Statistical Issues for Re-Identification in Network Data

Justin Vastola[1] Kobi Abayomi[1] Shawndra Hill[2]

1:Georgia Institute of Technology 2:University of Pennsylvania

Motivation

Repetitive Subscription Fraud

Telecommunications Industry

- Many people can't pay their bills, yet they still want telephone service.

Name	**Ted Hanley**
Address	**14 Pearl Dr** **St Peters, MN**
Balance	**$208.00**
Disconnected	**2/19/04 (nonpayment)**

Name	**Debra Handley**
Address	**14 Pearl Dr** **St Peters, MN**
Balance	**$142.00**
Connected	**2/22/04**

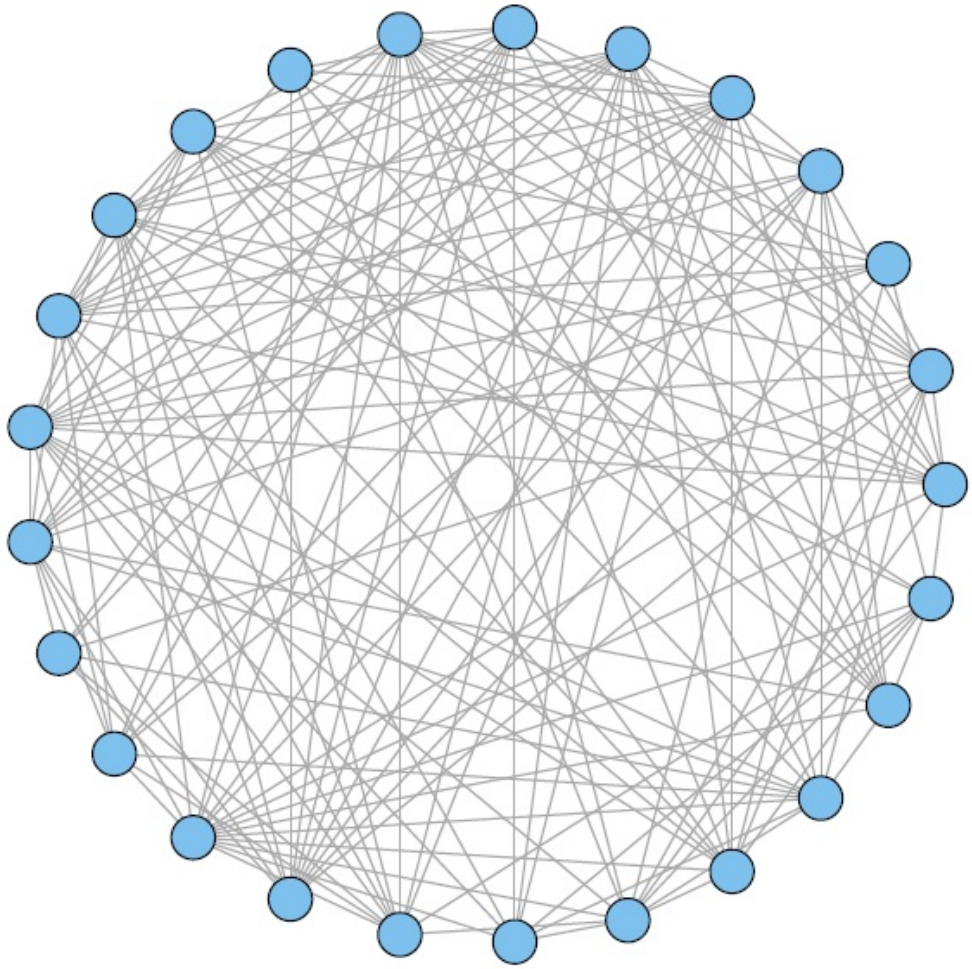
Name	**Elizabeth Harmon**
Address	**APT 1045** **4301 ST JOHN RD** **SCOTTSDALE, AZ**
Balance	**$149.00**
Disconnected	**2/19/04 (nonpayment)**

Name	**Elizabeth Harmon**
Address	**180 N 40TH PL** **APT 40** **PHOENIX, AZ**
Balance	**$72.00**
Connected	**1/31/04**

GOAL: Catch people who are taking part in fraudulent activities.

Motivation Continued

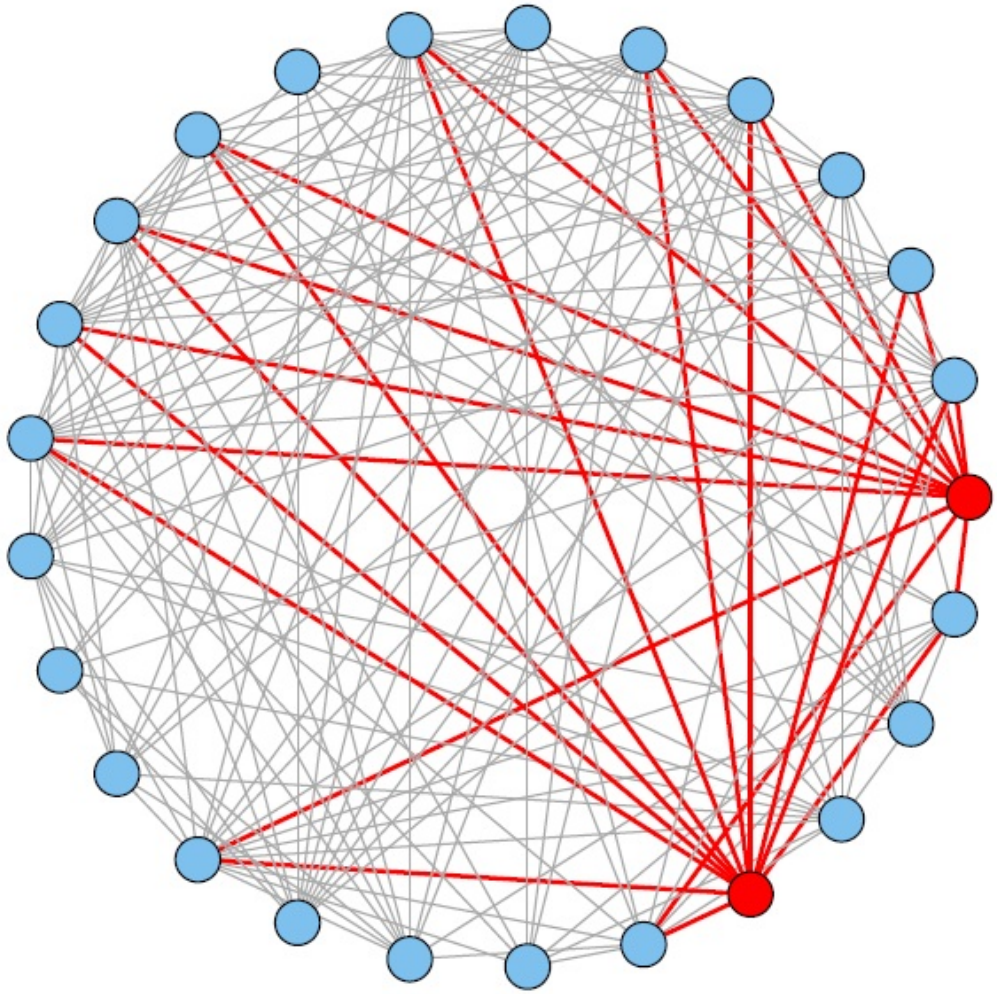
We observe a network with nodes representing phone users and edges representing a phone call between two uses.



Can you find the matching phone user?

Motivation Continued

The red nodes represent users who have the same call history.



Brief Background

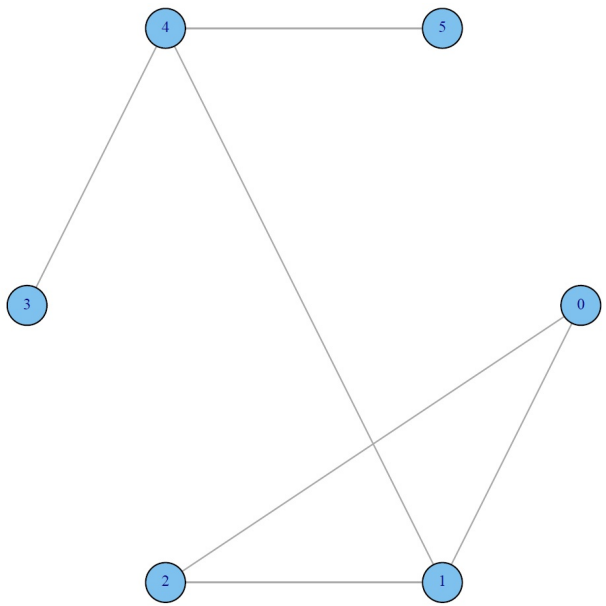
Networks

- A *network* $G = (V, E)$ is a mathematical structure consisting of *nodes* V and *edges* E.
- Often easier to consider a network based on an *adjacency matrix*, $A = [a_{ij}]_{i,j=1}^{n}$, where $a_{ij} = 1$ if an edge exists between nodes i and j and $a_{ij} = 0$ if no such edge exists.
- The degree of node i, denoted d_i, is the number of edges connected to i.

Brief Background

Networks

- A *network* $G = (V, E)$ is a mathematical structure consisting of *nodes* V and *edges* E.
- Often easier to consider a network based on an *adjacency matrix*, $A = [a_{ij}]_{i,j=1}^{n}$, where $a_{ij} = 1$ if an edge exists between nodes i and j and $a_{ij} = 0$ if no such edge exists.
- The degree of node i, denoted d_i, is the number of edges connected to i.



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$d_4 = 3$$

Brief Background Continued

Erdös-Rényi Networks proposed in 1959

- Each edge is place independently with probability p
- Known not to represent reality very well
- Very well studied

Brief Background Continued

Erdös-Rényi Networks proposed in 1959

- Each edge is place independently with probability p
- Known not to represent reality very well
- Very well studied

Watts-Strogatz Small-World Networks proposed in 1998

- Begin with a circle of nodes with each node is connected to its k neighbors on each side. Each edge is uniformly rewired with probability p.
- Incorporate high levels of clustering and short path length for $p \in [.001, .1]$

Brief Background Continued

Erdös-Rényi Networks proposed in 1959

- Each edge is place independently with probability p
- Known not to represent reality very well
- Very well studied

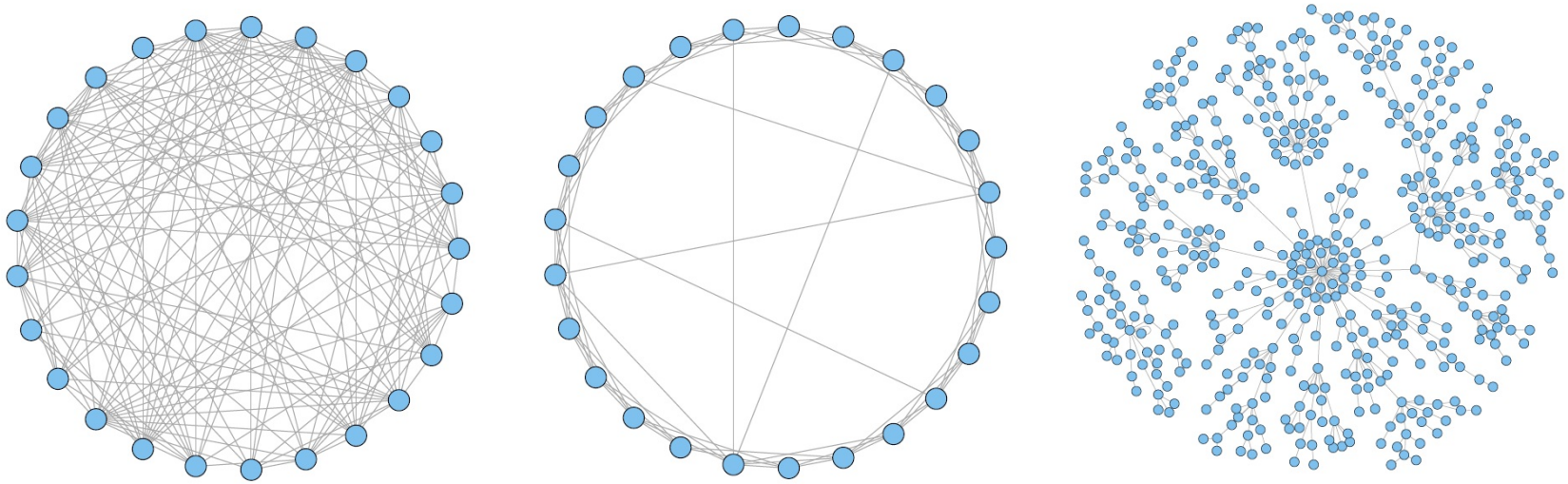
Watts-Strogatz Small-World Networks proposed in 1998

- Begin with a circle of nodes with each node is connected to its k neighbors on each side. Each edge is uniformly rewired with probability p.
- Incorporate high levels of clustering and short path length for $p \in [.001, .1]$

Barabási-Albert Scale-Free Networks proposed in 1999

- At each time step, a node enters the graph and connects to m nodes with probabilities proportional to node degree.
- The distribution of the degree decays slowly with $f_d \propto d^{-\alpha}$ for some constant α.
- Attribute scale-free phenomena to preferential attach and network growth

Brief Background Continued



Plots of Erdös-Rényi, Watts-Strogatz, and Barabási-Albert networks

Re-identification

- The *signature* of node i, denoted $\sigma(i)$, is the entity in which the node represents.
- As people interact with each other, they leave behind characteristic patterns of (a likely unique) behavior.
- The *re-identification problem* is the process of identifying two nodes have the same signature through this characteristic behavior.

Re-identification

- The *signature* of node i, denoted $\sigma(i)$, is the entity in which the node represents.
- As people interact with each other, they leave behind characteristic patterns of (a likely unique) behavior.
- The *re-identification problem* is the process of identifying two nodes have the same signature through this characteristic behavior.

Previous Work

- C. Cortes , D. Pregibon, and C. Volinsky. Computational Methods for Dynamic Graphs. Journal of Computational and Graphical Statistics, 12:950–970, 2003.
- S. Hill , D. K. Agarwal , R. Bell, C. Volinsky. Building an effective representation for dynamic networks. Journal of Computational and Graphical Statistics, 15(3):584–608, 2006.
- S. Hill and A. Nagle. Social Network Signatures: A Random Graph Approximation for Re-Identification and Experimental Results. In Proceedings of Computational Aspects of Social Networks, pp. 23–33, 2009.

Methodology

- Let G be an observed network from a family of networks $\mathcal{G}_\theta$, e.g., Erdös-Rényi, small world, or scale free networks, where θ is a parameter that completely characterizes $\mathcal{G}_\theta$.

Methodology

- Let G be an observed network from a family of networks $\mathcal{G}_\theta$, e.g., Erdös-Rényi, small world, or scale free networks, where θ is a parameter that completely characterizes $\mathcal{G}_\theta$.
- Let A be the corresponding adjacency matrix of G_θ.

Methodology

- Let G be an observed network from a family of networks $\mathcal{G}_\theta$, e.g., Erdös-Rényi, small world, or scale free networks, where θ is a parameter that completely characterizes $\mathcal{G}_\theta$.
- Let A be the corresponding adjacency matrix of G_θ.
- The *overlap score*, i.e., the number of neighbors that nodes i and j share is

$$S_\theta(i,j) = \langle \mathbf{a_i}, \mathbf{a_j} \rangle,$$

where $\langle \cdot, \cdot \rangle$ stands for the usual dot product and $\mathbf{a_i}$ stands for the i^{th} row of A. When $i \neq j$, we call S_θ the *non-match score*.

Methodology Continued

- Since the graph constructions are random, $S(i, j)$ is a random variable.

Methodology Continued

- Since the graph constructions are random, $S(i, j)$ is a random variable.
- Derive the distribution of $S(i, j)$, denoted F_θ, based on the algorithm in which the graph arises. In particular, $S(i, j) \sim F_\theta$.

Methodology Continued

- Since the graph constructions are random, $S(i,j)$ is a random variable.
- Derive the distribution of $S(i,j)$, denoted F_θ, based on the algorithm in which the graph arises. In particular, $S(i,j) \sim F_\theta$.
- For $i \neq j$, calculate $s(i,j)$. If this observed value is unusually large based on F_θ, we conclude that the *signatures* (identities) of i and j are the same.

Methodology Continued

- Since the graph constructions are random, $S(i,j)$ is a random variable.
- Derive the distribution of $S(i,j)$, denoted F_θ, based on the algorithm in which the graph arises. In particular, $S(i,j) \sim F_\theta$.
- For $i \neq j$, calculate $s(i,j)$. If this observed value is unusually large based on F_θ, we conclude that the *signatures* (identities) of i and j are the same.
- Formally, for all $i \neq j$, we perform the hypothesis test

$$H_0 : \sigma(i) \neq \sigma(j) \text{ vs. } H_1 : \sigma(i) = \sigma(j).$$

- We address the multiple hypothesis testing problem via controlling the false discovery rate (FDR).

Score Distributions: Erdös-Rényi Networks

To derive the *non-match score* distribution for nodes i and j,$i \neq j$, we consider the construction of the network.

Construction

1. Start with network consisting of n nodes and 0 edges.
2. For each pair of nodes (i, j), $i \neq j$, place an edge between them with probability p.
3. Once every pair of nodes is considered exactly once, the construction ends.

Score Distributions: Erdös-Rényi Networks Continued

Each scalar product in the *score*, i.e, dot product between two rows of A, can be viewed as a Bernoulli trial.

Score Distributions: Erdös-Rényi Networks Continued

Each scalar product in the *score*, i.e, dot product between two rows of A, can be viewed as a Bernoulli trial.

Example

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{aligned} s(1,2) &= \langle \mathbf{a_1}, \mathbf{a_2} \rangle \\ &= (0,1,1,0,0,0)(1,0,1,0,1,0)^T \\ &= 0 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 \\ &+ \quad 0 \cdot 0 + 0 \cdot 1 + 0 \cdot 0 \\ &= 1 \end{aligned}$$

Score Distributions: Erdös-Rényi Networks Continued

Each scalar product in the *score*, i.e, dot product between two rows of A, can be viewed as a Bernoulli trial.

Example

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{aligned} s(1,2) &= \langle \mathbf{a_1}, \mathbf{a_2} \rangle \\ &= (0,1,1,0,0,0)(1,0,1,0,1,0)^T \\ &= 0 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 \\ &+ \quad 0 \cdot 0 + 0 \cdot 1 + 0 \cdot 0 \\ &= 1 \end{aligned}$$

- The probability of success for each of these "trials" is

$$\mathbb{P}\{a_{i,i^*} \cdot a_{j,i^*} = 1\} = \mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\} = p^2$$

Score Distributions: Erdös-Rényi Networks Continued

Each scalar product in the *score*, i.e, dot product between two rows of A, can be viewed as a Bernoulli trial.

Example

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{aligned} s(1,2) &= \langle \mathbf{a_1}, \mathbf{a_2} \rangle \\ &= (0,1,1,0,0,0)(1,0,1,0,1,0)^T \\ &= 0 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 \\ &+ \quad 0 \cdot 0 + 0 \cdot 1 + 0 \cdot 0 \\ &= 1 \end{aligned}$$

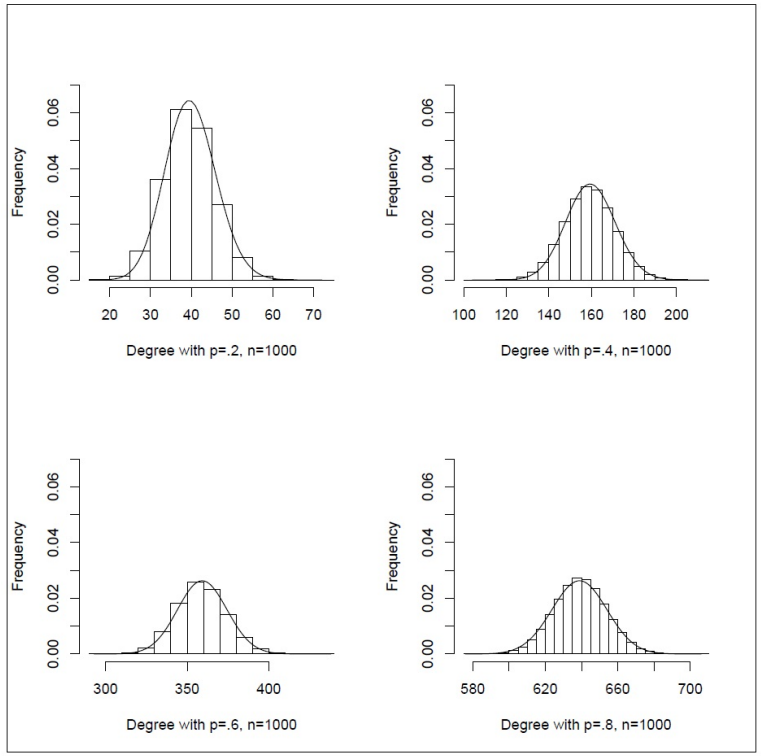
- The probability of success for each of these "trials" is

$$\mathbb{P}\{a_{i,i^*} \cdot a_{j,i^*} = 1\} = \mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\} = p^2$$

- The *non-match score* distribution is, therefore,

$$S_p(i,j) \sim \text{Bin}(n-2, p^2).$$

Score Distributions: Erdös-Rényi Networks Continued



Plots of non-match score distributions for Erdös-Rényi networks with varying p.

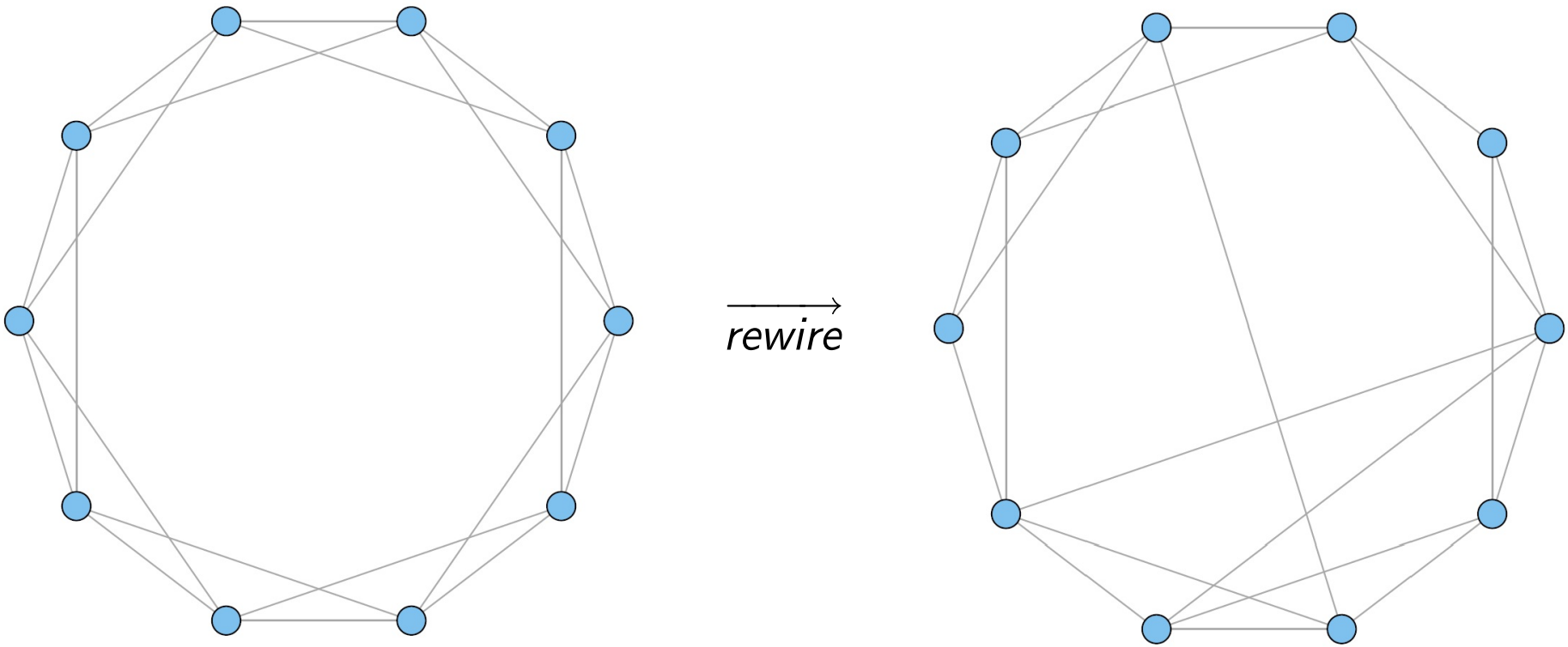
Score Distributions: Watts-Strogatz Small-world Networks

To derive the *non-match score* distribution for nodes i and j,$i \neq j$, we consider the construction of the network.

Construction

1. Start with circle of n nodes each connected to its k closest nodes on the left and right, giving a total of $2k$ edges for each node.
2. For each existing edge, decide to "rewire" it with probability p.
   - To rewire an edge, disconnect the edge from its right endpoint (node that lies clockwise). Then select one node based on a uniform distribution to reconnect to the right side of the edge to.
3. The algorithm ends when all original edges have been considered for rewiring exactly once.

Score Distributions: Watts-Strogatz Small-world Networks Continued



Score Distributions: Watts-Strogatz Small-world Networks Continued

To derive the *non-match score* distribution for nodes i and j,$i \neq j$, we need to consider three cases before rewiring.

1. nodes i and j are both unconnected to node i^*, i.e., $a_{i,i^*} = a_{j,i*} = 0$;
2. node i and j are both connected to node i^*, i.e., $a_{i,i^*} = a_{j,i^*} = 1$;
3. node i is connected to node i^* and node j is unconnected from i^*, i.e., $a_{i,i^*} = 1$, and $a_{j,i*} = 0$.

Score Distributions: Watts-Strogatz Small-world Networks Continued

To derive the *non-match score* distribution for nodes i and j,$i \neq j$, we need to consider three cases before rewiring.

1. nodes i and j are both unconnected to node i^*, i.e., $a_{i,i^*} = a_{j,i*} = 0$;
2. node i and j are both connected to node i^*, i.e., $a_{i,i^*} = a_{j,i^*} = 1$;
3. node i is connected to node i^* and node j is unconnected from i^*, i.e., $a_{i,i^*} = 1$, and $a_{j,i*} = 0$.

Example

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

For $i = 1$ and $j = 2$:

1. case 1 occurs 2 times ;
2. case 2 occurs 1 time;
3. case 3 occurs 3 times.

Score Distributions: Watts-Strogatz Small-world Networks Continued

The number of times each case occurs is:

$$n_1 = \begin{cases} n - 2k - ||i,j|| - 1, & \text{if } 1 \leq ||i,j|| \leq k \\ n - 2k - ||i,j|| + 1, & \text{if } k < ||i,j|| \leq 2k \\ n - 4k, & \text{if } ||i,j|| > 2k \\ 0, & \text{otherwise.} \end{cases}$$

$$n_2 = \begin{cases} 2k - ||i,j|| - 1, & \text{if } 1 \leq ||i,j|| \leq k \\ 2k - ||i,j|| + 1, & \text{if } k < ||i,j|| \leq 2k \\ 0, & \text{otherwise} \end{cases}$$

$$n_3 = \begin{cases} 2||i,j|| + 2, & \text{if } 1 \leq ||i,j|| \leq k \\ 2||i,j|| - 2, & \text{if } k < ||i,j|| \leq 2k \\ 4k, & \text{if } ||i,j|| > 2k \\ 0, & \text{otherwise.} \end{cases}$$

Note, each n_i depends on the distance between each pair of nodes before rewiring. The distance between nodes i and j is denoted $||i,j||$.

Score Distributions: Watts-Strogatz Small-world Networks Continued

- As before, each scalar product in the *score*, can be viewed as a Bernoulli trial were a success is defined by $a_{i,i^*} \cdot a_{j,i^*} = 1$, or $a_{i,i^*} = a_{j,i^*} = 1$
- We need to consider how a "success" arises for each case in order to derive the *non-match score* distribution.

Case 1

After the rewiring process, $a_{ii^*} = 1$ if

- one of the k edges considered for rewiring from node i is rewired to node i^*
- one of the k edges considered for rewiring from node i^* is rewired to node i

Case 1

After the rewiring process, $a_{ii^*} = 1$ if

- one of the k edges considered for rewiring from node i is rewired to node i^*
- one of the k edges considered for rewiring from node i^* is rewired to node i

The two cases combined can be described as a random variable $X \sim \text{Bin}(2k, p/n)$.

Case 1

After the rewiring process, $a_{ii^*} = 1$ if

- one of the k edges considered for rewiring from node i is rewired to node i^*
- one of the k edges considered for rewiring from node i^* is rewired to node i

The two cases combined can be described as a random variable $X \sim \text{Bin}(2k, p/n)$.

$$\mathbb{P}\{a_{i,i^*} = 1\} = \mathbb{P}\{X \geq 1\} = 1 - \mathbb{P}\{X = 0\} = 1 - (1 - p/n)^{2k}$$

$$f_1 \sim \text{Bin}(n_1, [1 - (1 - p/n)^{2k}]^2)$$

Case 2

After the rewiring process, $a_{ii^*} = 0$ if

- edge (i, i^*) is removed and not replaced while none of the $2k - 1$ edges remaining for rewiring connect nodes i and i^*.

Case 2

After the rewiring process, $a_{ii^*} = 0$ if

- edge (i, i^*) is removed and not replaced while none of the $2k - 1$ edges remaining for rewiring connect nodes i and i^*.

This event occurs with probability

$$\mathbb{P}\{a_{i,i^*} = 1\} = 1 - p(\frac{n-1}{n})(1 - p/n)^{2k-1}.$$

Case 2

After the rewiring process, $a_{ii^*} = 0$ if

- edge (i, i^*) is removed and not replaced while none of the $2k - 1$ edges remaining for rewiring connect nodes i and i^*.

This event occurs with probability

$$\mathbb{P}\{a_{i,i^*} = 1\} = 1 - p(\frac{n-1}{n})(1 - p/n)^{2k-1}.$$

$$f_2 \sim \text{Bin}(n_2, [1 - p(\tfrac{n-1}{n})(1 - p/n)^{2k-1}]^2)$$

Case 3

After the rewiring process, $a_{ii^*} = 1$ if the above two cases hold appropriately. Therefore,

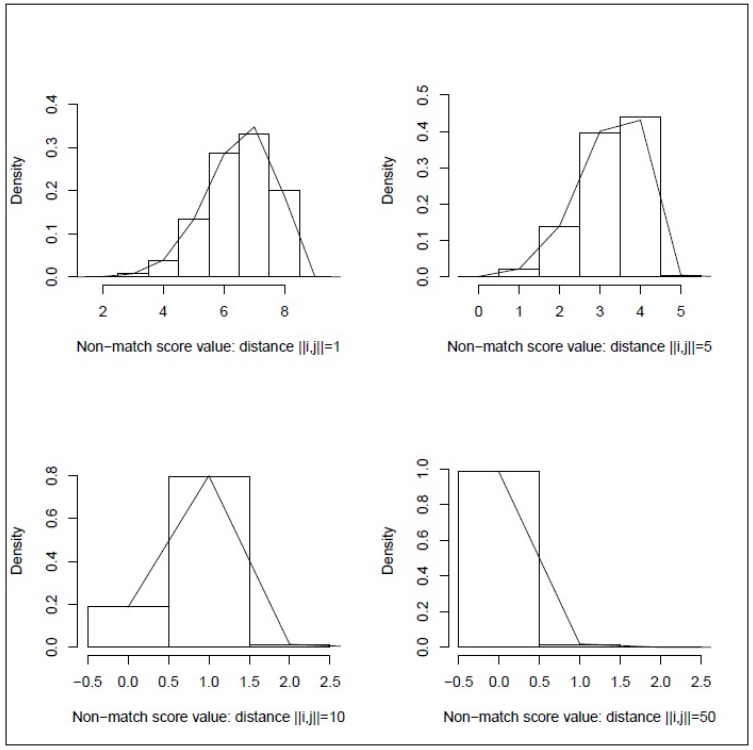
$$f_3 \sim \text{Bin}(n_3, [1 - p(\tfrac{n-1}{n})(1 - p/n)^{2k-1}][1 - (1 - p/n)^{2k}])$$

Score Distributions: Watts-Strogatz Small-world Networks Continued

The *non-match score* distribution is the convolution of random variables $X_1 \sim f_1, X_2 \sim f_2$, and $X_3 \sim f_3$,i.e.,

$$\begin{aligned}
\mathbb{P}\{S_\theta(i,j) = z\} &= \sum_{y=0}^{z} \binom{n_3}{z-y} p_3^{y-z}(1-p_3)^{n_3-(z-y)} \chi\{z-y \le n_3\} \\
&\times \sum_{x=0}^{y} \binom{n_1}{x} p_1^{x}(1-p_1)^{n_1-x} \chi\{x \le n_1\} \\
&\times \binom{n_2}{y-x} p_2^{y-x}(1-p_2)^{n_2-(y-x)} \chi\{y-x \le n_2\}
\end{aligned}$$

Score Distributions: Watts-Strogatz Networks Continued



Plots of non-match score distributions for Watts-Strogatz small-world networks

Score Distributions: Watts-Strogatz Networks Continued

- Typically, the graph will be unlabeled, however, the distributions we derived depend on the labeling of the network through the distances between nodes

Score Distributions: Watts-Strogatz Networks Continued

- Typically, the graph will be unlabeled, however, the distributions we derived depend on the labeling of the network through the distances between nodes
- To get around needing the labels, we view the data as coming from a mixture of all the *non-match score* distributions over all the possible distances between nodes.

Mixture Model

- Consider $s_1, \ldots, s_{\lfloor n/2 \rfloor}$, where s_i denotes the non-match score distribution for two nodes distance i apart.

Mixture Model

- Consider $s_1, \ldots, s_{\lfloor n/2 \rfloor}$, where s_i denotes the non-match score distribution for two nodes distance i apart.
- Let $\alpha_1, \ldots, \alpha_{\lfloor n/2 \rfloor} \in \mathbb{R}$ be the mixing parameters such that $0 \leq \alpha_i \leq 1$ for all i and $\sum_{i=1}^{\lfloor n/2 \rfloor} \alpha_i = 1$.

Mixture Model

- Consider $s_1, \ldots, s_{\lfloor n/2 \rfloor}$, where s_i denotes the non-match score distribution for two nodes distance i apart.
- Let $\alpha_1, \ldots, \alpha_{\lfloor n/2 \rfloor} \in \mathbb{R}$ be the mixing parameters such that $0 \leq \alpha_i \leq 1$ for all i and $\sum_{i=1}^{\lfloor n/2 \rfloor} \alpha_i = 1$.
- The mixture distribution is

$$f_{mix} = \sum_{i=1}^{\lfloor n/2 \rfloor} \alpha_i s_{||\cdot,\cdot||=i}.$$

Mixture Model

- Consider $s_1, \ldots, s_{\lfloor n/2 \rfloor}$, where s_i denotes the non-match score distribution for two nodes distance i apart.
- Let $\alpha_1, \ldots, \alpha_{\lfloor n/2 \rfloor} \in \mathbb{R}$ be the mixing parameters such that $0 \leq \alpha_i \leq 1$ for all i and $\sum_{i=1}^{\lfloor n/2 \rfloor} \alpha_i = 1$.
- The mixture distribution is

$$f_{mix} = \sum_{i=1}^{\lfloor n/2 \rfloor} \alpha_i s_{||\cdot,\cdot||=i}.$$

- The α_i's are actually known, so all we need to estimate is the parameters p and k as before.

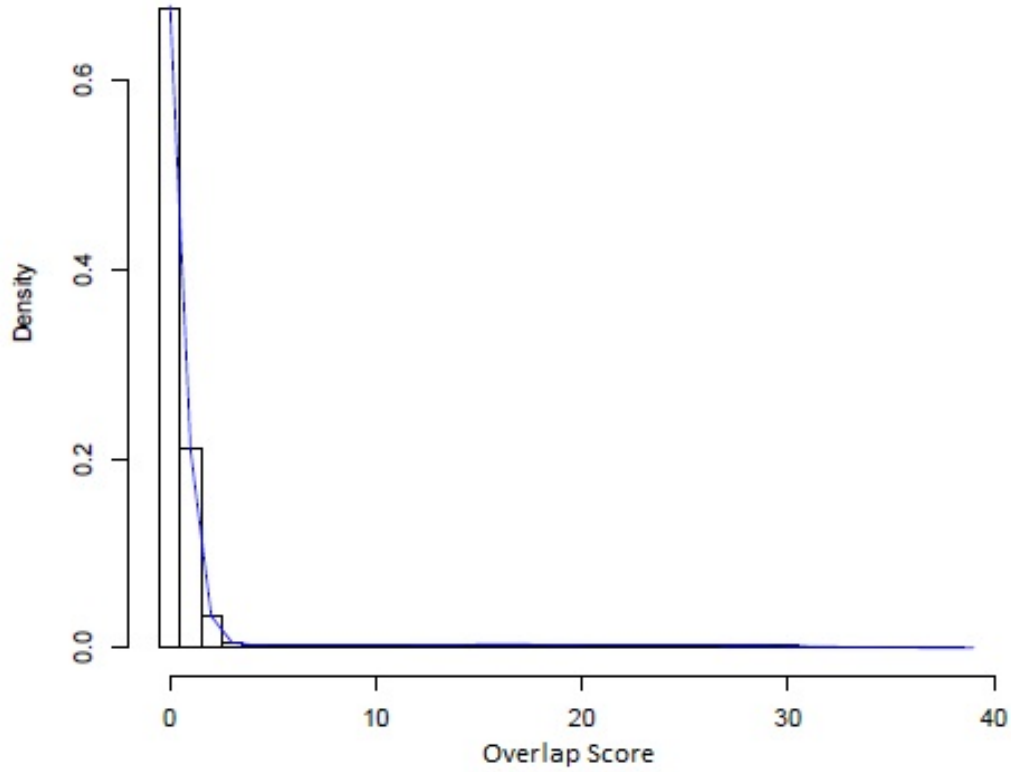
Mixture Model

- Consider $s_1, \ldots, s_{\lfloor n/2 \rfloor}$, where s_i denotes the non-match score distribution for two nodes distance i apart.
- Let $\alpha_1, \ldots, \alpha_{\lfloor n/2 \rfloor} \in \mathbb{R}$ be the mixing parameters such that $0 \leq \alpha_i \leq 1$ for all i and $\sum_{i=1}^{\lfloor n/2 \rfloor} \alpha_i = 1$.
- The mixture distribution is

$$f_{mix} = \sum_{i=1}^{\lfloor n/2 \rfloor} \alpha_i s_{||\cdot,\cdot||=i}.$$

- The α_i's are actually known, so all we need to estimate is the parameters p and k as before.
  - For instance, if n is even, $\alpha_i = \frac{2}{n-1}$ for $i = 1, \ldots, \lfloor n/2 \rfloor - 1$ and $\alpha_{\lfloor n/2 \rfloor} = \frac{1}{n-1}$.

Score Distributions: Watts-Strogatz Networks Continued



Plots of non-match score mixture distribution for Watts-Strogatz small-world networks

Score Distributions: Barabási-Albert Scale-free Networks

Barabási and Albert provide a method of constructing scale-free networks based on growth and preferential attachment, however, their description is imprecise. Bollobás and Riordan remedy this issue by precisely specifying the model of Barabási and Albert.

Construction

1. Start with an initial graph with one vertex and one loop.
2. Let $d_{n,i}$ denote the degree of node i when the size of the graph is n. At each time step add node n together with a single edge between nodes n and i, where i is randomly chosen with

$$\mathbb{P}(i = s) = \begin{cases} d_{n-1,s}/(2n-1), & 1 \leq s \leq n-1 \\ 1/(2n-1), & s = n. \end{cases}$$

3. Stop the algorithm once the desired number of nodes is reached.

Score Distributions: Barabási-Albert Scale-free Networks

Barabási and Albert provide a method of constructing scale-free networks based on growth and preferential attachment, however, their description is imprecise. Bollobás and Riordan remedy this issue by precisely specifying the model of Barabási and Albert.

Construction

1. Start with an initial graph with one vertex and one loop.
2. Let $d_{n,i}$ denote the degree of node i when the size of the graph is n. At each time step add node n together with a single edge between nodes n and i, where i is randomly chosen with

$$\mathbb{P}(i = s) = \begin{cases} d_{n-1,s}/(2n-1), & 1 \leq s \leq n-1 \\ 1/(2n-1), & s = n. \end{cases}$$

3. Stop the algorithm once the desired number of nodes is reached.

This network can be generalized to add m edges with the entrance of each new node.

Score Distributions: Barabási-Albert Scale-free Networks Continued

The two key elements of this construction are growth and preferential attachment.

Score Distributions: Barabási-Albert Scale-free Networks Continued

The two key elements of this construction are growth and preferential attachment.

- Growth: nodes are added to the network over time.

Score Distributions: Barabási-Albert Scale-free Networks Continued

The two key elements of this construction are growth and preferential attachment.

- Growth: nodes are added to the network over time.
- Preferential attachment: the probability that a new edge is attached to a node is proportional to the nodes degree.

Score Distributions: Barabási-Albert Scale-free Networks Continued

Preferential attachment introduces a dependence in the network which is not present in the Erdös-Rényi or Watts-Strogatz networks. In particular,

$$\mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\} \neq \mathbb{P}\{a_{i,i^*} = 1\}\mathbb{P}\{a_{j,i^*} = 1\}$$

Score Distributions: Barabási-Albert Scale-free Networks Continued

Preferential attachment introduces a dependence in the network which is not present in the Erdös-Rényi or Watts-Strogatz networks. In particular,

$$\mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\} \neq \mathbb{P}\{a_{i,i^*} = 1\}\mathbb{P}\{a_{j,i^*} = 1\}$$

Moreover, we must consider the probabilities that nodes i and j are both connected to node i^* in the following cases:

1. $i^* < i < j$
2. $i < i^* < j$
3. $i < j < i^*$

Preferential attachment introduces a dependence in the network which is not present in the Erdös-Rényi or Watts-Strogatz networks. In particular,

$$\mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\} \neq \mathbb{P}\{a_{i,i^*} = 1\}\mathbb{P}\{a_{j,i^*} = 1\}$$

Moreover, we must consider the probabilities that nodes i and j are both connected to node i^* in the following cases:

1. $i^* < i < j$
2. $i < i^* < j$
3. ~~$i < j < i^*$~~

We consider the case when $m = 1$, making the third case irrelevant.

Score Distributions: Barabási-Albert Scale-free Networks Continued

We condition on networks at previous time periods to derive the following.

Score Distributions: Barabási-Albert Scale-free Networks Continued

We condition on networks at previous time periods to derive the following.

Case 1

$$\begin{aligned}\mathbb{P}\{a_{i^*i}=1, a_{i^*j}=1\} &= \frac{4i^*+2}{(2j-1)(4(i^*)^2-1)}\prod_{s=i+1}^{j-1}\left(\frac{2s}{2s-1}\right) \\ &= \left(\frac{4i^*+2}{(2j-1)(4(i^*)^2-1)}\right)\left(\frac{4^{j-i-1}(j-1)!^2(2i)!}{(2j-2)!(i)!^2}\right)\end{aligned}$$

Case 2

$$\begin{aligned}\mathbb{P}\{a_{i^*i}=1, a_{i^*j}=1\} &= \frac{1}{(2i^*)(2j-1)}\prod_{s=i}^{j-1}\left(\frac{2s}{2s-1}\right) \\ &= \frac{1}{(2i^*)(2j-1)}\left(\frac{4^{j-i}(j-1)!^2(2i-2)!}{(2j-2)!(i-1)!^2}\right)\end{aligned}$$

Score Distributions: Barabási-Albert Scale-free Networks Continued

- Let $X_{i,j}^{i^*}$ denote the Bernoulli random variable with success probability

$$p_{i,j}^{i^*} := \mathbb{P}\left\{a_{i^*i} = 1, a_{i^*j} = 1\right\},$$

  i.e., the random variable representing whether or not nodes i and j are both connected to node i^*.

Score Distributions: Barabási-Albert Scale-free Networks Continued

- Let $X_{i,j}^{i^*}$ denote the Bernoulli random variable with success probability

$$p_{i,j}^{i^*} := \mathbb{P}\left\{a_{i^*i} = 1, a_{i^*j} = 1\right\},$$

  i.e., the random variable representing whether or not nodes i and j are both connected to node i^*.

- The non-match score distribution for specified nodes i and j is the convolution of random variables $X_{i,j}^{i^*}$ as i^* ranges from 1 to n,i.e.,

$$Z = \sum_{i^*=1}^{j-1} X_{i,j}^{i*}.$$

Score Distributions: Barabási-Albert Scale-free Networks Continued

- Let $X_{i,j}^{i^*}$ denote the Bernoulli random variable with success probability

$$p_{i,j}^{i^*} := \mathbb{P}\left\{a_{i^*i} = 1, a_{i^*j} = 1\right\},$$

  i.e., the random variable representing whether or not nodes i and j are both connected to node i^*.

- The non-match score distribution for specified nodes i and j is the convolution of random variables $X_{i,j}^{i^*}$ as i^* ranges from 1 to n,i.e.,

$$Z = \sum_{i^*=1}^{j-1} X_{i,j}^{i*}.$$

Unlike the previous network constructions, these Bernoulli random variables are dependent.

The dependence in the Bernoulli random variables is that the sum is restricted to be one. Thus, Z is a Bernoulli random variable with probability of a success

$$\mathbb{P}\{Z=1\} = \mathbb{E}\left[\sum_{i^*=1}^{j-1} X_{i,j}^{i^*}\right] = \sum_{i^*=1}^{j-1} p_{i,j}^{i^*}$$

The dependence in the Bernoulli random variables is that the sum is restricted to be one. Thus, Z is a Bernoulli random variable with probability of a success

$$\mathbb{P}\{Z=1\} = \mathbb{E}\left[\sum_{i^*=1}^{j-1} X_{i,j}^{i^*}\right] = \sum_{i^*=1}^{j-1} p_{i,j}^{i^*}$$

The *non-match score* distribution for nodes i and j is

$$\mathbb{P}\{S_1(i,j)=s\} = \begin{cases} 1-\sum_{i^*=1}^{j-1} p_{i,j}^{i^*}, & \text{if } s=0 \\ \sum_{i^*=1}^{j-1} p_{i,j}^{i^*}, & \text{if } s=1. \end{cases}$$

Estimation

- Problems arise in parameter estimation for networks due to the dependency in the data. For example, the likelihood of degrees is no longer the product of the observed marginal degree distributions.

Estimation

- Problems arise in parameter estimation for networks due to the dependency in the data. For example, the likelihood of degrees is no longer the product of the observed marginal degree distributions.
- Many types of parameter estimation in networks are derived from the sampling scheme in which the data was collected.

Estimation

- Problems arise in parameter estimation for networks due to the dependency in the data. For example, the likelihood of degrees is no longer the product of the observed marginal degree distributions.
- Many types of parameter estimation in networks are derived from the sampling scheme in which the data was collected.
- To get around this dependency and since we want the estimation to be valid no matter the sampling scheme, we use method of moments estimators.

Estimation: Erdös-Rényi Networks

- A family of Erdös-Rényi networks with known order n, along with the match and non-match scores distributions are completely characterized by parameter p.

Estimation: Erdös-Rényi Networks

- A family of Erdös-Rényi networks with known order n, along with the match and non-match scores distributions are completely characterized by parameter p.
- Let $\mathcal{E}$ denote the total number of edges in the observed network.

Estimation: Erdös-Rényi Networks

- A family of Erdös-Rényi networks with known order n, along with the match and non-match scores distributions are completely characterized by parameter p.
- Let $\mathcal{E}$ denote the total number of edges in the observed network.
- Since are $\binom{n}{2}$ possible edges each placed independently with probability p

$$\mathbb{E}[\mathcal{E}] = \frac{n(n-1)}{2} p.$$

Estimation: Erdös-Rényi Networks

- A family of Erdös-Rényi networks with known order n, along with the match and non-match scores distributions are completely characterized by parameter p.
- Let $\mathcal{E}$ denote the total number of edges in the observed network.
- Since are $\binom{n}{2}$ possible edges each placed independently with probability p

$$\mathbb{E}[\mathcal{E}] = \frac{n(n-1)}{2} p.$$

- Therefore, a moments estimator of p is

$$\hat{p} = \frac{2\mathcal{E}^{obs}}{n(n-1)},$$

where $\mathcal{E}^{obs}$ is the observed number of edges.

Estimation: Watts-Strogatz Small-world Networks

- For known n, the parameters p and k completely characterize this family of networks
- It was shown that $\mathbb{E}[\overline{d}] = 2k$, where $\overline{d}$ is the average degree. Thus,

$$\hat{k} = \frac{\overline{d}}{2}$$

Estimation: Watts-Strogatz Small-world Networks

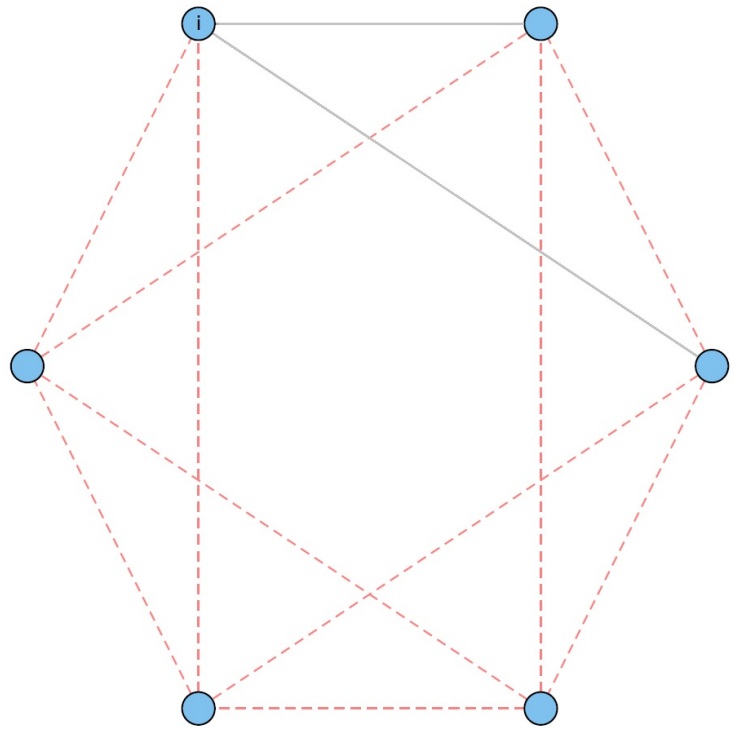
- For known n, the parameters p and k completely characterize this family of networks
- It was shown that $\mathbb{E}[\overline{d}] = 2k$, where $\overline{d}$ is the average degree. Thus,

$$\hat{k} = \frac{\overline{d}}{2}$$

- Estimating p is more complicated.

Estimation: Watts-Strogatz Small-world Networks Continued

- Let t_i denote the total number of triads centered at node i
- Let t_i^{fixed} and t_i^{var} be the number of triads centered at i that always exist and the number of triads centered at i that vary based on the rewiring process, respectively.



Estimation: Watts-Strogatz Small-world Networks Continued

To calculate the expected value of t^{var} we consider the following.

- Let X_i be a random variable denoting the number of edges that are connected to node i initially that are not rewired to a different edge,

$$\Rightarrow X_i \sim \text{Bin}(k, 1-p).$$

Estimation: Watts-Strogatz Small-world Networks Continued

To calculate the expected value of t^{var} we consider the following.

- Let X_i be a random variable denoting the number of edges that are connected to node i initially that are not rewired to a different edge,

$$\Rightarrow X_i \sim \text{Bin}(k, 1-p).$$

- Let Y_i denote the number of edges that are not initially connected to node i that are rewired to node i,

$$\Rightarrow Y_i \sim \text{Bin}\left((n-2)k, p/n\right).$$

Estimation: Watts-Strogatz Small-world Networks Continued

$$
\begin{aligned}
\mathbb{E}[t_i] &= \mathbb{E}[t_i^{fixed}] + \mathbb{E}[t_i^{var}] \\
&= \mathbb{E}[t_i^{fixed}] + \sum_{a=1}^{k} \sum_{b=1}^{(n-2)k} \mathbb{E}[t_i^{var} | X_i = a, Y_i = b] \mathbb{P}[X_i = a, Y_i = b] \\
&= \mathbb{E}[t_i^{fixed}] + \sum_{a=1}^{k} \sum_{b=1}^{(n-2)k} \mathbb{E}[t_i^{var} | X_i = a, Y_i = b] \mathbb{P}[X_i = a] \mathbb{P}[Y_i = b] \\
&= \sum_{l=1}^{2k-1} l + \left(\sum_{a=1}^{k} \sum_{b=1}^{(n-2)k} (a+b)k + \sum_{c=1}^{a+b-1} c \right) \mathbb{P}[X_i = a] \mathbb{P}[Y_i = b] \\
&= \frac{k(k-1)}{2} + \left(\sum_{a=1}^{k} \sum_{b=1}^{(n-2)k} (a+b)k + \frac{(a+b)(a+b-1)}{2} \right) \\
&\times \mathbb{P}[X_i = a] \mathbb{P}[Y_i = b],
\end{aligned}
$$

Estimation: Watts-Strogatz Small-world Networks Continued

The total expected number of triad, $t_{tot} = \sum_{i=1}^{n} t_i = n \cdot t_i$, is

$$\begin{aligned} \mathbb{E}[t_{tot}] &= n\frac{k(k-1)}{2} + n\left(\sum_{a=1}^{k}\sum_{b=1}^{(n-2)k}(a+b)k + \frac{(a+b)(a+b-1)}{2}\right) \\ &\times \mathbb{P}[X_i = a]\mathbb{P}[Y_i = b]. \end{aligned}$$

Replace $\mathbb{E}[t_{tot}]$ and k with $\frac{\overline{d}}{2}$ with the observed number of total triads, and solve numerically for $\hat{p}$

Estimation: Barabási-Albert Scale-free Networks

- The scale-free model considered here has only one parameter to be estimated, m.
- We only consider the case when $m = 1$, so no estimation is needed
- If we observe data and need to estimate m, contrary results arise. There are at least two ways to estimate m:

Estimation: Barabási-Albert Scale-free Networks

- The scale-free model considered here has only one parameter to be estimated, m.
- We only consider the case when $m = 1$, so no estimation is needed
- If we observe data and need to estimate m, contrary results arise. There are at least two ways to estimate m:
  1. taking m as the minimum degree;

Estimation: Barabási-Albert Scale-free Networks

- The scale-free model considered here has only one parameter to be estimated, m.
- We only consider the case when $m = 1$, so no estimation is needed
- If we observe data and need to estimate m, contrary results arise. There are at least two ways to estimate m:
  1. taking m as the minimum degree;
  2. using the total number of edges to find m.

Estimation: Barabási-Albert Scale-free Networks

- The scale-free model considered here has only one parameter to be estimated, m.
- We only consider the case when $m = 1$, so no estimation is needed
- If we observe data and need to estimate m, contrary results arise. There are at least two ways to estimate m:
  1. taking m as the minimum degree;
  2. using the total number of edges to find m.
- This complication in parameter estimating m stems from the highly simplistic nature of the model.

Hypothesis Testing

We use a methodology by Benjamini and Yekutieli discussed in:

The Control of the False Discovery Rate in Multiple Testing Under Dependency.

Hypothesis Testing

We use a methodology by Benjamini and Yekutieli discussed in:

The Control of the False Discovery Rate in Multiple Testing Under Dependency.

1. Calculate the p-values for each of the m tests giving $p_1, \ldots, p_m$.

Hypothesis Testing

We use a methodology by Benjamini and Yekutieli discussed in:

The Control of the False Discovery Rate in Multiple Testing Under Dependency.

1. Calculate the p-values for each of the m tests giving $p_1, \ldots, p_m$.
2. Order the p-values giving $p_{(1)}, \ldots, p_{(m)}$.

Hypothesis Testing

We use a methodology by Benjamini and Yekutieli discussed in:

The Control of the False Discovery Rate in Multiple Testing Under Dependency.

1. Calculate the p-values for each of the m tests giving $p_1, \ldots, p_m$.
2. Order the p-values giving $p_{(1)}, \ldots, p_{(m)}$.
3. Define $k = \max \left\{ i : p_{(i)} \leq \frac{i}{m\left(\sum_{i=1}^{m} 1/i\right)} q \right\}$, and reject $H^0_{(1)}, \ldots, H^0_{(k)}$.

Hypothesis Testing

We use a methodology by Benjamini and Yekutieli discussed in:

> The Control of the False Discovery Rate in Multiple Testing Under Dependency.

1. Calculate the p-values for each of the m tests giving $p_1, \ldots, p_m$.
2. Order the p-values giving $p_{(1)}, \ldots, p_{(m)}$.
3. Define $k = \max \left\{ i : p_{(i)} \leq \frac{i}{m\left(\sum_{i=1}^{m} 1/i\right)} q \right\}$, and reject $H^0_{(1)}, \ldots, H^0_{(k)}$.

Theorem

The above procedure always controls the FDR at level less that or equal to $\frac{m_0}{m} q$, *where* m_0 *is the number of true null hypothesis.*

Simulation Results

Erdös-Rényi Networks

$n = 1000, p = .2$	FDR=0.00615; TPR=1
$n = 1000, p = .5$	FDR=0.00501; TPR=1
$n = 1000, p = .8$	FDR=0.00570; TPR=1

Simulation Results

Erdös-Rényi Networks

$n = 1000, p = .2$	FDR=0.00615; TPR=1
$n = 1000, p = .5$	FDR=0.00501; TPR=1
$n = 1000, p = .8$	FDR=0.00570; TPR=1

Watts-Strogatz Small-World Networks

$n = 1000, k = 5, p = .001$	FDR=0.00100; TPR=0.99996
$n = 1000, k = 5, p = .01$	FDR=0.00108; TPR=0.99894
$n = 1000, k = 5, p = .1$	FDR=0.00082; TPR=0.94882

Simulation Results

Erdös-Rényi Networks

$n = 1000, p = .2$	FDR=0.00615; TPR=1
$n = 1000, p = .5$	FDR=0.00501; TPR=1
$n = 1000, p = .8$	FDR=0.00570; TPR=1

Watts-Strogatz Small-World Networks

$n = 1000, k = 5, p = .001$	FDR=0.00100; TPR=0.99996
$n = 1000, k = 5, p = .01$	FDR=0.00108; TPR=0.99894
$n = 1000, k = 5, p = .1$	FDR=0.00082; TPR=0.94882

Barabási-Albert Scale-free Networks

- For $m = 1$ this method has trouble identifying two nodes as the same.
- The problem arises since the degree of most of the nodes is just 1, which is not a rare value for the overlap score.

Conclusions and Future Work

- Contributions
  - Our approach offers a general and effective framework to answer the question, do two nodes represent the same identity.

Conclusions and Future Work

- Contributions
  - Our approach offers a general and effective framework to answer the question, do two nodes represent the same identity.
  - Using the *overlap score*, we can estimate re-identification performance for a class of networks without performing pairwise comparisons to build models, greatly reducing the complexity compared to existent methods.

Conclusions and Future Work

- Contributions
  - Our approach offers a general and effective framework to answer the question, do two nodes represent the same identity.
  - Using the *overlap score*, we can estimate re-identification performance for a class of networks without performing pairwise comparisons to build models, greatly reducing the complexity compared to existent methods.
- Future Work
  - Consider what happens when a complete network is not observed or a person changes his or her behavior.

Conclusions and Future Work

- Contributions
  - Our approach offers a general and effective framework to answer the question, do two nodes represent the same identity.
  - Using the *overlap score*, we can estimate re-identification performance for a class of networks without performing pairwise comparisons to build models, greatly reducing the complexity compared to existent methods.
- Future Work
  - Consider what happens when a complete network is not observed or a person changes his or her behavior.
  - Consider different measures of similarity and/or attributes associated with each node.

Conclusions and Future Work

- Contributions
  - Our approach offers a general and effective framework to answer the question, do two nodes represent the same identity.
  - Using the *overlap score*, we can estimate re-identification performance for a class of networks without performing pairwise comparisons to build models, greatly reducing the complexity compared to existent methods.
- Future Work
  - Consider what happens when a complete network is not observed or a person changes his or her behavior.
  - Consider different measures of similarity and/or attributes associated with each node.
  - Determine theoretical properties of the estimators or likelihood based methods for estimation and hypothesis testing.

Thank You

Thank You

Questions?