

Statistics for Re-identification in Social Networks

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

August 17, 2010

Abstract

This paper investigates statistical re-identification methodology for social network data: data that can be mapped to a graph of nodes and edges. These graphs may be viewed, as random objects, as instantiations of a network model. We specifically address the reidentification problem as a testing procedure on the inner product space of the observed graph, for a particular model. We illustrate the methodology on three well known networks—Erdős-Rényi random networks, Watts-Strogatz small world networks, and Barabási-Albert scale free networks.

1 Introduction

Complex networks exist extensively in nature and society. Applications range from describing the regulatory behavior among genes to the discovery relationships in social networks. Traditionally, these graphs have been modeled by the random graphs of Erdős and Rényi (ER) [5], however, the complex structures arising from real world networks reach beyond the descriptive capabilities of ER models. Watts and Strogatz [7] and Barabási and Albert [1] introduced small-world and scale-free network structures, respectively, to overcome some of the downfalls of the ER models. Extensions and hybrids of these models exist. While many model representations are stochastic in nature, “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” [6].

Statistical inference methodology is present (though) infrequently in network literature. Most work has not been done for the three graphs above and focuses more on simulation techniques as opposed to a rigorous statistical study. Exponential random graph models (ERGMs) were created to reasonably represent network data while allowing for precise statistical inference. Our goal is not to study ERGMs but to provide a statistical treatment of ER, small-world, and scale-free networks for the purpose of re-identification.

2 Methodology

Much network research today focuses on the generation of networks, both stochastically or deterministically, in a quest to describe well known phenomena like the internet or gene structure. A specific rewiring algorithm generates Watts and Strogatz’s “small worlds”, while preference and growth construct Barabasi and Albert’s scale free networks. Our approach, though not new, provides a foundation from which inferential procedures take root from these generative algorithms. We view individual graphs G_θ as draws from a family of graphs \mathcal{G}_θ based on a specified probability distribution \mathbb{P}_θ in the same way that a random variable X is drawn from a sample space χ based on the probability distribution F_X . Network distributions \mathbb{P}_θ are typically elusive, usually implicitly arising from their algorithmic construction rather than a complete, explicit specification. Under this formulation, G_θ is a random object for which we can adopt traditional statistical techniques for analyzing its structure.

From G_θ , we define statistics $\eta(G_\theta)$ in the same way that we define statistics on random samples $X \sim F_X$. Example network statistics include the clustering coefficient, diameter, girth, etc. In this paper, we consider an *overlap score* statistic as follows. Let G_θ be a graph from a family \mathcal{G}_θ with corresponding adjacency matrix $A_\theta = [a_{ij}]_{i,j=1}^n$, where $a_{ij} = 1$ if an edge between i and j exists and 0 if no edge exists. The *overlap score* is

$$\begin{aligned}\eta(G_\theta) &= \mathbf{S}_\theta(\mathbf{a}_i, \mathbf{a}_j) = \langle \mathbf{a}_i, \mathbf{a}_j \rangle \\ &= \# \{i^* : a_{i,i^*} = a_{j,i^*} = 1, i^* = 1, \dots, n\},\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ stands for the usual dot product. When $i = j$, we call $\mathbf{S}_\theta(\cdot)$ the *match score*, and when $i \neq j$, we call $\mathbf{S}_\theta(\cdot)$ 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 is, naturally extending to hypothesis testing.

Network construction algorithms are sufficient for finding the limiting score distributions, denoted F_{S_θ} . Therefore, given a network construction, statements about the likelihood of an

observation can be made. Hypothesis tests formalize such statements.

For the re-identification problem, we are interested in whether or not two nodes have the same *signature*, i.e., represent the same entity. 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 the phone user. Denote the signature of node i by $\sigma(i)$. The non-match score gives a measure of similarity between two nodes, in turn, providing a quantitative measure of the similarity between nodes. Under the assumption that nodes i and j have different signatures, $\mathbf{S}_\theta(\mathbf{a}_i, \mathbf{a}_j) \sim F_{S_\theta}$. We can then test $H_0 : \sigma(i) = \sigma(j)$ vs. $H_1 : \sigma(i) \neq \sigma(j)$, $i = 1, \dots, n$. Theoretically, F_{S_θ} is known completely, but in practice, θ has to be estimated from data. Where necessary, we will provide estimation techniques.

3 Score Distributions

We apply the methodology in section 2 to the three most prevalent network structures—ER, small-world, and scale-free. For each structure, we derive *match* and *non-match score* distributions from the constructions of family \mathcal{G}_θ .

3.1 Erdős-Rényi Networks

An Erdős-Rényi random network is a class of random networks, \mathcal{G}_θ , in which a fixed number of nodes, n , is populated based on independent Bernoulli trials with probability of success, p . Given n , the parameter $\theta = p$ completely defines \mathcal{G}_θ . The degree, and thus, *match score*, distribution is $\mathbf{S}_p(\mathbf{a}_i, \mathbf{a}_i) \sim \text{Bin}(n - 1, p)$.

To derive the *non-match score* distribution for nodes i and $j, i \neq j$, we consider the construction of the network. 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 Erdős-Rényi 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 $i^*, i^* \in \{1, \dots, n\} \setminus \{i, j\}$, is

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

since the two connections are made independently with probability p . We restrict node i^* from equalling 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, $\mathbf{S}_p(\mathbf{a}_i, \mathbf{a}_j) \sim \text{Bin}(n - 2, p^2)$. Figure 1 shows our theoretical distribution plotted against simulated data for varying p .

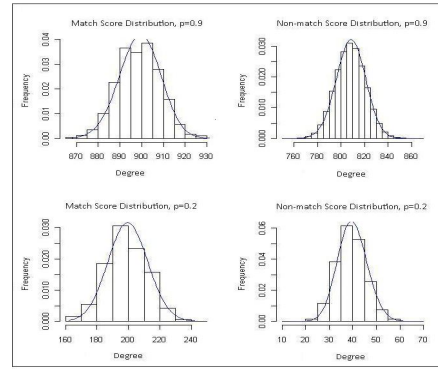


Figure 1: Plots of theoretical score distributions versus simulated distributions for Erdős-Rényi networks

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}_θ can be generated as a rewiring of a $2k$ connected lattice. As described in Watts and Strogatz [7], the lattice is algorithmically rewired with probability parameter p . Here, $\theta = (p, k)$.

The *match score* has the following distribution:

$$\mathbf{S}_{p,k}(\mathbf{g}_i, \mathbf{g}_i) \stackrel{d}{=} \sum_{n=0}^{\min(d-k,k)} \binom{k}{n} (1-p)^n p^{k-n} \frac{(kp)^{d-k-n}}{(d-k-n)!} e^{-pk},$$

the degree distribution derived by Barrat and Weigt [3].

The small world characterization yields a limiting distribution for the *non-match score* distribution. Consider the graph 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 whether or not i and j are connected pre-rewiring, which occurs in three ways. 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 a score, each case arises, and 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. We show that the final distribution is

$$\begin{aligned} \mathbf{S}_{p,k}(\mathbf{a}_i, \mathbf{a}_j) &\stackrel{d}{=} \sum_{y=0}^z f_3(z-y) \sum_{x=0}^y f_1(x) f_2(y-x) \\ &= \sum_{y=0}^z \binom{n_3}{z-y} p_3^{y-z} (1-p_3)^{n_3-(z-y)} \\ &\times \sum_{x=0}^y \binom{n_1}{x} p_1^x (1-p_1)^{n_1-x} \\ &\times \binom{n_2}{y-x} p_2^{y-x} (1-p_2)^{n_2-(y-x)}, \end{aligned}$$

where each n_i and p_i , $i = 1, 2, 3$, depends on the pre-wiring cases. Figure 2 shows the theoretical distribution plotted against simulated histograms for varying cases.

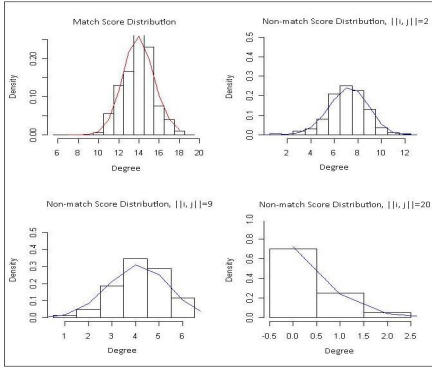


Figure 2: Plots of theoretical score distributions versus simulated distributions for Watts-Strogatz networks

3.3 Scale-Free Networks

Barabási and Albert [2] provided a method of constructing scale-free networks based on growth and preferential attachment, however, their description is imprecise. Bollobás and Riordan [4] remedy this issue by precisely specifying the model of Barabási and Albert. A result summarizing the complete degree distribution—and thus, the match score distribution—is not known. Since our methodology only requires the non-match score, this fact isn’t restricting.

For the non-match score distribution, we first consider the base case when $m = 1$. 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. As 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, the independence is erroneous for the scale-free network model. In particular, this dependence structure requires the consideration of whether $i < j < k$ or $j < i < k$.

For the first case, Bollobás and Riordan show that $\mathbb{P}\{a_{i,j} = 1, a_{i,k} = 1\} = O(i^{-1}(jk)^{-1/2})$ in [4]. It is not too difficult to show based on their work that

$$\mathbb{P}\{g_{i,j} = 1, g_{i,k} = 1\} = \frac{4i + 2}{(2k - 1)(4i^2 - 1)} \prod_{s=j+1}^{k-1} \left(\frac{2s}{2s - 1}\right).$$

Consider the second case when $1 \leq j < i < k$. By modifying the work of Bollobás and Riordan, we calculate

$$\mathbb{P}\{g_{i,j} = 1, g_{i,k} = 1\} = \frac{1}{(2i)(2k - 1)} \prod_{s=j}^{k-1} \left(\frac{2s}{2s - 1}\right).$$

For each combination of i, j , and k , we have a different partial non-match score distributions. We will denote each distribution by f_{ijk} , where $f_{ijk} \sim \text{Bernoulli}(p_{jk}^i)$ and p_{jk}^i denotes the probability that both nodes j and k are connected

to node i . Let X_{ijk} denote a random variable drawn from distribution f_{ijk} . Fix $j = j^*$ and $k = k^*$. The non-match score distribution, similar the that of the small world networks, is a convolution of the of the random variables $X_{ij^*k^*}$, i.e., the distribution of $Z = \sum_i X_{ij^*k^*}$ —a distribution not easily written out.

4 Hypothesis Testing and Parameter Estimation

Making decisions about a persons signature is formalized through statistical hypothesis testing via tests of the form $H_0 : \sigma(i) = \sigma(j)$ vs. $H_1 : \sigma(i) \neq \sigma(j)$. For each network type, the essence of the testing procedure is identical, however, the assumed underlying model yields its own complications to overcome, namely, parameter estimation. We cannot appropriately address hypothesis testing without discussing parameter estimation.

Erdős-Renyí graphs, as the simplest graph, are relatively easy to deal with. One parameter, p , needs to be estimated to completely characterize the nonmatch score distribution. On the other hand, the Watts-Strogatz model introduces different parameters. Beyond p , k and the pre-wiring ordering of the nodes must be estimated. Barabási and Albert’s model requires the estimation of m and the order in which the vertices enter. With proper estimates, we test whether two nodes in a graph have the same signature.

References

- [1] Albert-Lázló Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [2] Albert-Lázló Barabási, Erzsébet Ravasz, and Tamás Vicsek. Deterministic scale-free networks. *Physica A*, 299:559–564, 2001.
- [3] A. Barrat and M. Weigt. On the properties of small-world network models. *Europ.Phys.J.B*, 13:547, 2000.
- [4] Béla Bollobás and Oliver Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, 2004.
- [5] P. Erdős and A. Rényi. On random graphs. *Publ. Math. Debrecen*, 209:290–297, 1959.
- [6] Eric Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, New York, New York, 2009.
- [7] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.