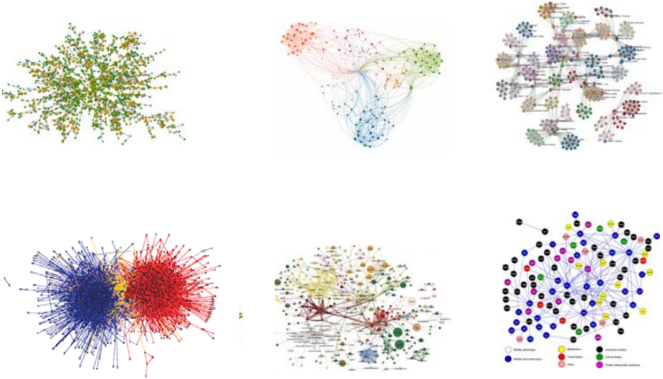
Dense graph limits under respondent-driven sampling

Siva Athreya

Indian Statistical Institute, Bangalore, India

Results: Joint work with Adrian Röllin.

Types of network



Food

Political

Social

Linkedin

Protien

Professional

Source: IUPUI Network Sampling course

Techniques to Analyse Networks

Network Characteristics:

- Distribution: degree, clustering coefficient, hop-plot.
- Diameter of network, max-k-core.
- singular values of the adjacency matrix.
- Subgraph counts

Sampling from Network:

- Uniform selection of nodes.
- Breadth-first or Depth first.
- Respondent driven sampling.

Hypothesis Testing:

- Random Graph models.
- Large Sample Theory.
- Estimation.
- Level of Significance.

Question of Interest

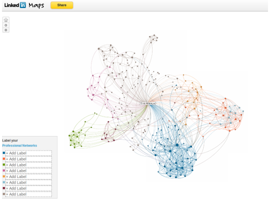
- Assume N to be very large. Let G_N be a graph on N vertices.
- We sample n vertices from N of them.
- we construct G_n a sub-graph of G_N.

Question: How close is G_n to G_N ?

Close: measure of distance, estimate bias, Testing.

Hypothesis Testing:

H_0: Graph is a linkedin network



H_1: Graph is a facebook network

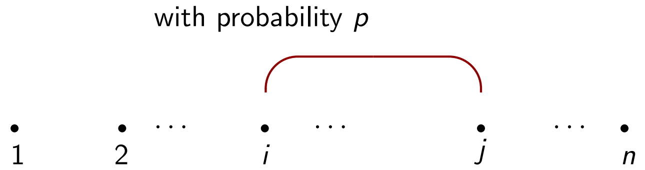
A Method: Respondent Driven Sampling

- Used to obtain estimates about properties of so-called "hidden" or "hard-to-reach" populations.
- Start with a convenient sample of participants.
- Ask the participants for *referrals* among their peers.
- Iterate this process. Use "backbone" to reconstruct the network.

Bias ? , Issues with under-estimating sample variance

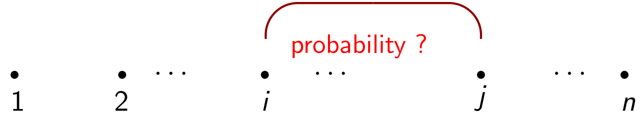
Reconstruct Examples: Erdös-Renyi $G(n, p)$.

- Consider vertex set $\{1, 2, \ldots, n\}$.
- connect an edge with probability p.



Reconstruct Examples: Two Groups.

How do you "reconstruct" a graph ?



δ between group connections	β within group 2 connections
α within group 1 connections	δ between group connections

Re-construction : Two groups

Vertices:

- Consider vertex set $\{1, 2, \ldots, n\}$.
- Assign Random labels $U_1, U_2, \ldots U_n$ from $[0, 1]$ to each vertex.

Edges:

connect two vertices i and j, with probability α or β or δ

independently of all the other edges.

Use labels $U_1, U_2, \ldots U_n$ to define groups

Reconstruction of finite graphs "many" groups from Graphon

Vertices:

- Consider vertex set $\{1, 2, \ldots, n\}$.
- Assign labels $U_1, U_2, \ldots U_n$ from $[0, 1]$ to each vertex.

Edges: Take a function $\kappa : [0, 1]^2 \to [0, 1]$ and

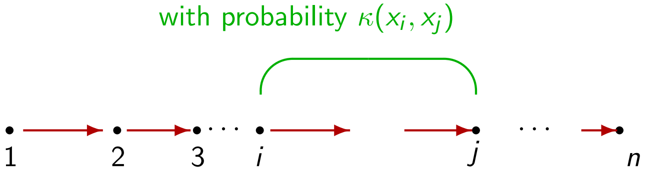
connect two vertices i and j, with probability $\kappa(U_i, U_j)$

independently of all the other edges.

Use U_i to define groups, κ to estimate the probability of connection among groups.

RDS: $G(x, H, \kappa)$

- Consider vertex set $\{1, 2, \ldots, n\}$ with labels $\{x_1, x_2, \ldots, x_n\}$.
- Given a referral backbone H.
- Add extra edges with probability $\kappa(x_i, x_j)$.



Limit Theorem: Statistical Analysis

Do above reconstruction using random labels.

Question: What happens for large n ?

Answers: Dense Graph Theory. Graphon.
Measure of Distance: Subgraph counts

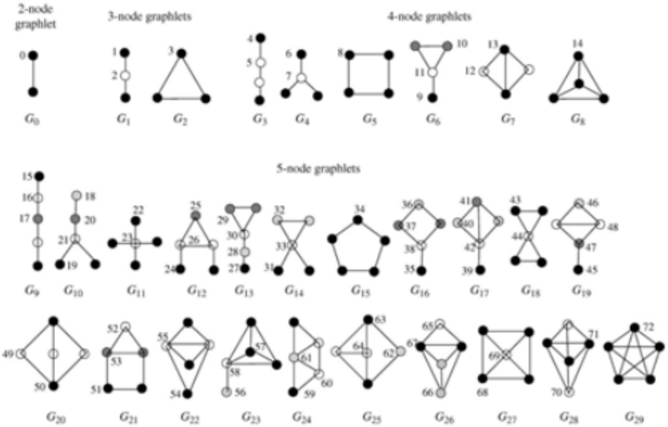
Limiting/Reconstructed graph: Transformed. What is transformation ?

Subgraph Density

- Let G be a graph on n vertices.
- Let F be a (small) graph on k vertices.
- Define subgraph density

$$t(F, G) = \frac{\text{copies of } F \text{ in } G}{\text{copies of } F \text{ in complete graph } K_n}.$$

Sub-graph count (fingerprints)



Source: IUPUI Network Sampling course

Aim and Result

- **Motivation:** provide a rigorous framework for Respondent Driven Sampling on dense graphs.
- **Theorem :** (in words)

  Limit of a dense graph sequence constructed via R.D.S., where the sequence of the vertex-sets is ergodic, can be expressed as a transformation of the original graph.

Result

Theorem: *Let* $X_n = (X_{n,1}, \dots, X_{n,n})$, $n \geq 1$, *be a triangular array of random variables taking values in* $[0,1]$. *Assume that there is a non-atomic probability measure* π *on* $[0,1]$ *such that, for all bounded, measurable functions* f,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(X_{n,i}) = \int_0^1 f(x)d\pi(x)$$

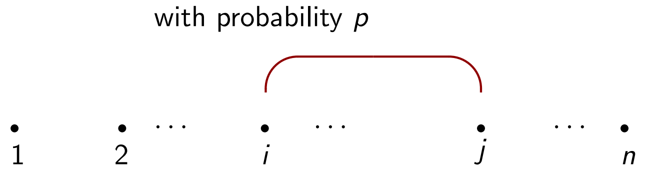
almost surely. Let $(H_n)_{n\geq 1}$ *be a sequence of graphs, where graph* H_n *has vertex set* $\{1,\dots,n\}$, *and* $\#\text{edges}(H_n) = o(n^2)$. *Then, with* $\tau(x) := \pi([0,x])$,

$$d_{sub}(G(X_n, H_n, \kappa), G(n, \kappa^\tau)) \to 0$$

almost surely.

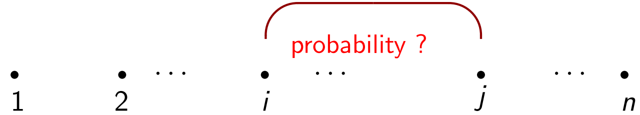
Reconstruct Examples: Erdös-Renyi $G(n,p)$.

- Consider vertex set $\{1, 2, \ldots, n\}$.
- connect an edge with probability p.



Reconstruct Examples: Two Groups.

How do you "reconstruct" a graph ?



δ between group connections	β within group 2 connections
α within group 1 connections	δ between group connections

Re-construction : Two groups

Vertices:

- Consider vertex set $\{1, 2, \ldots, n\}$.
- Assign Random labels $U_1, U_2, \ldots U_n$ from $[0, 1]$ to each vertex.

Edges:

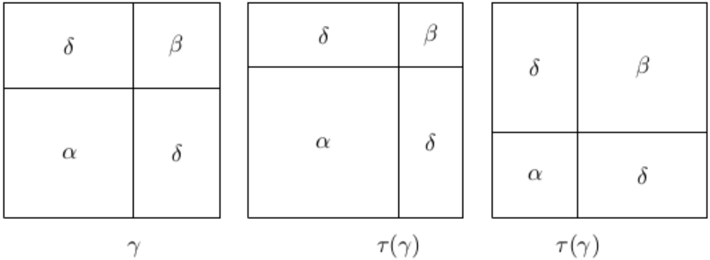
connect two vertices i and j, with probability α or β or δ

independently of all the other edges.

Use labels $U_1, U_2, \ldots U_n$ to define groups

κ and κ^τ

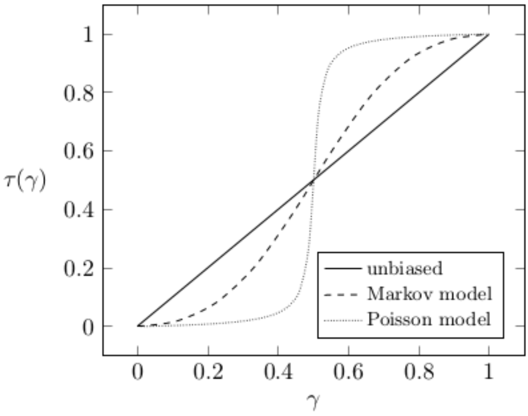
Let $\tau : [0,1] \to [0,1]$ and define $\kappa^\tau(x,y) = \kappa(\tau^{-1}(x), \tau^{-1}(y))$.



Labelling procedures from κ

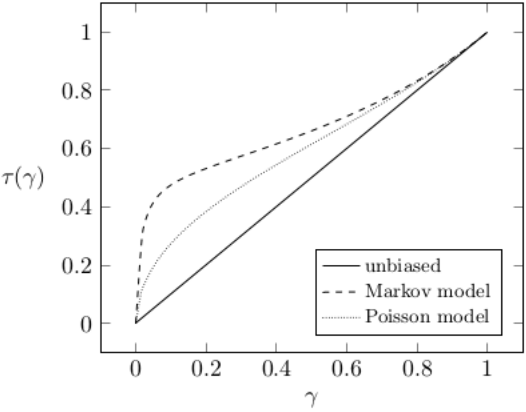
- take labels to be in $[0, 1]$.
- M: from label x_i to vertex i give label x_{i+1} to vertex $i+1$ proportional to $\kappa(x_{i+1}, dy)$.
- P: from label x_i to vertex i , get Poisson number of labels with intensity measure $\lambda\kappa(x_i, y)dy$

Bias: M vs P: $\kappa = \begin{array}{|c|c|}\hline 0.005 & 0.2 \\ \hline 0.2 & 0.005 \\ \hline\end{array}$

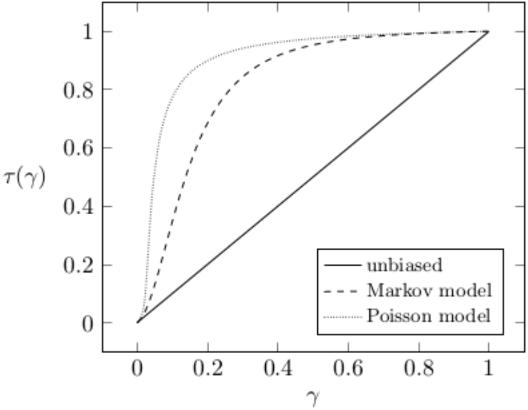


Bias: M vs P:

$$\kappa = \begin{array}{|c|c|} \hline 0.2 & 0.005 \\ \hline 0.2 & 0.2 \\ \hline \end{array}$$



Bias: M vs P: $\kappa = \begin{array}{|c|c|} \hline 0.005 & 0.005 \\ \hline 0.2 & 0.005 \\ \hline \end{array}$



Remarks

- Falls in Dense Graph Limits
- Similar approach for Sparse Graph limits.
- Very different approaches available for very sparse graphs.

Thank You.