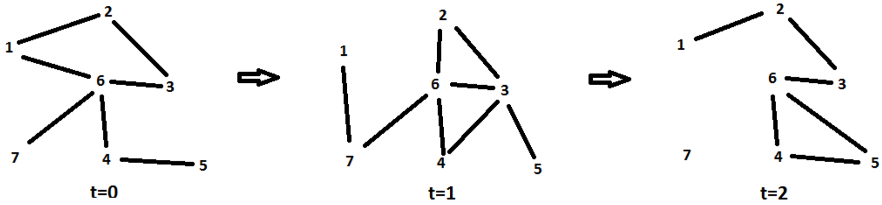


Basic assumptions:

$(\mathbf{Y}(t))_{t\geq 0}$ *is a Markov process satisfying*

- ***exchangeability*: for all $\sigma : \mathbb{N} \rightarrow \mathbb{N}$, $(\mathbf{Y}^{\sigma}(t))_{t\geq 0}$ has same finite-dimensional distributions as Y.**
- projective Markov property*:* $(\mathbf{Y}(t)|_{[n]})_{t\geq 0}$ *is a Markov chain for every* $n = 1, 2, \ldots$.

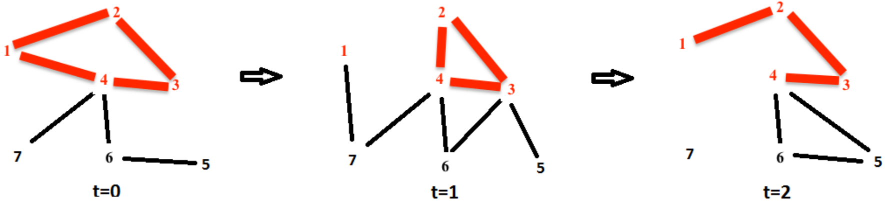
Time-varying network models



Basic assumptions:

$(\mathbf{Y}(t))_{t \geq 0}$ *is a Markov process satisfying*

- ***exchangeability*: for all $\sigma : \mathbb{N} \to \mathbb{N}$, $(\mathbf{Y}^{\sigma}(t))_{t \geq 0}$ has same finite-dimensional distributions as Y.**
- projective Markov property*: $(\mathbf{Y}(t)|_{[n]})_{t \geq 0}$ is a Markov chain for every $n = 1, 2, \ldots$.*



Basic assumptions:

$(\mathbf{Y}(t))_{t\geq 0}$ *is a Markov process satisfying*

- exchangeability*: for all* $\sigma : \mathbb{N} \to \mathbb{N}$*,* $(\mathbf{Y}^{\sigma}(t))_{t\geq 0}$ *has same finite-dimensional distributions as* $\mathbf{Y}$*.*
- ***projective Markov property*****:** $(\mathbf{Y}(t)|_{[n]})_{t\geq 0}$ **is a Markov chain for every** $n = 1, 2, \ldots$**.**

Rewiring processes (Crane, 2015, 2017)

Let $\phi : [0,1] \times [0,1] \to [0,1] \times [0,1]$ be a "Markovian graphon":

Write $\phi(u,v) = (\phi_0(u,v), \phi_1(u,v))$ for $(u,v) \in [0,1] \times [0,1]$.

Rewiring process:
Make a transition $G \mapsto G'$ by generating a transition probability for each pair of vertices:

$$\begin{pmatrix} 1-\phi_0(U_i,U_j) & \phi_0(U_i,U_j) \\ 1-\phi_1(U_i,U_j) & \phi_1(U_i,U_j) \end{pmatrix}$$

"Rewire" every edge according to transition probability determined by $\phi(U_i, U_j)$.

Theorem (Crane (2017))

Every exchangeable, projective Markov chain is a rewiring process.

- *H. Crane. (2015). Time-varying network models.* Bernoulli.
- *H. Crane. (2017). Exchangeable graph-valued Feller processes.* Probability Theory and Related Fields.
- *H. Crane. (2017+). Combinatorial Lévy processes.* Annals of Applied Probability.

High Level Takeaways

sampling	$\rightarrow$	*statistical units*	$\rightarrow$	*network rep.*	$\rightarrow$	*model*
selection		*vertices*		*graph*		*ERGM*
edge sampling		*edges*		*edge-labeled network*		*SBM*
snowball		*ego-network*				*graphon*
traceroute		*paths*				
$\vdots$		$\vdots$				$\vdots$

1. Sampling scheme a key part of modeling.

Sampling scheme often depends on the network itself.

2. Edge sampling, snowball sampling, traceroute sampling all depends on network

Sampling affects invariance principles of model $\rightarrow$ affects inference.

3. Vertex exchangeability, edge exchangeability, relative exchangeability, relational exchangeability

How are network dynamics affected by sampling? Wide open.

Further reading

Network Analysis:

- E. Kolaczyk. (2009). *Statistical Analysis of Network Data: Methods and Models*.

Network Modeling:

- H. Crane and W. Dempsey. (2015). A framework for statistical network modeling. arXiv:1509.08185.
- H. Crane. Principles of Statistical Network Modeling (tentative title). Forthcoming, 2018.
- Goldenberg, Zheng, Fienberg and Airoldi. (2009). A survey of statistical network models. arXiv:0912.5410.

Exchangeability:

- D.J. Aldous. (1983). *Exchangeability and Related Topics*.
- H. Crane and W. Dempsey. (2016). Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*.
- H. Crane and H. Towsner. (2015). Relative exchangeability.

Dynamic networks:

- H. Crane. (2015). Time-varying network models. *Bernoulli*.
- H. Crane. (2017). Exchangeable graph-valued Feller processes. *Probability Theory and Related Fields*.
- H. Crane. (2017+). Combinatorial Lévy processes. *Annals of Applied Probability*.