# STATISTICS FOR RE-IDENTIFICATION IN NETWORK MODELS 

By Justin Vastola ${ }^{\dagger}$, Kobi Abayomi ${ }^{\dagger}$ and Shawndra Hill ${ }^{*, \ddagger}$<br>Georgia Institute of Technology ${ }^{\dagger}$ and University of Pennsylvania ${ }^{\ddagger}$


#### Abstract

Re-identification is the process of matching records or behaviors that belong to the same individual, sometimes when the individual is acting anonymously. The ability to re-identify individuals from their social network behavior-their interactions with others on a social network, has many real-world implications in areas such as fraud detection, online target marketing, and author attribution. We considered statistics for re-identification in social network data from three popular network models: Erdös-Renyi, Small World and Scale Free. Many researchers who have worked on these statistical physics models - while cognizant of the inherent stochasticity of the problem have inadequately addressed statistical estimation and inference. We view re-identification, in this setting, as hypotheses tests of network similarity modulo a network data model. 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. As humans generate more and more data in their daily lives, they leave behind massive trails of information that reveal their unique behavioral characteristics. Using this data, it is possible to identify each individual. Methods similar to those that rely on biometric data, using fingerprints and retinal scans to identify people, are needed to identify individuals through their behavioral data, including social-network data.

Problems of statistical re-identification in networks have been identified in the literature regarding various real-world applications. For example, a repetitive fraud problem has been identified by telecommunications firms; an individual perpetrates a fraud, his account is disconnected for non-payment, and then that same individual signs up for another account, using a different alias, perpetrates the same type of fraud, has his account disconnected again, and so on. The problem for the firm is to identify the user accounts as fraudulent as early as possible by linking new accounts to known fraudulent ones employing user behavior, for example similarities in call behavior, used

[^0]to create behavioral signatures [Hill et al. (2006); Cortes, Pregibon and Volinsky (2003)]. Network signatures have also been shown to be effective in completing the re-identification task of identifying online users based on their browsing behavior to provide better targeted marketing and advertising [Hill et al. (2006)]. In addition, social network signatures have been used for author attribution of written documents, where the identities of authors of articles can be inferred based on the authors they citations [Hill and Provost (2003)]. The applications of re-identification are vast, ranging from protecting the privacy of personal records to asymmetric threat detection for national security. Therefore, developing reliable methods for re-identification of users is an important problem.

In this paper, we take a network-based approach to the re-identification task. Network data models [Goldenberg et al. (2009)] are applied in fields as diverse as genomic regulation in biology [Kanehisa and Goto (2000)] and the social geography of weblogs [Gopal (2007)]. Traditionally, these networks have been modeled by the random networks of Erdös and Rényi (ER) [Erdös and Rényi (1959)]. More recently, more complex structures have been suggested by Watts and Strogatz's (WS) [Watts and Strogatz (1998)] and Barabási and Albert's [Barabási and Albert (1999)] small-world and scale-free network structures (SF), respectively.

Each of these models is random - the models are generators for network structures under stochasticity - and each of the models may be 'fit', in the sense that each specifies parameters that 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" [Kolaczyk (2009)]. Commonly, attention is focused on characterizing degree distributions or diameters [Newman, Barabsi and Watts (2006)] as opposed to parameter estimation or goodness-of-fit.

This paper presents a precise statistical method for solving 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 means identifying people by their network connections or relational patterns. For example, re-identification in social networks has been implemented by Cortes, Pregibon and Volinsky (2003) and Hill et al. (2006) for telecommunications fraud detection. Their approaches focus on performing re-identification on a dynamic network through multiple tuning parameters and similarity scores while ignoring any specified network models. In addition, Hill and Nagle (2009) provided an approach to
re-identification by using a normal approximation for distributions of similarity. Our work extends the extant research on this network-re-identification task.

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 [Newman, Barabsi and Watts (2006)]. Statistical inference methods and properties have been infrequently addressed in relation to 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, Raftery and Handcock (2002) developed a latent space modelbased 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 provide information about our investigation of statistical inference for the ER, WS and SF models for network data through a consideration of the re-identification problem. We illustrate, in particular, a method for considering each of the graphs as probability models and data 'fit' to a model as a random instantiation. We approach the problem by exploiting the algorithmic specifications of each of the graphs, as set out in the literature [Barabási and Albert (1999); Erdös and Rényi (1959); Watts and Strogatz (1998)]. Our contribution is placing these models within a statistical framework and describing the of inferential - estimation and hypothesis testing - procedures that are meaningful for observed data.

This paper is a reconsideration of the re-identification problem using an exact statistical characterization for the ER, WS and SF models. In this paper, we address re-identification, beginning with 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 and Pentland (2006) and the infamous Enron emails.

The rest of the article is organized in the following way. In Section 2, we present our general methodology; Section 3 contains the details about score distributions, followed by a discussion of parameter estimation in Section 4. We address hypothesis testing in Section 5. Results from 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 $F_{\theta}$ in the same way that a random variable $X$ is an instantiation of $X(\omega)$ (drawn from a sample space $\Omega$ ) based on a (not the same) probability distribution $F_{\theta}$, say. The distribution of $\mathcal{G}, F_{\theta}$, 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 $F_{\theta}$ 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 $F_{\theta}$

Formally, let $\mathcal{G}_{\theta}$ be a family of random network structures indexed by a vector valued parameter $\theta \in \Theta$. This is: $\mathcal{G}_{\theta}=\left\{G \sim F_{\theta} \mid \theta \in \Theta\right\}$. 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 verticesdenoted $v_{i}$ for all $i \in 1 \ldots \#\{V\}$-and to denote an edge between vertices $v_{i}$ and $v_{j}$ by $e_{i j}$. 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,

$$
A=\left[\begin{array}{c}
\mathbf{a}_{1} \\
\vdots \\
\mathbf{a}_{\mathbf{n}}
\end{array}\right]=\left[\begin{array}{ccc}
a_{11} & \ldots & a_{1 n} \\
& \ldots & \\
a_{n 1} & \ldots & a_{n n}
\end{array}\right]
$$

where

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

This is: $A(G)$ - the adjacency matrix of $G$. The sum of the $i^{\text {th }}$ 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 [Hill and Nagle (2009)]. 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:

$$
\begin{equation*}
\eta(A) \equiv S(i, j)=\left\langle\mathbf{a}_{\mathbf{i}}, \mathbf{a}_{\mathbf{j}}\right\rangle=\#\left\{i^{*}: a_{i, i^{*}}=a_{j, i^{*}}=1, i^{*}=1, \ldots, n\right\} \tag{2.1}
\end{equation*}
$$

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 nonmatch 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

$$
\begin{equation*}
H_{0}: \sigma(i)=\sigma(j) \text { vs. } H_{a}: \sigma(i) \neq \sigma(j), i=1, \ldots, n \tag{2.2}
\end{equation*}
$$

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 $\eta(A)=\left\langle\mathbf{a}_{\mathbf{i}}, \mathbf{a}_{\mathbf{j}}\right\rangle$.

For example, the probability that two nodes share an edge can be deduced from the the rewiring scheme discussed by Watts and Strogatz (1998). 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, $F_{\theta}$ can be written down exactly: in practice, of course, $\theta$ has to be estimated from data. In a hypothesis testing framework, (2.2) 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

$$
\begin{equation*}
1-F_{\theta}(s)=\operatorname{Pr}_{\theta}(S(i, j)>s) \tag{2.3}
\end{equation*}
$$

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: $\theta$, 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 (2.1) 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 reidentification 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 $\mathbb{P}\left\{a_{i, i^{*}} \cdot a_{j, i^{*}}=1\right\}=$ $\mathbb{P}\left\{a_{i, i^{*}}=a_{j, i^{*}}=1\right\}$ 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, SW and SF, don't have this accommodating property.

The probability that nodes $i$ and $j$ share a common edge $i^{*}$, $i^{*} \in\{1, \ldots, n\} \backslash\{i, j\}$, 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, $S(i, j) \sim \operatorname{Bin}\left(n-2, p^{2}\right)$. Plots of the empirical match and non-match score distributions along with the theoretical distributions derived above are shown in figure 1.


Fig 1. These plots show the fits (solid lines) of theoretical non-match score distributions to empirical data (histograms) 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 $2 k$ connected lattice. As described in Watts and Strogatz Watts and Strogatz (1998), 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, nonrewired $2 k$ 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 $\mathbb{P}\left\{a_{i, i^{*}}=a_{j, i^{*}}=1\right\}$ 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 $\|i, j\|$, $\left\|i, i^{*}\right\|$ and $\left\|j, i^{*}\right\|$ are both less than $k$. The number of times this case arises is

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

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

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

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 $2 k$ edges as $2 k$ 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., $X \sim \operatorname{Bin}(2 k, p / n)$. Then,

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

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

$$
f_{1} \sim \operatorname{Bin}\left(n_{1},\left[1-(1-p / n)^{2 k}\right]^{2}\right)
$$

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

$$
n_{2}=\left\{\begin{array}{lc}
2 k-\|i, j\|-1, & \text { if } 1 \leq\|i, j\| \leq k \\
2 k-\|i, j\|+1, & \text { if } k<\|i, j\| \leq 2 k \\
0, & \text { otherwise }
\end{array}\right.
$$

times. Post rewiring, $a_{i, i^{*}}=0$ if edge $\left(i, i^{*}\right)$ is removed and not replaced while none of the $2 k-1$ edges remaining for rewiring connect nodes $i$ and $i^{*}$. Edge $\left(i, i^{*}\right)$ is removed and not replaced with probability $p\left(\frac{n-1}{n}\right)$, while none of the $2 k-1$ edges is rewired forming edge $\left(i, i^{*}\right)$ with probability $\mathbb{P}\{Y=0\}$, where $Y \sim \operatorname{Bin}(2 k-1, p / n)$. Post rewiring, $\mathbb{P}\left\{a_{i, i^{*}}=0\right\}=p\left(\frac{n-1}{n}\right)(1-p / n)^{2 k-1}$ implying $\mathbb{P}\left\{a_{i, i^{*}}=1\right\}=1-p\left(\frac{n-1}{n}\right)(1-p / n)^{2 k-1}$. The partial non-match score distribution is

$$
f_{2} \sim \operatorname{Bin}\left(n_{2},\left[1-p\left(\frac{n-1}{n}\right)(1-p / n)^{2 k-1}\right]^{2}\right) .
$$

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

$$
n_{3}= \begin{cases}2\|i, j\|+2, & \text { if } 1 \leq\|i, j\| \leq k \\ 2\|i, j\|-2, & \text { if } k<\|i, j\| \leq 2 k \\ 4 k, & \text { if }\|i, j\|>2 k \\ 0, & \text { otherwise. }\end{cases}
$$

Without loss of generality, assume $\left\|i, i^{*}\right\|>k$ and $\left\|j, i^{*}\right\| \leq k$. After the rewiring process, we have $\mathbb{P}\left\{a_{i, i^{*}}=1\right\}=1-p\left(\frac{n-1}{n}\right)(1-p / n)^{2 k-1}$, and $\mathbb{P}\left\{a_{j, i^{*}}=1\right\}=1-(1-p / n)^{2 k}$. 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

$$
f_{3} \sim \operatorname{Bin}\left(n_{3},\left[1-p\left(\frac{n-1}{n}\right)(1-p / n)^{2 k-1}\right]\left[1-(1-p / n)^{2 k}\right]\right) .
$$

The complete non-match score distribution is found by taking the convolution of $f_{1}, f_{2}$, and $f_{3}$. In particular,

$$
\begin{aligned}
S(i, j) & \stackrel{d}{=} \sum_{y=0}^{z} f_{3}(z-y) \sum_{x=0}^{y} f_{1}(x) f_{2}(y-x) \\
& =\sum_{y=0}^{z}\binom{n_{3}}{z-y} p_{3}^{y-z}\left(1-p_{3}\right)^{n_{3}-(z-y)} \mathbb{I}\left\{z-y \leq n_{3}\right\} \\
& \times \sum_{x=0}^{y}\binom{n_{1}}{x} p_{1}^{x}\left(1-p_{1}\right)^{n_{1}-x} \mathbb{I}\left\{x \leq n_{1}\right\} \\
& \times\binom{ n_{2}}{y-x} p_{2}^{y-x}\left(1-p_{2}\right)^{n_{2}-(y-x)} \mathbb{I}\left\{y-x \leq n_{2}\right\},
\end{aligned}
$$

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.


Fig 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.
3.3. Scale-Free Networks. Scale-free networks are networks whose degree distribution follows a power law. Barabási and Albert (1999) provided a method of constructing scale-free networks based on growth and preferential attachment, however, their description is imprecise. Bollobás and Riordan
(2004) 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, $G_{1,1}$, 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

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

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 $\mathbb{P}\left\{a_{i, i^{*}}=a_{j, i^{*}}=1\right\}=\mathbb{P}\left\{a_{i, i^{*}}=1\right\} \mathbb{P}\left\{a_{j, i^{*}}=1\right\}$ 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., $\mathbb{P}\left\{a_{j, i}=1, a_{k, i}=1\right\}$, 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
$1 \leq i<j<k$,

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

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

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

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

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

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

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

yielding a final non-match score distribution of

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

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 (1997) 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 (2009) 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 $\hat{p}=\frac{\mathcal{E}^{\text {obs }}}{\binom{n}{2}}$, where $\mathcal{E}^{o b s}$ 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 (2000) showed that the mean behavior of the degree of any particular node is $2 k$, i.e., $\mathbb{E}\left[d_{i}\right]=2 k$. It is easy to see this relationship without the technical derivations. Noting that $\mathbb{E}[\bar{d}]=\mathbb{E}\left[d_{i}\right]$, 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 $t_{i}^{\text {fixed }}$ and $t_{i}^{v a r}$ 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, $X_{i} \sim \operatorname{Bin}(k, 1-p)$ and $Y_{i} \sim \operatorname{Bin}((n-2) k, p / n)$. Then we have

$$
\begin{aligned}
\mathbb{E}\left[t_{i}\right]= & \mathbb{E}\left[t_{i}^{f i x e d}\right]+\mathbb{E}\left[t_{i}^{v a r}\right] \\
= & \mathbb{E}\left[t_{i}^{f i x e d}\right]+\sum_{a=1}^{k} \sum_{b=1}^{(n-2) k} \mathbb{E}\left[t_{i}^{v a r} \mid X_{i}=a, Y_{i}=b\right] \mathbb{P}\left[X_{i}=a, Y_{i}=b\right] \\
= & \mathbb{E}\left[t_{i}^{f i x e d}\right]+\sum_{a=1}^{k} \sum_{b=1}^{(n-2) k} \mathbb{E}\left[t_{i}^{v a r} \mid X_{i}=a, Y_{i}=b\right] \mathbb{P}\left[X_{i}=a\right] \mathbb{P}\left[Y_{i}=b\right] \\
= & \sum_{l=1}^{2 k-1} l+\left(\sum_{a=1}^{k} \sum_{b=1}^{(n-2) k}(a+b) k+\sum_{c=1}^{a+b-1} c\right) \mathbb{P}\left[X_{i}=a\right] \mathbb{P}\left[Y_{i}=b\right] \\
= & \frac{k(k-1)}{2}+\left(\sum_{a=1}^{k} \sum_{b=1}^{(n-2) k}(a+b) k+\frac{(a+b)(a+b-1)}{2}\right) \\
& \times \mathbb{P}\left[X_{i}=a\right] \mathbb{P}\left[Y_{i}=b\right],
\end{aligned}
$$

where

$$
\mathbb{P}\left[X_{i}=a\right]=\binom{k}{a} p^{a}(1-p)^{k-a} \mathbb{I}\{a \leq k\}
$$

and

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

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

$$
\begin{aligned}
\mathbb{E}\left[t_{t o t}\right] & =n \frac{k(k-1)}{2} \\
& +n\left(\sum_{a=1}^{k} \sum_{b=1}^{(n-2) k}(a+b) k+\frac{(a+b)(a+b-1)}{2}\right) \mathbb{P}\left[X_{i}=a\right] \mathbb{P}\left[Y_{i}=b\right] .
\end{aligned}
$$

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

$$
\begin{aligned}
\hat{p}= & \arg \min _{0 \leq p \leq 1}\left\{\left\lvert\, t_{t o t}^{o b s}-\left(n \frac{k(k-1)}{2}+n\left(\sum_{a=1}^{k} \sum_{b=1}^{(n-2) k}(a+b) k\right.\right.\right.\right. \\
& \left.\left.\left.+\frac{(a+b)(a+b-1)}{2}\right) \times \mathbb{P}\left[X_{i}=a\right] \mathbb{P}\left[Y_{i}=b\right]\right) \mid\right\} .
\end{aligned}
$$

Barrat and Weigt (2000) 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

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

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 $1 / \mathrm{n}$ 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 nonmatch 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

$$
f_{m i x}=\sum_{i=1}^{\lfloor n / 2\rfloor} \alpha_{i} s_{i} .
$$

The $\alpha_{i}$ '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 reidentification problem. This result is not a problem, however, since the reidentification 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 $m n$ 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

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

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 (1995) and extended by Benjamini and Yekutieli (2001). Recall,

$$
F D R=\mathbb{E}\left(\left.\frac{R_{\text {false }}}{R} \right\rvert\, R>0\right),
$$

where $R$ is the number of rejections among $m$ tests and $R_{\text {false }}$ 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 $\gamma$ 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 $k=\max \left\{i: p_{(i)} \leq\left(\frac{i}{m}\right) \gamma\right\}$; and reject $H_{(1)}^{0}, \ldots, H_{(k)}^{0}$. In their work, they show that the FDR will be controlled at the level $\gamma$ 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 $\theta$ 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 (1998). In particular, we fix $k=5$ and vary $p$ over the set $\{.001, .01, .1\}$. The results are shown below.

Erdös-Rényi Networks

| $n=1000, p=.2$ | $\mathrm{FPR}=0.00615 ; \mathrm{TPR}=1$ |
| :---: | :---: |
| $n=1000, p=.5$ | $\mathrm{FPR}=0.00501 ; \mathrm{TPR}=1$ |
| $n=1000, p=.8$ | $\mathrm{FPR}=0.00570 ; \mathrm{TPR}=1$ |

$$
\begin{aligned}
& \text { Watts-Strogatz Networks } \\
& \begin{array}{|l|l|}
\hline n=1000, k=5, p=.001 & \text { FPR }=0.00100 ; \text { TPR }=0.99996 \\
\hline n=1000, k=5, p=.01 & \text { FPR }=0.00108 ; \mathrm{TPR}=0.99894 \\
\hline n=1000, k=5, p=.1 & \text { FPR }=0.00082 ; \mathrm{TPR}=0.94882 \\
\hline
\end{array}
\end{aligned}
$$

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 $\mathrm{TPR} \approx .5$-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 publicly available datasets: the Reality Mining Group at MIT which was introduced by Eagle and Pentland (2006) 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, Goodreau and Handcock (2008) 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 3.

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

| Erdös-Rényi | $p=.0411$ | $\mathrm{TPR}=.4167, \mathrm{FPR}=.00469$ |
| :---: | :---: | :--- |
| Watts-Strogatz | $p=.9326, k=1$ | $\mathrm{TPR}=.4167, \mathrm{FPR}=.00469$ |
| Barabási-Albert | $m=1$ | $\mathrm{TPR}=.7083, \mathrm{FPR}=.02152$ |

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


FIG 3. This plot shows the call patterns in the restricted Reality Mining dataset.
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 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 4. The table below shows the parameter


FIG 4. This plot shows the email patterns of 144 top Enron executives.
estimates for each of the network constructions along its associated FPR and TPR.

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

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.
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 [Hill and Nagle (2009)].

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?
We have derived overlap-based 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.

## References.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. Science 286 509-512.
Barrat, A. and Weigt, M. (2000). On the properties of small-world network models. Europ.Phys.J.B 13547.
Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B (Methodological) 57 289-300.
Benjamini, Y. and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics 29 1165-1188.
Bollobás, B. and Riordan, O. (2004). The Diameter of a scale-free random graph. Combinatorica 24 5-34.
Cortes, C., Pregibon, D. and Volinsky, C. (2003). Computational methods for dynamic graphs. Journal of Computational E Graphical Statistics 12 950-970.
Eagle, N. and Pentland, A. (2006). Reality Mining: Sensing Complex Social Systems. Personal and Ubiquitous Computing 10 255-268.
Erdös, P. and Rényi, A. (1959). On random graphs. Publ. Math. Debrecen 209 290-297.
Goldenberg, A., Zheng, A. X., Fienberg, S. E. and Airoldi, E. M. (2009). A Survey of Statistical Network Models. Foundations and Trends in Machine Learning 2 129-233.
Gopal, S. (2007). The Evolving Social Geography of Blogs. Berlin:Springer.
Hill, S. and Nagle, A. (2009). Social Network Signatures: A Framework for Reidentification in Networked Data In Proceedings of the Internatinal Conference on Computational Aspects of Social Networks. IEEE Press, Fontainebleau, France.

Hill, S. and Provost, F. (2003). The myth of the double-blind review?: author identification using only citations. SIGKDD Explor. Newsl. 5 179-184.
Hill, S. B., Agarwal, K. K., Bell, R. and Volinsky, C. (2006). Building an effective representation for dynamic networks. Journal of Computational \& Graphical Statistics 15 584-608.
Hoff, P., Raftery, A. and Handcock, M. (2002). Latent Space Approaches to Social Network Analysis. Journal of the American Statistical Association 97 1090-1098.
Hunter, D., Goodreau, S. and Handcock, M. (2008). Goodness of Fit of Social Network Models. Journal of the American Statistical Association 103 248-258.
Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28 27-30.
Kolaczyk, E. (2009). Statistical Analysis of Network Data: Methods and Models. Springer, New York, New York.
McKay, B. D. and Wormald, N. C. (1997). The Degree Sequence of a Random Graph. 1. The Models. Random Stuctures and Algorithms 11 97-117.

Newman, M., Barabsi, A. L. and Watts, D. J. (2006). The Structure and Dynamics of Networks. Princeton University Press.
Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature 393 440-442.
J. Vastola
K. Abayomi

Georgia Institute of Technology
765 Ferst Drive
Atlanta, Georgia, 30332
USA
E-MAIL: jvastola@gatech.edu kobi@gatech.edu
кооregatecn.eau
S. Hill

University of Pennsylvania
3730 Walnut Street, Suite 500
Philadelphia, PA 19104
USA
E-mAIL: shawndra@wharton.upenn.edu


[^0]:    *Supported in part by the Office of Naval Research (ONR) MURI grant number N00014-08-1-0747

