Statistics for Re-identification in Network Models

Justin Vastola
Georgia Institute of Technology
765 Ferst Drive
Atlanta, Georgia, 30332
email: jvastola@gatech.edu

Kobi Abayomi
Georgia Institute of Technology
765 Ferst Drive
Atlanta, Georgia, 30332
email: kobi@gatech.edu

Shawndra Hill
University of Pennsylvania
3730 Walnut Street, Suite 500
Philadelphia, PA 19104
shawndra@wharton.upenn.edu

January 5, 2011

Abstract

We consider statistics for re-identification for three popular network data models: Erdös-Renyi, Small World and Scale Free. We view re-identification, in this setting, as hypotheses tests of network similarity modulo a network data model. This problem has many implications in areas such as fraud detection and author attribution. Much of the prior work on these statistical physics models — while cognizant of the inherent stochasticity of the problem — has inadequately addressed statistical estimation and inference. In this paper we offer a formal statistical framework for re-identification, using first principles and the algorithmic specification of these models. Using our framework, we illustrate the method and its performance on three network data examples: simulations, the Enron emails, and a telecommunications dataset.

1 Introduction

Network data models [11] are applied in fields as diverse as genomic regulation in biology [1] and the social geography of weblogs [12]. Traditionally, these graphs have been modeled by the random graphs of Erdös and Rényi (ER) [10]. More recently, more complex structures have been suggested by Watts and Strogatz's (WS) [20] and Barabási and Albert's [2] small-world and scale-free network structures (SF), respectively.

Each of these models are random — the models are generators for network structures under stochasticity — and each of the models may be 'fit', in the sense that each specify parameters which are dependent upon the observed networked data. In fact: "the structural analysis of network graphs has traditionally been treated primarily as a descriptive task, as opposed to an inferential task, and the tools commonly used for such purposes derive largely from areas outside of 'mainstream' statistics" [17]. Commonly, attention

is focused on characterizing degree distributions or diameters [19] as opposed to parameter estimation or goodness-of-fit.

Much of the work on the Watts-Strogatz and Scale-Free models, in particular, has focused on their construction: design algorithms, settings where they might arise, and descriptive properties such as the clustering coefficient or average geodesic length [19]. Statistical inference methods and properties have been infrequently addressed for these models, principally because the model specifications resist inferential approaches. Alternately, p_1 and p^* models –those that impose a distribution from an exponential family for the adjacency matrix–allow, in theory, straightforward application of typical statistical inference procedures. Additionally, Hoff et al [15] developed a latent space model–a model based on an established latent geometry on nodes–to achieve statistical flexibility in accurately describing a network. These types of models have been used for inference on complex networked data, in contrast to the relative lack of inferential results for WS and SF type networks.

In this paper we investigate statistical inference for the ER, WS and SF type models for network data through consideration of the re-identification problem. We illustrate, in particular, a way of considering each of the graphs as probability models and data 'fit' to a model as a random instantiation. We approach the problem by exploiting algorithmic specifications of each of the graphs, as set out in the literature [2, 10, 20]. Our contribution is the placement of these models within a statistical framework and the description of inferential — estimation and hypothesis testing — procedures that are meaningful for observed data.

1.1 Re-identification in Network Data

This paper develops a precise statistical method for the re-identification problem — the classification or location of nodes in a network that represent the same identity, i.e., have the same *signatures*. In social network data, this problem is often to identify people by their network connections or relational patterns. Re-identification in social networks has been implemented by Cortes et al [8] and Hill et al [14]. Their approaches focus on re-identification on a dynamic network through multiple tuning parameters and similarity scores while ignoring any specified network model. In addition, Hill and Nagle [13] provide an approach to re-identification by using a normal approximation for distributions of similarity.

This paper is a reconsideration of the re-identification problem — under an exact statistical characterization — for the ER, WS and SF models. In this paper we address re-identification from a definition of the similarity 'score'. This approach offers distributional results and consequently yields score distributions, parameter estimation, and hypothesis testing methods in a proper statistical setting. We present our findings on simulated network data, the Reality Mining dataset of Eagle et al [9] and the infamous Enron emails.

The rest of the article is organized as follows: in section 2, we present our general methodology; section 3 contains the details for score distributions followed by the discussion of parameter estimation in section

4; hypothesis testing is addressed in section 5; and results for simulated and real world data are found in sections 6 and 7, respectively. A discussion with conclusions is in section 8.

2 Methodological Overview

We view individual networks G as draws from a family of networks $\mathcal{G}$ based on a specified probability distribution F_θ in the same way that a random variable X is an instantiation of $X(\omega)$ (drawn from a sample space Ω) based on a (not the same) probability distribution F_θ, say. The distribution of $\mathcal{G}$, F_θ, say, may be ethereal: implicitly, we consider the distribution of a graph as arising from an algorithmic construction in lieu than of a complete, explicit specification. In fact, the difficulty in characterizing F_θ lies in mapping the graphical object to an observable datum.

We choose to elide this vagueness by focusing on a particular representation of a random graphical object and then considering its statistical properties. Under this formulation, $\mathcal{G}_\theta$ is a random object — with G the instantiation. We choose a particular representation of this instantiated network and then adopt traditional statistical techniques for analyzing its structure by considering its distribution, or functions of it, via the algorithmic specification of F_θ

Formally, let $\mathcal{G}_\theta$ be a family of random network structures indexed by a vector valued parameter $\theta \in \Theta$. This is: $\mathcal{G}_\theta = \{G \sim F_\theta | \theta \in \Theta\}$. A realization of $\mathcal{G}_\theta$ is a network, G, consisting of vertices $v \in V$ and edges $e \in E$, where V and E are the vertex and edge sets, respectively.

When speaking of vertices and edges, it is convenient to index the vertices–denoted v_i for all $i \in 1 \ldots \#\{V\}$–and to denote an edge between vertices v_i and v_j by e_{ij}. For simplicity, we will often refer to a vertex v_i just by i, its index number only. A convenient representation of a network of order n is in terms of its $n \times n$ adjacency matrix,

$$A = \begin{bmatrix} \mathbf{a_1} \\ \vdots \\ \mathbf{a_n} \end{bmatrix} = \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ & \ldots & \\ a_{n1} & \ldots & a_{nn} \end{bmatrix},$$

where

$$a_{ij} = \begin{cases} 1, & \text{if edge exists between } i \text{ and } j \\ 0, & \text{otherwise.} \end{cases}$$

This is: $A(G)$ - the adjacency matrix of G. The sum of the i^{th} row (or column) of A gives the degree of node i–the number of edges connected to i–denoted d_i. Notice that $A = A(G)$ is a particular choice of representation of the random graphical object G. This allows us to be very general when considering random graphical object, and very specific when referring to the matrix-valued random variable A.

Using A, we look to define statistics $\eta(A) = \eta(A(G))$ (in the same way that we define statistics, $\eta(X)$, on random samples $X \sim F$) that we can use to address the re-identification problem.

2.1 Re-identification in a Network Model

For the re-identification problem, we are interested in whether or not two nodes have the same *signature*, i.e., represent the same entity [13]. Consider, for example, a telecommunications network. A person may appear more than once via a cell phone number and a land line. The *signature* of the cell phone and land line numbers is an identifier for the phone user. In our setup, phone user i is node i in a network model. We denote the signature of node i by $\sigma(i)$.

In the network model, given two nodes i and j, we are interested in the 'similarity' between i and j, *a fortiori*, the distribution of this similarity. We consider the *overlap score* statistic:

$$\eta(A) \equiv S(i,j) = \langle \mathbf{a_i}, \mathbf{a_j} \rangle = \#\left\{ i^* : a_{i,i^*} = a_{j,i^*} = 1, i^* = 1, \ldots, n \right\} \tag{1}$$

with $\langle \cdot, \cdot \rangle$ the usual dot product as our measure of similarity. When $i = j$, we call $S(i,i)$ the *match score*, and when $i \neq j$, we call $S(i,j)$ the *non-match score*. The match score is nothing more than the degree distribution of a particular node. The non-match score, on the other hand, measures how many edges two particular nodes share. If (limiting) distributions can be calculated for both scores, then the likelihood of an observed score can be known, and we can frame similarity via hypothesis testing.

Another way to say this is to suppose distribution $S(i,j) \sim F_\theta$. We could then suggest

$$H_0 : \sigma(i) = \sigma(j) \; vs. \; H_a : \sigma(i) \neq \sigma(j), \; i = 1, \ldots, n \tag{2}$$

as a re-identification test (or tests) for person/node j. A straightforward way to conduct the test is with the observed value of $S(i,j)$ as the estimator for $\sigma(i)$. Our approach is to notice that network construction algorithms — for the Erdös-Rènyi, Watts-Strogatz, and Barabási-Albert scale-free models — are sufficient for finding limiting score distributions for this particular choice of statistic $\eta(A) = \langle \mathbf{a_i}, \mathbf{a_j} \rangle$.

For example, the probability that two nodes share an edge can be deduced from the the rewiring scheme discussed by Watts and Strogatz [20]. We use these probabilities to completely specify the non-match distributions (section 3 provides details). Therefore, given a network construction, statements about the likelihood of an observation can be made. Hypothesis tests formalize such statements.

2.2 Distributions for Similarity

Theoretically, F_θ can be written down exactly: in practice, of course, θ has to estimated from data. In a hypothesis testing framework, (2) is a particular choice of tests for identifying person/node j in an observed

network of n nodes. For a test statistic, we may consider the observed value of $S(i,j)$ and thus

$$1 - F_\theta(s) = Pr_\theta(S(i,j) > s) \tag{3}$$

as a choice for the p-value of the test *under a particular setting for the parameters of the score (similarity) distribution.* Here lies the richness of the problem: θ, for network models that are not completely random (i.e. models other than the ER graphs), includes the labellings — orderings — of the nodes in the graph. This means that the observed data must be used to estimate node labels — observation 1 is node i, observation 2 is node $i+1$, say — concurrently with similarity scoring. For non-random models, in fact, the distributions of this scoring statistic (1) are completely dependent upon the choice of labels, which we properly consider as additional model parameters.

For the match score distribution this distinction is trivial: the observed scoring statistic is merely the observed degree of the node and re-identification is null. We attack the problem for the non-match score distributions from first principles: we generate the distributions via the algorithmic constructions of the network models and then consider hypothesis tests for re-identification that allow us to elide the dependency between labellings and non-match scoring.

3 Score Distributions

We consider three well known network structures — ER, small-world, and scale-free. For each structure, we derive its non-match score distribution by simply following the algorithmic constructions of the family $\mathcal{G}_\theta$ that generate each type of model.

3.1 Erdös-Rényi Networks

An ER network is constructed by independently placing edges with probability p among n nodes. Each pair of nodes is considered once and an edge is placed between them with probability p, yielding $\binom{n}{2}$ independent Bernoulli trials. To derive the non-match score distribution for nodes i and j,$i \neq j$, we consider this construction. Notice, first, that each scalar product in the dot product via the score can be viewed as a Bernoulli trial. Each scalar product is between two elements of the adjacency matrix, resulting in scalar products equal to either a 0 or 1. The probability of success for each of these "trials" is $\mathbb{P}\{a_{i,i^*} \cdot a_{j,i^*} = 1\} = \mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\}$ when testing if nodes i and j share an edge with node i^*. Therefore, the score distribution, both match and non-match, is the sum of Bernoulli random variables. For Erdös-Renyí networks, these random variables are independent and identically distributed. The other two networks don't have this accommodating property.

The probability that nodes i and j share a common edge i^*, $i^* \in \{1, \ldots, n\} \setminus \{i, j\}$, is

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

since the two connections are made independently with probability p. We restrict node i^* from equaling i and j because the network construction doesn't allow loops. Thus, $a_{i,i} = a_{j,j} = 0$ with probability 1, and the probability of interest stated above is always 0 for these two cases. The non-match score distribution is, therefore, $S(i, j) \sim \text{Bin}(n - 2, p^2)$. Plots of the empirical match and non-match score distributions along with the theoretical distributions derived above are shown in figure 1.

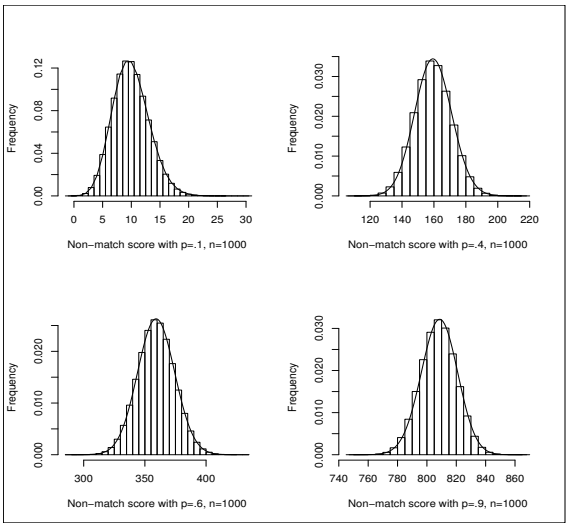

Figure 1: These plots show the fits of theoretical non-match score distributions to empirical data for Erdös-Rényi Networks of size $n = 1000$ with varying p.

3.2 Small-World Networks

Small-world networks, namely, those constructed by Watts and Strogatz, were developed to incorporate high levels of clustering and small distances between most nodes–both properties found in real world data. A family of small-world networks $\mathcal{G}_\theta$ can be generated as a rewiring of a $2k$ connected lattice. As described in Watts and Strogatz [20], the lattice is iteratively rewired with probability parameter p.

This characterization yields a limiting distribution for the non-match score distribution. Consider the network generated from an initial, non-rewired $2k$ connected lattice. Post rewiring, the probability that an edge is shared by nodes i and j is dependent upon the following pre-rewiring cases:

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 $a_{j,i*} = 0$.

We denote the limiting distribution of $\mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\}$ for each case by f_i, $i = 1, 2, 3$–the *partial non-match score* distributions. When computing the overlap/similarity score $S(i, j)$, each of case $i = 1, 2, 3$ may arise. The total number of matching edges is the sum of the number of edges arising from each case. Thus, we are interested in the distribution of $Z = X_1 + X_2 + X_3$, where X_1, X_2, X_3 are random variables from distributions $f_1, f_2,$ and f_3, respectively.

Consider case 1 first. Prior to rewiring, nodes i and j are unconnected to node i^*. Denoting the lattice distance between nodes i and j by $||i, j||$, $||i, i^*||$ and $||j, i^*||$ are both less than k. The number of times this case arises is

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

When calculating $\mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\}$, we first note by the independence of the rewiring process,

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

Now assume $a_{i,i^*} = 0$ pre-rewiring. Post rewiring $a_{i,i^*} = 1$ if: $(i.)$ one of the k edges considered for rewiring from node i is rewired to node i^*, or $(ii.)$ one of the k edges considered for rewiring from node i^* is rewired to node i. View the rewiring of the $2k$ edges as $2k$ independent Bernoulli trials with success probability p/n, where a success is defined as a rewiring that results in a_{i,i^*}=1. Let X denote the number of success, i.e., $X \sim \text{Bin}(2k, p/n)$. Then,

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

implying $\mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\} = (1 - (1 - p/n)^{2k})^2$. Note, we do not need to address whether or not node i or i^* is rewired first by assuming the possibility of multigraphs. The limiting distribution of the partial non-match score is completely specified as

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

In case 2, nodes i and j are connected to node i^*, i.e., $||i, i^*||$ and $||j, i^*||$ are both less than k. This case occurs

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

times. Post rewiring, $a_{i,i^*} = 0$ if edge (i, i^*) is removed and not replaced while none of the $2k - 1$ edges remaining for rewiring connect nodes i and i^*. Edge (i, i^*) is removed and not replaced with probability

$p(\frac{n-1}{n})$, while none of the $2k-1$ edges is rewired forming edge (i, i^*) with probability $\mathbb{P}\{Y=0\}$, where $Y \sim \text{Bin}(2k-1, p/n)$. Post rewiring, $\mathbb{P}\{a_{i,i^*}=0\} = p(\frac{n-1}{n})(1-p/n)^{2k-1}$ implying $\mathbb{P}\{a_{i,i^*}=1\} = 1 - p(\frac{n-1}{n})(1-p/n)^{2k-1}$. The partial non-match score distribution is

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

The number of occurrences of the last case, when $||i, i^*|| > k$ and $||i, i^*|| \leq k$, or vice versa, is

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

Without loss of generality, assume $||i, i^*|| > k$ and $||j, i^*|| \leq k$. After the rewiring process, we have $\mathbb{P}\{a_{i,i^*}=1\} = 1 - p(\frac{n-1}{n})(1-p/n)^{2k-1}$, and $\mathbb{P}\{a_{j,i^*}=1\} = 1-(1-p/n)^{2k}$. The product of these two probabilities is the probability of nodes i and j both being connected to node i^*. The partial non-match score distribution is

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

The complete non-match score distribution is found by taking the convolution of f_1, f_2, and f_3. In particular,

$$\begin{aligned} S(i,j) &\stackrel{d}{=} \sum_{y=0}^{z} f_3(z-y) \sum_{x=0}^{y} f_1(x) f_2(y-x) \\ &= \sum_{y=0}^{z} \binom{n_3}{z-y} p_3^{y-z} (1-p_3)^{n_3-(z-y)} \mathbb{1}\{z-y \leq n_3\} \\ &\times \sum_{x=0}^{y} \binom{n_1}{x} p_1^x (1-p_1)^{n_1-x} \mathbb{1}\{x \leq n_1\} \binom{n_2}{y-x} p_2^{y-x} (1-p_2)^{n_2-(y-x)} \mathbb{1}\{y-x \leq n_2\}, \end{aligned}$$

where $p_1, p_1,$ and p_3 are the probabilities of success for $f_1, f_2,$ and f_3, respectively. Figure 2 shows plots of empirical histograms against the theoretical distributions derived above.

3.3 Scale-Free Networks

Scale-free networks are networks whose degree distribution follows a power law. Barabási and Albert [3] provided a method of constructing scale-free networks based on growth and preferential attachment, however, their description is imprecise. Bollobás and Riordan [7] remedy this issue by precisely specifying the model of Barabási and Albert.

Denote a network with of size n–or equivalently, the network at time $n = t$–by $G_{m,n}$, where m is the number of edges introduced at each time step. The construction follows. Start with an initial network, $G_{1,1}$,

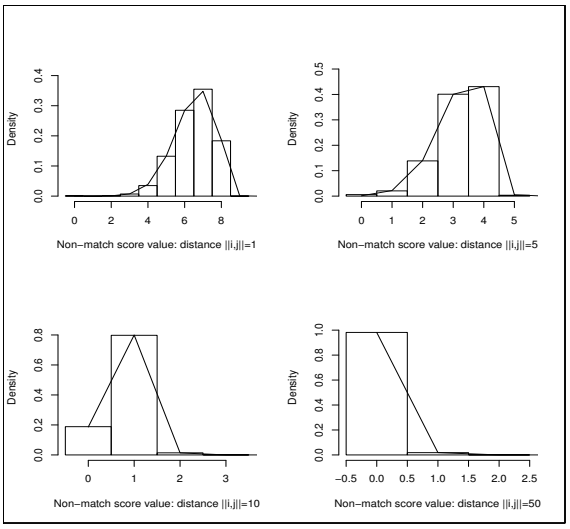

Figure 2: These plots show empirical histograms of the match and non-match score distributions for the small-world networks of Watts and Strogatz along with the theoretical fits. The sample size, rewiring probability, and initial connectivity are $n = 1000$, $p = 0.1$, and $k = 5$, respectively.

with one vertex and one loop. Let $d_{n,i}$ denote the degree of node i when the size of the network 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

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

Each entering vertex is connected to an existing vertex proportional to its degree. We will denote the case when the entering node j is attached to existing node i by $h_j = i$. Note, when node n enters, it has degree one before being connected to an additional node. The authors generalize this construction to adding $m \geq 1$ edges at each time step. Each edge is connected to an existing node one at a time taking into account the degrees added after each connection. In this case, node n has degree m when it enters.

For the non-match score distribution, we first consider the base case when $m = 1$, for each case where $m > 1$ is derived from this base case. An important distinction between the scale-free network model and the two previously considered models is that the preferential attachment scheme introduces a dependency between whether or not two nodes share an edge. Where $\mathbb{P}\{a_{i,i^*} = a_{j,i^*} = 1\} = \mathbb{P}\{a_{i,i^*} = 1\}\,\mathbb{P}\{a_{j,i^*} = 1\}$ for the ER and small-world network models, here the independence assumption does not hold. In particular, these probabilities are dependent upon the order of the addition of the node in the wiring algorithm.

When calculating the probability that nodes j and k are both connected to node i, i.e., $\mathbb{P}\{a_{j,i} = 1, a_{k,i} = 1\}$, we have to consider the following four orderings of nodes i, j, and k:

- $1 \leq i < j < k$
- $1 \leq i = j < k$
- $1 \leq j < i < k$
- $1 \leq j < k < i$

We first note that the last case is irrelevant when $m = 1$ because only one edge is connected from node i; i cannot be connected to both nodes j and k. This case is relevant, however, when $m > 1$. Following in the footsteps of Bollobás and Riordan — i.e., by conditioning on the network at previous time steps–for each case, we derive the following success probabilities. When $1 \leq i < j < k$,

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

When $1 \leq i = j < k$, we get the following probability:

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

Lastly, when $1 \leq j < i < k$, by modifying the work of Bollobás and Riordan, we calculate

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

Let $X_{j,k}^i$ denote the Bernoulli random variable with success probability $p_{j,k}^i := \mathbb{P}\left\{a_{i,j} = 1, a_{i,k} = 1\right\}$, i.e., the random variable representing whether or not nodes j and k are both connected to node i. The non-match score distribution is the sum of these random variables from $i = 1, \ldots, k-1$ for $i \neq j$. Unlike in the small-world scenario, these Bernoulli random variables are dependent, where the dependence arises not from the preferential attachment, but from the parameter m which restricts the maximum value of the overlap score. In the case we are considering, $m = 1$, the overlap score is restricted to be a Bernoulli random variable. Denote this random variable, the non-match score, by $Z = \sum_{i=1}^{k-1} X_{j,k}^i$, assuming the $k > j$. The probability of success is

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

yielding a final non-match score distribution of

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

We have only considered the case when $m = 1$. In principle, similar logic will lead us to distributions for arbitrary m, but the actually derivations for $m > 1$ rapidly increase in difficulty, leaving the appropriate distributions to be calculated empirically via simulations.

4 Parameter Estimation

Complications arise in parameter estimation in network data due to the dependent nature of the data. In particular, the dependence among the degrees of each node makes finding likelihood distributions difficult since the likelihood is no longer the product of the marginal distributions. For the ER network model, estimating its sole parameter p is not burdensome, yet developing estimators in WS and SF network models poses some complications.

See Mckay and Wormald [18] for a conversation about of the dependency of the degrees in an Erdös-Rényi network. Other styles of estimators, such as Horvitz-Thompson estimators, heavily depend on the sampling schemes to estimate certain parameters along with the accuracy of each estimate. See chapter 5 of Kolaczyk [17] for illustrations.

4.1 Estimation in 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. Unlike the other two network models, this is observable quantity is random. Since the placement of each edge is independent of the placement of all other edges and the probability of placing each edge is identical, the total number of edges in an ER network is simply a binomial random variable with $\binom{n}{2}$ trials and success probability p, reducing the parameter estimation in ER networks to estimation for a binomial random variable. In particular, the maximum likelihood estimate of p is $\hat{p} = \frac{\mathcal{E}^{obs}}{\binom{n}{2}}$, where $\mathcal{E}^{obs}$ is the observed number of edges. For this simple model, we have the luxury of knowing statistical properties of this estimator, i.e., limiting distribution, while as for the other network models, properties of their estimators are analytically elusive due to complex dependencies in the data.

4.2 Estimation in small-world networks

For the small-world network construction we consider, the parameter to estimate in order to completely specify $\mathcal{G}_\theta$ is $\theta = (k, p)$. To get around the dependency in the data, we consider estimates of k and p based

on the method of moments. Barrat and Weigt [4] showed that the mean behavior of the degree of any particular node is $2k$, i.e., $\mathbb{E}[d_i] = 2k$. It is easy to see this relationship without the technical derivations. Noting that $\mathbb{E}[\bar{d}] = \mathbb{E}[d_i]$, the method of moments estimator of k is $\hat{k} = \bar{d}/2$, where $\bar{d}$ is the observed average degree. To estimate p, we consider the total number of triads centered at node i, denoted t_i. Our definition of a triad centered at node i is a subnetwork consisting of 3 nodes connected by two edges with the degree of node i equal to 2. 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. Additionally, 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 and Y_i denote the number of edges that are not initially connected to node i that are rewired to node i. Then, $X_i \sim \text{Bin}(k, 1-p)$ and $Y_i \sim \text{Bin}((n-2)k, p/n)$. Then we have

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

where

$$\mathbb{P}[X_i = a] = \binom{k}{a} p^a (1-p)^{k-a} \mathbb{1}\{a \leq k\}$$

and

$$\mathbb{P}[Y_i = b] = \binom{(n-2)k}{b} \left(\frac{p}{n}\right)^b \left(1 - \frac{p}{n}\right)^{(n-2)k-b} \mathbb{1}\{b \leq (n-2)k\}$$

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

$$\mathbb{E}[t_{tot}] = n\frac{k(k-1)}{2} + n \left(\sum_{a=1}^{k} \sum_{b=1}^{(n-2)k} (a+b)k + \frac{(a+b)(a+b-1)}{2} \right) \mathbb{P}[X_i = a] \mathbb{P}[Y_i = b]. \tag{4}$$

To estimate p, plug $\hat{k}$ into equation (1) as well as replace $\mathbb{E}[t_{tot}]$ with the observed total number of triads which we denote by t^{obs}. Analytically solving the resulting equation is quite cumbersome, however, numerical solutions are very easy to come by. The estimate of p is:

$$\hat{p} = \arg\min_{0 \leq p \leq 1} \left\{ \left| t_{tot}^{obs} - \left(n\frac{k(k-1)}{2} + n \left(\sum_{a=1}^{k} \sum_{b=1}^{(n-2)k} (a+b)k + \frac{(a+b)(a+b-1)}{2} \right) \mathbb{P}[X_i = a] \mathbb{P}[Y_i = b] \right) \right| \right\}.$$

Barrat and Weigt [4] also argue that the mean behavior of the *clustering coefficient*–the ratio of the mean number of links between the neighbors of a vertex and the mean number of possible links between the neighbors of a vertex–asymptotically behaves like

$$cc = \frac{3(k-1)}{2(2k-1)}(1-p)^3.$$

We can use this asymptotic identity to estimate p, but it must be noted that the resulting estimator serves as an approximate method of moments estimator since terms of order 1/n are ignored. We propose the above estimator for its specificity.

We must emphasize that the non-match score distribution for each pair of nodes (i, j) is dependent on the lattice distance between the nodes through the parameters n_1, n_2, and n_3, meaning to properly specify the non-match score distribution, we essentially need the labels of the nodes. To the best of our knowledge, we don't know of any algorithm that will accurately estimate the node labels, and in fact, discovering the labels from an unlabeled graph may be an infeasible task with out any additional information. If we know the labels, then we can just apply the methodology to the above non-match score distribution, while on the other hand, when we have no knowledge of the node labels, we view the data–the non-match scores–as coming from a mixture distribution which can replace the non-match score distribution in our methodology. Formally, 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

$$f_{mix} = \sum_{i=1}^{\lfloor n/2 \rfloor} \alpha_i s_i.$$

The α_i's are actually known, so we are not introducing any additional parameters, leaving only p and k to be estimated as before. For instance, if n is even, $\alpha_i = \frac{2}{n-1}$ for $i = 1, \ldots, \lfloor n/2 \rfloor - 1$ and $\alpha_{\lfloor n/2 \rfloor} = \frac{1}{n-1}$.

Adopting the mixture distribution yields less power in the statistical tests for detecting an anomalous event–the event that two nodes have too many common neighbors–yet reduces the risk of type I error for the re-identification problem. This result is not a problem, however, since the re-identification problem is not to detect all anomalous event but the event that two nodes have the same behavior. For example, consider a WS network with $n = 200$ and $k = 7$. Two nodes, i and j, that are at opposite ends of the lattice, i.e., $||i, j|| = \lfloor n/2 \rfloor$, will be concluded to represent the same identity using the non-match score distribution $s_{\lfloor n/2 \rfloor}$ for an non-match score of 2. On average, however, nodes will have 14 neighbors (see section 4.2), so we don't want to conclude that two nodes have the same signature when they have only two common neighbors.

4.3 Estimation in 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 to characterize the family of scale free networks that we consider here. If we observe data and need to estimate m, contrary results arise. There are two methods for finding m: taking m as the minimum degree and using the total number of edges to find m. This network construction always gives a network with mn edges. Many real world networks will have a node with degree 1 at some point in its construction, yet the network will almost always have more than n edges. This complication in parameter estimating m stems from the highly simplistic nature of the model.

Similar to the WS model, the non-match score distributions depend on the labellings of the nodes–the time stamp when node enters the network in opposition to the positioning around a lattice. At a first glance, one may think that the preferential attachment characteristic would give the order in which the nodes enter–hence the labeling–since earlier nodes will typically have a higher degree than nodes that enter the network at a later time, yet this result is the case. Consider this likely scenario. When the fourth node enters into the network, suppose the degrees of nodes 1, 2, and 3 are 4, 1, and 1, respectively, the event that nodes 2 and 3 both connect to node 1. Node 4 is just as likely to attach to node 2 as it is to node 3, and if its edge links to node 3, than node 3 will have a higher 'preference' than node 2, likely to result in node three having a larger degree than node 2. Similar scenarios will play out as the network grows, and the likelihood that the node time stamps will be the same as the labeling induced from the ordered degrees is very small. As a result, developing even a consistent estimate for the labellings is likely impossible, where consistency here refers to asymptotic properties resulting from an infinite network.

Once again, we will consider a mixture distribution to eradicate the necessity of node labels. Let $s_{i,j}$ denotes the non-match score distribution for two nodes, i and j. Let $\alpha_{i,j} \in \mathbb{R}$ be the mixing parameters such that $0 \leq \alpha_{i,j} \leq 1$ for all (i, j) and $\sum_{\text{all } i,j} \alpha_{i,j} = 1$. The mixture distribution is

$$f_{mix} = \sum_{\text{all } i,j} \alpha_{i,j} s_{i,j},$$

where $\alpha_{i,j} = \frac{1}{\binom{n}{2}}$ for all (i, j).

5 Hypothesis Testing

A single hypothesis test $H_0 : \sigma(i) \neq \sigma(j)$ vs. $H_a : \sigma(i) = \sigma(j)$ is straightforward to perform, yet may not be very relevant to the re-identification problem since the knowledge of which two nodes to test is elusive. Instead, multitudes of hypothesis tests need to be performed, resulting in the need for a multiple testing procedure. The multiple testing procedure we choose to use is based on controlling the false discovery rate (FDR) using methods proposed by Benjamini and Hochberg [5] and extended by Benjamini and Yeuketieli [6].

Recall,

$$FDR = \mathbb{E}\left(\frac{R_{false}}{R}|R>0\right),$$

where R is the number of rejections among m tests and R_{false} is the number of false rejections. Controlling the FDR has advantages when considering applications of re-identification. In context of fraud detection, a company wants to minimize the number of false discoveries–the claim that a user is committing fraud when in fact is not–so as to not have to wrongfully accuse customers of fraud and to reduce the amount of time that a company representative has to investigate false fraud claims. The Benjamini-Hochberg procedure for controlling the significance level at γ is as follows: calculate the p-values for each of the m tests giving $p_1, \ldots, p_m$; order the p-values giving $p_{(1)}, \ldots, p_{(m)}$; define $k = \max\left\{i : p_{(i)} \leq (\frac{i}{m})\gamma\right\}$; and reject $H^0_{(1)}, \ldots, H^0_{(k)}$. In their work, they show that the FDR will be controlled at the level γ for independent hypothesis tests. The tests we are considering are not independent but have an ambiguous dependence inherited from each construction algorithm. Consider an ER network with the extreme case that two nodes i and j have an overlap score of $n-2$, i.e., each node is connected to every other node. The overlap score for node i^* and i is at least one less than the degree of i^*. It is clear that a dependence in the test statistics exist, yet the nature of the dependency is not well understood, leaving creating network specific tests to control the FDR as a highly arduous task. Benjamini and Yeuketieli, provide a method that allows for any general dependence amongst hypothesis tests. In particular, they show that by changing the definition of k as defined in the Benjamini-Hochberg procedure to $k = \max\left\{i : p_{(i)} \leq \frac{i}{m\left(\sum_{i=1}^{m} 1/i\right)}q\right\}$, the FDR will be controlled at a level less than or equal to $\frac{m_0}{m}q$, where m_0 is the number of true null hypothesis. The generality of the multiple testing offered by Benjamini-Yeuketieli procedure makes it an attractive testing procedure for the unknown dependence structure inherited by the network constructions.

6 Simulation Results

We first illustrate how our methodology for each network through simulations. For each construction, we simulate a network of size $n = 1000$ with specified parameter θ and measure performance based on the false positive rate (FPR) and the true positive rate (TPR). We replicate each network 100 times and average the FPR and TPR for a more accurate evaluation.

For ER networks, we vary p over the set $\{.2, .5, .8\}$, and for the WS networks, we study the cases that were highlighted in the seminal work of Watts and Strogatz [20]. In particular, we fix $k = 5$ and vary p over the set $\{.001, .01, .1\}$. The results are shown below.

Erdös-Rényi Networks

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

Watts-Strogatz Networks

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

Figure 3: False Positive and True Positive rates for node re-identification, Erdös Rényi and Watts-Strogatz Networks, using overlap score distribution.

Each method performs very well, with the TPR of the ER networks slightly higher than in the WS networks. This result is due to the fact that the clustering coefficient–thus, the propensity for two nodes to have more neighbors in common–in WS networks is higher than that of ER networks. However, even with the high clustering, our method performs very well for WS networks.

For the SF network, we ran simulations for $m = 1$, the case that we derived exactly. Our framework has trouble correctly identifying nodes representing the same signature–with the TPR $\approx .5$–due to the SF network property that many nodes will simply have degree m. For a small value of m, especially for $m = 1$, we know very little information about many nodes, making it hard to 're-identify' it.

7 Applications

We apply our methodology to two datasets: the Reality Mining Group at MIT which was introduced by Eagle et al [9] and the infamous Enron email dataset. Since each dataset doesn't have fraudulent activity in the form of two nodes representing the same entity, we treat each match score, i.e., the comparison of a node with itself, as a non-match score. After estimating the parameters, a decision needs to be made on which model is actually provides the best fit. Goodness-of-fit (GOF) has been addressed by Hunter et al [16] who offer a graphical procedure for GOF by plotting simulated distributions of the degree, edge-wise shared neighbors, and mean geodesic distance against the observed distributions for a given network dataset. Theoretical results for GOF and model selection procedures are uncharted due to the dependencies in the data which make traditional chi-square GOF tests and AIC and BIC model selection procedures irrelevant. We provide a brief discussion for each dataset on which model we feel is most appropriate.

7.1 Reality Mining Dataset

The Reality Mining dataset consists of 100 subjects–faculty and undergraduate and graduate students from MIT–whose telephone calls were being tracked. We consider a restricted dataset by only considering in-network calls and removing all loops, multiple edges, and isolated nodes, resulting in a network with 72 phone users. A plot of the network is show in Figure 4.

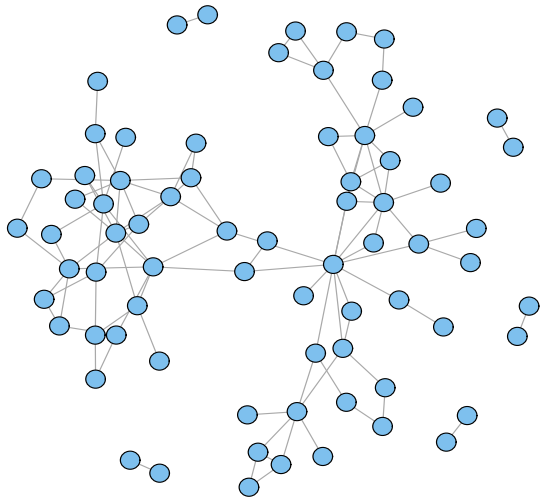

Figure 4: Network representation of the call patterns in the restricted Reality Mining dataset.

The table below shows the parameter estimates for each of the network constructions along its associated FPR and TPR.

Notice for the WS model, p is estimated to be close to 1, giving a highly random network that is essentially an imitation of an ER network. As a result, the ER and WS models yield the same performance based on the FPR and TPR. The performance for the SF network seems to be the best, but this performance is due to a model misspecification which leads to the rejection of the null hypothesis whenever two nodes have an overlap score greater than 1. One of the criterion for GOF proposed by Hunter et al is the edge-wise shared neighbors, which is similar to our overlap score. Based on this criterion, it is clear that the SF model is not an appropriate model since the many pairs of nodes have an overlap score greater than 1 which occurs with probability zero under the assumed model. None of the models performs adequately, which is due to the

Erdös-Rényi	$p = .0411$	TPR=.4167, FPR=.00469
Watts-Strogatz	$p = .9326, k = 1$	TPR=.4167, FPR=.00469
Barabási-Albert	$m = 1$	TPR=.7083, FPR=.02152

Figure 5: FPR and TPR for overlap score distribution re-identification on Reality Mining data

simplistic nature of the ER model and the rigidity of the WS and SF models.

7.2 Enron Emails

The Enron emails have been well studied in network literature. The data we consider consists of emails sent between 144 top executives, where multiple edges and loops have been removed. A plot of the dataset is shown in Figure 6. The table below shows the parameter estimates for each of the network constructions

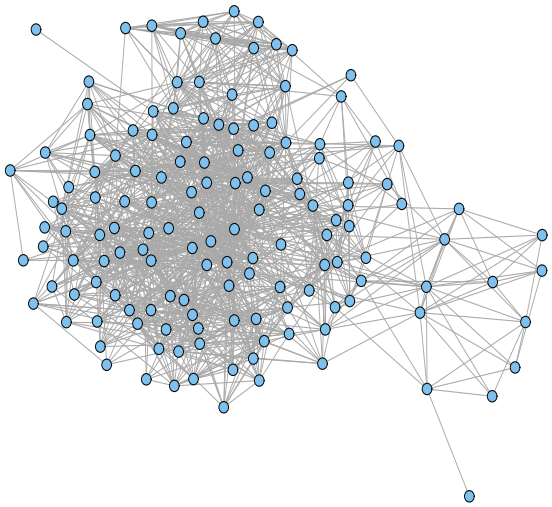

Figure 6: Network representation of the email patterns of 144 top Enron executives.

along its associated FPR and TPR.

The ER model performs the best, as it is the best approximation of the three network models. The WS model inherits trouble through the estimate of $k = 9$, when many nodes have degree less than 9. A WS model is restricted in such a way that all of its nodes must have degree greater than or equal to k. Similar to the Reality Mining dataset, the SF model is not appropriate since many pairs of nodes have an overlap greater than 1. As a result, the TPR is high at a cost of the FPR being high as well.

Erdös-Rényi	$p = .1312$	TPR=.8690, FPR=.1312
Watts-Strogatz	$p = .7480, k = 9$	TPR=.0069, FPR=0
Barabási-Albert	$m = 1$	TPR=.9793, FPR=.5575

Figure 7: FPR and TPR for overlap score re-identification on Enron email data.

8 Discussion and Conclusions

In this paper, we have explored three well known network models — ER, WS, and SF — in the context of the re-identification problem by considering a precise statistical treatment which has previously been ignored for WS and SF networks and only approximately considered for ER models [13].

This approach illustrates how statistical inference can conducted with only an algorithmic recipe for the construction of a network model. In a direct sense, this take suggests a sufficiency property for at least this (perhaps trivial) version of a similarity score. The broader specification of these physics-type network models as statistical objects is important:

Which network properties are 'wiring' algorithms sufficient for? Here we have derived overlap/similarity scores from standard 'wiring' procedures. Are these algorithms sufficient for other statistical functions on networked data, for other interesting hypothesis settings?

What constitutes a unique 'wiring' algorithm? These network models overlap in that it is possible that an observed network could have resulted through more than one algorithmic constructions, i.e., through restrictions on the parameter spaces, for example, an ER network and a WS network may be indistiguishable. In the same way that a characteristic function completely defines the probability distribution of a random variable, is there an analogue for these physics type models?

Estimation, Hypothesis Testing and Goodness of Fit: We have offered straightforward, almost *post hoc*, methods for estimating the network model parameters and testing the re-identification hypotheses. Is there a framework that can be deduced from a proper statistical specification of these models as random objects?

In a sense, we can think of these questions as placing these (physics type) models within a modeling framework. We comment that these models have arisen in the literature in ignorance of their statistical specification; hypothesis testing uses of these models in absence of this framework has relied on *ad hoc*, empirical methods. We have demonstrated a straightforward statistical characterization and illustrated its consequences for a version of the re-identification problem.

In this paper we have had to work around the dependence structure in these network models; principally because the observed degree distribution for a network cannot be treated as an independent sample. In particular, we have relied on parameter estimation and hypothesis testing procedures that do not depend, directly, on the statistical likelihood. This loses the convenience of asymptotic theory that maximum likelihood methods provide for hypothesis testing, model selection, and inference on parameter estimation.

The simplicity of the ER model and the rigidity of the WS and SF models complicates inference on real datasets. Although it is possible to estimate parameters, the appropriateness of the networks may be

limited, however, as our understanding of networks through generative algorithms increases, these procedures could be applied to better (statistical) specifications of network models for more accurate inferences on real networked data.

9 Acknowledgements

This research was supported in part by the Office of Naval Research.

References

[1] "kyoto encyclopedia of genes and genomes," available at http://www.genome.ad.jp/kegg/kegg1.html.

[2] Albert-Lázló Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] Albert-Lázló Barabási, Erzsébet Ravasz, and Tamás Vicsek. Deterministic scale-free networks. *Physica A*, 299:559–564, 2001.

[4] Alain Barrat and Martin Weigt. On the properties of small-world network models. *Europ.Phys.J.B*, 13:547, 2000.

[5] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300, 1995.

[6] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[7] Béla Bollobás and Oliver Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, 2004.

[8] Corinna Cortes, Daryl Pregibon, and Chris Volinsky. Computational methods for dynamic graphs. *Journal of Computational & Graphical Statistics*, 12(4):950–970, 2003.

[9] Nathan Eagle and Alex Pentland. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[10] Paul Erdös and Alfréd Rényi. On random graphs. *Publ. Math. Debrecen*, 209:290–297, 1959.

[11] Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2009.

[12] Sucharita Gopal. *The evolving social geography of blogs.* Berlin: Springer.

[13] Shawndra Hill and Akash Nagle. Social network signatures: A framework for re-identification in networked data. In *Proceedings of the Internatinal Conference on Computational Aspects of Social Networks*, Fontainebleau, France, 2009. IEEE Press.

[14] Shawndra B. Hill, Keepak K. Agarwal, Robert Bell, and Chris Volinsky. Building an effective representation for dynamic networks. *Journal of Computational & Graphical Statistics*, 15(3):584–608, 2006.

[15] Peter Hoff, Adrian Raftery, and Mark Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098, 2002.

[16] David Hunter, Steven Goodreau, and Mark Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258, 2008.

[17] Eric Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, New York, New York, 2009.

[18] Brendan D. McKay and Nicholas C. Wormald. The degree sequence of a random graph. 1. the models. *Random Stuctures and Algorithms*, 11:97–117, 1997.

[19] Mark Newman, Albert Lszl Barabsi, and Duncan J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[20] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.