On the Estimation and Prediction of Tie Strength in Social Networks

On the Estimation and Prediction of Tie Strength in Social Networks

A dissertation presented

by

Heather Mattie

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

April 2017

Dissertation Advisor: Professor Jukka-Pekka Onnela Heather Mattie

On the Estimation and Prediction of Tie Strength in Social Networks

Abstract

Humans interact with each other both online and in-person, forming and dissolving social ties throughout our lives. The flexible architecture of networks or graphs make them a useful paradigm for modeling these complex relationships at the individual, group, and population levels. Social networks have been shown to have a direct impact on public health from leveraging network properties to target highly connected individuals in public health interventions to finding that households that refuse to have their children vaccinated against polio have a disproportionate number of social ties to other vaccine-reluctant and vaccine-refusing households. Social network data has traditionally been collected from surveys, mostly capturing small, static network snapshots at one point in time. Dozens of different metrics have been created to quantify and study the structure of these simple networks. However, with the recent availability of increasingly rich, complex network data, limitations of these metrics have become increasingly clear. In the first chapter of this dissertation, we extend definitions of edge overlap, the proportion of friends two connected individuals share, to weighted and directed networks, and we present closed-form expressions for the mean and variance of each version for the classic Erdős-Rényi random graph and its weighted and directed counterparts. We apply these results to social network data collected in rural villages in India, and we use our analytical results to quantify the extent to which the average edge overlap in the empirical social networks deviates from that of corresponding random graphs. Finally, we carry out comparisons across attribute categories including sex, caste, and age, finding that women tend to form more tightly clustered friendship circles than men, where the extent of overlap depends on the nature of social interaction in question.

In social networks the notion of tie strength, and the factors that influence it, have received much attention in a myriad of disciplines for decades. With the internet and cellular phones providing additional avenues of communication, measuring and inferring tie strength has become much more complex. Measuring and predicting tie strength, and moreover, understanding the factors that drive tie strength, has been an expanding area of interest, with increasing utility and complexity in the digital age, i.e., the ever-increasing forms of communication via mobile phones and social media. Knowledge of the strength of a tie, as well as the social dynamics contributing to tie strength, has been shown to increase the accuracy of link prediction, enhance the modeling of the spread of disease and information, and lead to more targeted marketing. Numerous models incorporating indicators of tie strength have been proposed and used to quantify relationships in both online and offline social networks, and a standard set of structural network metrics have been applied to predominantly online social media sites to predict tie strength. The second chapter of this dissertation details tie strength prediction methodology. We introduce the concept of the "social bow tie" framework, which for any given network tie is a small subgraph of the network that consists of a collection of nodes and ties that surround the tie of interest, forming a topological structure that resembles a bow tie. We also define several intuitive and interpretable metrics that quantify properties of the bow tie which enable us to investigate associations between the strength of the "central" dyadic tie and properties of the bow tie. We combine the bow tie framework with machine learning to investigate what aspects of the bow tie are most predictive of tie strength in two very different types of social networks, a collection of medium-sized social networks from 75 rural villages in India and a nationwide call network of European mobile phone users. For two connected individuals, we find that the more their friendship circles overlap, the stronger the tie between them. Conversely, the more close-knit each individual's separate friendship network, the weaker the tie between them. Our findings also demonstrate that incorporating properties of the bow tie results in more accurate predictions of tie strength and a more nuanced understanding of the factors that are associated with it.

Missing data and non-response are common occurrences in, and great hindrances to, the analysis of social network data. While any kind of statistical analysis can be negatively affected by missingness, the effects can be even more detrimental in network data analysis due to the high sensitivity of missing data on network topology and the complexity of network surveys and data collection. Many imputation methods have been introduced in the classical statistics literature as a way to maintain power and sample size in the presence of missing data. However, the extension of these methods to the networks framework has been scarcely studied. The third chapter of this dissertation addresses the issue of missing data in statistical analyses of network data. We use Super Learner to impute both edge and nodal attributes of a nationwide call network of European mobile phone users with varying amounts of missingness. We impute the age, age category, and sex of individuals, and the total call duration and text message communication between two individuals over a three-month time period. We find that Super Learner performs better or as well as any individual learning algorithm alone for the imputation of each attribute, and that the amount of missingness does not significantly affect performance. Additionally, we find that the accuracy of age category imputation is sensitive to the choice of categorical thresholding. A thresholding scheme that results in approximately equal proportions of individuals in each category ensures a gain in age-stratified accuracy over the null accuracy of random assignment, but a lower overall accuracy when compared to thresholding resulting in imbalanced categories.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation is dedicated to Poppy and Grammy: my grandparents, Bud and Connie Leiva. I think of you and miss you both every day. You taught me how important education and knowledge are, but also how to be respectful, humble and kind. You supported me even before I was born and always made sure I had what I needed. You are the reason I have been so successful and have had such an incredible life and amazing opportunities offered to me. While I didn't become an engineer as you'd hoped, I hope that I still make you proud. I love you more than you'll know and will keep you in my heart forever.

I also want to thank my two biggest fans: momma and dada. You too taught me the importance of education, kindness and leading a balanced life. Your never-ending support and belief in my abilities has been invaluable, and I'm not sure I ever thanked you for that. Knowing that I always had someone I could call and who cared about me helped pull me through the toughest times. I love you around the world and back again and can't wait to celebrate with you.

Finally, I'd like to thank Ellie Dimopoulos. I literally wouldn't be here if you hadn't entered my life. You saw my potential the moment you met me and convinced me to apply. Then you sacrificed everything to ensure my success and completion of this degree. You'll never know the depth of my gratitude and I hope to spend the rest of my life showing you how much you changed my life for the better.

1

Generalizations of Edge Overlap to Weighted and Directed Networks

Abstract

With the increasing availability of behavioral data from diverse digital sources, such as social media sites and cell phones, it is now possible to obtain detailed information on the strength and directionality of social interactions in various settings. While most metrics used to characterize network structure have traditionally been defined for unweighted and undirected networks only, the richness of current network data calls for extending these metrics to weighted and directed networks. One fundamental metric, especially in social networks, is edge overlap, the proportion of friends shared by two connected individuals. Here we extend definitions of edge overlap to weighted and directed networks, and we present closed-form expressions for the mean and variance of each version for the classic Erdős-Rényi random graph and its weighted and directed counterparts. We apply these results to social network data collected in rural villages, and we use our analytical results to quantify the extent to which the average edge overlap in the empirical social networks deviates from that of corresponding random graphs. Finally, we carry out comparisons across attribute categories including sex, caste, and age, finding that women tend to form more tightly clustered friendship circles than men, where the extent of overlap depends on the nature of social interaction in question.

Term	Notation	Description
adjacency matrix	A	A square matrix whose elements A_{ij} have a value different from 0 if there is an edge from some node i to some node j. $A_{ij} = 1$ if the link is a simple connection (unweighted graph). $A_{ij} = w_{ij}$ when the link is assigned some kind of weight (weighted graphs). If the graph is undirected (links connect nodes symmetrically), A is symmetric.
degree	k_i	The number of nodes a node i is connected to
in-degree	k_i^{in}	In a directed network, the number of incoming edges to a node i
out-degree	k_i^{out}	In a directed network, the number of outgoing edges emanating from a node i
weight	w_{ij}	In a weighted network, weight assigned to an edge from some node i to some node j
strength	$s_i = \sum_{j=1}^{k_i} w_{ij}$	The sum of weights attached to ties belonging to some node i
Erdős-Rényi	$G(n, p)$	A random graph of n nodes and edges generated by connecting a pair of nodes with some
random graph model		probability p independently of all other edges
Call Detail Records	CDRs	Digital records of the attributes of a certain instance of a telecommunication transaction (such as the start time or duration of a call), but not the content.

Table 1.1: Networks Terminology and Notation

1.1 Introduction

Humans interact with each other both online and in-person, forming and dissolving social ties throughout our lives. The flexible architecture of networks or graphs make them a useful paradigm for modeling these complex relationships at the individual, group, and population levels. Social network nodes typically represent individuals, and edges the connections between individuals, such as friendships, sexual contacts, or cell phone calls. Social networks have been shown to have a direct impact on public health Christakis and Fowler (2007, 2008); Fowler and Christakis (2008a,b); Goodreau et al. (2009). For example, a recent study examined the social networks of households in Malegaon, India, finding that households that refuse to have their children vaccinated against polio have a disproportionate number of social ties to other vaccine-reluctant and vaccine-refusing households Onnela et al. (2016). Several studies have now successfully modeled the spread of epidemics through various populations, finding that different network structures have an effect on the potential efficacy of an intervention Banerjee et al. (2013); Valente (2005); VanderWeele (2011). Studies have also leveraged network properties to target highly connected individuals in public health interventions Kim et al. (2015). The structure of connections in contact networks have also been shown to affect statistical power in cluster randomized trials Banerjee et al. (2013); Staples et al. (2015). Additionally, new classes of connectivity-informed study designs for cluster randomized trials have been proposed recently, and these designs appear to simultaneously improve public health impact and detect intervention effects Banerjee et al. (2013); Harling and Onnela (2016); Kim et al. (2016).

There is also accumulating evidence that the habits of our friends influence our own behavior, such as the uptake of smoking or lifestyle choices that can lead to obesity Christakis and Fowler (2007, 2008); Fowler and Christakis (2008a,b). Moreover, electronic billing records have been used to study patient-physician interaction networks to learn about structural properties of these networks and how these properties are associated with the quality and cost of health care Kim et al. (2016); Landon et al. (2012); Sima et al. (2010).

Network structure can be studied at different scales ranging from local to global. Microscopic (local) structures include one or a few nodes, macroscopic (global) structures involve most to all nodes, and mesoscopic structures lie between the microscopic and macroscopic scales. It has been shown that the different structures are not independent Fortunato (2010). Specifically, several microscopic mechanisms are known to give rise to microsopic, mesoscopic, and macroscopic structure Bianconi et al. (2014); Fortunato (2010); Kumpula et al. (2007). For example, triadic closure, the process of getting to know a friend of a friend, can generate network communities Fortunato (2010); Kumpula et al. (2007); Porter et al. (2009). The term community here refers to a group of nodes that are densely connected to one another but only sparsely connected to the rest of the network. Community structure is of particular interest because most social networks have meaningful community structure that is related to their function. Communities also arise from humans forming tightly-knit groups through shared interests and similar characteristics, and they play an important role in the spread of disease and information Christakis and Fowler (2007, 2008); Fortunato (2010).

Social network data has traditionally been collected from surveys, mostly capturing small, static network snapshots at one point in time Wasserman and Faust (1994). Dozens of different metrics have been created to quantify and study the structure of these simple networks. However, with the recent availability of increasingly rich, complex network data, limitations of these metrics have become increasingly clear. For example, betweenness centrality, the number or proportion of all pairwise shortest paths in a network that pass through a specified node, is used quite broadly but becomes much more computationally demanding as the size of the network increases and, even more importantly, it is unclear how meaningful this metric is in very large social networks. Another example of a widely used metric is the clustering coefficient, which is defined as the fraction of paths of length two in the network that are closed, i.e., groups of three nodes where "the friend of my friend is also my friend" Watts and Strogatz (1998).

The clustering coefficient has subsequently been extend to weighted and directed networks Saramaki et al. (2007); Tore (2013). For the classic Erdős-Rényi random graph, the local clustering coefficient (the average clustering coefficient taken across all nodes in the network) asymptotically tends to p where p is the probability of forming a tie between any two nodes in the network Reinert (2012). Most social networks are more clustered than corresponding random networks Newman (2003, 2010). This observation is expected since people are more likely to become friends with others whom they meet through their current friends. While an expression has been derived for the mean of the local clustering coefficient, an expression for the variance has not been presented. Thus, classification of a given value for clustering as either high or low, and whether that value is statistically significant, is not currently possible and its value cannot be compared across networks.

The rest of this chapter is organized as follows. In Section 1.2, we introduce the microscopic metric known as edge overlap and define extensions of edge overlap for weighted and directed networks. We then present two closed-form expressions for the mean and variance of each version of edge overlap for the Erdős-Rényi random graph and its weighted and directed counterparts. We then demonstrate the accuracy of our mean and variance approximations through simulation. Finally, we apply our results to empirical social network data and quantify the difference in the observed average overlap to the value expected for a corresponding random graph. We present the results of our data analysis in Section 1.3 and discuss our conclusions in Section 1.4. Derivations and additional figures are presented in sections 1.5, 1.6, 1.7 and 1.8.

1.2 Methods

1.2.1 Overlap Extensions

A central microscopic metric, which captures the effect of triadic closure and is related to the clustering coefficient, is edge overlap, the proportion of common friends two connected individuals share. In mathematical terms, the overlap between two connected individuals i and j is defined as

$$o_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}} \tag{1.1}$$

where n_{ij} is the number of common neighbors of nodes i and j, and k_i (k_j) denotes the degree, or number of connections, node i (j) has. Note that the tie between nodes i and j is not included in the calculation; overlap for the edge (i, j) is defined only where $A_{ij} = 1$ and $k_1 + k_j > 2$. Currently, edge overlap is only defined for simple networks in which edges are both unweighted and undirected Onnela et al. (2007). Moreover, expressions for the mean and variance of edge overlap do not yet exist, making it hard to carry out statistical comparisons of this metric across networks, in particular networks of different sizes.

In a weighted network, each edge has a weight assigned to it. We define weighted overlap in Eq. (1.2) as the proportion of total weight associated with ties to common friends nodes i and j share, and denote it o_{ij}^W:

$$o_{ij}^W = \frac{\sum_{k=1}^{n_{ij}}(w_{ik} + w_{jk})}{s_i + s_j - 2w_{ij}}. \tag{1.2}$$

Here, n_{ij} is the number of common neighbors of nodes i and j, w_{ij} denotes the weight associated with the tie between nodes i and j, and s_i (s_j) denotes the strength of node i (j). According to the definition, the common friends of two connected individuals are first identified, the weights associated with these edges are summed together, and this sum is then divided by the combined strengths of the two nodes excluding the tie that connects them. The last step is intended to ensure consistency with the original version of edge overlap, i.e., the weight of the tie between the two individuals being considered is not included in the calculation of o_{ij}^W. Also, the metric is only defined for $w_{ij} > 0$ and for $s_i + s_j > 2w_{ij}$.

In a directed network, each edge has a direction associated with it. Thus, ties between nodes can be reciprocated, meaning that there can be an edge pointing from node i to j and another edge pointing from j to i. For directed networks, the concept of a 'common friend' of two individuals is ambiguous due to the directionality associated with the ties. We define a common friend as a node that creates a directed path of length two between the two nodes either from i to j, j to i, or both. Defining a common friend in this manner allows information to flow between i and j via a neighbor of both i and j. To illustrate this, let i and j be the two connected individuals of interest, and k a potential common friend. If there is a directed edge from i to k and a directed edge from k to j, then there is a path a length two from i to j through k, and k is considered a common friend. Using this criterion, we define directed overlap in Eq. (1.3) as the proportion of paths of length two between two connected individuals, and denote it o_{ij}^D:

$$o_{ij}^D = \frac{\sum_{k=1}^n (A_{ik}A_{kj} + A_{jk}A_{ki})}{\min(k_i^{\text{in}}, k_j^{\text{out}}) + \min(k_j^{\text{in}}, k_i^{\text{out}}) - 1}. \tag{1.3}$$

Here, A_{ij} is the (i, j) element of the directed adjacency matrix, k_i^{in} (k_j^{in}) denotes the in-degree of node i (j), k_i^{out} (k_j^{out}) denotes the out-degree of node i (j), and $\min(\cdot, \cdot)$ the minimum of the two arguments. We consider each edge separately, even in the case of unreciprocated edges, and again, the tie between nodes i and j is not included in the calculation. The metric is only defined if $\min(k_i^{\text{in}}, k_j^{\text{out}}) + \min(k_j^{\text{in}}, k_i^{\text{out}}) > 1$.

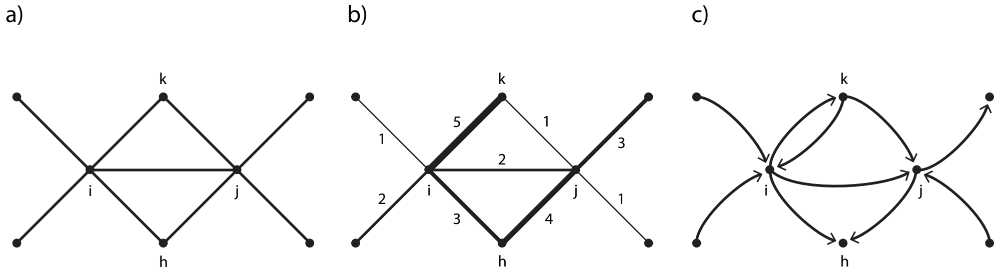


Figure 1.1: Schematics of edge overlap for (a) an unweighted network, (b) weighted network, and (c) directed network. Nodes are labeled with letters and weights are labeled with numbers.

1.2.2 Erdős-Rényi Random Graph Models

With the extensions of edge overlap defined above, one can easily compute the mean overlap (simple or weighted or directed) across all edges in the network. However, in order to make meaningful comparisons, such as to learn whether the observed value is small or large for the given network, or whether it represents a statistically significant deviation from what might be expected to occur at random, one needs to consider suitable null models and derive both the expected value and the variance of overlap under these null models. The Erdős-Rényi random graph model, often denoted $G(n, p)$, is the simplest model for generating random graphs Erdős and Rényi (1959). In this model, graphs are created by considering $\binom{n}{2}$ distinct pairs of n nodes and connecting each pair with probability p independently of all other dyads (node pairs). The random process can therefore be thought of as a series of Bernoulli trials or coin flips. Suppose we have a coin that lands on heads with probability p. If the coin flip results in heads, the two nodes are connected, otherwise, they are not. Note that here the number of edges is not fixed, but rather the probability of creating an edge.

The weighted random graph (WRG) is the weighted counterpart of the canonical Erdős-Rényi random graph Garlaschelli (2009). In this case, a network of n nodes is generated by selecting each pair of nodes and carrying out a series of independent Bernoulli trials for each pair with success probability p. This process is continued until the first failure is encountered, and every success preceding the failure adds a unit weight to the tie. Note that if the first Bernoulli trial is a failure, the two nodes will not be connected. We can again relate this to the tossing of a coin. If the coin lands on heads with probability p, the weight associated with an edge is given by the number of heads flipped until the first tails appears, and therefore tie weights are distributed according to the geometric distribution. This process is repeated for every distinct pair of nodes in the network.

The directed random graph is the directed version of the Erdős-Rényi random graph, and it is generated in a very similar manner as its canonical counterpart. For two nodes i and j, in a network of n nodes, an edge pointing from i to j is created with probability p and, likewise, an edge pointing from j to i is also connected independently with probability p Bollobás (1985); Erdős and Rényi (1959, 1960). In this case, in the coin analog of the model, we flip a coin twice for each pair of nodes, one flip for each direction. This process is repeated for every pair of nodes in the network.

1.2.3 Erdős-Rényi Overlap

In order to perform inference about overlap, i.e., to compare point estimates of overlap across networks, we need to know the mean and variance of each version of overlap under the null model in question. To fix our notation, we will let uppercase letters stand for random variables: K_i denotes the degree of node i, N_{ij} the number of common neighbors of nodes i

and j, S_i the strength of node i, W_{ij} the weight of the edge connecting nodes i and j, K_i^{in} the in-degree of node i, K_i^{out} the out-degree of node i, and A_{ij} the adjacency matrix element, where a nonzero (positive) value represents the existence of an edge between nodes i and j (binary in the case of unweighted graphs).

For the Erdős-Rényi random graph, a given node is connected to each of the remaining $n-1$ nodes with probability p, and its resulting degree can thus be viewed as a sum of independent Bernoulli trials. Therefore, as is well known, $K_i \sim \text{binomial}(n-1, p)$, which can be approximated by a Poisson(np) distribution for large n. For any pair of (connected) nodes, the probability of both nodes being connected to the same neighboring node, meaning that they have a common neighbor, is p^2 as each edge occurs independently of any others. Moreover, the total number of possible common friends two nodes can have is $n-2$. Thus, $N_{ij} \sim \text{binomial}(n-2, p^2)$, which can similarly be approximated by a Poisson(np^2) random variable for large n. With these definitions, the numerator of edge overlap is a Poisson random variable, and the denominator is the difference of two Poisson random variables, known as a Skellam random variable Skellam (1946). In this case, the denominator is a Skellam$(2np - 2 - np^2)$ random variable. We can now view overlap as a random variable as in Eqn. (1.4).

$$O_{ij} = \frac{N_{ij}}{(K_i - 1) + (K_j - 1) - N_{ij}} \tag{1.4}$$

Edge overlap is a ratio of two dependent random variables since the maximum number of possible common friends is bounded by the $\min(K_i, K_j)$. This dependency increases the difficulty of deriving exact expressions for the mean and variance of overlap. However, despite this dependence, we can approximate both the mean and variance in two different ways. The first approach observes the weakness of the dependence between the numerator and denominator and simply ignores it, defining the ratio as a function of independent random variables. Approximations for the mean and variance of the ratio are then derived using Taylor expansions of the function about the means of the random variables Elandt-Johnson and Johnson (1998); Stuart and Ord (1998). This results in Eqs. (1.5) and (1.6) (for details, see section 1.5.1).

$$\mathbb{E}[O_{ij}] = \frac{p}{2-p} \tag{1.5}$$

$$\mathrm{Var}(O_{ij}) = \frac{np^2}{(2np-2-np^2)^2} + \frac{n^2p^4(2np-2+np^2)}{(2np-2-np^2)^4}. \tag{1.6}$$

Our second approach incorporates results from Lin (2007), where the local clustering coefficient for an Erdős-Rényi random graph is also written as a ratio of dependent random variables with the intention of deriving its distribution. The dependency is eliminated by replacing the random variable in the denominator with its expectation, and this approximation turns the denominator into a constant. Thus, the distribution of the clustering coefficient is approximated by a scaled version of the random variable in the numerator. It is subsequently shown that this is a good approximation for the actual distribution. We adopt the same approach here, and approximate the distribution of edge overlap by replacing the denominator with its expectation. We then derive the mean and variance of O_{ij} using the distributional properties of the numerator. This results in the expressions in Eqs. (1.7) and (1.8) (for details, see section 1.6.1):

$$\mathbb{E}[O_{ij}] = \frac{p}{2-p} \tag{1.7}$$

$$\mathrm{Var}(O_{ij}) = \frac{np^2}{(2np-2-np^2)^2}. \tag{1.8}$$

Note that the expressions for the mean, Eqs. (1.5) and (1.7), are equivalent, but the expressions for the variance, Eqs. (1.6) and (1.8) differ, with the expression for Eq. (1.8) corresponding to the first term of Eq. (1.6).

We use the same two approaches for the weighted and directed cases. For the weighted Erdős-Rényi random graph (WRG), we first define the distributions of W_{ij} and S_i. Given how WRGs are constructed (as given above), the tie weights follow a geometric distribution, such that if an edge is placed between a pair of nodes with probability p, tie weight distribution will be $W_{ij} \sim$ geometric$(1-p)$. It then follows that node strength S_i is a sum of geometric random variables, i.e., is the sum of the weights of the ties that are adjacent to the given node, leading to $S_i \sim$ negative binomial$(n-1, 1-p)$ Garlaschelli (2009).

For the first approach, the numerator can be written as $\sum_{k=0}^{N_{ij}}(W_{ik} + W_{jk})$, where N_{ij} is again the number of common neighbors of nodes i and j, and is distributed as in the unweighted Erdős-Rényi random graph. Thus, the numerator is a sum of geometric random variables, where the number of summed variables is itself a random variable. Moreover, we must have $W_{ik} > 0$ and $W_{jk} > 0$ since a common neighbor of two nodes can only exist if both nodes are attached to the node in question (the common neighbor). To address this constraint, each of the random variables must first be transformed into zero-truncated geometric random variables, and their mean and variance altered correspondingly. We can now write weighted overlap as a random variable as in Eqn. (1.9).

$$O_{ij}^W = \frac{\sum_{k=1}^{N_{ij}}(W_{ik} + W_{jk})}{S_i + S_j - 2W_{ij}}. \tag{1.9}$$

Now hierarchical models can be used to find the mean and variance of the numerator, and these results combined with the mean and variance values of the denominator can be used to derive the expressions in Eqs. (1.10) and (1.11) (see section 1.5.2 for details):

$$\mathbb{E}[O_{ij}^W] \quad = \quad p \tag{1.10}$$

$$\mathrm{Var}(O_{ij}^W) \quad = \quad \frac{p+1}{n}. \tag{1.11}$$

The second approach again replaces the denominator with its expectation. The mean and variance derivations are then straightforward and result in the expressions in Eqs. (1.12) and (1.13). Again, the expressions for the mean are equivalent for both approaches, and the variance expressions are quite similar (See section 1.6.2 for details):

$$\begin{aligned}\mathbb{E}[O_{ij}^W] &= p \qquad (1.12)\\ \mathrm{Var}(O_{ij}^W) &= \frac{np^2(p+2)}{2(np-1)^2}. \qquad (1.13)\end{aligned}$$

The derivations for the directed Erdős-Rényi random graph are more complicated and do not have a closed form due to the minimum expressions in the denominator. Focusing on the numerator, each of the $A_{ik}A_{kj}$ and $A_{jk}A_{ki}$ terms is equal to one if and only if both adjacency matrix values are equal to 1, which happens with probability p^2 since each edge is independent. Thus, each of the terms is a Bernoulli(p^2) random variable, and the numerator consists of a sum of $2n$ independent Bernoulli random variables, meaning it is a binomial($2n, p^2$) random variable, which we will again approximate with a Poisson($2np^2$) random variable. The denominator includes the minimum of two identically distributed random variables K_i^{in} and K_i^{out}. Due to the definition given in Section 3.1, the in and out degrees of nodes i and j cannot equal 0, making them zero-truncated binomial($n-1, p$) random variables, which will also be approximated as zero-truncated Poisson(np) random variables since n is assumed to be large. We can now write directed overlap as a random variable as in Eqn. (1.14).

$$O_{ij}^D = \frac{\sum_{k=1}^n (A_{ik}A_{kj} + A_{jk}A_{ki})}{\min(K_i^{\text{in}}, K_j^{\text{out}}) + \min(K_j^{\text{in}}, K_i^{\text{out}}) - 1}. \qquad (1.14)$$

The mean and variance of the denominator can now be calculated and used to derive the expressions in Eqs. (1.15) and (1.16) Stuart and Ord (1998) (for details, see section 1.5.3):

$$\begin{aligned}
\mathbb{E}[O_{ij}^D] &= \frac{np^2}{e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 0.5} && (1.15)\\
\mathrm{Var}(O_{ij}^D) &= \frac{2n^2p^4}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 1)^2} && (1.16)\\
&+ \frac{\frac{32n^3p^5e^{np}}{e^{np}-1}\left[1-\frac{np}{e^{np}-1}\right]}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 1)^2}.
\end{aligned}$$

The second approach again replaces the denominator with its expectation, and the mean and variance derivations result in the expressions in Eqs. (1.17) and (1.18) (see section 1.6.3 for details). Again, the expressions for the mean are equivalent for both approaches, but note that the expression for the variance using the second approach in Eq. (1.18) is equivalent to the first term of the variance resulting from the first approach in Eq. (1.16).

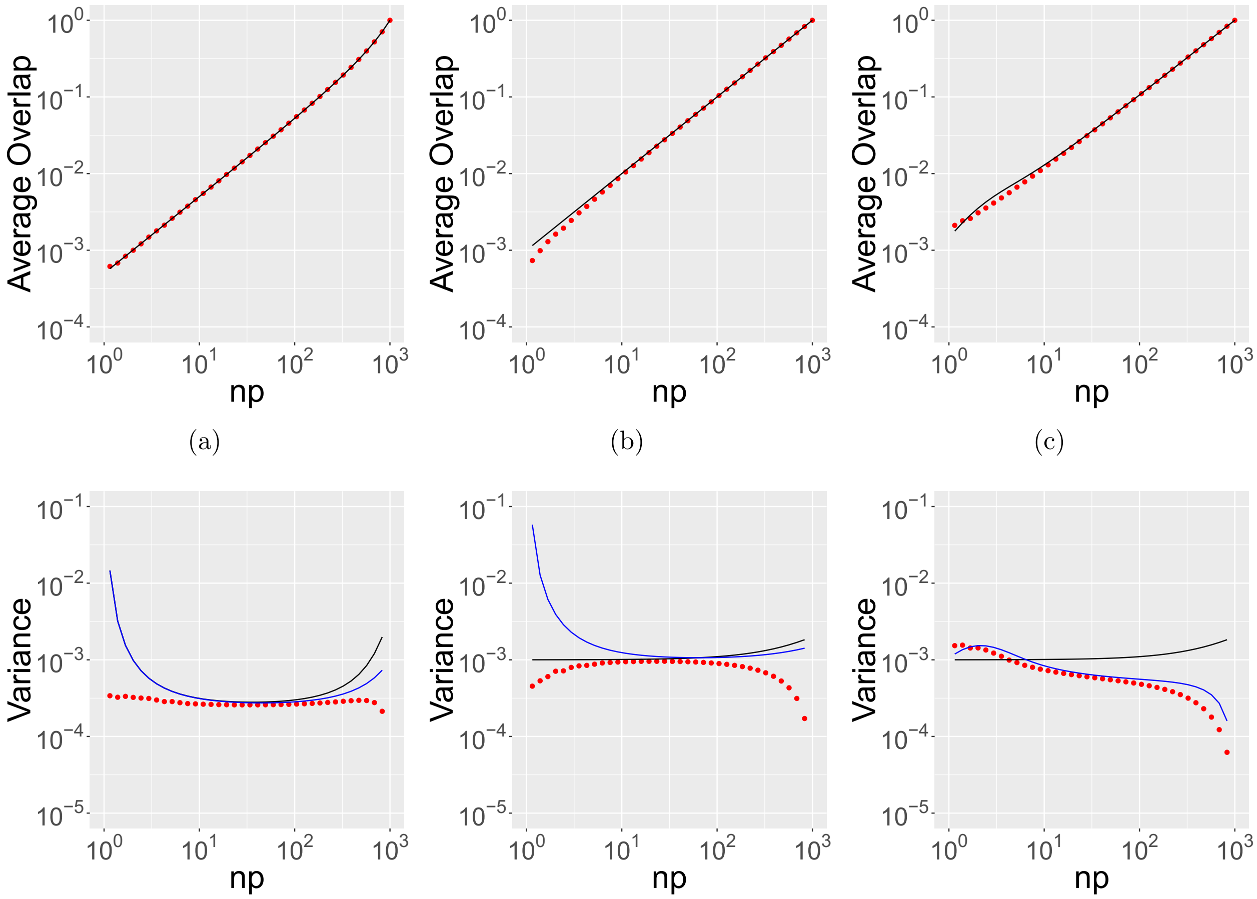
$$\begin{aligned}
\mathbb{E}[O_{ij}^D] &= \frac{np^2}{e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 0.5} && (1.17)\\
\mathrm{Var}(O_{ij}^D) &= \frac{2n^2p^4}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 1)^2} && (1.18)
\end{aligned}$$

1.2.4 Simulation Studies

We conducted simulation studies to evaluate the accuracy of the proposed mean and variance expressions for each version of Erdős-Rényi edge overlap. We simulated 5,000 realizations of networks with $n = 1,000$ nodes for various values of $p \in (0, 1)$. The mean and variance of edge overlap was calculated for each network realization, and those values subsequently averaged over all simulations. We considered values of $p > 1/n$, such that the resulting average degree $np > 1$, which ensures (asymptotically) that the graphs have non-vanishing largest connected components.

Figure 1.2 displays the simulation results and accuracy of our approximations. The top row contains the results for the mean unweighted overlap (Figure 1.2a), mean weighted overlap (Figure 1.2b) and mean directed overlap (Figure 1.2c). In each plot, the red dots represent the simulated results, black lines represent the theoretical values using the first approach and blue lines the second approach. Note that each expression for average overlap is equivalent for the two approaches, making only the black lines visible. The bottom row of panels shows the results for the variance of unweighted overlap (Figure 1.2d), weighted overlap (Figure 1.2e) and directed overlap (Figure 1.2f). In each plot, black lines represent the theoretical values using the first approach, blue lines the second approach, and the red dots the simulated values.

For each version of overlap, our theoretical approximations of the mean closely match the simulations, with the unweighted case being the best fit for all values of np. The approximations of the variance overall are not as accurate, where the accuracy of the fit depends on the value of np. In the unweighted case (Figure 1.2d), both theoretical approaches match the simulated values for average degree $np \geq 10$ until about $np = 100$. The first approximation then deviates from the simulated values, followed by the second approach deviating from them when $np \approx 300$. In the weighted case (Figure 1.2e) the first approximation is more accurate than the second for average degree less than or equal to about 30. The approaches are then equally precise until the average degree is approximately 170; after this point, the second approximation is closer to the simulated values. In the directed case (Figure 1.2f) the two approximations are equivalent and closely match the simulated values until the average degree reaches about 10. After that point, approach two is more accurate. Furthermore, in all cases, both approximations systematically overestimate variability. We stress that this overestimation leads to inflated standard errors and thus to conservative hypothesis tests, which is preferable over the opposite situation, i.e., having deflated standard errors and anti-conservative tests.



(a) (b) (c)

(d) (e) (f)

Figure 1.2: Simulation results for the mean (top row) and variance (bottom row) of each type of Erdős-Rényi overlap. The first column corresponds to the unweighted Erdős-Rényi overlap, the second column to the weighted Erdős-Rényi overlap and the third to the directed Erdős-Rényi overlap case. The top row plots (a), (b) and (c) plot the average overlap on the y-axis and average degree (np) on the x-axis. The red dots represent values from the simulations, and the black line represents the theoretical outcome using approach 1 and the blue line represents the theoretical outcome using approach 2. Note that the blue lines are completly covered by the black lines since the values for average overlap are the same for both approaches. The bottom row plots (d), (e) and (f) plot the variance of edge overlap on the y-axis and average degree (np) on the x-axis. In each plot, the red dots represent values from the simulations, the black line represents the theoretical outcome using approach 1 and the blue line represents the theoretical outcome using approach 2.

1.2.5 Data Analysis

As an application of our derivations to analysis of empirical social networks, we used social network data collected in 2006 from 75 villages housed in 5 districts in rural southern Karnataka, India, all within 3 hours driving distance from Bangalore (Figure 1.3) Banerjee et al. (2013). The data were collected as part of a study that examined how participation in a microfinance program diffuses through social networks. First, a baseline survey was conducted in all 75 villages. The survey consisted of a village questionnaire, a full census that collected data on all households in the villages, and a detailed follow-up survey fielded to a subsample of individuals. The village questionnaire collected data on village leadership, the presence of pre-existing non-governmental organizations (NGOs) and savings self-help groups and various geographical features of the area. The household census gathered demographic information, GPS coordinates of each household and data on a variety of amenities for every household in each village (roof type, latrine type, and access to electric power). The individual surveys were administered to a random sample of villagers in each village and were stratified by religion and geographic sub-location. Over half of the households in each stratification were sampled, yielding a sample of about 46% of all households per village. The individual questionnaire asked for information including age, sub-caste, education, language, native home, and occupation of the person. Additionally, the survey included social network data along 12 dimensions: friends or relatives who visit the respondent's home, friends or relatives the respondent visits, any kin in the village, nonrelatives with whom the respondent socializes, those who the respondent receives medical advice from, who the respondent goes to pray with, from whom the respondent would borrow money, to whom the respondent would lend money, from whom the respondent would borrow or to whom the respondent would lend material goods, from whom the respondent gets advice, and to whom the respondent gives advice.

The median pairwise distance between villages was 46km and the number of cross-village ties was minimal, allowing the villages to be regarded as independent networks. The villages were linguistically homogeneous but had variability in caste. Each village contained

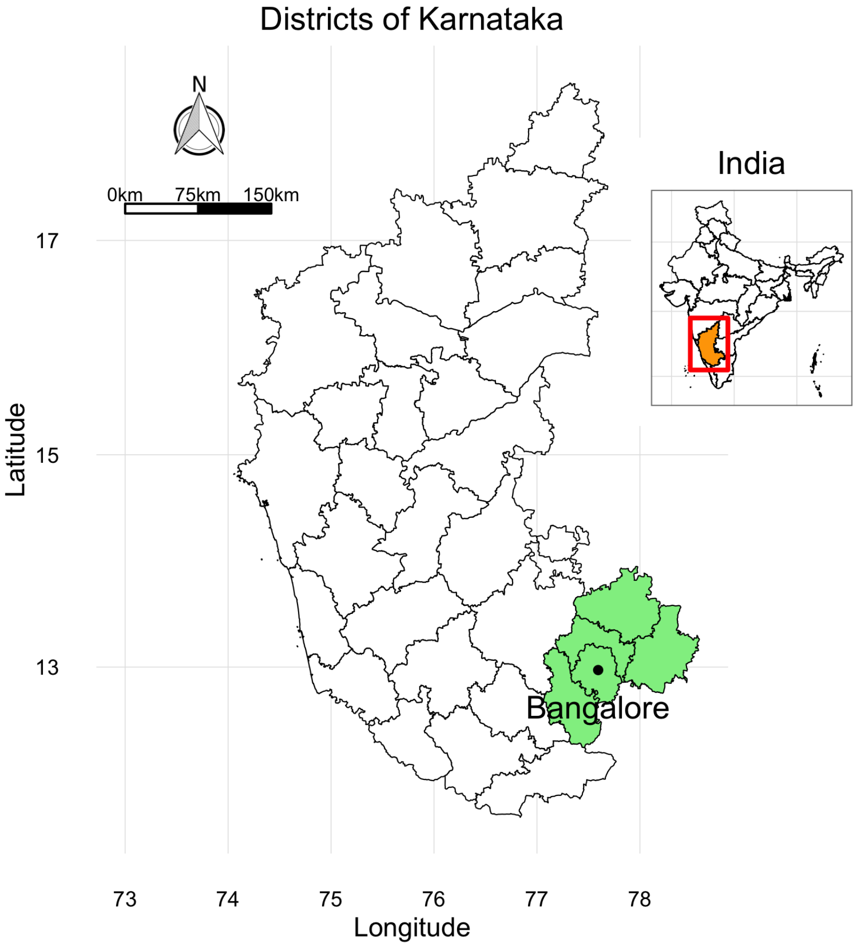


Figure 1.3: A map of the districts of Karnataka, India. The five districts colored in green house all of the villages included in the data set. The districts included are Bangalore, Bangalore Rural, Kolar, Ramanagara and Chikballapura Hijmans (2009); Mukerjee (2013).

anywhere from 354 to 1775 residents, with a total population of 69,441 people in the 75 villages combined. The number of edges across all social networks totaled 2,361,745 which included 480 self-loops and 6,402 isolated dyads. The average degree was 6.79 (standard deviation of 4.03), and the average number of connected components was 17.99 per village. Among the respondents for whom covariate data was collected via the individual surveys, 55.4% were female and 44.6% were male. The mean age was 39 years with a range of 10 to 99 years. Four different castes were represented: scheduled caste, scheduled tribe, general caste, and OBC ("other backward castes"), with a majority of respondents members of the general and OBC castes (≈ 69.5%) Banerjee et al. (2013).

Label	Type of social interaction
1	The respondent borrows money from this individual
2	The respondent gives advice to this individual
3	The respondent helps this individual make a decision
4	The respondent borrows kerosene or rice from this individual
5	The respondent lends kerosene or rice to this individual
6	The respondent lends money to this individual
7	The respondent obtains medical advice from this individual
8	The respondent engages socially with this individual
9	The respondent is related to this individual
10	The respondent goes to temple with this individual
11	The respondent has visited this individual's home
12	The respondent has been invited to this individual's home

Table 1.2: The types of social interactions recorded for individuals in each village.

We first calculated the average unweighted overlap for each type of social relationship (labeled 1-12, see Table 1.2) for each village by treating all ties as undirected and by removing all self-loops since they do not contribute to edge overlap (Figure 1.7). Then we standardized each average overlap by subtracting the expected mean and dividing by the standard deviation under the null; the results from the unweighted Erdős-Rényi overlap derivations using the first approach discussed above (Figure 1.8 in section 1.7). We stratified edges according to the availability of nodal attributes (since not all villagers completed an individual survey), sex, caste and age. Here we detail our results from stratifying by sex with Figures 1.4 and 1.5 showing raw and standardized overlap for female-female (F/F), male-male (M/M) and male-female (M/F) ties. For details and figures of stratification by attribute availability, age and caste, see section 1.7.

We next collapsed the twelve unweighted networks into one weighted network. Specifically, the weight of a tie between two individuals corresponds to the number of types of social relationships they are engaged in with each other. For example, if individual i borrows money from, gives advice to and goes to temple with individual j, the weight of the (undirected) tie between i and j would be equal to 3. Similar to the unweighted networks, we stratified the weighted networks by nodal attributes, including the presence or absence of attribute information, sex, caste and age. Figure 1.6 shows the distributions of raw and standardized weighted overlap for F/F, M/M and M/F ties. See section 1.7 for figures stratified by attribute availability, caste and age.

1.3 Results

Here we detail our observations of the figures in the previous section where overlap is stratified by sex. For explanations about the figures detailing stratification by attribute information, caste and age, see section 1.8. In Figure 1.4, the median average unweighted overlap is the largest for F/F ties, followed by M/F ties and then M/M ties. There is a clear separation in the values of average overlap between F/F and M/M ties with no overlap in values for interaction types 1, 2, 3, 4, 5, 6, 8, and 11. For more details, see Table **??** in section 1.7. This suggests that women in these villages tend to form 'cliques', tighter friendship circles where most individuals interact with each other more regularly and intensely than others in the same setting, much more than men for every type of social interaction. This kind of social development is quite common among females and has been studied in the social sciences Hwong et al. (2016a,b). However, this trend could also be due to the significant difference in the average degree for males and females across the villages (Figure 1.21 in section 1.7). The degrees of two attached nodes directly effects the value of overlap; it is easier for pairs of nodes with smaller degrees to have a higher value of overlap due to the smaller number of neighbors they need to have in common. The values of average overlap

for the M/F ties are closer to the values for F/F ties than M/M ties and their distributions tend to have smaller variance compared to the other types of ties. This suggests that individuals who have mixed-sex social ties typically have more friends in common than individuals who are part of a M/M social tie. Interestingly, when the average overlap values are standardized, which effectively adjusts for differences in average degree, M/F and M/M ties have much more similar values and are still well below the F/F ties values. The only exceptions are for interaction types 9 and 10 where the F/F and M/M ties have comparable values. All

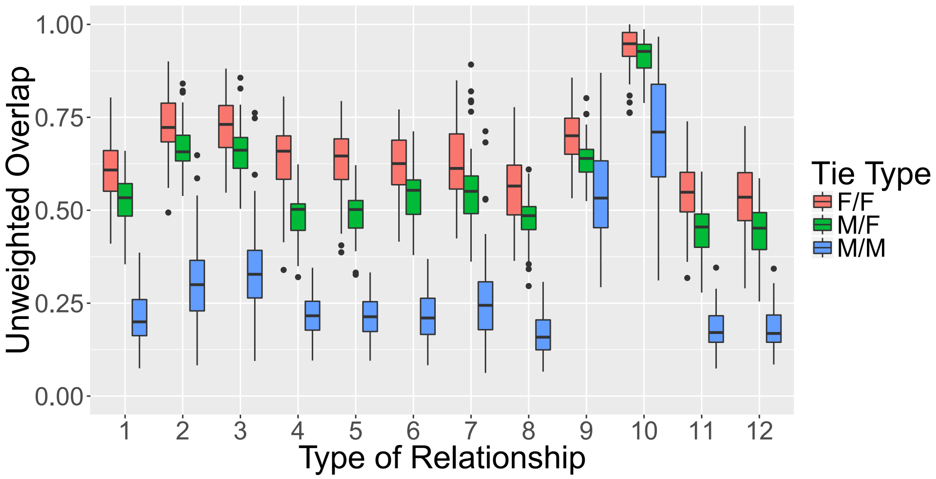


Figure 1.4: Distribution of average unweighted overlap for each village for each type of social interaction stratified by sex. A female individual is labeled with an 'F' and a male individual is labeled with an 'M'. We stratified the edges by sex, and labeled an edge between two female individuals as 'F/F', an edge between two male individuals as 'M/M', and an edge between a female individual and a male individual as 'M/F'. The y-axis represents the proportion of average edge overlap and the x-axis represents the type of social interaction.

values are significantly higher than expected under the null, which is not surprising.

Figure 1.6 shows that when ties are aggregated across interaction types, the values of average weighted overlap for F/F and M/F ties are very similar. The distribution for F/F ties has larger values and more variation, but its median is almost equivalent to that of the M/F ties distribution. It can also be seen that the values for average weighted overlap are much smaller for M/M ties; in fact there is no overlap in values between the M/M ties and the F/F and M/F ties. This again points to females having the tendency to create social 'cliques' more often than males. This trend is also seen when all values are standardized (Figure 1.6b). Again, all values are significantly higher than expected for each type of tie, as we would expect from Figure 1.5 above.

1.4 Conclusions and Discussion

In this paper we introduced extensions of edge overlap for weighted and directed networks. We also used the classic Erdős-Rényi random graph and its weighted and directed counterparts to define a null model and derive approximations for the expected mean and variance of edge overlap for each type of graph. Edge overlap can be standardized using these approxima-

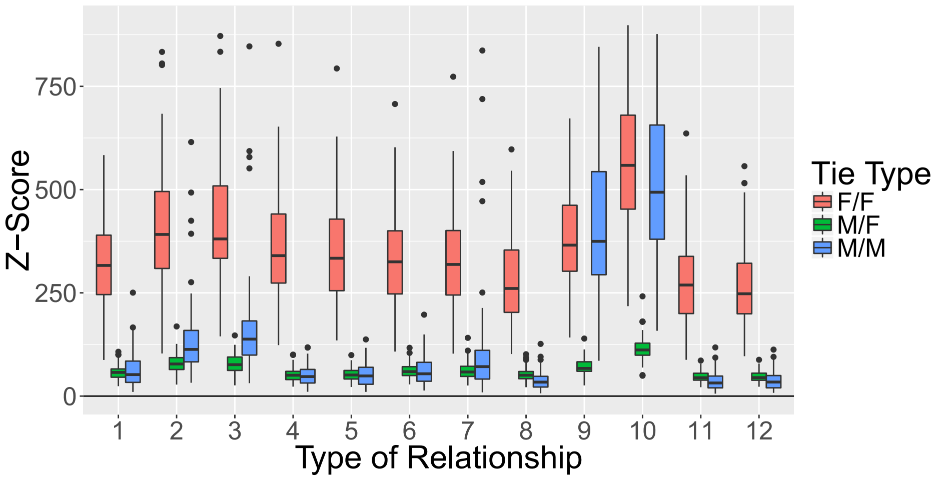
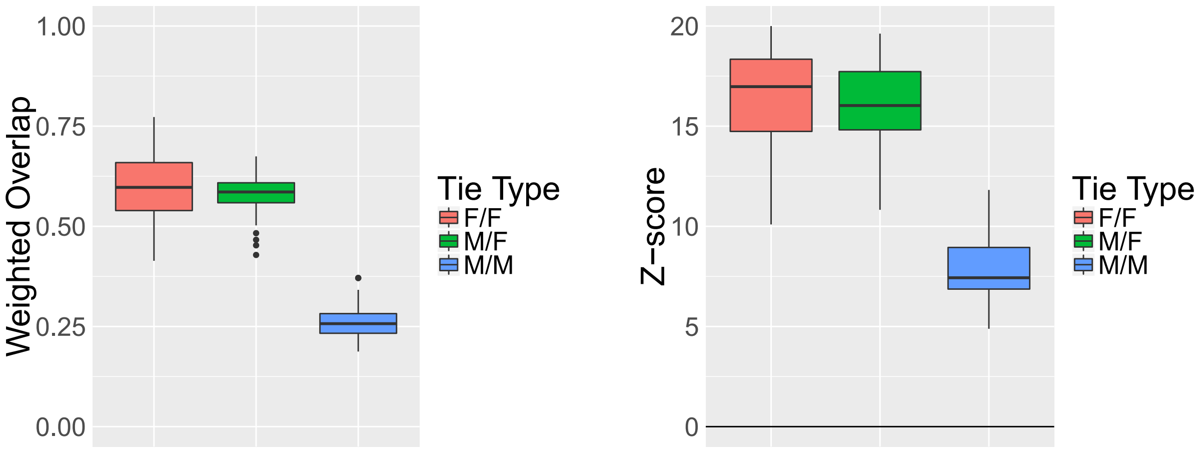


Figure 1.5: Distribution of standardized unweighted overlap for each village for each type of social interaction stratified by sex. A female individual is labeled with an 'F' and a male individual is labeled with an 'M'. We stratified the edges by sex, and labeled an edge between two female individuals as 'F/F', an edge between two male individuals as 'M/M', and an edge between a female individual and a male individual as 'M/F'. The y-axis represents the standardized value, also known as the Z-score, and the x-axis represents the type of social interaction.

tions allowing its comparison across networks of different size. We used these approximations in a data analysis of the social networks of 75 villages in rural India. We found that overall, the average proportion of overlap was much higher than expected under the null for each type of social interaction, especially when the social activity was going to temple together.

We also found that there is a marked difference in the amount of overlap between female-female ties and male-male ties, with female-female ties consistently achieving much higher values of overlap. This could be a consequence of two types of mechanisms; the average degrees of males versus females and the tendency of women forming friendship 'cliques' with other women much more frequently than men forming the same types of friendship circles with other men. We found that in this case, men have a significantly higher degree than women across all networks. Whichever mechanism is at work here, this structural information could lead to an alternative method of eliciting social network data to optimize diffusion or intervention strategies based on the type of tie.

While our work generalizes a central microscopic network metric, making it more broadly applicable, there are limitations to our work. The Erdős-Rényi random graph model is a simple and somewhat naive null model in the context of social networks. This model does



(a)

(b)

Figure 1.6: Distribution of average weighted overlap (a) and standardized weighted overlap (b) stratified by sex. A female individual is labeled with an 'F' and a male individual is labeled with an 'M'. We stratified the edges by sex, and labeled an edge between two female individuals as 'F/F', an edge between two male individuals as 'M/M', and an edge between a female individual and a male individual as 'M/F'. The y-axis in (a) represents the proportion of average weighted edge overlap, and the y-axis in (b) represents the standardized value, also known as the Z-score.

not preserve the degree distribution and is relatively easy to reject. An alternative would be to derive these expressions for the configuration model, which does preserve the degree distribution. However, deriving the mean and variance under the configuration model null model would be considerably more difficult. Another limitation with our mean and variance approximations is the ignoring of the correlations that are present among the random variables in the overlap expressions. In each version of overlap, the number of common neighbors is constrained by the degree of the edge-sharing nodes, making the numerator dependent upon the denominator. While our approximations are quite precise for the majority of values of mean degree, they could be improved if the correlation were also included in the approximations.

1.5 Approach 1 Mean and Variance Derivations

1.5.1 Original Erdős-Rényi Overlap

Edge overlap is considered a random variable with mean and variance (See Eq. (1.19)). We first define the distributions of the variables used to define overlap (denoted by uppercase letters) and then proceed to approximate its mean and variance. For each approximation, we assume n is large.

$$O_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}} \Rightarrow \frac{N_{ij}}{K_i + K_j - 2 - N_{ij}} = \frac{N_{ij}}{H_{ij}} \tag{1.19}$$

Suppose we have an Erdős-Rényi random graph with n nodes and connection probability p. The probability that both i and j are connected to a common neighbor k is equal to p^2, and the total number of possible common neighbors is equal to $n-2$. Thus, the distribution of the number of common neighbors, N_{ij}, is a binomial random variable with $n-2$ trials and connection probability p^2. For large n, this can be approximated with a Poisson(np^2) distribution. Similarly, the probability that one node is connected to another is p, and each node has a total of $n-1$ other nodes it could connect to. Thus, the degree distribution, K_i, is also a binomial random variable with $n-1$ trials and probability p. This can also be approximated by a Poisson(np^2) for large n. Using the Possion approximations, the denominator becomes the difference between two Poisson random variables, $H_{ij} = (K_i + K_j - 2)$ and N_{ij}, which is a Skellam random variable Skellam (1946). Table 1.3 summarizes these distributions.

Variable	Distribution	Mean	Variance
N_{ij}	Poisson(np^2)	np^2	np^2
K_i, K_j	Poisson($2np-2$)	$2np-2$	$2np-2$
H_{ij}	Skellam($2np-2-np^2$)	$2np-2-np^2$	$2np-2+np^2$

Table 1.3: The distribution, mean and variance for each random variable included in Erdős-Rényi overlap.

Edge overlap is the ratio of two random variables and its mean and variance can be approximated using a Taylor series expansion Elandt-Johnson and Johnson (1998); Stuart and Ord (1998). The general form of a first order Taylor series expansion for a function $g(x) = g(x_1, x_2, \ldots, x_k)$ about $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is

$$g(x) = g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(x_i - \theta_i) + O(n^{-1}) \tag{1.20}$$

where $g'(x)$ denotes the derivative of $g(x)$. Here, the function is the ratio of N_{ij} over H_{ij}. Define $g(N_{ij}, H_{ij}) = \frac{N_{ij}}{H_{ij}}$ where H_{ij} has no mass at 0. This assumption is assured by the constraints defined in the Methods section of the paper. Equation (1.21) shows the Taylor series expansion for $g(N_{ij}, H_{ij})$ about the mean, $\theta = (\mathbb{E}(N_{ij}), \mathbb{E}(H_{ij}))$.

$$\begin{aligned} g(N_{ij}, H_{ij}) &= g(\theta) + \sum_{i=1}^{2} g_i'(\theta)(x_i - \theta_i) + O(n^{-1}) \qquad (1.21) \\ &= g(\theta) + g'_{N_{ij}}(\theta)(N_{ij} - \theta_{N_{ij}}) + g'_{H_{ij}}(\theta)(H_{ij} - \theta_{H_{ij}}) + O(n^{-1}) \\ &= g(\theta) + g'_{N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij})) + g'_{H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij})) + O(n^{-1}) \end{aligned}$$

Using the above approximation, the expectation of the ratio, $\mathbb{E}[g(N_{ij}, H_{ij})]$, can be derived as in Eq. (1.22).

$$\begin{aligned} \mathbb{E}[g(N_{ij}, H_{ij})] &= \mathbb{E}[g(\theta) + g'_{N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij})) \qquad (1.22) \\ &+ g'_{H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij})) + O(n^{-1})] \\ &= \mathbb{E}[g(\theta)] + \mathbb{E}[g'_{N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij}))] + \mathbb{E}[g'_{H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij}))] \\ &= \mathbb{E}[g(\theta)] + g'_{N_{ij}}(\theta)\mathbb{E}[N_{ij} - \mathbb{E}(N_{ij})] + g'_{H_{ij}}(\theta)\mathbb{E}[H_{ij} - \mathbb{E}(H_{ij})] \\ &= \mathbb{E}[g(\theta)] + 0 + 0 \approx g(\mathbb{E}(N_{ij}), \mathbb{E}(H_{ij})) = \frac{\mathbb{E}(N_{ij})}{\mathbb{E}(H_{ij})} \\ &= \frac{np^2}{2np - 2 - np^2} \approx \frac{p}{2-p} \end{aligned}$$

Using the definition of variance and the result that $\mathbb{E}[g(N_{ij}, H_{ij})] \approx g(\theta)$ from Eq. (1.22), the variance of $g(N_{ij}, H_{ij})$ can be first approximated by Eq. (1.23).

$$\begin{aligned} \mathrm{Var}(g(N_{ij}, H_{ij})) &= \mathbb{E}\left\{[g(N_{ij}, H_{ij}) - \mathbb{E}(g(N_{ij}, H_{ij}))]^2\right\} \qquad (1.23) \\ &\approx \mathbb{E}\left\{[g(N_{ij}, H_{ij}) - g(\theta)]^2\right\} \end{aligned}$$

Using the first order Taylor expansion for $g(N_{ij}, H_{ij})$ from Eq. (1.21), we have

$$\begin{aligned} \mathrm{Var}(g(N_{ij}, H_{ij})) &\approx \mathbb{E}\{[g(\theta) + g'_{N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij})) \qquad (1.24) \\ &+ g'_{H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij})) - g(\theta)]^2\} \\ &= \mathbb{E}\left\{[g'_{N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij})) + g'_{H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij}))]^2\right\} \\ &= \mathbb{E}[{g'_{N_{ij}}}^2(\theta)(N_{ij} - \mathbb{E}(N_{ij}))^2 + {g'_{H_{ij}}}^2(\theta)(H_{ij} - \mathbb{E}(H_{ij}))^2 \\ &+ 2g'_{N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij}))g'_{H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij}))] \\ &= {g'_{N_{ij}}}^2(\theta)\mathrm{Var}(N_{ij}) + {g'_{H_{ij}}}^2(\theta)\mathrm{Var}(H_{ij}) \\ &+ 2g'_{N_{ij}}(\theta)g'_{H_{ij}}(\theta)\mathrm{Cov}(N_{ij}, H_{ij}). \end{aligned}$$

In this case, $g(N_{ij}, H_{ij}) = \frac{N_{ij}}{H_{ij}}$, $g'_{N_{ij}} = \frac{1}{H_{ij}}$, $g'_{H_{ij}} = \frac{-N_{ij}}{H_{ij}^2}$, and $\theta = (\mathbb{E}(N_{ij}), \mathbb{E}(H_{ij}))$, $g'_{N_{ij}}(\theta)g'_{H_{ij}}(\theta) = \frac{-\mathbb{E}(N_{ij})}{\mathbb{E}^3(H_{ij})}$, ${g'_{N_{ij}}}^2(\theta) = \frac{1}{\mathbb{E}^2(H_{ij})}$, ${g'_{H_{ij}}}^2(\theta) = \frac{\mathbb{E}^2(N_{ij})}{\mathbb{E}^4(H_{ij})}$. Placing these expressions into

(1.24) we have that

$$\begin{aligned} \mathrm{Var}(g(N_{ij}, H_{ij})) &\approx \frac{\mathrm{Var}(N_{ij})}{\mathbb{E}^2(H_{ij})} + \frac{\mathbb{E}^2(N_{ij})\mathrm{Var}(H_{ij})}{\mathbb{E}^4(H_{ij})} - 2\frac{\mathrm{Cov}(N_{ij}, H_{ij})\mathbb{E}(N_{ij})}{\mathbb{E}^3(H_{ij})} \qquad (1.25) \\ &= \frac{\mathbb{E}^2(N_{ij})}{\mathbb{E}^2(H_{ij})}\left[\frac{\mathrm{Var}(N_{ij})}{\mathbb{E}^2(N_{ij})} + \frac{\mathrm{Var}(H_{ij})}{\mathbb{E}^2(H_{ij})} - 2\frac{\mathrm{Cov}(N_{ij}, H_{ij})}{\mathbb{E}(N_{ij})\mathbb{E}(H_{ij})}\right] \\ &= \frac{np^2}{(2np-2-np^2)^2} + \frac{n^2p^4(2np-2+np^2)}{(2np-2-np^2)^4} - 2\frac{np^2\mathrm{Cov}(N_{ij}, H_{ij})}{(2np-2-np^2)^3}. \end{aligned}$$

Note that $\mathrm{Cov}(N_{ij}, H_{ij}) > 0$ since $N_{ij} \not\perp\!\!\!\perp H_{ij}$. The value for the covariance could be simulated, but for simplicity we choose to ignore this dependence and include only the first two terms of (1.25) in the variance approximation.

A second order Taylor series expansion can be used as a more precise approximation of the mean. The second order Taylor expansion for the overlap ratio is

$$\begin{aligned} g(N_{ij}, H_{ij}) &= g(\theta) + g'_{N_{ij}}(\theta)(N_{ij} - \theta_{N_{ij}}) + g'_{H_{ij}}(\theta)(H_{ij} - \theta_{H_{ij}}) \qquad (1.26) \\ &+ \frac{1}{2}g''_{N_{ij}N_{ij}}(\theta)(N_{ij} - \theta_{N_{ij}})^2 + \frac{1}{2}g''_{H_{ij}H_{ij}}(\theta)(H_{ij} - \theta_{Hij})^2 \\ &+ g''_{N_{ij}H_{ij}}(\theta)(N_{ij} - \theta_{Nij})(H_{ij} - \theta_{Hij}) + O(n^{-1}) \\ &= g(\theta) + g'_{N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij})) + g'_{H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij})) \\ &+ \frac{1}{2}g''_{N_{ij}N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij}))^2 + \frac{1}{2}g''_{H_{ij}H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij}))^2 \\ &+ g''_{N_{ij}H_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij}))(H_{ij} - \mathbb{E}(H_{ij})) + O(n^{-1}). \end{aligned}$$

Thus, a better approximation of $E(g(N_{ij}, H_{ij}))$ about $\theta = (\mathbb{E}(N_{ij}), \mathbb{E}(H_{ij}))$ is

$$\begin{aligned}
\mathbb{E}[g(H_{ij}, N_{ij})] &= \mathbb{E}[g(\theta) + g'_{N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij})) + g'_{H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij})) \qquad (1.27)\\
&+ \frac{1}{2}g''_{N_{ij}N_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij}))^2\frac{1}{2}g''_{H_{ij}H_{ij}}(\theta)(H_{ij} - \mathbb{E}(H_{ij}))^2\\
&+ g''_{N_{ij}H_{ij}}(\theta)(N_{ij} - \mathbb{E}(N_{ij}))(H_{ij} - \mathbb{E}(H_{ij})) + O(n^{-1})]\\
&= \mathbb{E}[g(\theta) + \frac{1}{2}\left\{g''_{N_{ij}N_{ij}}(\theta)\mathrm{Var}(N_{ij}) + g''_{H_{ij}H_{ij}}(\theta)\mathrm{Var}(H_{ij})\right\}]\\
&+ g''_{N_{ij}H_{ij}}(\theta)\mathrm{Cov}(N_{ij}, H_{ij}) + O(n^{-1})].
\end{aligned}$$

For $g(N_{ij}, H_{ij}) = \frac{N_{ij}}{H_{ij}}$, $g''_{N_{ij}N_{ij}} = 0$, $g''_{N_{ij}H_{ij}} = \frac{-1}{H_{ij}^2}$, $g''_{H_{ij}H_{ij}} = \frac{2N_{ij}}{H_{ij}^3}$. Plugging these expressions into (1.27) results in Eq. (1.28).

$$\begin{aligned}
\mathbb{E}[g(N_{ij}, H_{ij}))] &= \frac{\mathbb{E}(N_{ij})}{\mathbb{E}(H_{ij})} + \frac{\mathrm{Var}(H_{ij})\mathbb{E}(N_{ij})}{\mathbb{E}^3(H_{ij})} - \frac{\mathrm{Cov}(N_{ij}, H_{ij})}{\mathbb{E}^2(H_{ij})} \qquad (1.28)\\
&= \frac{np^2}{2np - 2 - np^2} + \frac{(2np - 2 + np^2)(np^2)}{(2np - 2 - np^2)^3} - \frac{\mathrm{Cov}(N_{ij}, H_{ij})}{(2np - 2 - np^2)^2}\\
&= \frac{p}{2 - p} + \frac{(2np - 2 + np^2)(np^2)}{(2np - 2 - np^2)^3} - \frac{\mathrm{Cov}(N_{ij}, H_{ij})}{(2np - 2 - np^2)^2}.
\end{aligned}$$

Again, $\mathrm{Cov}(N_{ij}, H_{ij}) > 0$ since $N_{ij} \not\perp\!\!\!\perp H_{ij}$. The value for the covariance could be simulated, but for simplicity we chose to ignore this dependence and only include the first two terms of (1.28) in the approximation of the mean.

1.5.2 Weighted Erdős-Rényi Overlap

Now suppose we introduce weights to the network edges and construct a WRG with n nodes. The weighted Erdős-Rényi overlap can be written as in Eq. (1.29). N_{ij} is again the of the number of common neighbors nodes i and j share, W_{ij} is the weight of the tie between nodes i and j and $S_i(S_j)$ is the strength of node $i(j)$. The ratio is denoted as V_{ij} over M_{ij}. We again define the distribution of each of the random variables in the expression and then use the Taylor series expansion approximation outlined in the previous section to derive expressions for the mean and variance of weighted overlap.

$$O_{ij}^{W} = \frac{\sum_{k=1}^{n_{ij}}(w_{ik} + w_{jk})}{s_i + s_j - 2w_{ij}} \Rightarrow \frac{\sum_{k=1}^{N_{ij}}(W_{ik} + W_{jk})}{S_i + S_j - 2W_{ij}} = \frac{V_{ij}}{M_{ij}} \tag{1.29}$$

For each pair of nodes, an edge is created between them with probability p, and a unit weight is added to that edge again with probability p until the first ‘failure’. This describes a geometric distribution, meaning $W_{ij} \sim$ geometric$(1-p)$. However, to ensure the existence of overlap, we are assuming that the values of all weights are > 0. Consequently, W_{ij} is a zero-truncated geometric$(1-p)$. The strength of a node is the sum of the weights associated with the edges between that node and all other nodes in the network. Thus, the strength of any node is the sum of $n-1$ geometric random variables, meaning $S_i \sim$ negative binomial$(n-1, 1-p)$ Garlaschelli (2009). Regardless of the weight of the edge, the probability of an edge existing between nodes i and j is equal to p. Therefore, the distribution of N_{ij} is identical to that described in the previous section; a binomial$(n-2, p^2)$. This can again be approximated by a Poisson(np^2) distribution for large n.

Focusing on the numerator, V_{ij} is a sum of zero-truncated geometric random variables where the number of variables summed is itself a random variable. More specifically, V_{ij} is a negative binomial random variable with a parameter that depends on the value of N_{ij}. We use hierarchical models to calculate the mean (Eq. (1.30)) and variance (Eq. (1.31)) of V_{ij}.

$$\begin{aligned} E[V_{ij}] &= E[E[V_{ij}|N_{ij}]] = E\left[\frac{2N_{ij}}{(1-p)}\right] \qquad (1.30) \\ &= \frac{2}{(1-p)}E[N_{ij}] = \frac{2np^2}{(1-p)} \end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(V_{ij}) &= E[\mathrm{Var}(V_{ij}|N_{ij})] + \mathrm{Var}(E[V_{ij}|N_{ij}]) \\
&= E\left[\frac{2pN_{ij}}{(1-p)^2}\right] + \left(\mathrm{Var} \quad \frac{2N_{ij}}{(1-p)}\right) \\
&= \left[\frac{2p}{(1-p)^2}\right] E[N_{ij}] + \left[\frac{2}{(1-p)}\right]^2 \mathrm{Var}(N_{ij}) \\
&= \frac{2np^2(p+2)}{(1-p)^2}
\end{aligned} \tag{1.31}$$

The distribution of M_{ij} is more convoluted. In fact, it is unknown, and its mean and variance must be calculated directly (Eqs. (1.32) and (1.33)). Table 1.4 summarizes all of these distributions.

$$\begin{aligned}
\mathbb{E}[M_{ij}] &= \mathbb{E}[S_i] + \mathbb{E}[S_j] - \mathbb{E}[2W_{ij}] \\
&= \frac{(n-1)p}{(1-p)} + \frac{(n-1)p}{(1-p)} - \frac{2}{(1-p)} \approx \frac{2np-2}{(1-p)}
\end{aligned} \tag{1.32}$$

$$\begin{aligned}
\mathrm{Var}(M_{ij}) &= \mathrm{Var}(S_i) + \mathrm{Var}(S_j) + \mathrm{Var}(2W_{ij}) \\
&= \frac{(n-1)p}{(1-p)^2} + \frac{(n-1)p}{(1-p)^2} - \frac{4p}{(1-p)^2} = \frac{2np}{(1-p)^2}
\end{aligned} \tag{1.33}$$

Now that the mean and variance of the numerator and denominator have been defined, the mean and variance of weighted overlap can be approximated. Define $g(V_{ij}, M_{ij}) = \frac{V_{ij}}{M_{ij}}$. Using the same equations introduced in the previous section, we have

$$\mathbb{E}[g(V_{ij}, M_{ij})] \approx g(\mathbb{E}(V_{ij}), \mathbb{E}(M_{ij})) = \frac{\mathbb{E}(V_{ij})}{\mathbb{E}(M_{ij})} = \frac{np^2}{np-1} \approx p \tag{1.34}$$

Variable	Distribution	Mean	Variance
W_{ij}	Zero-truncated Geometric$(1-p)$	$\frac{1}{(1-p)}$	$\frac{1}{(1-p)^2}$
S_i, S_j	Negative Binomial$(n-1, 1-p)$	$\frac{(n-1)p}{(1-p)}$	$\frac{(n-1)p}{(1-p)^2}$
N_{ij}	Poisson(np^2)	np^2	np^2
V_{ij}	Negative Binomial	$\frac{2np^2}{(1-p)}$	$\frac{2np^2(p+2)}{(1-p)^2}$
M_{ij}	Unknown	$\frac{2np-2}{(1-p)}$	$\frac{2np}{(1-p)^2}$

Table 1.4: The distribution, mean and variance for each random variable included in weighted Erdős-Rényi overlap.

$$\begin{aligned} \mathrm{Var}(g(V_{ij}, M_{ij})) &\approx \frac{\mathbb{E}^2(V_{ij})}{\mathbb{E}^2(M_{ij})}\left[\frac{\mathrm{Var}(V_{ij})}{\mathbb{E}^2(V_{ij})} + \frac{\mathrm{Var}(M_{ij})}{\mathbb{E}^2(M_{ij})} - 2\frac{\mathrm{Cov}(V_{ij}, M_{ij})}{\mathbb{E}(V_{ij})\mathbb{E}(M_{ij})}\right] \qquad (1.35) \\ &= p^2\left[\frac{p+2}{2np^2} + \frac{1}{2np} - \frac{(1-p)^2\mathrm{Cov}(V_{ij}, M_{ij})}{2np^2(np-1)}\right] = \frac{p+1}{n}. \end{aligned}$$

Note that $\mathrm{Cov}(V_{ij}, M_{ij}) > 0$ since $V_{ij} \not\perp\!\!\!\perp M_{ij}$. The value for the covariance could be simulated, but for simplicity we chose to ignore this dependence and do not include the covariance term in the final approximation.

Again, a second order Taylor series expansion can be used as a more precise approximation of the mean. Using the same equations introduced in the previous section, the second order Taylor approximation for the weighted overlap mean is

$$\begin{aligned} \mathbb{E}[g(V_{ij}, M_{ij}))] &= \frac{\mathbb{E}(V_{ij})}{\mathbb{E}(M_{ij})} + \frac{\mathrm{Var}(M_{ij})\mathbb{E}(V_{ij})}{\mathbb{E}^3(M_{ij})} - \frac{\mathrm{Cov}(V_{ij}, M_{ij})}{\mathbb{E}^2(M_{ij})} \qquad (1.36) \\ &\approx p + \frac{n^2p^3}{(np-1)^3} - \frac{(1-p)^2\mathrm{Cov}(V_{ij}, M_{ij})}{4(np-1)^2}. \end{aligned}$$

Again, $\mathrm{Cov}(V_{ij}, M_{ij}) > 0$ since $V_{ij} \not\perp\!\!\!\perp M_{ij}$. The value for the covariance could be simulated, but for simplicity we chose to ignore this dependence and only include the first two terms of Eq. (1.36) in the approximation of the mean.

1.5.3 Directed Erdős-Rényi Overlap

Now suppose we introduce directionality to the network edges and construct a directed random graph with n nodes and connection probability p. The directed Erdős-Rényi overlap can be written as Eq. (1.37). A_{ij} is the adjacency matrix value from node i to node j. If $A_{ij} = 1$, there is a directed edge from i to j. K_i^{in} and K_i^{out} denote the in and out-degree distributions of node i, respectively. Note that because K_i^{in} and K_i^{out} are identically distributed for each node i, $\min(k_i^{\text{in}}, k_j^{\text{out}}) = \min(k_j^{\text{in}}, k_i^{\text{out}})$, and w.l.o.g., we write their sum as $2\min(K_j^{\text{in}}, K_i^{\text{out}})$. We denote the numerator and denominator using D_{ij} and C_{ij} respectively. Again we define the distribution of each of the random variables in the expression and then use the Taylor series expansion approximation outlined in the previous sections to derive expressions for the mean and variance of directed overlap. However, directed version derivations are more complicated and do not have a closed form due to the minimum expressions in the denominator.

$$O_{ij}^D = \frac{\sum_{k=1}^n (A_{ik}A_{kj} + A_{jk}A_{ki})}{\min(k_i^{\text{in}}, k_j^{\text{out}}) + \min(k_j^{\text{in}}, k_i^{\text{out}}) - 1} \Rightarrow \frac{D_{ij}}{2\min(K_j^{\text{in}}, K_i^{\text{out}}) - 1} = \frac{D_{ij}}{C_{ij}} \tag{1.37}$$

Focusing on the numerator, each of the $A_{ik}A_{kj}$ and $A_{jk}A_{ki}$ terms is equal to one if and only if both adjacency matrix values are equal to 1, which happens with probability p^2 since each generation of an edge is independent. Thus, each of the terms is a Bernoulli(p^2) random variable, and the numerator consists of a sum of $2n$ Bernoulli random variables, meaning it is a binomial$(2n, p^2)$ random variable. For large n, this can be approximated by a Poisson$(2np^2)$ distribution.

The denominator includes the minimum of two, identically distributed random variables, K_i^{in} and K_i^{out}. Due to the constraint of existence mentioned in section 3.1 above, the in and out degrees of nodes i and j can not equal 0, making them zero-truncated binomial$(n-1, p)$ random variables. We again approximate this with a zero-truncated Poisson(np) distribution. The distribution of the minimum of two Poisson random variables is unknown. However, an expression for the exact mean (Eq. (1.38)) and an upper bound for the variance (Eq. (1.39)) can be derived Papadatos (1995). We denote the minimum of two random variables as $K_{(1)}$

Variable	Distribution	Mean	Variance
$A_{ik}A_{kj}$	Bernoulli(p^2)	p^2	$p^2(1-p^2)$
D_{ij}	Poisson($2np^2$)	$2np^2$	$2np^2$
$K_i^{\text{in}}, K_i^{\text{out}}$	Zero-truncated Poisson(np)	$\frac{npe^{np}}{e^{np}-1}$	$\frac{npe^{np}}{e^{np}-1}\left[1-\frac{np}{e^{np}-1}\right]$
$K_{(1)}$	Unknown	$e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2$	$\frac{2npe^{np}}{e^{np}-1}\left[1-\frac{np}{e^{np}-1}\right]$
C_{ij}	Unknown	$2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2-1$	$\frac{8npe^{np}}{e^{np}-1}\left[1-\frac{np}{e^{np}-1}\right]$

Table 1.5: The distribution, mean and variance for each random variable included in directed Erdős-Rényi overlap.

and K_{in}^i, K_{out}^i as simply K_i. Table 1.5 summarizes these random variables.

$$\begin{aligned}
\mathbb{E}[K_{(1)}] &= \sum_{k=1}^{(n-1)} P(K_{(1)} \geq k) = \sum_{k=1}^{(n-1)} P(K_1 \geq k)^2 \\
&= \sum_{k=1}^{(n-1)} \left[\sum_{j=k}^{(n-1)} P(K_i = j)\right]^2 = e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2
\end{aligned} \tag{1.38}$$

$$\text{Var}(K_{(1)}) = 2\text{Var}(K_i) = \frac{2npe^{np}}{e^{np}-1}\left[1-\frac{np}{e^{np}-1}\right] \tag{1.39}$$

Now that the mean and variance of the numerator and denominator have been defined, the mean and variance of directed overlap can be approximated. Define $g(D_{ij}, C_{ij}) = \frac{D_{ij}}{C_{ij}}$. Using the same equations introduced in the previous section, we have

$$\begin{aligned}
\mathbb{E}[g(D_{ij}, C_{ij})] &\approx g(\mathbb{E}(D_{ij}), \mathbb{E}(C_{ij})) \\
&= \frac{\mathbb{E}(D_{ij})}{\mathbb{E}(C_{ij})} = \frac{np^2}{e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 0.5}
\end{aligned} \tag{1.40}$$

$$\begin{aligned}
\mathrm{Var}(g(D_{ij},C_{ij})) &\approx \frac{\mathbb{E}^2(D_{ij})}{\mathbb{E}^2(C_{ij})}\left[\frac{\mathrm{Var}(D_{ij})}{\mathbb{E}^2(D_{ij})}+\frac{\mathrm{Var}(C_{ij})}{\mathbb{E}^2(C_{ij})}-2\frac{\mathrm{Cov}(D_{ij},C_{ij})}{\mathbb{E}(D_{ij})\mathbb{E}(C_{ij})}\right] \qquad (1.41)\\
&= \frac{2n^2p^4}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2-1)^2}\\
&+ \frac{\frac{32n^3p^5e^{np}}{e^{np}-1}\left[1-\frac{np}{e^{np}-1}\right]}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2-1)^2}\\
&- \frac{4np^2\mathrm{Cov}(D_{ij},C_{ij})}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2-1)^3}.
\end{aligned}$$

Note that $\mathrm{Cov}(D_{ij},C_{ij}) > 0$ since $D_{ij} \not\perp C_{ij}$. The value for the covariance could be simulated, but for simplicity we chose to ignore this dependence and do not include the covariance term in the final approximation.

Again, a second order Taylor series expansion can be used as a more precise approximation of the mean. Using the same equations introduced in the previous section, the second order Taylor approximation for the directed overlap mean is

$$\begin{aligned}
\mathbb{E}[g(D_{ij},C_{ij}))] &= \frac{\mathbb{E}(D_{ij})}{\mathbb{E}(C_{ij})}+\frac{\mathrm{Var}(C_{ij})\mathbb{E}(D_{ij})}{\mathbb{E}^3(C_{ij})}-\frac{\mathrm{Cov}(D_{ij},C_{ij})}{\mathbb{E}^2(C_{ij})} \qquad (1.42)\\
&= \frac{np^2}{e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2-0.5}\\
&+ \frac{\frac{16n^2p^3e^{np}}{e^{np}-1}\left[1-\frac{np}{e^{np}-1}\right]}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2-1)^3}\\
&- \frac{\mathrm{Cov}(D_{ij},C_{ij})}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2-1)^2}.
\end{aligned}$$

Again, $\mathrm{Cov}(D_{ij},C_{ij}) > 0$ since $D_{ij} \not\perp C_{ij}$. The value for the covariance could be simulated, but for simplicity we chose to ignore this dependence and only include the first two terms of Eq. (1.42) in the approximation of the mean.

1.6 Approach 2 Mean and Variance Derivations

1.6.1 Original Erdős-Rényi Overlap

Again, suppose we have an Erdős-Rényi random graph with n nodes and connection probability p. Edge overlap is again viewed as a random variable with the same distributions for the numerator and denominator defined in the first approach described in section A.3. The expectation of the denominator is equal to $(2np - 2 - np^2)$, and we can rewrite O_{ij} as Eq. (1.43).

$$O_{ij} = \frac{N_{ij}}{H_{ij}} \approx \frac{N_{ij}}{\mathbb{E}[H_{ij}]} = \frac{1}{2np - 2 - np^2} N_{ij} \tag{1.43}$$

The distribution of overlap is now a scaled version of the distribution of N_{ij}, making it a scaled Poisson(np^2) random variable and its mean (Eq. (1.44)) and variance (Eq. (1.45)) can be easily derived.

$$\mathbb{E}[O_{ij}] = \frac{1}{2np - 2 - np^2}\mathbb{E}[N_{ij}] = \frac{np^2}{2np - 2 - np^2} \approx \frac{p}{2-p} \tag{1.44}$$

$$\text{Var}(O_{ij}) = \frac{1}{(2np - 2 - np^2)^2}\text{Var}(N_{ij}) = \frac{np^2}{(2np - 2 - np^2)^2} \tag{1.45}$$

Note that the mean is equivalent to the mean derived in the first approach while the variance is equal to the first term of the variance derived in the first approach. Additionally, there can not be a second order approximation of the mean using this approach since a Taylor expansion has not been used.

1.6.2 Weighted Erdős-Rényi Overlap

Now suppose we have a WRG with n nodes and connection probability p, and weighted overlap is again viewed as a random variable with the same distributions for the numerator and denominator defined in section A.2. The expectation of the denominator is equal to $\frac{2(n-1)p-2}{(1-p)}$, and we can rewrite O_{ij}^W as Eq. (1.46).

$$O_{ij}^W = \frac{V_{ij}}{M_{ij}} \approx \frac{V_{ij}}{\mathbb{E}[M_{ij}]} = \frac{(1-p)}{2np-2}V_{ij} \tag{1.46}$$

The distribution of weighted overlap is now a scaled version of the distribution of V_{ij}, making it a scaled Compound Poisson random variable. The mean (Eq. (1.47)) and variance (Eq. (1.48)) are now easily derived.

$$\mathbb{E}[O_{ij}^W] = \frac{(1-p)}{2np-2}\mathbb{E}[V_{ij}] = \frac{(1-p)}{2np-2}\left(\frac{2np^2}{1-p}\right) \approx p \tag{1.47}$$

$$\begin{aligned} \mathrm{Var}(O_{ij}^W) &= \frac{(1-p)^2}{(2np-2)^2}\mathrm{Var}(V_{ij}) \qquad (1.48) \\ &= \frac{(1-p)^2}{(2np-2)^2}\left(\frac{2np^2(p+2)}{(1-p)^2}\right) \\ &\approx \frac{np^2(p+2)}{2(np-1)^2} \end{aligned}$$

Note that the mean is equivalent to the mean derived in the first approach while the variance is equal to the first term of the variance derived in the first approach. Additionally, there can not be a second order approximation of the mean using this approach since a Taylor expansion has not been used.

1.6.3 Directed Erdős-Rényi Overlap

Now suppose we have a directed Erdős-Rényi random graph with n nodes and connection probability p, and directed overlap is again viewed as a random variable with the same distributions for the numerator and denominator defined in section A.3. The expectation of the denominator is equal to $2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 1$, and we can rewrite O_{ij}^D as Eq. (1.49).

$$O_{ij}^D = \frac{D_{ij}}{C_{ij}} \approx \frac{D_{ij}}{\mathbb{E}[C_{ij}]} = \frac{1}{2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 1}D_{ij} \tag{1.49}$$

The distribution of directed overlap is now a scaled version of the distribution of D_{ij}, making it a scaled Poisson$(2np^2)$ random variable. The mean (Eq. (1.50)) and variance (Eq. (1.51)) are now easily derived.

$$\begin{aligned}
\mathbb{E}[O_{ij}^D] &= \frac{1}{2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 1}\mathbb{E}[D_{ij}] \qquad (1.50)\\
&= \frac{np^2}{e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 0.5}
\end{aligned}$$

$$\begin{aligned}
\text{Var}(O_{ij}^D) &= \frac{1}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 1)^2}\text{Var}(D_{ij}) \qquad (1.51)\\
&= \frac{2np^2}{(2e^{-2np}\sum_{k=1}^{(n-1)}\left[\sum_{j=k}^{(n-1)}\frac{(np)^j}{j!}\right]^2 - 1)^2}
\end{aligned}$$

Note that the mean is equivalent to the mean derived in the first approach while the variance is equal to the first term of the variance derived in the first approach. Additionally, there can not be a second order approximation of the mean using this approach since a Taylor expansion has not been used.

1.7 Additional Analysis

As was stated in the Data Analysis section (section 2.5) of this paper, we calculated the average unweighted and weighted overlap for each type of social relationship for each village before and after stratification by attribute availability, sex, caste and age. The figures and conclusions regarding stratification by sex are included in sections 2.5 and 3 of the paper. Here, we provide the figures and details of overlap before stratification and after stratifying by attribute availability, caste and age.

Figures 1.7 and 1.8 show the distributions of raw and standardized unweighted overlap for all edges in the network before stratification. Similarly, figure 1.9 shows the distributions of raw and standardized weighted overlap for all edges in the network regardless of nodal attribute information.

We first stratified edges according to the availability of nodal attributes due to the fact that not all villagers completed an individual survey. We labeled nodes with attribute information available 'A' (for attribute) and nodes without attribute information available 'U' (for unknown). Raw and standardized unweighted average overlap for each of the 12 social interactions were calculated separately for ties with both nodes having attribute information (A/A ties), neither node having attribute information (U/U ties), and one node having attribute information and the other not (A/U ties). See Figures 1.10 and 1.11. Figure 1.12 shows the distributions of raw and standardized weighted overlap for A/A, U/U and A/U ties after collapsing the twelve unweighted networks into one weighted network.

It is worth noting a subtle caveat to our method of calculating the standardized overlap values after stratification related to the presence or absence of attribute data. To illustrate, suppose we have a network of 20 people. If no attribute information is available, all of the edges are interchangeable and the total number of possible edges in the network is the usual $\binom{n}{2} = \binom{20}{2}$. Now suppose we introduce attribute information to all of the nodes and label 5 of them male and 15 of them female. The total number of possible male-male ties is $\binom{5}{2}$, the total number of possible female-female ties is $\binom{15}{2}$, and the total number of possible male-female ties is $\binom{20}{2} - \binom{5}{2} - \binom{15}{2} = 75$. We could then use this information to update the denominator of the connection probability for the null model for each type of edge. However, now suppose we only have attribute information for half of the network, say 2 males and 8 females. We could again calculate the total number of possible edges for each type of tie, but we would then be ignoring the contribution of the edges connected to nodes without attribute information. Additionally, after stratification, we calculate overlap for each eligible edge regardless of the neighbors of the nodes attached to the edge having attribute information. To overcome this dilemma, we chose to use $\binom{n}{2}$ as the total number of possible edges in all calculations, regardless of the type of tie. If one did wish to use the the nodal attribute information to update the denominator of the connection probability for each specific type of tie, one could use induced subgraphs. Specifically, if a subgraph included only the nodes with attribute information available, and only the edges connecting two nodes with attribute information, then one could proceed with calculations as in the case where every node had attribute information. We chose to not use these induced subgraphs since they ignore all edges attached to nodes without attribute information, which made up over half of the nodes in each network in this case.

We next stratified by caste membership. Due to the low number of respondents who were members of the scheduled tribe, general caste or scheduled caste, we grouped members of these castes into one caste category and labeled them ‘Other’. Members in the OBC caste were labeled ‘OBC’. The distributions of the raw and standardized average overlap stratified by caste are shown in Figures 1.13 and 1.14. Edges between two individuals in the ‘Other’ caste are labeled as ‘Other’, edges between two individuals in the OBC caste are labeled ‘OBC’, and edges between one individual in the ‘Other’ caste and one individual in the OBC caste are labeled ‘Mixed’. Figure 1.15 shows the distributions of raw and standardized weighted overlap stratified by caste.

Finally, we stratified by age. Similar to caste membership, age was categorized into 4 approximately equally sized groups; 10-29 years, 30-39 years, 40-49 years and 50-99 years. The distributions of the raw and standardized average overlap stratified by age are shown in Figures 1.17 and 1.18. Each age category contains edges connecting two nodes belonging to the same age category. Figure 1.16 shows the distributions of raw and standardized weighted overlap stratified by age.

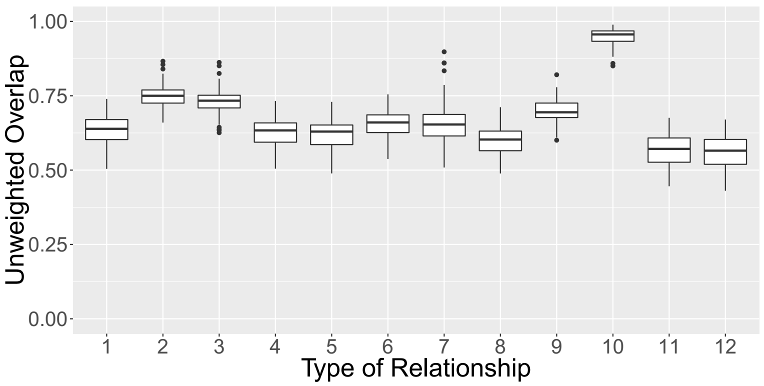


Figure 1.7: Distribution of average unweighted overlap for each type of social interaction. The average overlap was calculated for each type of interaction for each of the 75 villages. The y-axis represents the proportion of average edge overlap and the x-axis represents the type of social interaction. See Table 1.2 above for full descriptions interaction types.

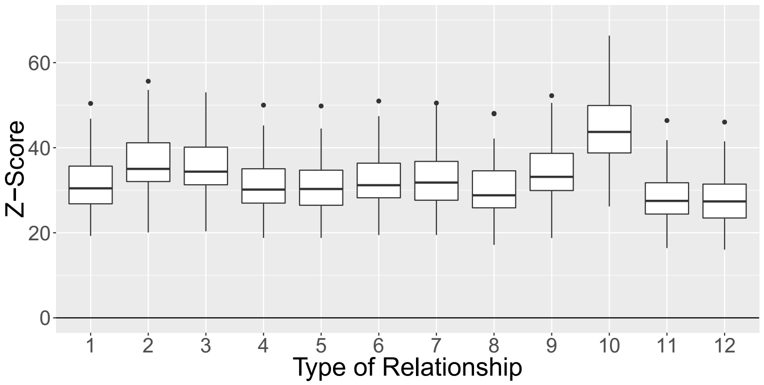
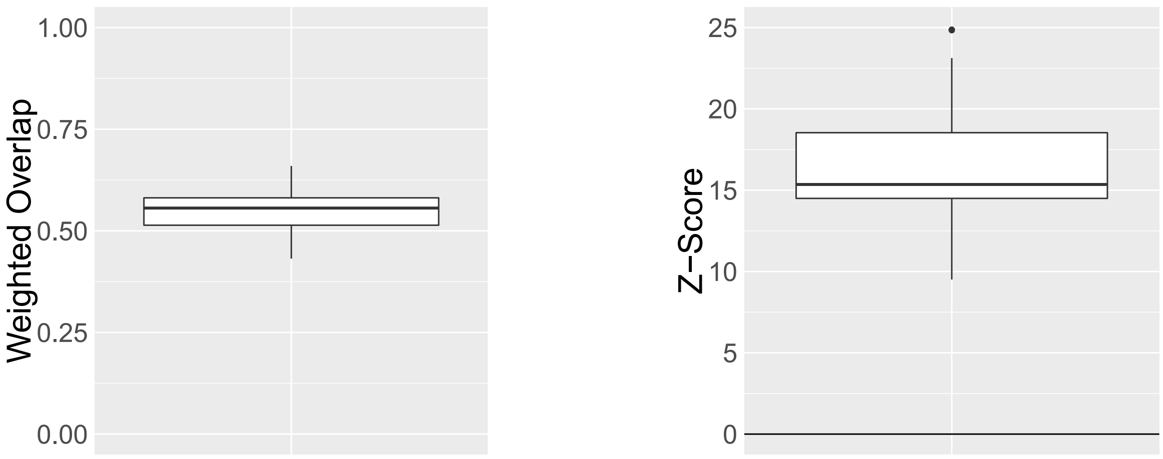


Figure 1.8: Distribution of standardized unweighted overlap for each village for each type of social interaction. Using the approximations from Approach 1, each standardized value was calculated by first subtracting the expected mean overlap under the null from the observed average overlap (the values in Figure 1.7), and then dividing that value by the expected standard deviation under the null. The y-axis represents the standardized value, also known as the Z-score, and the x-axis represents the type of social interaction. See Table 1.2 above for full descriptions interaction types.



(a)

(b)

Figure 1.9: Distribution of average weighted overlap (a) and standardized weighted overlap (b) for all villages.

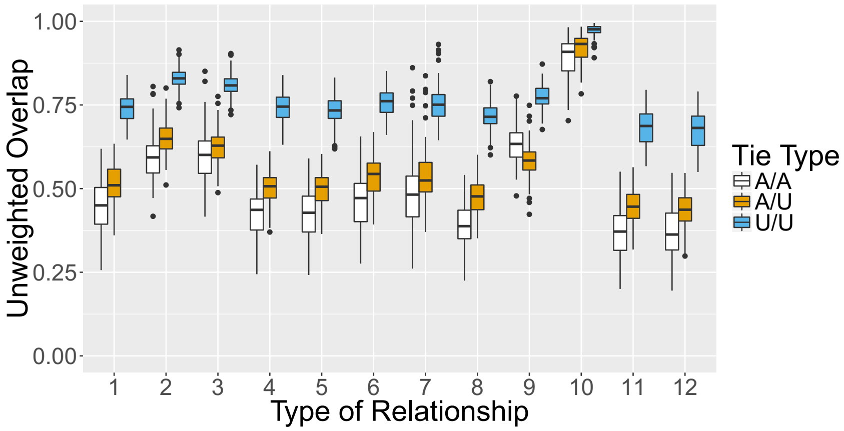


Figure 1.10: Distribution of average unweighted overlap for each village for each type of social interaction stratified by the presence or absence of nodal attributes.

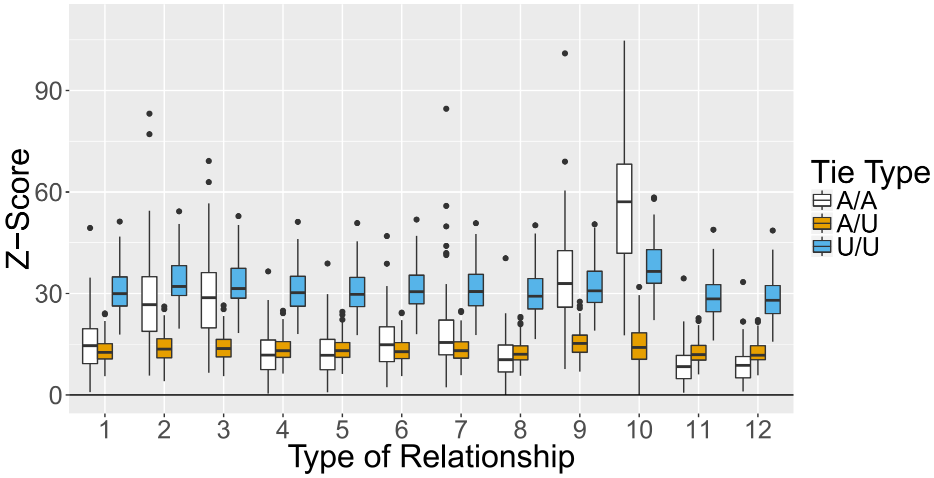


Figure 1.11: Distribution of standardized unweighted overlap for each village for each type of social interaction stratified by the presence or absence of nodal attributes.

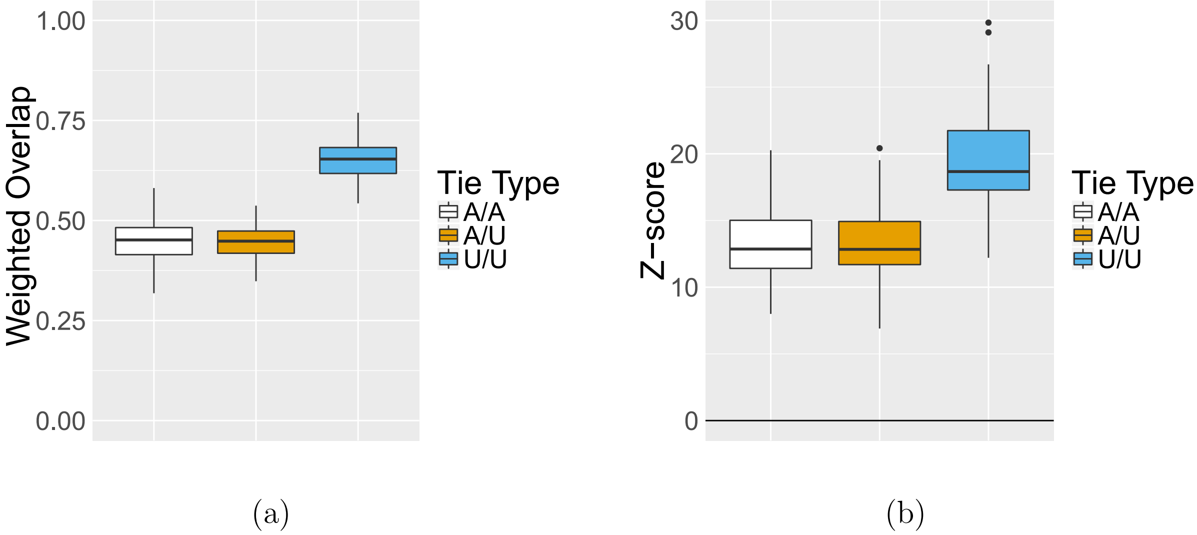


Figure 1.12: Distribution of average weighted overlap (left) and standardized weighted overlap (right) stratified by the presence or absence of nodal attributes.

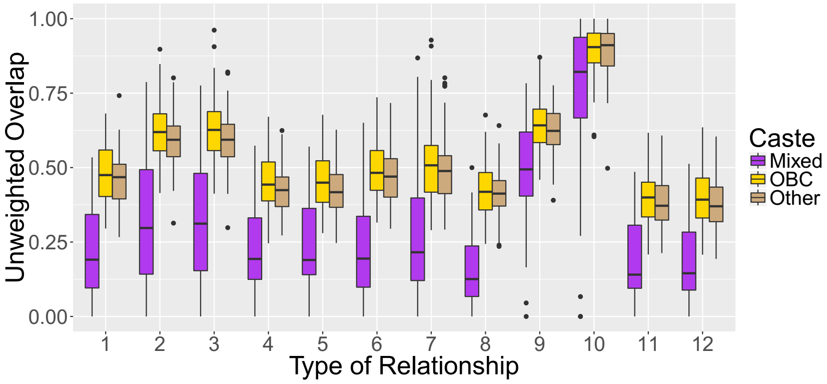


Figure 1.13: Distribution of average unweighted overlap for each village for each type of social interaction stratified by caste. We stratified the edges by caste and labeled an edge between two individuals in the OBC caste 'OBC', edges between two individuals in the Scheduled Caste, Scheduled Tribe or General caste as 'Other', and edges between two individuals in different castes as 'Mixed'. The y-axis represents the proportion of average edge overlap and the x-axis represents the type of social interaction. See Table 1.2 above for full descriptions interaction types.

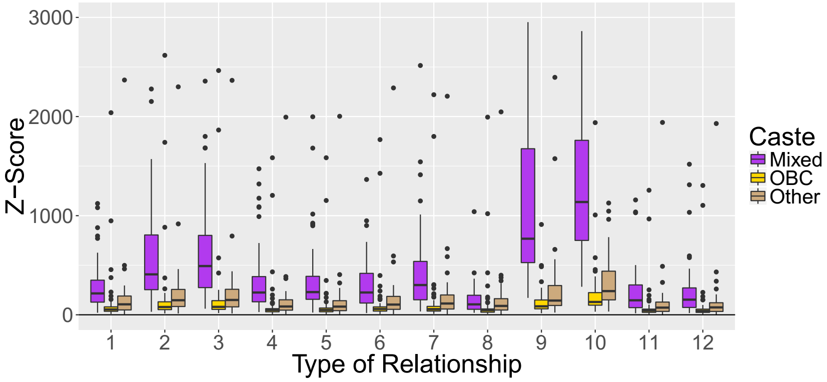
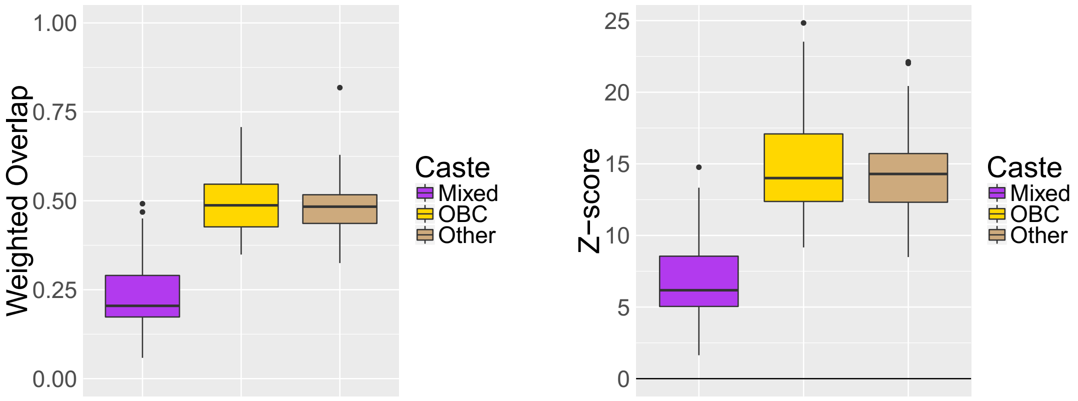
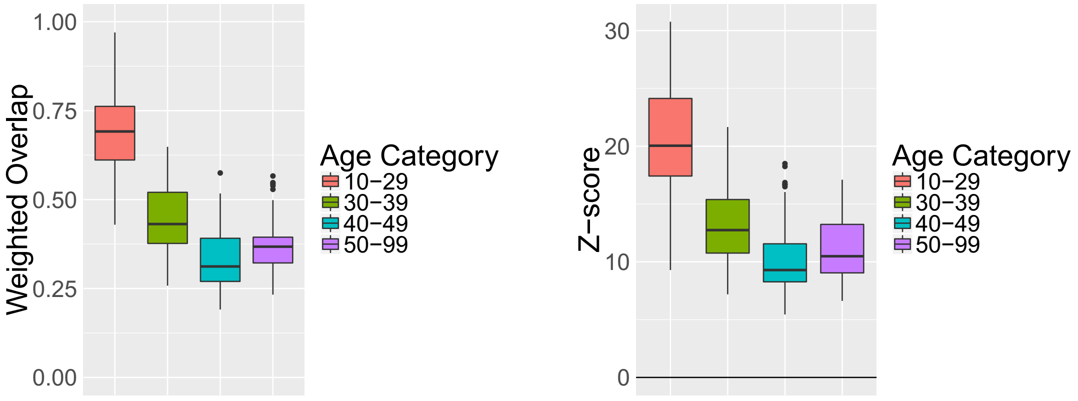


Figure 1.14: Distribution of standardized unweighted overlap for each village for each type of social interaction stratified by caste.



(a) (b)

Figure 1.15: Distribution of average weighted overlap (a) and standardized weighted overlap (b) stratified by caste.



(a) (b)

Figure 1.16: Distribution of raw weighted overlap (a) and standardized weighted overlap (b) stratified by age.

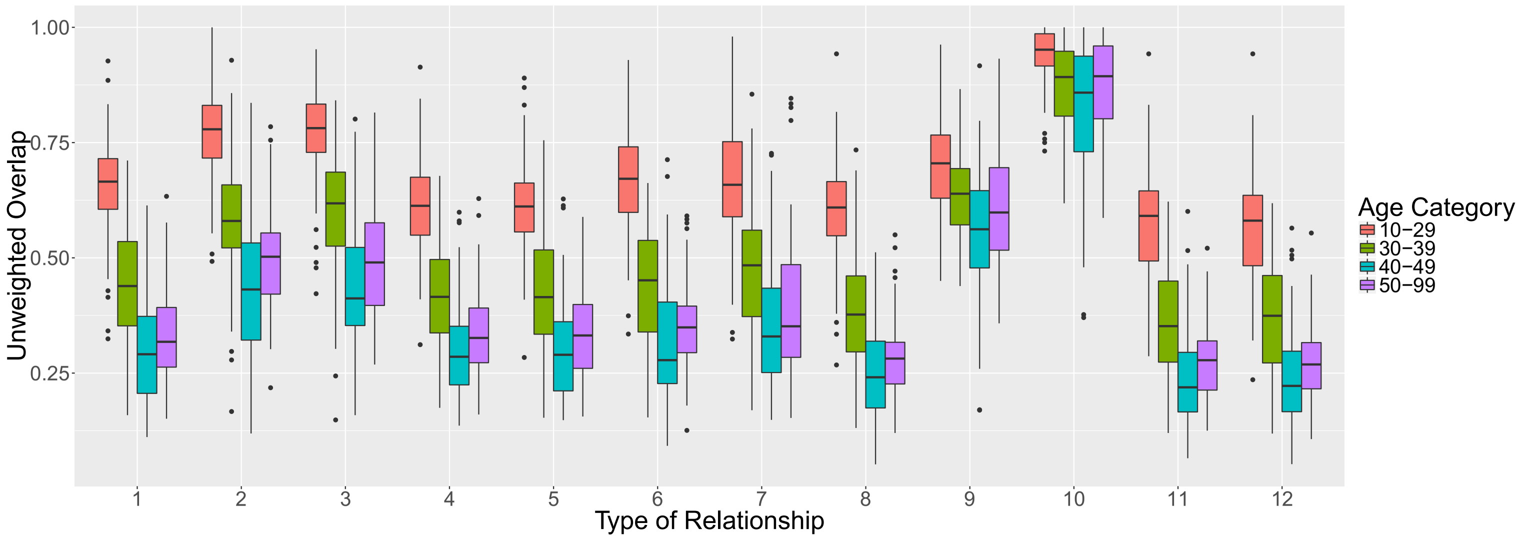


Figure 1.17: Distribution of average unweighted overlap for each village for each type of social interaction stratified by age. We stratified the edges by age category and labeled an edge between two individuals in the 10-29 age group as ‘10-29’, two individuals in the 30-39 age group as ‘30-39’, two individuals in the 40-49 age group as ‘40-49’, two individuals in the 50-99 age group as ‘50-99’. The y-axis represents the proportion of average edge overlap and the x-axis represents the type of social interaction. See Table 1.2 above for full descriptions interaction types.

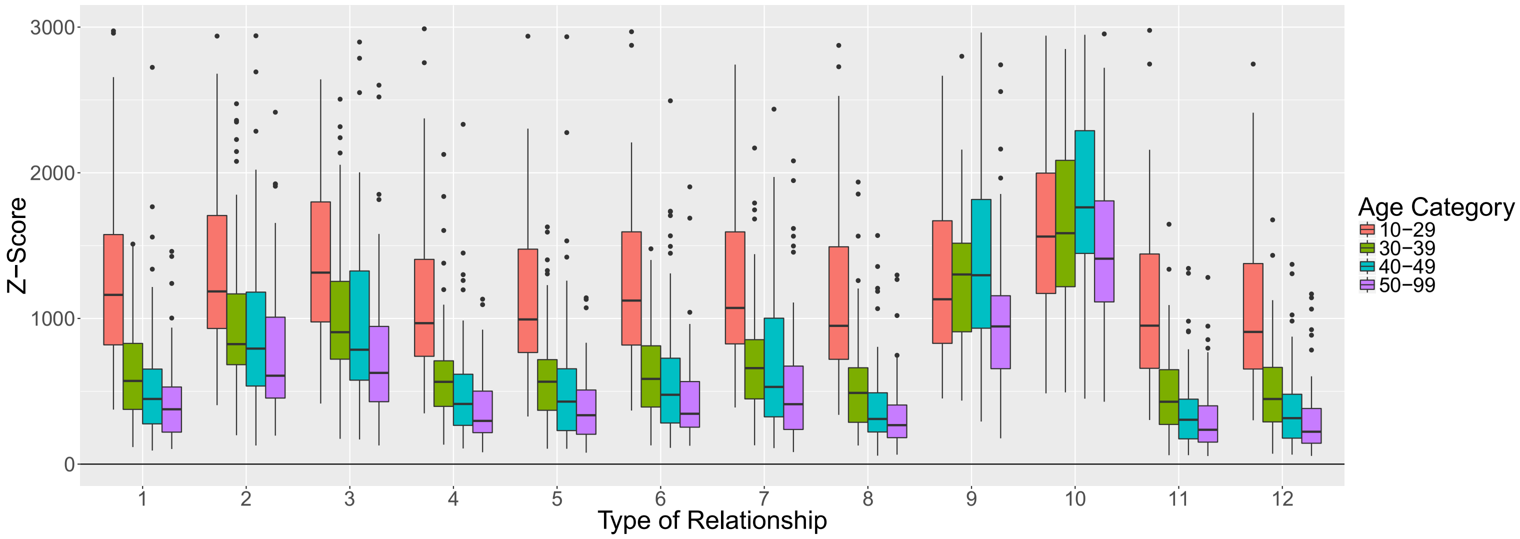


Figure 1.18: Distribution of standardized unweighted overlap for each village for each type of social interaction stratified by age.

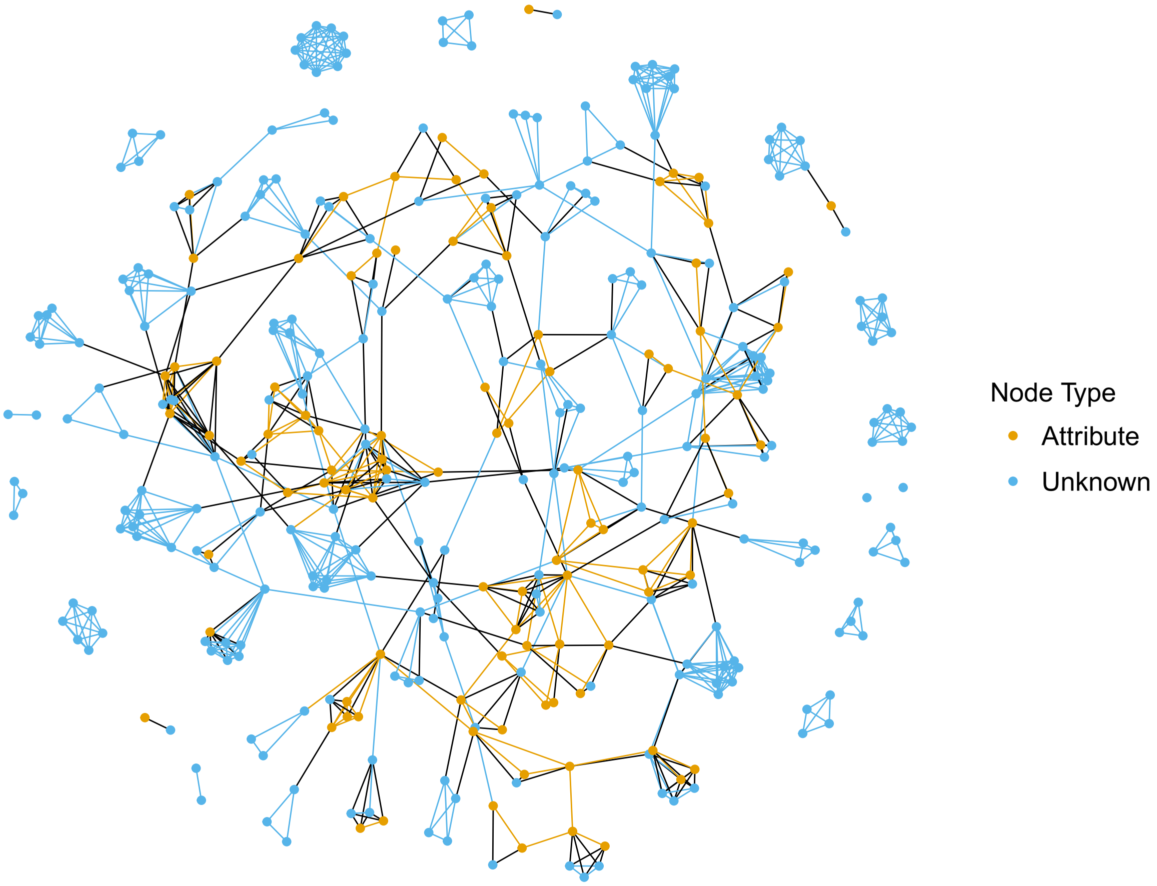


Figure 1.19: Visualization of the interaction type 2 (the respondent gives advice to this individual) network in village 10, stratified by sex. Individuals with attribute data available are colored orange and individuals without attribute information available are colored blue. An edge between two individuals with attribute information is colored orange, an edge between two individuals without attribute information is colored blue and an edge between one individual with attribute information and one individual without is colored black.

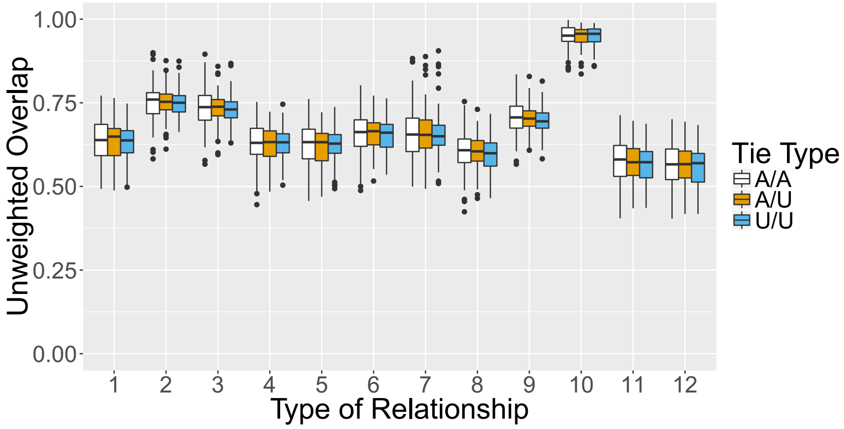


Figure 1.20: Distribution of average unweighted overlap for each village for each type of social interaction stratified by the presence or absence of nodal attributes, after randomization of attributes.

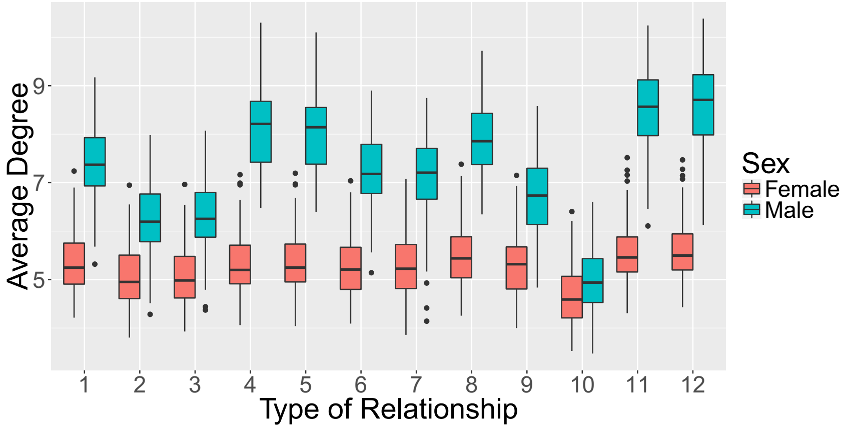


Figure 1.21: Distributions of average degree stratified by sex for each type of social interaction. Average degree is plotted on the y-axis, and type of social interaction is represented on the x-axis.

1.8 Additional Results

Figure 1.7 illustrates the average raw unweighted overlap for each type of social interaction for each village. Each distribution is fairly normally distributed with the exception of interaction types 2, 7 and 10. Each distribution also showcases minimal variance and medians above 0.5. It is also clear that the values of average overlap for social interaction type 10 are very large and could indicate the importance of attending temple among these villages. Figure 1.8 shows the distributions of the standardized unweighted overlaps. Clearly, every value of average unweighted overlap is significantly larger than expected under the null of a random network; the minimum values for each type of interaction never fall below 10 standard deviations from the mean, and the maximum value is greater than 60 standard deviations from the mean. Again, the values from interaction type 10 are among the largest values, suggesting that villagers who attend temple together have a significantly higher proportion of mutual friends compared to other types of interaction and the null model. Values significantly higher than expected under the null are not unusual since social networks are known to have a larger amount of clustering compared to random graphs due to different social mechanisms that drive the formation of clustered ties. Additionally, the Erdős-Rényi random graph model is the simplest null model and is easily rejected when analyzing empirical social networks. The distribution of average weighted overlap (Figure 1.9a) is normally distributed with a mean of 0.548 and standard deviation of 0.046. Each village's average weighted overlap is significantly different from what is expected under each corresponding null value (Figure 1.9b). This is expected given the values in Figure 1.8 for each type of social interaction are also significantly higher than expected, and that humans do not typically create friendships randomly.

Figure 1.10 shows a clear pattern in average unweighted overlap when stratified by the existence of attribute information. For every type of social interaction, the median average overlap for U/U ties is the largest, followed by A/U ties, and finally A/A ties. The one exception is for interaction type 9 (the respondent is related to this individual) where A/A ties have a larger median value than A/U ties. In most relationships, the median average overlap is 50% higher among U/U ties compared to A/A and A/U ties. When standardized (Figure 1.11), the values for the A/A and A/U ties are very similar, except for interaction types 2, 9 and 10 where the values for the A/A ties are much higher. For more details see Figure **??**. The values for the U/U ties are still significantly higher than the other types of ties, with the exception of interaction types 9 and 10 where they are quite similar to A/A ties, with none of the values falling within 15 standard deviations from the mean. Surprisingly, there are several values for the A/A and A/U ties that are not statistically significantly different from the mean of the null model. Such a discrepancy in the values of overlap (both raw and standardized) suggests that individuals who were not sampled to complete an individual survey form more tightly-knit groups and point to a possible sampling bias. Villagers were randomly sampled to complete an individual survey after stratifying by religion and geographic sub-location. However, as in most attribute information collection, the structure of the network was not taken into account when sampling individuals to administer the survey to. This leads to a loss of information for significant parts of the network which could include much of the network's community structure and inhibit analysis of the network (See Figure 1.19). If attribute information were truly randomly sampled, we would expect to see very similar values of overlap for each type of tie, as in Figure 1.20 in section 1.7, where we randomly assigned attribute information to individuals in each village and calculated average overlap again. This could point to a potential bias in the sampling of villagers for completing individual surveys, and could indicate that overlap would be a useful metric to include in a sampling scheme or for recognizing sampling bias for network data. The average weighted overlap stratified by node attribute information (Figure 1.12a) follows, not surprisingly, the same pattern as its unweighted counterpart (Figure 1.10). The values of average weighted overlap are extremely similar for A/A and A/U edges, and the values for U/U nodes are significantly larger. All of the values of weighted overlap are significant (Figure 1.12b), which is again expected from the unweighted distributions in Figure 1.11.

The median average unweighted overlap is quite similar for the OBC and 'Other' caste categories across all interaction types (Figure 1.13). The 'Mixed' category has the least amount of overlap across all interaction types, except type 10 which again suggests the importance of going to temple together. It isn't surprising that the 'Mixed' category would have the lowest values of overlap; most social interaction is done among members of the same caste. The standardized overlap plot (Figure 1.14) shows an interesting pattern. The 'Mixed' category has the most significant values as well as the only non-significant values. The values that did not reach significance are not surprising due to the low amount of social activity across castes. The significant values could be due to those individuals having lower degree or the small amount of cross-caste ties. The OBC and 'Other' distributions are much more similar to each other and less significant overall with the exception of several outliers. The distributions of average weighted overlap follow a similar pattern as seen in the unweighted case; the OBC and 'Other' categories values are comparable and significantly higher than the 'Mixed' category values (Figure 1.15a). This pattern holds after standardization (Figure 1.15b), which is a departure from what was seen in the unweighted case (Figure 1.14). All values are significantly higher than expected under the null except for a few values in the 'Mixed' category. This could again be due to the small number of cross-caste ties in those villages.

Figure 1.17 shows a distinct pattern in the median unweighted overlap values for all interaction types when stratified by age. The 10-29 year age category contains the highest values of overlap, followed by the 30-39 age group, then the 50-99 age group and finally the 40-49 age group. The differences in values across age categories are minimal for interaction type 10, which once again suggests that regardless of category, individuals of similar ages who attend temple together have a high proportion of friends in common. Interestingly, a slightly different pattern is observed when overlap is standardized (Figure 1.18). Except for types 9 and 10, the median values of overlap decrease by age category. All categories are comparable for types 9 and 10. All of the standardized values are again significantly larger than expected under the null. Figure 1.16a showcases the same pattern among the age categories for the average weighted overlap as seen in the unweighted plot in Figure

1.17. The median value for average weighted overlap is highest among the 10-29 age group, the 30-39 age group is the second largest, followed by the 50-99 age group, and finally the 40-49 age group. The same trend holds for the standardized values, and all of the values are significantly larger than the expected value under the null hypothesis (Figure 1.16b).

2

The Social Bow Tie

Abstract

In social networks the notion of tie strength, and the factors that influence it, have received much attention in a myriad of disciplines for decades. With the internet and cellular phones providing additional avenues of communication, measuring and inferring tie strength has become much more complex. Numerous models incorporating indicators of tie strength have been proposed and used to quantify relationships in both online and offline social networks, and a standard set of structural network metrics have been applied to predominantly online social media sites to predict tie strength. Here, we introduce the concept of the "social bow tie" framework, which for any given network tie is a small subgraph of the network that consists of a collection of nodes and ties that surround the tie of interest, forming a topological structure that resembles a bow tie. We also define several intuitive and interpretable metrics that quantify properties of the bow tie which enable us to investigate associations between the strength of the "central" dyadic tie and properties of the bow tie. We combine the bow tie framework with machine learning to investigate what aspects of the bow tie are most predictive of tie strength in two very different types of social networks, a collection of medium-sized social networks from 75 rural villages in India and a nationwide call network of European mobile phone users. For two connected individuals, we find that the more their friendship circles overlap, the stronger the tie between them. Conversely, the more close-knit each individual's separate friendship network, the weaker the tie between them.

Our findings also demonstrate that incorporating properties of the bow tie results in more accurate predictions of tie strength and a more nuanced understanding of the factors that are associated with it.

2.1 Introduction

The strength of any kind of relationship between two individuals lies on a spectrum. People in general have a close relationship with only a few friends or family members, a somewhat weaker tie with a larger group of individuals with whom they interact less frequently, and an even weaker connection with a large number of casual acquaintances. This tradeoff between tie strength and the number of people a person is connected to through his or her ties was elegantly captured by Dunbar Dunbar (1992). Measuring and predicting tie strength, and moreover, understanding the factors that drive tie strength, has been an expanding area of interest, with increasing utility and complexity in the digital age, i.e., the ever-increasing forms of communication via mobile phones and social media. Knowledge of the strength of a tie, as well as the social dynamics contributing to tie strength, has been shown to increase the accuracy of link prediction, enhance the modeling of the spread of disease and information, and lead to more targeted marketing Li et al. (2013); Linyuan and Tao (2010); Sá and Prudêncio.

Several indicators of tie strength have been proposed, perhaps most notably by Mark Granovetter in his seminal work The Strength of Weak Ties Granovetter (1973). Granovetter differentiated between strong and weak ties and proposed the weak ties hypothesis: the stronger the tie between any two people, the higher the fraction of friends they have in common Granovetter (1973). Much of the current methodology centered on tie strength has stemmed from Granovetter's weak ties hypothesis and his proposed four dimensions of tie strength: the amount of time spent interacting with someone, the level of intimacy, the level of emotional intensity, and the level of reciprocity. More recently, three additional dimensions of tie strength have been proposed: 1) emotional support Marsden and Campbell (1984); Wellman and Wortley (1990), 2) structural variables, i.e. network topology Ellison et al. (2007); Lin et al. (1981); Xiang et al. (2010), and 3) social distance, i.e. the difference in socioeconomic status, education level, political affiliation, race, and gender He et al. (2006); Lin et al. (1981). These categories have facilitated the definition and quantification of numerous possible predictors of tie strength; some generalizable to any network, and some specific to a limited number of social networks. Of importance to this analysis is a corresponding perspective outlined by Elizabeth Bott Bott (1957) that suggests that the degree of clustering in an individual's network has the potential to draw them away from a dyadic tie if there are not mutual ties.

Initially, highly generalizable similarity indices such as the number of common neighbors two nodes share, preferential attachment, and path distance were used to infer tie strength. These metrics were most commonly used for link prediction and were shown to provide some information regarding tie strength Linyuan and Tao (2010); Pappalardo et al. (2012). However, it was quickly discovered that the addition of nodal attributes and other metrics not solely based on network topology greatly enhanced the measurement and prediction of tie strength Kahanda and Neville; Luarn and Chiu (2015). Gilbert and Karahalios defined indicators of tie strength specific to a network of Facebook users and built a predictive model that achieved 85% accuracy for binary tie strength (weak vs. strong) classification Gilbert and Karahalios (2009). They found that the act of communicating once leads to a significant increase in tie strength, and that educational difference plays a role in determining tie strength. Pappalardo et al. introduced a measure of tie strength using multiple online social networks and found that the strength of a tie is related to the number of interactions between the two individuals Pappalardo et al. (2012). In addition, several studies have shown that frequent communication, both online and offline, is positively related to tie strength Marsden and Campbell (1984); Wiese et al..

While previous studies have provided advances and valuable insights, they suffer from a binary definition of tie strength (weak vs strong), low diversity in the types of social networks studied (the vast majority being social media sites), and non-representative samples. In this work, we propose a decomposition of a social network into an ensemble of interconnected "social bow ties," constellations consisting of nodes and ties that surround each network tie. We call any such subgraph a "social bow tie" because the topological structure that surrounds each tie resembles a bow tie. We also introduce several simple metrics that quantify properties of the bow tie. Further, we use random forests and linear regression to build models that predict categorical and continuous measures of tie strength from different properties of the bow tie, including nodal attributes (covariates) of the nodes included in the bow tie. We apply our framework to two social networks, a collection of 75 social networks from the villages of Karnataka, India, and a call network of European mobile phone subscribers. We find that the bow tie framework contributes to more accurate predictions of tie strength and provides insights on which metrics are the most informative of tie strength. Specifically, we find that the larger the proportion of shared friends, the stronger the tie, and the more clustered the individual friendship circles (consisting of non-overlapping friends), the weaker the tie. Consequently, these findings provide evidence to support both the weak ties hypothesis and a generalized version of the Bott hypothesis Bott (1957).

This paper is organized as follows. In Section 2.2 we describe the construction of each social network, and our measures of tie strength. We then introduce our framework for tie strength prediction and detail our method in Section 2.3. Model selection and results are presented in Section 2.4, and we discuss our conclusions in Section 2.5. Additional figures are included in Section 2.6.

2.2 Data Description

We analyzed two social network data sets. The first data set is social network data collected in 2006 from 75 villages located in 5 districts in rural southern Karnataka, India. The data were collected through household and individual surveys as part of a study by Banerjee et al. Banerjee et al. (2013). Of relevance for this study, the survey included social network data along 12 dimensions: friends or relatives who visit the respondent's home, friends or relatives the respondent visits, any kin in the village, non-relatives with whom the respondent socializes, those from whom who the respondent receives medical advice, with whom who the respondent goes to temple to pray, from whom the respondent would borrow money, to whom the respondent would lend money, from whom the respondent would borrow material goods from, to whom the respondent would lend material goods, from whom the respondent gets advice, and to whom the respondent gives advice. It is worth noting that these forms of interaction are largely face-to-face, unlike the mediated material from the call detail records (CDRs) described below. Additionally, a proportion of villagers were given individual surveys that recorded age and sex, among other attributes.

For this data set, we define the strength of a tie as the number of distinct types of social relationships reported to exist between the two individuals. For example, if individual i borrows money from individual j and in addition gives advice to individual j, the weight of the (undirected) tie between i and j would be equal to 2. If i and j also attend temple together, their tie strength would be 3 and so on, with a minimum strength of 1 and a maximum strength of 12 for any tie. Note that a tie strength of 0 implies that the two individuals are not connected by any kind of social tie. We denote the strength of a tie between individuals i and j as w_{ij}. Because we ignore the directionality of ties, our definition of tie strength is symmetric.

The second data set consists of call detail records (CDRs) from a mobile phone provider in an undisclosed European country where 68% of citizens own a smartphone and 85% own a cellular phone. The data examined here span a period of three months in 2013, and each record consists of the following daily aggregate communication summaries for pairs of individuals: the date, anonymized caller ID, anonymized callee ID, daily call duration (in minutes), daily number of calls, daily number of text messages (SMS), and daily number of multimedia messages (MMS). Age, sex, and billing zip codes were available for a large majority (72.3%) of individuals.

An undirected, weighted call network was created from the records by first summing the call durations between any two individuals over the three-month period. If two individuals spoke on the phone at least once during the period, we connected them with an edge of strength w_{ij}, where the value of edge strength was set to the total amount of time spent on the phone with one another. Since tie strength is defined in terms of absolute time, it does not take into account the total amount of time each individual spends on the phone, which makes it somewhat difficult to quantify the relative strength of ties since the strength of a tie is not measured on the same scale either for individuals or pairs of individuals. We therefore normalized tie strength and represent it with two measurements: one that represents tie strength from the perspective of individual i, and one that represents tie strength from the perspective of individual j. Specifically, for each tie, the first measurement of tie strength is the total call duration (w_{ij}) divided by the total time individual i spends on the phone s_i, the strength of node i. similarly, the second measurement of tie strength is the total call duration divided by the total time individual j spends on the phone s_j, the strength of node j. Dividing total call duration by the strength of each focal node results in a consistent definition of tie strength. We denote these new tie strength measurements as y_{ij} and y_{ji}. We created another summary measure of tie strength by taking the average of y_{ij} and y_{ji}, and we denote this $z_{ij} = \frac{1}{2}(y_{ij} + y_{ji})$.

2.3 Methods

To introduce the "bow tie" structure, consider a weighted social network G, which may be directed or undirected, and consider a tie with weight w_{ij} that connects two individuals i and j. We call these two individuals the focal nodes of the bow tie. We use the term focal tie to refer to the tie that links them. We start by partitioning i's friends and j's friends into three disjoint sets. Group i, denoted g_i, contains the nodes that are connected to only i; group j, denoted g_j, contains nodes that are connected to only j; and group ij, denoted g_{ij} , contains nodes that are connected to both $i$and j. These three groups jointly make up the shared and non-shared friends of i and j. We call this structure the ij bow tie. Formally, the groups g_i, g_j and g_{ij} are induced subgraphs, where the node sets that induce them are the neighbors of i, the neighbors of j, and the common neighbors of i and j, respectively. The bow tie ij, denoted by G_{ij} , is the subgraph that is induced by the union of all neighbors of i and j. Note that G_{ij} is more than the sum of g_i, g_j and g_{ij} : in addition to containing the same set of nodes and ties as those subgraphs do, it also contains the inter-group ties among this set of nodes, i.e., the ties linking nodes across g_i, g_j and g_{ij} . Important to our analysis below is the hierarchical structure of the bow tie: at the upper level of hierarchy we have the bow tie G_{ij}; at the intermediate level, we have the three groups, g_i, g_j and g_{ij}; and at the lowest level we have the nodes and ties from which each group is composed. A simple example of the bow tie structure surrounding nodes i and j is shown in Figure 2.1. The localized nature of the bow tie framework gives rise to several topological metrics that can be used to predict tie strength. We include unweighted Onnela et al. (2007) and weighted Mattie and Onnela (2017) edge overlap, which we denote *overlap* and *woverlap*, respectively. Metrics based on customized versions of the clustering coefficients of i and j are used, where the calculation of a clustering coefficient is limited to the non-shared friends of each node, i.e., for node i, the nodes and edges in g_i are used to calculate the clustering coefficient of i, and similarly, g_j is used for node j. We denote the sum and absolute difference of these quantities as *ccSum* and *ccDiff* for the unweighted clustering coefficients, and *wccSum* and *wccDiff* for the weighted clustering coefficients. Here, we use the definition of weighted clustering coefficient provided by Saramki et. al. Saramaki et al. (2007). Specifically, the weights of ties are considered and the metric reflects how large triangle weights are compared to a network maximum. Other predictors include the sum and absolute difference in the degrees of i and j (*kSum* and *kDiff*), the sum and absolute difference in the strengths of i and j (*sSum* and *sDiff*), the number of nodes and edges in g_{ij} (n_{ij} and e_{ij}), and the sum and absolute difference in the number of nodes and the number of edges in g_i and g_j (*nSum*, *nDiff*, *eSum* and *eDiff*). Predictors created from the attributes of i and j include the sum and absolute difference in the ages of i and j (*AgeSum* and *AgeDiff*), the paired sex category (male-male, male-female, female-female) denoted *SexPair*, and an indicator if i and j have the same billing zip code, denoted *ZipPair*. See Table 2.1 for a detailed description of each variable.

To predict tie strength and study how it is associated with different metrics, we used regression as well as Random Forest (RF) regression and classification Breiman (2001). For the India social network, tie strength is discrete with $w_{ij} \in \{1, \ldots, 12\}$. Thus, the weight of a

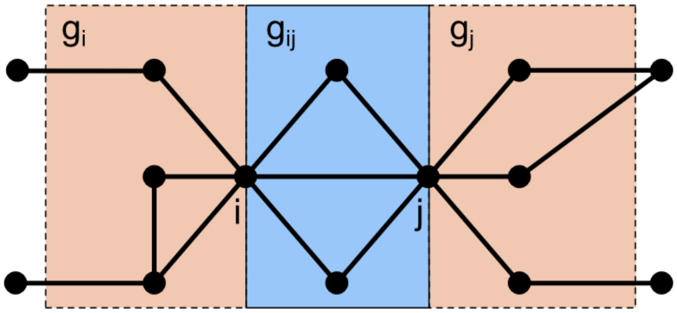


Figure 2.1: A simple example of the social bow tie. The center panel highlights the nodes and edges that comprise the overlapping friendship circle of nodes i and j, denoted g_{ij}. The left and right panels contain the individual (non-overlapping) social circles of the focal nodes, denoted g_i for node i and g_j for node j.

tie can be viewed as a categorical outcome, allowing RF classification and Poisson regression to be used to predict tie strength, or as continuous with RF regression used for prediction. For the CDR call network, tie strength is most naturally treated as a continuous variable, and we used RF regression and linear regression to predict both measures of tie strength.

In addition to ordinary least squares (OLS) regression, least absolute shrinkage and selection operator (LASSO) and ridge regression were used to fit more parsimonious and interpretable models as well as increase prediction accuracy. Before using LASSO and ridge regression, all data was centered around the mean and 10-fold cross validation was performed to select the best tuning parameters; denoted λ^L for LASSO and λ^R for ridge regression. For RF classification, the number of trees used was 200, and the maximum number of features (covariates) considered when splitting a node was $\sqrt{n}$ where n is the total number of features. For RF regression, 200 trees were used and the maximum number of features considered when splitting a node was n.

Nodal attributes were expected to be informative of tie strength and were therefore included in the models. All missing attributes were imputed for each data set. Because both age and sex were recorded for a subset of the villagers in India, full attribute information was available for some individuals but completely missing for others. This resulted in three types of ties: those with complete attribute information available for both individuals, those

Variable Name	Description
kSum	Sum of the degrees of i and j $(k_i + k_j)$
kDiff	Absolute difference in the degrees of i and j $(\lvert k_i - k_j \rvert)$
sSum	Sum of the strengths of i and j $(s_i + s_j)$
sDiff	Absolute difference in the strengths of i and j $(\lvert s_i - s_j \rvert)$
ccSum	Sum of the clustering coefficients of i and j
ccDiff	Absolute difference in the clustering coefficients of i and j
wccSum	Sum of the weighted clustering coefficients of i and j
wccDiff	Absolute difference in the weighted clustering coefficients of i and j
AgeSum	Sum of the ages of i and j
AgeDiff	Absolute difference in the ages of i and j
SexPair	Categorical variable; Male-Male, Male-Female, Female-Female
ZipPair	Indicator if i and j have the same billing zip code
overlap	Unweighted overlap of edge between i and j
woverlap	Weighted overlap of edge between i and j
n_{ij}	Number of common friends of i and j
e_{ij}	Number of edges among the common friends of i and j

Table 2.1: Descriptions of predictors of tie strength used in the analyses.

with attribute information available for only one individual, and those missing all attribute information. To use all available information, imputation was performed in three stages for this data set. We first imputed individual i's sex and age if individual j's attributes were known, using j's attributes to infer i's attributes. We then imputed individual j's attributes if individual i's attributes were known, similarly using i's attributes to infer j's attributes. Finally, attributes for both i and j were imputed if neither individual's attributes were known. In each stage, RF classification was used to impute sex and RF regression was used to impute age. Individuals in the CDR call network could have any combination of age, sex and billing zip code information missing. We again used RF classification to impute sex and RF regression to impute age. Because of the abundance of billing zip code possibilities, rather than imputing billing zip code directly, we created a paired billing zip code dichotomous variable equal to 1 if the two focal nodes had the same billing zip code and 0 if they did not. We then used RF classification to impute paired billing zip code.

2.4 Results

2.4.1 India Social Network

The India network contained 69,444 nodes, of which 16,984 (24.5%) had full attribute information available, and 294,778 edges after the removal of isolated ties. Of these, 37,714 (12.8%) edges were between two individuals with complete attribute information available, 107,739 (36.5%) were between one individual with and one individual without attribute information available, and 149,492 (50.7%) were between two individuals without any attribute information available. We discovered tie strength had a bimodal distribution with $\approx 46\%$ of ties having a strength of 12. This was due to the fact that the majority ($\approx 96\%$) of ties between individuals living in the same household had a weight of 12. We decided to exclude ties between individuals from the same household and only included cross-household ties. This resulted in a Poisson distribution of tie tie strength.

RF regression and classification were used to fit three models both before and after nodal attribute imputation, where ties with complete attribute information available were included in the analysis before imputation and all ties were included after imputation. Model 1 is the full model and includes all covariates described in 2.1 with the exception of *ZipPair* since it is specific to the CDR data set; Model 2 includes all covariates except weighted overlap; and Model 3 includes all covariates except unweighted overlap. It has been shown that categorical predictors do not need to be split into multiple dichotomous covariates (referred to as dummy variables) when implementing RF if there are a small number of them and their cardinality is low Breiman (2001); Hastie et al. (2001). Therefore, the variable *SexPair* was not split into two separate dummy variables due to its low cardinality and it being the single categorical predictor. Accuracy was measured as the residual, the absolute difference between empirical tie strength (w_{ij}) and predicted tie strength ($\hat{w}_{ij}$). Figure 2.2(a) shows the accuracy of RF regression and classification after imputation for Model 3. RF regression predicted tie strength exactly with 52.9% accuracy, and RF classification with 71.6% accuracy. Within one unit of tie strength, an accuracy of 70.1% and 80.8% was achieved by RF regression and classification, respectively. The accuracy of all three models using RF regression and classification, both before and after imputation, are shown in Figures 2.4(a) and 2.5(a) in section 2.6.1. Imputation boosts regression accuracy by approximately 15% and classification accuracy by approximately 10%. Furthermore, Model 3 outperforms models 1 and 2.

Feature importance for each of the three models after imputation for both RF regression and classification is shown in Figure 2.2(b)-(d). The horizontal bars represent how informative the predictor is with a longer bar meaning more informative. The black vertical line represents the value of an equilibrium or null importance if every predictor were equally informative. For both classification and regression, weighted overlap is the most informative variable in models 1 and 3, and the sum of the clustering coefficients is the most informative in model 2, followed by unweighted overlap. These results provide evidence that the proposed indicators of tie strength in the weak ties and Bott hypotheses (the overlap of friendship circles and the amount of clustering in the non-overlapping friendship circles) are predictive of tie strength. Feature importance plots using RF regression and classification before and after imputation are shown in Figures 2.4(b)-(d) and 2.5(b)-(d) in the supplementary material. The results before and after imputation are quite similar for both regression and classification.

Poisson regression was used to model the associations between tie strength and each of the predictors, and the coefficients of significant predictors with magnitude greater than ± 0.2 are reported in Table 2.3. The predictors with the largest magnitudes include $woverlap$, $ccSum$, and MF . Weighted overlap is positively associated with tie strength, illustrating the greater the proportion of strength among overlapping friends of the focal nodes, the stronger the tie between the focal nodes, and showing evidence to support Granovetter's hypothesis. The sum of the clustering coefficients of the focal nodes is positively associated with tie strength, meaning tie strength decreases as the amount of clustering in the non-overlapping friendship circles increases. This provides quantitative evidence of Bott's hypothesis in a novel population. Finally, the predictor MF is positively associated with tie strength, indicating that on average, MF ties are stronger than MM ties; the reference group.

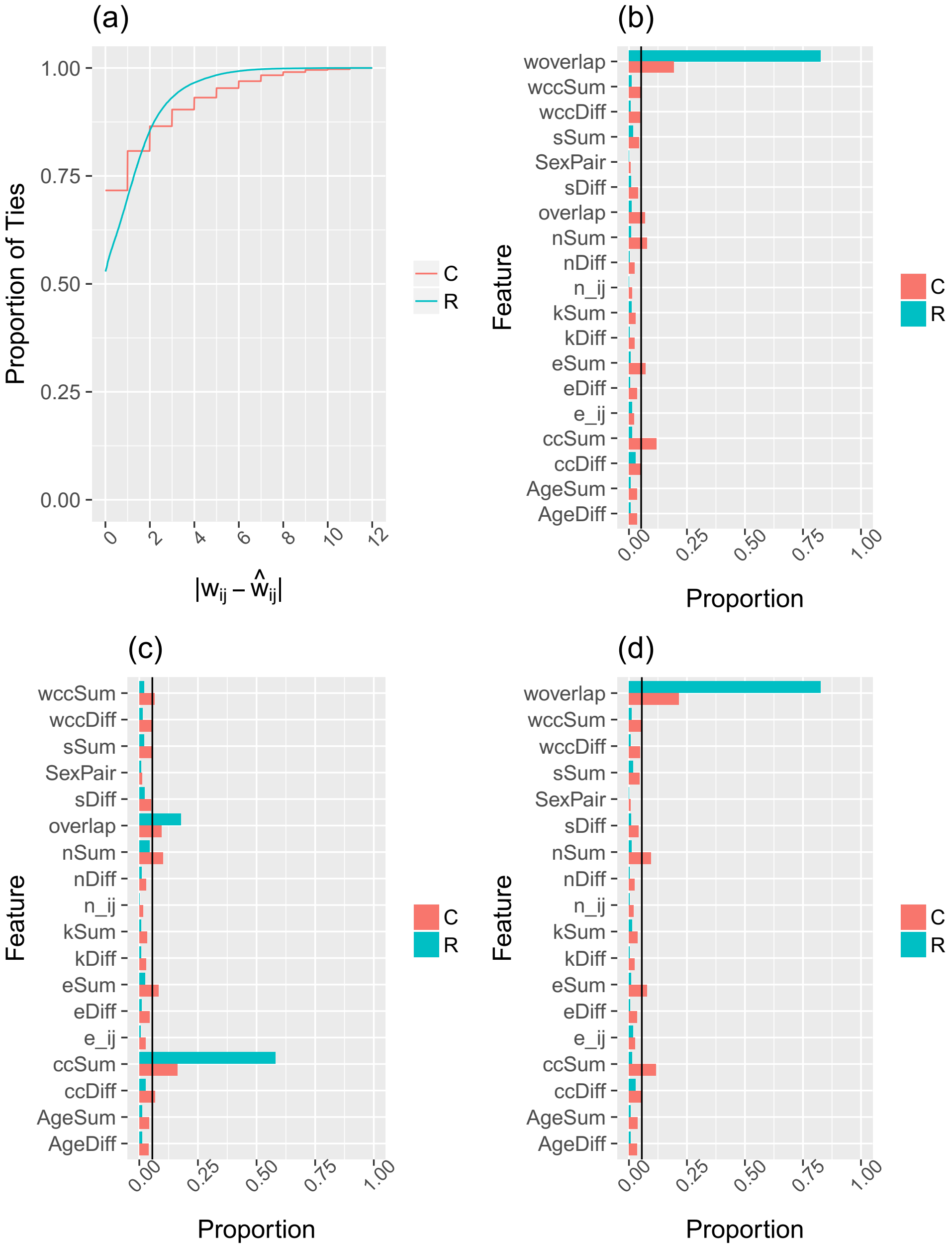


Figure 2.2: Accuracy and feature importance plots for the India social network. Accuracy, measured as the absolute difference between empirical tie strength (w_{ij}) and predicted tie strength ($\hat{w}_{ij}$), for Model 3 using both RF regression (R) and classification (C) after imputation is shown in (a). Feature importance using RF regression and classification after imputation are shown for Model 1 (b), Model 2 (c) and Model 3 (d). The horizontal bars represent how informative the predictor is with a longer bar meaning more informative. The black vertical line represents the value of an equilibrium or null importance if every predictor were equally informative.

Model	Predictors	Coefficient	Adjusted R^2
A	*woverlap*	4.91	0.8021
	ccDiff	0.78	
	wccSum	-0.22	
	n_{ij}	-1.38	
B	*woverlap*	4.61	0.8001
	ccDiff	0.77	
	wccSum	-0.28	
	n_{ij}	-1.01	
C	*woverlap*	4.82	0.8002
	ccDiff	0.79	
	wccSum	-0.47	
	n_{ij}	-1.28	

Table 2.2: Poisson regression results for the India social network. Predictors, coefficients and the adjusted R^2 value is reported.

2.4.2 CDR Call Network

The CDR call network contained 2,276,495 nodes and 12,345,848 edges. Age was available for 89.25% of the individuals and had a mean of 48.2 (sd = 18.2) years. Of the 89.03% of individuals whose sex was recorded, 52.51% were male. Billing zip code was available for 99.35% of individuals. Due to the large size of the network, a random sample of 500,000 edges was drawn. After the removal of isolated ties, a total of 496,941 edges remained. Full attribute information was available for both focal nodes for 359,367 (72.3%) edges.

Similar to the India data set, three models were fit with RF regression both before and after nodal attribute imputation for each measure of tie strength and are denoted Models 1-3. Figure 2.3(a) shows the accuracy for RF regression after imputation for all three models and each measure of tie strength. The difference in accuracy for all models is very minimal and only one curve is visible for each tie strength measure. Within 0.05 units (a 5% difference between empirical and predicted tie strength), an accuracy of 61% was achieved for normalized tie strength, and 56.7% for averaged tie strength. Within 0.1 units, an accuracy of 76.5% was achieved for normalized tie strength and 77.3% for averaged tie strength. Accuracy for all models before and after imputation are shown in Figures B1(a) and B2(a).

Imputation has a smaller impact on accuracy for this data set in all cases.

Feature importance for each of the three models after imputation is shown in Figure 2.3(b)-(d). The black vertical line represents the value of importance if every predictor were equally informative. The most informative predictors in each model are *sSum*, *sDiff*, *nSum* and *kSum*, with *woverlap* and *AgeSum* slightly more informative than the null importance value in models 1 and 3. This suggests focal node strength, degree and number of non-overlapping friends are the aspects of the bow tie most predictive of tie strength in this network. Feature importance plots for all models and all tie strength measures before and after imputation are presented in Figures B2(b)-(d) and B3(b)-(d) in the supplementary material.

For each measure of tie strength, three different models, denoted Models A - C, were fit using linear regression methods following imputation. Model A denotes the full model that was fit using OLS regression. Model B was fit using LASSO and Model C using ridge regression. Because the distributions of normalized and averaged tie strength are highly skewed for this data set (see Figure 2.6 in section 2.6.2), we first log-transformed each measure of tie strength and then centered them around the mean. All predictors were standardized (centered around the mean with unit variance) before fitting models B and C. Implementing LASSO and ridge regression require the selection of tuning parameters that determine the extent of shrinkage administered when calculating coefficient estimates. As the tuning parameter approaches 0, the corresponding coefficient estimates match the OLS estimates. In this extreme, the amount of bias is minimal, if nonexistent, but the amount of variance is comparatively high. As the tuning parameter is increased, the values of the coefficients decrease and approach 0 once the tuning parameter is sufficiently large. In this extreme, bias is increased but variance in the estimates is decreased. The optimal choice for a tuning parameter balances the amount of bias and variance and can be selected via cross-validation. We performed 10-fold cross validation to select values of the tuning parameters λ^L and λ^R. The values of the LASSO coefficients as a function of λ^L and, as a more interpretable measure, the l_1 penalty $\|\hat{\beta}^L_\lambda\|/\|\hat{\beta}\|_1$ which represents the amount of

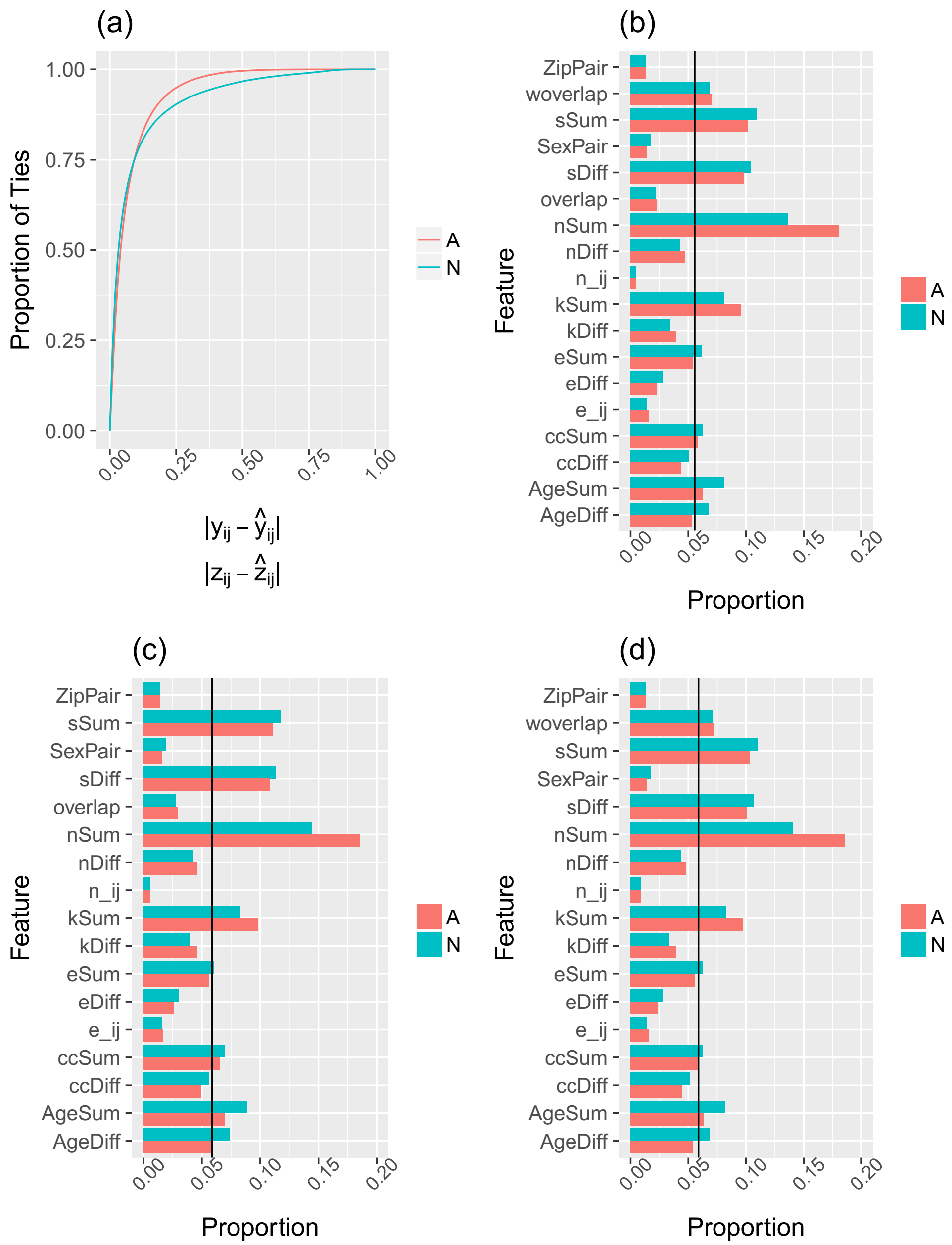


Figure 2.3: Accuracy and feature importance plots for the CDR call network with normalized (N) and averaged (A) tie strengths. Accuracy, measured as the absolute difference between empirical tie strength (y_{ij}, z_{ij}) and predicted tie strength $(\hat{y}_{ij}, \hat{z}_{ij})$, for all three models using RF regression after imputation is shown in (a). Note that only one curve is visible for each strength measure since the accuracy of all three models is indistinguishable. Feature importance using RF regression after imputation are shown for Model 1 (b), Model 2 (c) and Model 3 (d). The horizontal bars represent how informative the predictor is with a longer bar meaning more informative. The black vertical line represents the value of an equilibrium or null importance if every predictor were equally informative.

shrinkage, are shown in Figures 2.7 and 2.9 in section 2.6.2. The values of the ridge regression coefficients as a function of λ^R and the l_2 penalty $\|\hat{\beta}_\lambda^R\|/\|\hat{\beta}\|_2$ are shown in Figures 2.8 and 2.10 in section 2.6.2. Significant predictors, their coefficients, adjusted R^2 values and the values of the tuning parameters for models B and C are presented in Table 2.3.

For normalized tie strength, λ^R was sufficiently large such that no shrinkage was implemented, and the estimated ridge regression coefficients are equivalent to the OLS estimates. The amount of LASSO shrinkage was approximately 12%, resulting in slightly different coefficient estimates. In all models, *overlap*, *kDiff*, *sSum*, *ccDiff* and *ZipPair* were significantly associated with tie strength. Edge overlap is positively associated with tie strength in all models, showing that as the proportion of common friends two individuals share increases, so does the strength of the tie between the two individuals, supporting Granovetter's hypothesis. Tie strength is negatively associated with *kSum* and *sSum*. This suggests that as the focal nodes expand their social circles and the time spent interacting with friends, the weaker the tie between them; an appearance of Dunbar's tradeoff. The positive association between *ZipPair* and tie strength implies having the same billing zip code increases the strength of a tie and could suggest a geographical impact on tie strength. Analogous to the India social network, *ccDiff* is positively associated with tie strength in this call network, and the more dissimilar the non-overlapping clustering coefficients of the focal nodes, the stronger their tie. Lastly, the R^2 values for these models are on the lower side (0.112 on average). This could be due to the network being constructed with phone-based communication rather than face-to-face interactions among highly clustered villagers. Furthermore, quantifying tie strength for CDR data is currently still rather ambiguous; the operationalization of using communication as a proxy for tie strength has not yet been validated Wiese et al.. An alternate measure of tie strength may increase the R^2 values.

2.5 Discussion

In this work, we introduce the social bow tie; a novel framework we use to perform a comprehensive analysis of the association between network structure and tie strength. Our

		Normalized Strength (y_{ij})			Averaged Strength (z_{ij})		
Model	Predictor	λ	Coeff.	Adj. R^2	λ	Coeff.	Adj. R^2
A	*overlap*	-	0.27	0.116	-	0.27	0.117
	kDiff		-0.35			-0.35	
	sSum		-0.25			-0.25	
	sDiff		-			-0.20	
	ccDiff		0.29			0.29	
	ZipPair		0.23			0.23	
B	*overlap*	0.01	0.21	0.115	0.022	-	0.110
	kDiff		-0.33			-0.21	
	sSum		-0.25			-0.39	
	ccDiff		0.23			0.24	
	ZipPair		0.23			0.23	
C	*overlap*	10^3	0.27	0.116	10^3	0.28	0.100
	kDiff		-0.35			-0.27	
	sSum		-0.25			-0.49	
	sDiff		-			0.31	
	ccDiff		0.29			0.36	
	ZipPair		0.23			0.24	

Table 2.3: Poisson regression results for the India social network. Predictors, coefficients and the adjusted R^2 value is reported.

framework decomposes a social network into a collection of nodes and ties immediately surrounding each network tie. This utilization of local structure produces easily interpretable metrics that quantify social perspectives of tie strength and allows for analyses that are computationally feasible for networks of any size. Through machine learning and regression methods including LASSO and ridge regression, we determine which properties of the bow tie structure are the most predictive of tie strength in two different types of social networks; a contact network of Indian villagers and a nationwide call network of European mobile phone users.

Overall, following Granovetter, we find that the more friends two individuals share, the stronger their tie. Following Bott, the more tightly-knit their individual social circles, the weaker their tie. Furthermore, we find Dunbar's tradeoff between tie strength and the size of an individual's social circle present in both social networks. In addition, we find that the bow tie framework provides metrics that predict tie strength with high accuracy for both networks.

In future work, it would be interesting to apply the bow tie framework to a social network of married couples. In this case the dominant strong tie has properties that are not seen in more casual social ties, namely the individuals constitute a particularly strongly defined social institution that has both emotional (romantic attachment) as well as structural (e.g. common responsibility for children and common ownership of capital investments such as a home) elements that provide it resiliency. This would enable testing of the original version of Bott's hypothesis, rather than a generalized form as we present here. It would also be interesting to test if the strength of in-person ties behaves similarly for the mobile phone call network.

2.6 Additional Figures

2.6.1 India Social Network

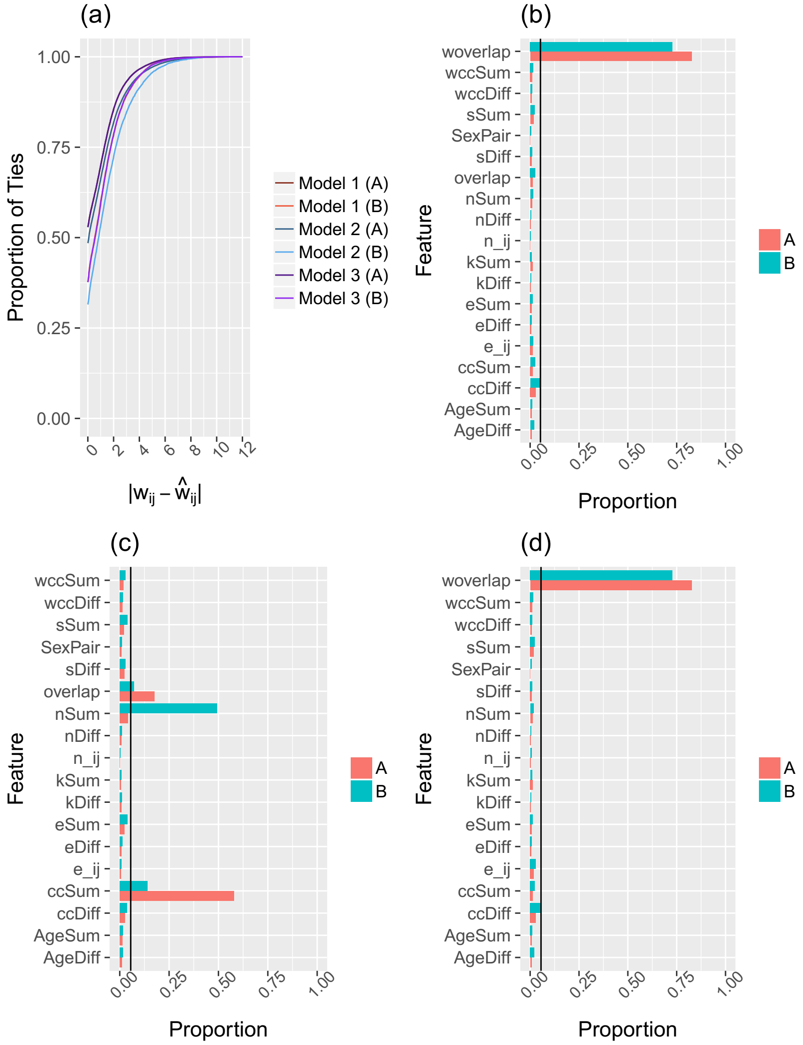


Figure 2.4: Accuracy and feature importance plots for the India social network. Accuracy, measured as the absolute difference between empirical tie strength (w_{ij}) and predicted tie strength ($\hat{w}_{ij}$), for all three models using RF regression before (B) and after (A) imputation is shown in (a). Feature importance using RF regression before and after imputation are shown for Model 1 (b), Model 2 (c) and Model 3 (d).

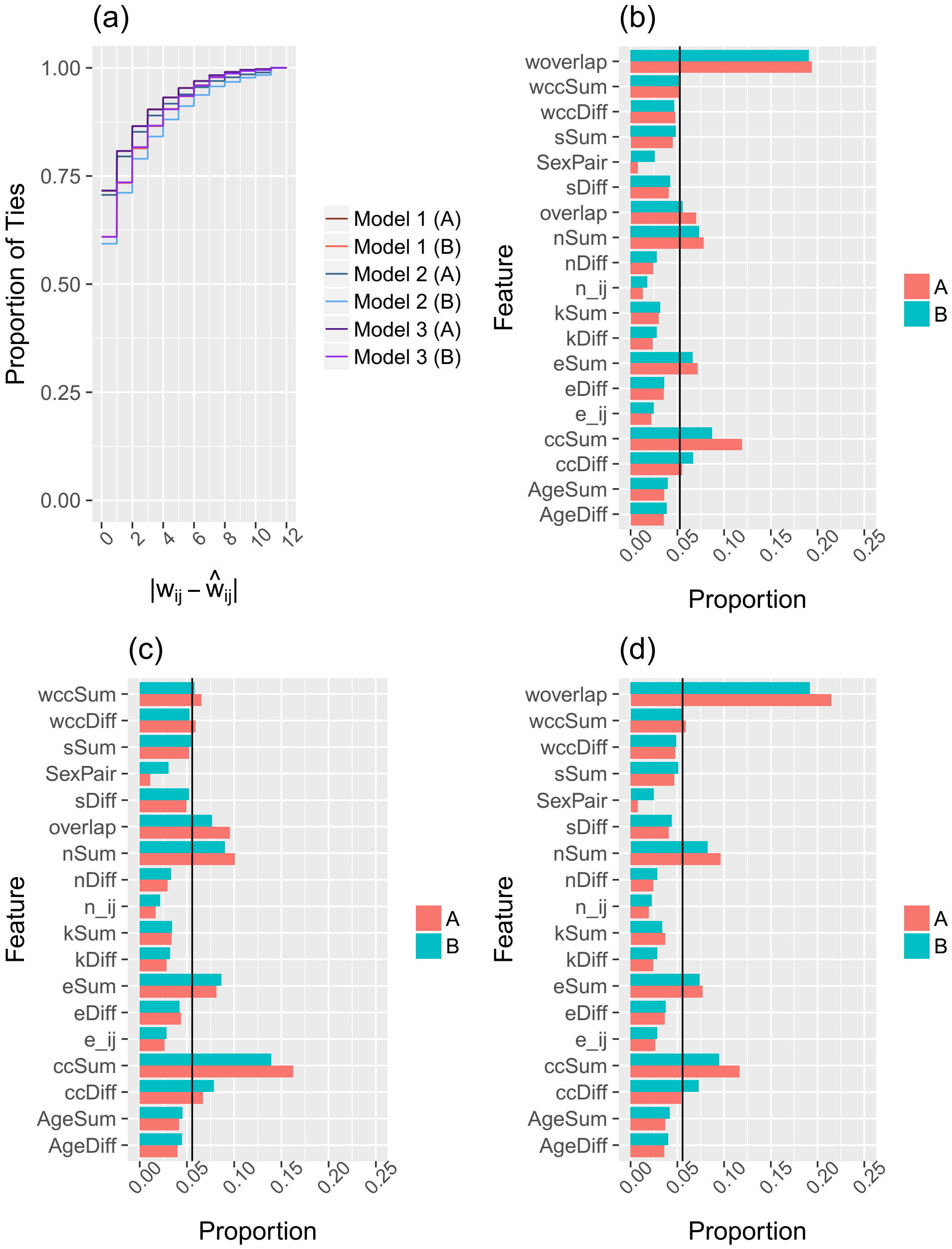


Figure 2.5: Accuracy and feature importance plots for the India social network. Accuracy, measured as the absolute difference between empirical tie strength (w_{ij}) and predicted tie strength ($\hat{w}_{ij}$), for all three models using RF classification before (B) and after (A) imputation is shown in (a). Feature importance using RF regression before and after imputation are shown for Model 1 (b), Model 2 (c) and Model 3 (d).

2.6.2 CDR Call Network

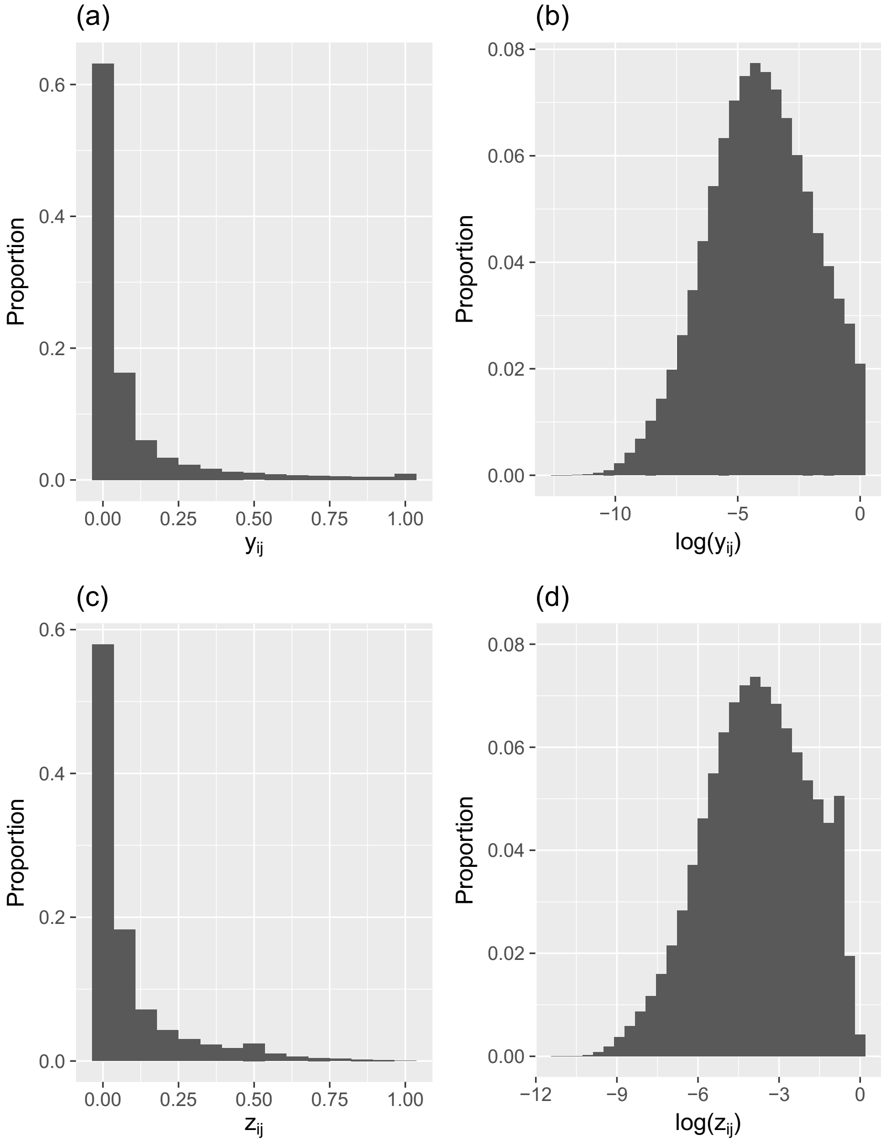
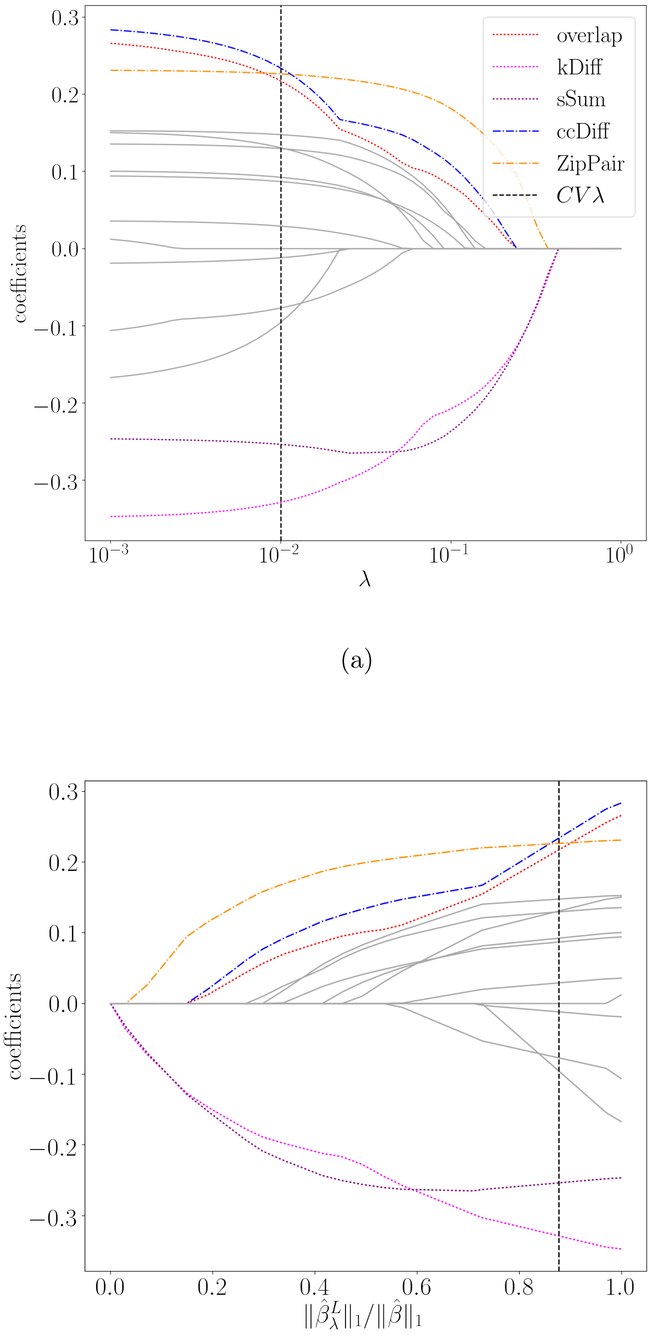


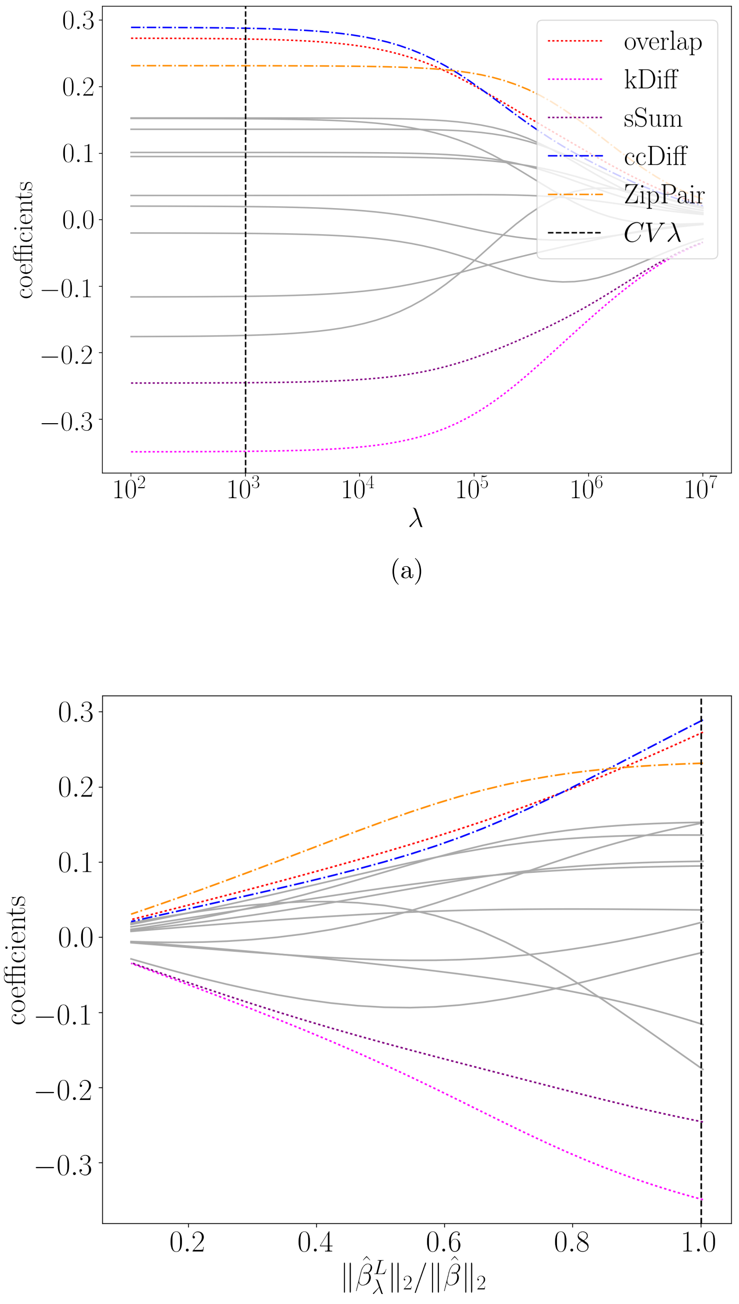
Figure 2.6: Distributions of (a) normalized call duration y_{ij}, (b) the natural log of y_{ij}, (c) averaged call duration z_{ij}, and (d) the natural log of z_{ij}.



(a)

(b)

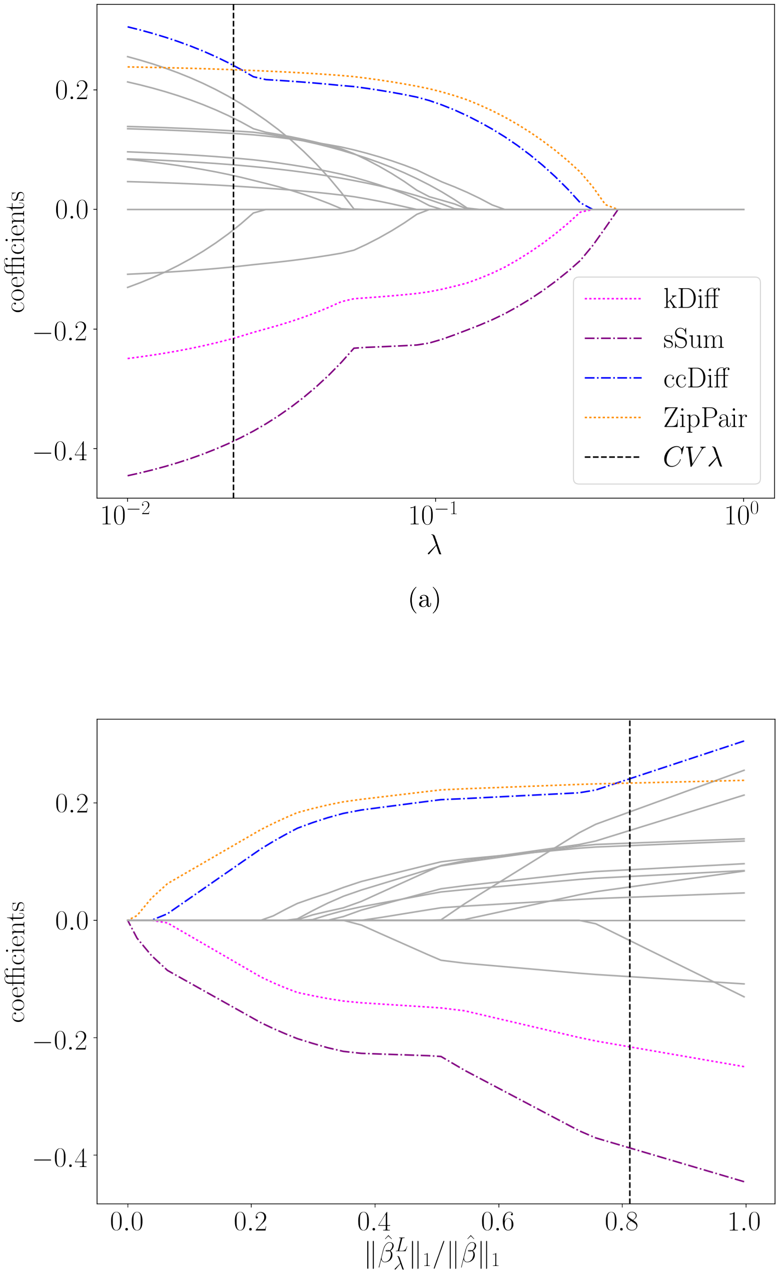
Figure 2.7: The standardized LASSO coefficients as a function of λ^L (a) and $\|\hat{\beta}_L\|/\|\hat{\beta}\|_1$ (b) using 10-fold cross validation for the CDR normalized tie strength data set after imputation. Each line represents a different predictor. The dashed black line indicates the value of chosen via cross validation.



(a)

(b)

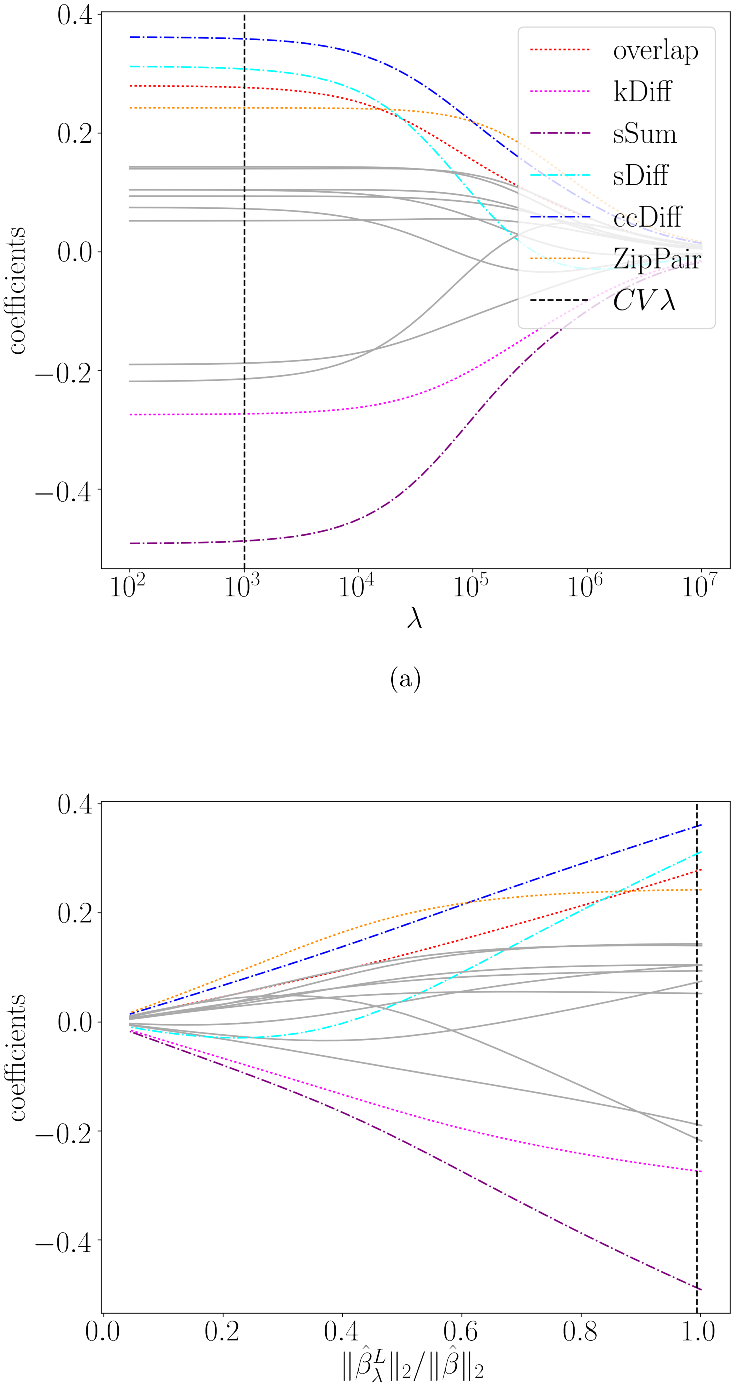
Figure 2.8: The standardized ridge regression coefficients as a function of λ^R (a) and $\|\hat{\beta}_L\|/\|\hat{\beta}\|_2$ (b) using 10-fold cross validation for the CDR normalized tie strength data set after imputation.



(a)

(b)

Figure 2.9: The standardized LASSO coefficients as a function of λ^L (a) and $\|\hat{\beta}_L\|/\|\hat{\beta}\|_1$ (b) using 10-fold cross validation for the CDR averaged tie strength data set after imputation. Each line represents a different predictor. The dashed black line indicates the value of chosen via cross validation. The colored lines represent the predictors significantly different than 0.



(a)

(b)

Figure 2.10: The standardized ridge regression coefficients as a function of λ^R (a) and $\|\hat{\beta}_L\|/\|\hat{\beta}\|_2$ (b) using 10-fold cross validation for the CDR averaged tie strength data set after imputation.

3

Imputation in Social Networks Using Super Learner

Abstract

Missing data and non-response are common occurrences in, and great hindrances to, the analysis of social network data. While any kind of statistical analysis can be negatively affected by missingness, the effects can be even more detrimental in network data analysis due to the high sensitivity of missing data on network topology and the complexity of network surveys and data collection. Many imputation methods have been introduced in the classical statistics literature as a way to maintain power and sample size in the presence of missing data. However, the extension of these methods to the networks framework has been scarcely studied. Here we use Super Learner to impute both edge and nodal attributes of a nationwide call network of European mobile phone users with varying amounts of missingness. We impute the age, age category, and sex of individuals, and the total call duration and text message communication between two individuals over a three-month time period. We find that Super Learner performs better or as well as any individual learning algorithm alone for the imputation of each attribute, and that the amount of missingness does not significantly affect performance. Additionally, we find that the accuracy of age category imputation is sensitive to the choice of categorical thresholding. A thresholding scheme that results in approximately equal proportions of individuals in each category ensures a gain in age-stratified accuracy over the null accuracy of random assignment, but a lower overall accuracy when compared to thresholding resulting in imbalanced categories.

3.1 Introduction

Standard statistical methods are designed to analyze complete data sets where all data has been observed. However, missing data and non-response are common occurrences in social network data. Network data collection can be time consuming and complex, resulting in missing nodes (individuals) and edges (social connections between individuals). One of the main causes of missingness in social networks is non-response, of which there are two types: unit non-response, where a node and all of its ties and attributes are missing, and item non-response, where data is missing for particular ties or nodal attributes Huisman (2007). While missingness hampers all statistical analyses, its impact can be substantially greater for network data analyses since network structure and metrics are extremely sensitive to missingness Burt (1987)Borgatti and Molina (2003).

The impact of not only missing data, but the mechanism driving the missingness, i.e., how missing variables are related to the underlying values of the variables in the data set, were largely ignored until Rubin Rubin formalized missingness theory in 1976. The most restrictive, and most commonly assumed mechanism in practice, is missing completely at random (MCAR). In this scenario, the probability of missingness does not depend on the values of missing or observed data. Missing at random (MAR) is less restrictive and assumes missingness only depends on the observed data. A simple example being male patients refusing to answer survey questions about depression, but their non-response not being related to the level of their depression. The least restrictive assumption is that data is not missing at random (NMAR), and missingness could also depend on the data not observed. A common example of this mechanism is patients dropping out of a study due to a higher severity of the disease being studied. Ignoring the cause of missingness or choosing the incorrect missingness mechanism, results in invalid inference by introducing significant bias and attenuation of regression coefficients. Perhaps even more detrimental to analysis, is removing all missing data and analyzing only 'complete cases', which results in a loss of information and statistical power due to a decrease in sample size, and the diminished ability to report conclusions about the target population being studied Little and Rubin (2002). Methods for data analysis in the presence of missingness have been created and fall into three main categories: likelihood-based methods, inverse probability weighting of complete cases, and imputation. The focus of our work pertains to imputation.

Data imputation is the process of replacing missing data with plausible estimates. This results in a complete data set with maintained sample size and no loss of information, which can be analyzed using standard statistical techniques Little and Rubin (2002). Furthermore, if the observed data provides information on the missing data mechanism, better predictions of missing values can be obtained Little and Rubin (2002). Methods for selecting the best value to substitute fall into main two categories; single and multiple imputation. Single imputation methods for networks include reconstruction Stork and Richards (1992), where the missing part of the network is reconstructed using the observed relations (ties) of the missing actors, unconditional mean, i.e., imputing the density of a network as the probability of imputing an edge, preferential attachment, where the probability that an individual i will be connected to another individual j is proportional to the degree of j, and hot-deck imputation where the attributes and structural properties of completely observed nodes are used as 'donors' and replace missing nodes and edges. Initially, imputation was conducted using only topological metrics as predictors. More recently, it was found that including the attributes of nodes as predictors can increase the accuracy of both imputation and link prediction, and several studies have since used a combination of structural metrics and nodal attributes, as well as more advanced algorithms for imputation Gong (2014) Brea et al. (2014) Dong et al. (2014) Zamal et al. (2012) Gong and Liu (2016). Brea et al. presented the reaction-diffusion algorithm for the imputation of age of mobile phone users in Mexico Brea et al. (2014). Age was partitioned into four categories and the probability of being assigned to each category was updated at each time step with topological metrics and nodal attributes impacting the probabilities. At the last time step, an individual was assigned the age category with the highest probability. Dong et al. were the first to introduce an algorithm for imputing two attributes (age and sex) simultaneously; the Double Dependent-

Variable Factor Graph Model Dong et al. (2014). They use homophily in conjunction with the maximization of objective functions to impute sex and categorical age for individuals in a network created from phone call and SMS messages. While both of these more recent algorithms show significant gains in accuracy, they are limited to binary and categorical dependent variables.

In this work we employ the loss-based supervised learning method Super Learner to impute binary, categorical, and continuous edge and nodal attributes of a nationwide call network of European mobile phone users, and study how the amount of missingness affects our results. Our analysis includes predictors that are novel to network attribute imputation, and the optimal combination of a collection of prediction algorithms for the imputation of each attribute. We find that Super Learner predicts attributes with greater or as much accuracy as any of the other learning algorithms alone for each attribute. Our analysis also provides evidence that the accuracy of imputation of age category is highly sensitive to category thresholding.

The rest of this paper is as follows. In section 2 we describe the data set used for analysis. Section 3 details the attributes imputed, the predictors chosen for imputation, the different amounts and mechanism of missingness introduced, and the learning algorithms included in the Super Learner library. Our results are presented in Section 4 and discussed in Section 5.

3.2 Data Description

The data set used consists of call detail records (CDRs) from a mobile phone provider in an undisclosed European country where 68% of citizens own a smartphone and 85% own a cellular phone. The records span a period of three months in 2013, and each record consists of the following daily aggregate communication summaries for pairs of individuals: the date, anonymized caller ID, anonymized callee ID, daily call duration (in minutes), daily number of calls, daily number of text messages (SMS), and daily number of multimedia messages (MMS). Age, sex, and billing zip codes were available for a large majority (72.3%) of individuals. An undirected, weighted call network was created from the records by first summing the call durations and number of text messages between any two individuals over the three-month period. If two individuals spoke on the phone or communicated via text message at least once during the period, we connected them with an edge. Each edge is characterized by two measures of strength; w_{ij}^{CD}, the total amount of time spent on the phone with each other, and w_{ij}^{SMS}, the total number of text messages shared.

3.3 Methods

Imputation was performed for nodal attributes age (continuous and categorical) and sex (binary), as well as the edge attributes total call duration between two individuals, and total SMS communication between two individuals. We first introduce the variables used for the imputation of age and sex for an individual i. Because age and sex are nodal attributes, we used node-level metrics for imputation, with many based on the assumption of homophily. The metrics included i's degree, clustering coefficient, call duration strength, i.e., the total time i spends on the phone, SMS strength, i.e., the total number of SMS messages sent or received by i, the average degree of i's friends, the average time i's friends spend on the phone, the average number of SMS messages sent or received by i's neighbors, the proportion of node i's friends that are male, the proportion of friends that are female, the mean age of i's friends and its standard deviation, the median age of i's friends, and the entropy of node i's call duration strength and SMS strength. Here, entropy measures how an individual distributes their total strength, either call duration or SMS messages, among their neighbors, where the value is maximized when strength is evenly distributed among all neighbors. See Table3.1 for a full description of all variables. Note that for each individual with missing attributes, we assume neither their age or sex is known. Therefore, the age of an individual i is not used to impute the sex of individual i, and similarly the sex of an individual is not used to impute their age or age category.

Similar metrics were chosen for the imputation of attributes of an edge between two individuals i and j.The predictors chosen include standard network metrics as well as novel metrics introduced in Mattie et al. (2017). The standard metrics include the sum and absolute difference in the degrees of i and j, unweighted Onnela et al. (2007) and weighted Mattie and Onnela (2017) edge overlap. The metrics proposed by Mattie et al. (2017) include the sum and absolute difference in customized versions of the clustering coefficients of i and j, the number of friends i and j share, and the number of edges among the shared friends. Predictors created from the attributes of i and j include the sum and absolute difference in the ages of i and j, and an indicator of the paired sex category, i.e., male-male, male-female, female-female. See Table3.2 for a full description of all variables.

The imputation of each attribute was performed using Super Learner with 10-fold cross-validation. As required, we re-scaled all predictors (centered around the mean with unit variance), and created indicator variables for all categorical variables (often referred to as dummy variables) Polley and van der Laan (2010). Then, each sample was split into 10 groups of 100 observations. We then randomly removed attributes from $p\%$ of the observations with $p \in \{5, 10, 20, 25, 50\}$, simulating a MCAR missing data mechanism. Three libraries of learning algorithms were created; one for the imputation of binary variables, one for categorical variables, and one for continuous variables. The binary value library included random forest (RF), logistic regression (LR), Naive Bayes (NB), k-nearest neighbors (KNN), least absolute shrinkage and selection operator (LASSO), neural networks (NNET) and support vector machines (SVM). The categorical variables library included RF, conditional mean (MEAN), multinomial logistic regression (MLR) and multinomial log-linear models via neural networks, also labeled NNET. The algorithms used for continuous attribute values included mean imputation (MEAN), ridge regression (RIDGE), general linear models (GLM), linear regression (LM), SVM, NNET, LASSO, and RF. The number of trees used when implementing RF was 1000, and the number of predictors considered when splitting a node was 5. A value of $k = 10$ was used for KNN, and the shrinkage parameters for LASSO and RIDGE were selected using 10-fold cross-validation. Accuracy of imputation was measured as the absolute residual for continuous variables (age, w_{ij}^{CD} and w_{ij}^{SMS}), and precision, recall, F-measure and % hits were used for discrete variables (sex, age category). The accuracy of Super Learner, along with all of the other learning algorithms, was quantified

Variable Name	Description
k	Degree of node i
sCD	The call duration strength of node i
$sSMS$	The SMS strength of node i
cc	The clustering coefficient of node i
$avgk$	The average degree of node i's neighbors
$avgCD$	The average call duration strength of node i's friends
$avgSMS$	The average SMS strength of node i's friends
$pMale$	The proportion of node i's friends that are male
$pFemale$	The proportion of node i's friends that are female
$avgAge$	The average age of node i's friends
$sdAge$	The standard deviation of node i's friends' ages
$medAge$	The median age of node i's friends
$entCD$	The entropy of node i's call duration strength, $\sum_i p_i^{CD} f(p_i^{CD})$
$entSMS$	The entropy of node i's SMS strength, $\sum_i p_i^{SMS} f(p_i^{SMS})$

Table 3.1: Variables used in the imputation of nodal attributes.

Variable Name	Description
$kSum$	Sum of the degrees of i and j $(k_i + k_j)$
$kDiff$	Absolute difference in the degrees of i and j $(\lvert k_i - k_j \rvert)$
$ccSum$	Sum of the clustering coefficients of i and j
$ccDiff$	Absolute difference in the clustering coefficients of i and j
$AgeSum$	Sum of the ages of i and j
$AgeDiff$	Absolute difference in the ages of i and j
MM	Indicator that i and j are both male
MF	Indicator that i and j are opposite sexes
FF	Indicator that i and j are both female
$overlap$	Unweighted overlap of edge between i and j
$woverlap$	Weighted overlap of edge between i and j
n_{ij}	Number of common friends of i and j
e_{ij}	Number of edges among the common friends of i and j
*CD**	The total call duration between i and j. *Used only for the imputation of SMS
*SMS**	The total number of text message communication between i and j. * Used only for the imputation of *CD*.

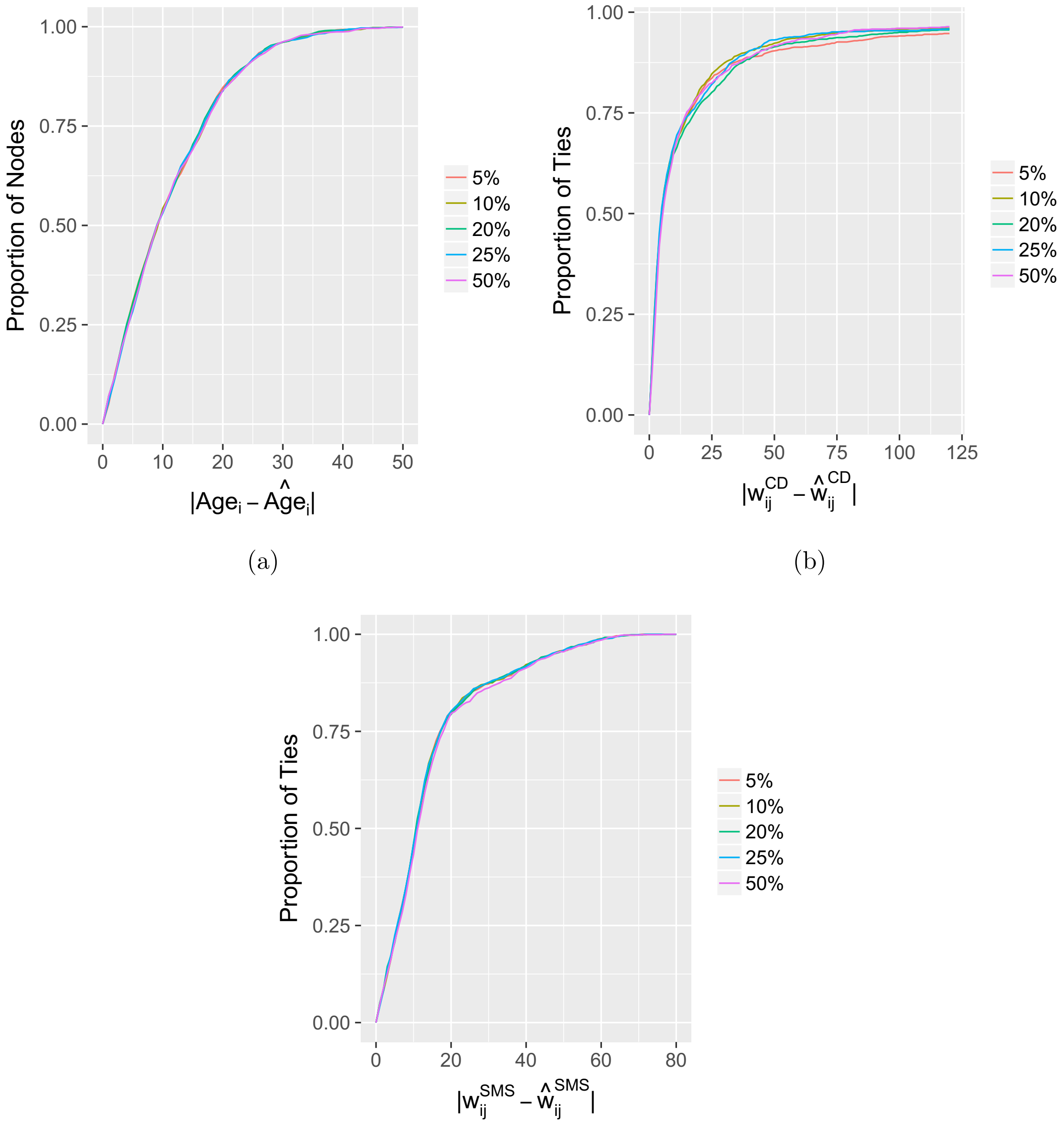
Table 3.2: Variables used in the imputation of edge attributes.

using the cross-validated risk estimate.

3.4 Results

The call network contained 2,276,495 nodes and 12,345,848 edges. Age was available for 89.25% of the individuals and had a mean of 48.2 (sd = 18.2) years. Of the 89.03% of individuals whose sex was recorded, 52.51% were male. Due to the large size of the network, three random samples were drawn; two for edge attribute imputation and one for nodal attribute imputation. The first sample was for call duration w_{ij}^{CD} imputation and consisted of 1,000 randomly drawn edges with call duration greater than 0 and full attribute information available for all nodes. The average call duration was 154 (sd = 290.63) minutes. Due to the skewness of the distribution, w_{ij}^{CD} was log-transformed before imputation. Similarly, the second sample was for SMS imputation, and consisted of 1,000 randomly drawn edges with w_{ij}^{SMS} greater than 0 and full attribute information available for all nodes. The average number of SMS messages was 36.62 (sd = 20.39). This distribution was also skewed, and w_{ij}^{SMS} was also log-transformed before imputation. The third sample was used for both age and sex imputation and contained 1,000 randomly sampled nodes such that the nodes sampled, as well as all of their neighbors, had full attribute information, i.e., age and sex, available. Age was normally distributed with a mean of 47.2 (sd = 20.1) years. The average clustering coefficient was 0.23, 47.2% of individuals were male, and the average degree was equal to 10.89. Note that obtaining samples without any missing data ensured knowledge of the 'ground truth' needed for quantifying the accuracy of our method.

Imputation accuracy results for age, total call duration w_{ij}^{CD} and total text message communication w_{ij}^{SMS} are shown in Figure3.1. In each case, the proportion of missing data has no significant impact on performance. Age within 17 years is predicted with an accuracy of 75%. Total call duration within 13 minutes and total SMS communication within 18 messages can be predicted with 75% accuracy. Table 3 shows the F-measure, precision, and recall results for sex and two different categorizations of age. The first categorization follows that of Brea et al. (2014), and divides age into four categories: < 25, $25 - 34$, $35 - 50$,



(a)

(b)

(c)

Figure 3.1: Imputation accuracy plots for a) age, b) age category, c) sex, d) total call duration (w_{ij}^{CD}) and e) total text message communication (w_{ij}^{SMS}). Accuracy is measured as the absolute residual for continuous variables (age (years), w_{ij}^{CD} (minutes per 90 days), and w_{ij}^{SMS} (text messages per 90 days)). F-measure, recall and precision are presented as accuracy measures for discrete variables (sex (M/F) and age category (< 25, $25-34$, $35-50$, 50+).

Attribute	F-measure	Precision	Recall
Sex	0.63	0.67	0.64
AgeCat$_1$	0.34	0.21	0.86
AgeCat$_2$	0.34	0.21	0.87

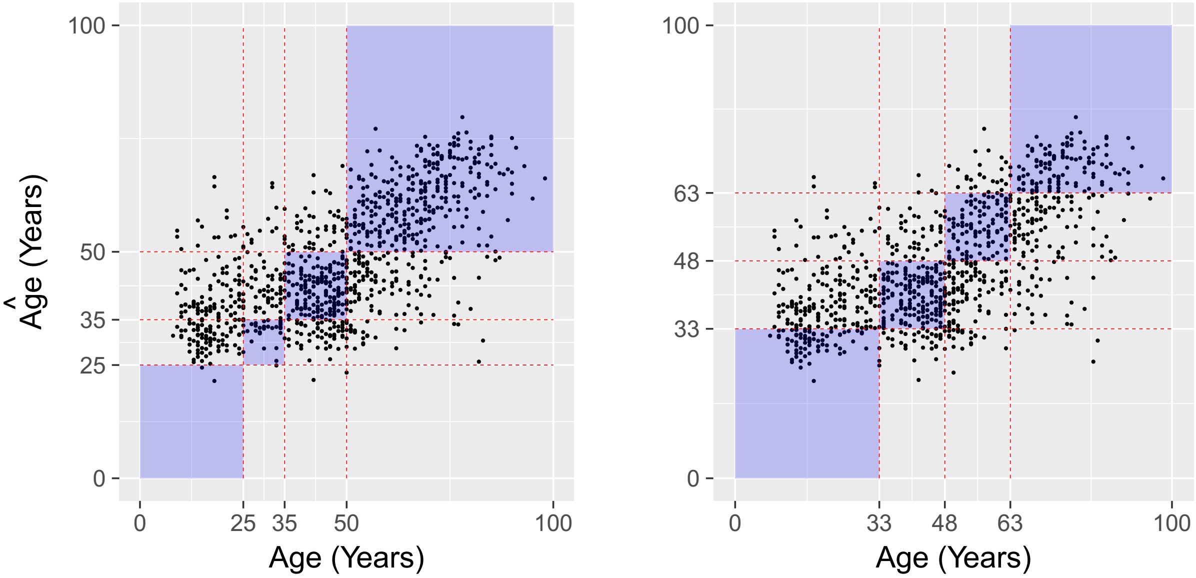
Table 3.3: Accuracy in terms of F-measure, precision and recall for sex (M/F), and both categorizations of age. AgeCat$_1$ refers to the < 25, $25-34$, $35-50$, $50+$ categorization, and AgeCat$_2$ refers to the < 33, $33-47$, $48-63$, $63+$ categorization.

		Learning Algorithm					
Category	% Population	SL$_1$	SL$_2$	RF	NNET	MLR	MEAN
< 25	17.7%	41%	1.1%	41.2 %	39.5%	30%	0%
$25-34$	9.7%.	4%	29.9%	6.2%	2.1%	0%	0%
$35-50$	29.3%	60%	62.8%	60%	45.4%	51.2%	0%
50+	43.3 %	81.5%	70.7%	81%	82.9%	83.8%	100%
Overall		60.6%	52.1%	60.6%	56.4%	56.6%	43.3%
< 33	24.9%	59%	30%	57%	55.8 %	55.4 %	2.8%
$33-47$	25%	44%	66%	47.2%	36.8%	38%	15.6%
$48-63$	25.4%	41%.	46.5%	36.2%	31.9%	32.7%	68.5%
63+	24.7 %	69%	46.5%	70.4%	74.9%	74.5%	0%
Overall		52.9%	49%	52.6%	49.7%	50%	22%

Table 3.4: Accuracy of age category imputation overall, and stratified by category for both categorizations. The category thresholds of the first group match those inBrea et al. (2014). The thresholds of the second group ensure an approximately equal proportion of the population in each category. SL$_1$ denotes the Super Learner imputation of age category directly, and SL$_2$ denotes the categorization of the Super Learner imputation of continuous age. The individual learning algorithms include random forest (RF), multinomial log-linear models via neural networks (NNET), multinomial logistic regression (MLR) and conditional mean imputation (MEAN).

and 50+. We denote this categorization AgeCat$_1$. However, our call network consists of a majority of older individuals, and these category thresholds create a heavy imbalance in the proportion of individuals in each category. Here, 17.7% of individuals are < 25 years old, 9.7% are between 25 and 34, 29.3% are between 35 and 50, and 43.3% are over 50 years old. Therefore, we categorized age again using thresholds that ensured an approximately equal number of individuals in each category. The second set of categories are < 33, $33-47$, $48-63$ and 63+, which we denote AgeCat$_2$. The overall accuracy (% hits) and accuracy stratified by age category for Super Learner and the individual learning algorithms are shown in Table 4. The proportion of missing data again had no significant effect on the results and the results with $p = 10\%$ are shown. In addition to the accuracy of direct age category imputation using Super Learner, we also calculated the accuracy of first imputing a continuous value of age using Super Learner, and subsequently categorizing those values using the two categorizations mentioned above. We denote the direct imputation of age category using Super Learner as SL_1, and the indirect imputation of age category via categorization of a continuous imputation of age using Super Learner as SL_2. With the first categorization, the 50+ age category is accurately classified 81.5% of the time; a 56.5% increase in accuracy if age category were assigned randomly, and a 38.2% increase if age were set to the most probable age category (50+). For the 35-50 year old age category, a gain of 35% was achieved compared to randomly assigning age category. There was a gain of 16% over random assignment for the < 25 age category, but a significant loss in accuracy (21%) for the $25 - 34$ age group. However, this is likely partially due to the high imbalance in the proportion of individuals in each category since there is a significant gain in accuracy for the $25 - 34$ group when the proportions of individuals in each category are more evenly distributed. There is an interesting tradeoff in overall accuracy versus age-stratified accuracy using the two different thresholds for categorization. A more balanced thresholding (AgeCat_2) guarantees age-stratified accuracy will be higher for every category than if age category were randomly assigned. However, the overall accuracy for a more balanced categorization is lower than for the unbalanced (AgeCat_1). Furthermore, the accuracy of the most populous category in an unbalanced categorization is much higher than any one of the categories of a balanced categorization. Additionally, in this data set, categorizing an imputed value of continuous age (method SL_2) does not increase overall accuracy, but does illustrate the sensitivity of age-stratified accuracy. This sensitivity and variation in accuracy is visualized in Figure3.2. The blue boxes indicate where the categorization of a continuous value of age would be accurate. If all of the black dots fell into the shaded regions, the algorithm would achieve 100% accuracy.

The algorithm weights chosen by Super Learner for each attribute are shown in Table3.5.



(a)

(b)

Figure 3.2: Visual representation of the relationship between thresholds of categorizing imputations of continuous age and accuracy. In each plot, the x-axis represents an individual's actual age, and the y-axis their predicted age using Super Learner. The red dashed lines denote the age category thresholds for each categorization. Plot (a) is the < 25, $25 - 34$, $35 - 50$, and 50+ categorization and plot (b) is the < 33, $33 - 47$, $48 - 63$, and 63+ categorization. The blue boxes shade the accuracy area, i.e., if the actual age category matches the predicted age category. If all dots were contained in the blue boxes, age category imputation would be 100% accurate.

Random forest was given the majority of weight for the imputation of each attribute, reaching a maximum of 0.86 for continuous age. Logistic regression was given a significant weight of 0.37 for the imputation of sex. Neural networks was also given significant weight for the equal thresholding of age category (AgeCat$_2$) and call duration w_{ij}^{CD}. Note that the RF weights were similar for both categorizations of age, but the second algorithm given positive weight differed; MLR for AgeCat$_1$ and NNET for AgeCat$_2$. The graphical results of the accuracy of Super Learner for $p = 5\%$ missingness are shown in Figure 3. As expected, Super Learner performs as well or better than the most accurate algorithm van der Laan (2007)Polley and van der Laan (2010). Random forest was the most accurate algorithm for imputing age, neural networks was the most accurate for total call duration w_{ij}^{CD}, and Super Learner was the most accurate for total SMS communication w_{ij}^{SMS}.

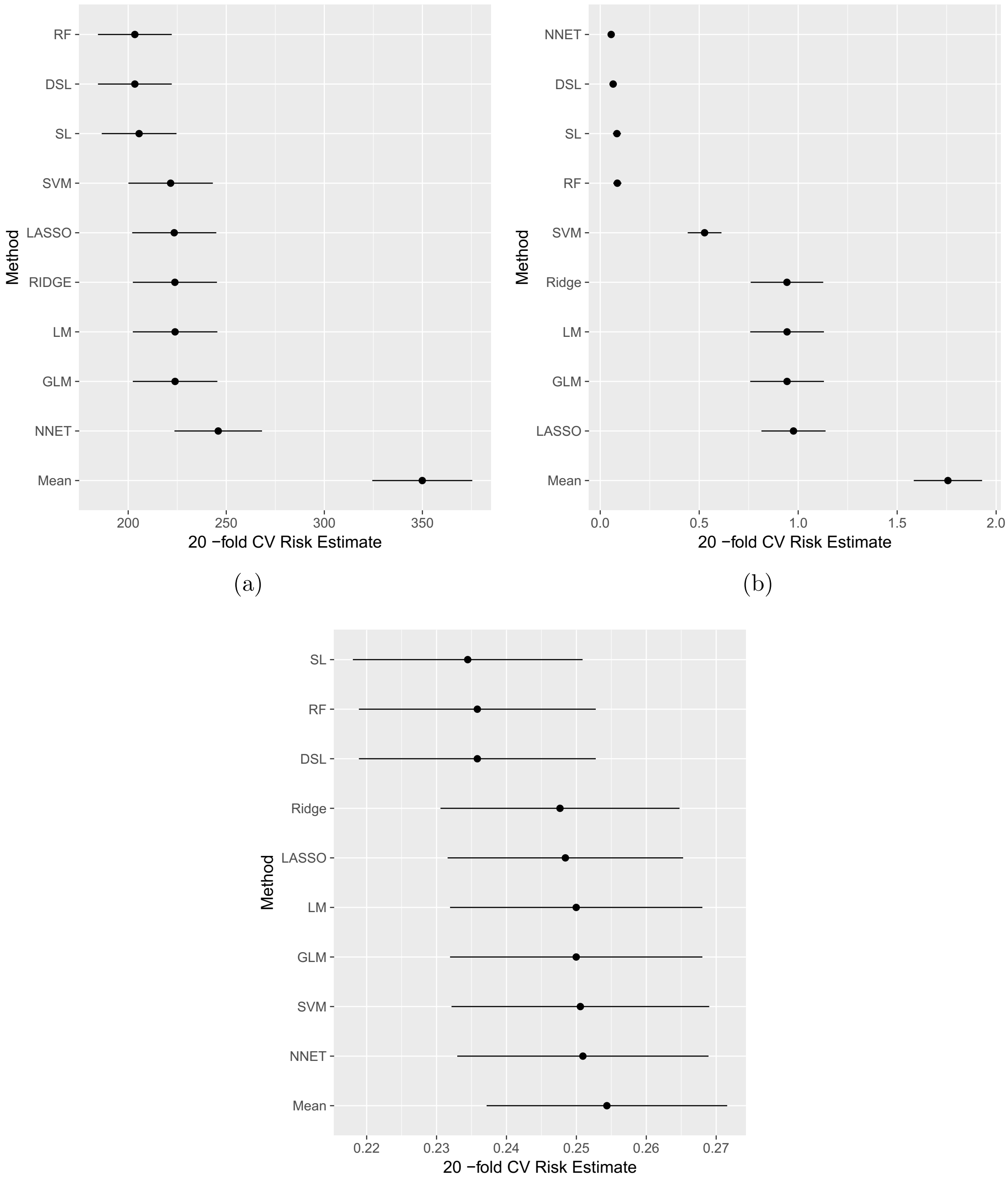
	Learning Algorithm											
Attr.	RF	NNET	LASSO	SVM	KNN	NB	LR	LM	RIDGE	MEAN	GLM	MLR
Age	0.86	0.01	0	0.05	-	-	-	0	0.08	0	0	-
AgeCat$_1$	0.90	0	-	-	-	-	-	-	-	0	-	0.10
AgeCat$_2$	0.78	0.22	-	-	-	-	-	-	-	0	-	0
Sex	0.55	0.03	0.009	0.04	0	0.001	0.37	-	-	-	-	-
w_{ij}^{CD}	0.476	0.412	0.07	0	-	-	-	0.002	0.013	0.027	0	-
w_{ij}^{SMS}	0.66	0.131	0	0	-	-	-	0	0.2	0	0	-

Table 3.5: Average learning algorithm weights for Super Learner. The algorithms include random forest (RF), neural networks (NNET), least absolute shrinkage and selection operator (LASSO), support vector machines (SVM), k-nearest neighbors (KNN), Naive Bayes (NB), logistic regression (LR), linear regression (LM), ridge regression (RIDGE), mean imputation (MEAN), generalized linear models (GLM) and multinomial logistic regression (MLR). AgeCat$_1$ references the < 25, $25 - 34$, $35 - 50$, and 50+ categorization, and AgeCat$_2$ the < 33, $33 - 47$, $48 - 63$, and 63+ categorization.

3.5 Discussion

In this work we perform network attribute imputation through the implementation of the loss-based supervised learning method Super Learner. We define libraries of learning algorithms for binary, categorical, and continuous responses as well as novel attribute predictors based on network topology and homophily and apply our method to a nationwide call network of European mobile phone users. We introduce different amounts of missingness ranging from 5% to 50%, and impute the age, age category and sex of individuals, as well as the total call duration and text message communication over a three-month period between two individuals. We find that Super Learner performs as well as or better than any one learning algorithm alone for the imputation of each attribute. We also find that the overall and age-stratified accuracy of age category imputation is sensitive to category thresholding. Specifically, if there is a roughly equal number of individuals who belong in each category, accuracy stratified by age will be greater than if age category were randomly assigned. However, overall accuracy is increased if there is an imbalance in the proportion of membership in each category. Interestingly, in either case, Super Learner outperforms the reaction-diffusion algorithm presented by Brea et al Brea et al. (2014); the first categorization of age (the imbalanced thresholding) yields an accuracy increase of over 10%, and the second (balanced) categorization an increase of 3%. While this is not a completely accurate comparison due to the differences in data sets and age distributions in our study and the study by Brea et al., our results seem to be promising and it would be worth applying their algorithm to our data set for a more precise comparison.

In future work we would like to test if accuracy is altered by choosing different ensemble methods when implementing Super Learner. Here we chose the most common method, non-negative least squares, but could use non-negative log likelihood instead. Additionally, it would be interesting to see how altering the missing-data mechanism to MAR or NMAR impacts our results. Huisman implemented all three mechanisms on a small social network of teenage girls and found that the mechanisms produce some differences in results Huisman (2007). Specifically, he found that transitivity and assortativity were highly sensitive to the missingness mechanism. We would also like to apply our method to different data sets to see if the results depend on the type of network studied.



(a)

(b)

(c)

Figure 3.3: Accuracy of Super Learner measured as cross-validation (CV) risk for continuous attributes a) age, b) total call duration, and c) total SMS communication. The black dots represent mean CV risk and the horizontal lines span ±2 standard deviations. Super Learner performs better than or as well as all other algorithms.

References

Banerjee, A., Chandrasekhar, A., Duflo, E., and Jackson, M. (2013). The diffusion of microfinance. *Science*, 341.

Bianconi, G., Darst, R., Iacovacci, J., and Fortunato, S. (2014). Triadic closure as a basic generating mechanism of communities in complex networks. *Physics Review*, 90.

Bollobás, B. (1985). *Random Graphs.* Academic Press.

Borgatti, S. and Molina, J. (2003). Ethical and strategic issues in organizational social network analysis. *Journal of Applied Behavioral Science*, 39:337–349.

Bott, E. (1957). *Family and Social Network: Roles, Norms and External Relationships in Ordinary Urban Families.* Abingdon: Routledge.

Brea, J., Burroni, J., Minnoni, M., and Sarraute, C. (2014). Harnassing mobile phone social network topology to infer users demographic attributes. *ACM.*

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Burt, R. (1987). A note on missing network data in the general social survey. *Social Networks*, 9:63–73.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.*, 357:370–379.

Christakis, N. A. and Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *N. Engl. J. Med.*, 358:2249–2258.

Dong, Y., Yang, Y., Tang, J., and Chawla, N. (2014). Inferring user demographics and social strategies in mobile social networks. *ACM*.

Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493.

Elandt-Johnson, R. and Johnson, N. (1998). *Survival Models and Data Analysis*. John Wiley Sons.

Ellison, N. B., Steinfield, C., and Lampe, C. (2007). The benefits of facebook 'friends:' social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168.

Erdős, P. and Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae*, 6:290–297.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5.

Fortunato, S. (2010). Community detection in graphs. *Physics*, 486:75–174.

Fowler, J. H. and Christakis, N. A. (2008a). Dynamic spread of happiness in a large scale network: longitudinal analysis over 20 years in the framingham heart study. *BMJ*, 337.

Fowler, J. H. and Christakis, N. A. (2008b). Estimating peer effects on health in social networks. *J. Health Econ.*, 27:1400–1405.

Garlaschelli, D. (2009). The weighted random graph model. *New Journal of Physics*, 11.

Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. *ACM*, pages 211–220.

Gong, N. e. a. (2014). Jointly predicting links and inferring attributes using social-attribute network (san). *ACM Trans. Intell. Syst. Technol.*, 5:27:1–27:20.

Gong, N. Z. and Liu, B. (2016). You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016.*, pages 979–995.

Goodreau, S., Kitts, J., and Morris, M. (2009). Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *EPL*, 87.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, 78:1360–1380.

Harling, G. and Onnela, J.-P. (2016). Impact of degree truncation on the spread of a contagious process on networks.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer New York Inc., New York, NY.

He, J., Chu, W. W., and Liu, Z. (2006). Inferring privacy information from social networks. *Intelligence and Security Informatics*, pages 154–165.

Hijmans, R. (2009). Global administrative areas: Boundaries without limits. `http://gadm.org/country`. Accessed: 2016-05-12.

Huisman, M. (2007). Imputation of missing network data: Some simple procedures. Sunbelt XXVII International Sunbelt Social Network Conference.

Hwong, A., Onnela, J., Kim, D., Stafford, D., Hughes, D., and Christakis, N. (2016a). Not created equal: Sex differences in the network-based diffusion of public health interventions.

Hwong, A., Staples, P., and Onnela, J. (2016b). Simulating network-based public health interventions in low-resource settings.

Kahanda, I. and Neville, J. Using transactional information to predict link strength in online social networks. In *Proceedings of the Third International ICWSM Conference*, pages 74–81.

Kim, D., Hwong, A., Stafford, D., Hughes, D., O'Malley, A., Fowler, J., and Christakis, N. (2015). Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 386.

Kim, D., O'Malley, A., and Onnela, J.-P. (2016). The social geography of american medicine.

Kumpula, J., Onnela, J.-P., Saramaki, J., Kaski, K., and Kertesz, J. (2007). Emergence of communities in weighted networks. *Physical Review Letters*, 99.

Landon, B., Keating, N., Barnett, M., Onnela, J.-P., Paul, S., O'Malley, A., Keegan, T., and Christakis, N. (2012). Variation in patient-sharing networks of physicians across the united states. *JAMA*, 308:265–273.

Li, N., Feng, X., Ji, S., and Xu, K. (2013). *Modeling Relationship Strength for Link Prediction*, pages 62–74.

Lin, K. (2007). Motif counts, clustering coefficients and vertex degrees in models of random networks.

Lin, N., Vaughn, J., and Ensel, W. (1981). Social resources and occupational status attainment. *Social Forces*, 59(4):1163–1181.

Linyuan, L. and Tao, Z. (2010). Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters)*, 89(1):18001.

Little, R. and Rubin, D. (2002). *Statistical Analysis With Missing Data*, volume 17. Wiley Series in Probability and Statistics.

Luarn, P. and Chiu, Y.-P. (2015). Key variables to predict tie strength on social network sites. *Internet Research*, 25:218–238.

Marsden, P. V. and Campbell, K. E. (1984). Measuring tie strength. *Social Forces*, 63(2):482–501.

Mattie, H., Engø-Monsen, K., Ling, R., and Onnela, J.-P. (2017). The social bow tie. Unpublished manuscript.

Mattie, H. and Onnela, J.-P. (2017). Edge overlap in weighted and directed social networks. Unpublished manuscript.

Mukerjee, P. (2013). Vizualyse. `http://visual.yantrajaal.com/2015/05/using-r-for-maps-of-india-state.html`. Accessed: 2016-05-12.

Newman, M. (2003). Properties of highly clustered networks. *Phys. Rev. E*, 68.

Newman, M. (2010). *Networks: An Introduction.* Oxford University Press.

Onnela, J.-P., Landon, B., Kahn, A., Ahmed, D., Verma, H., O'Malley, A., Bahl, S., Sutter, R., and Christakis, N. (2016). Polio vaccine hesitancy in the networks and neighborhoods of malegaon, india. *Social Science and Medicine*.

Onnela, J.-P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., Kertesz, J., and Barabasi, A.-L. (2007). Structure and tie strengths in mobile communication networks. *PNAS*, 104:7332–7336.

Papadatos, N. (1995). Maximum variance of order statistics. *Ann. Inst. Statist. Math*, 47:185–193.

Pappalardo, L., Rossetti, G., and Pedreschi, D. (2012). "how well do we know each other?" detecting tie strength in multidimensional social networks.

Polley, E. and van der Laan, M. J. (2010). Super Learner In Prediction.

Porter, M., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, 56:1082 – 1166.

Reinert, G. (2012). Probability and statistics for network analysis. University Lecture.

Rubin, D. Inference and missing data. *Biometrika*, 63:581–592.

Sá, H. R. d. and Prudêncio, R. B. C. Supervised link prediction in weighted networks. In *The 2011 International Joint Conference on Neural Networks*, pages 2281–2288.

Saramaki, J., M. Kivela, J.-P. O., Kaski, K., and Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E.*

Sima, C., Panageas, K., Heller, G., and Schrag, D. (2010). Analytical strategies for characterizing chemotherapy diffusion with patient-level population-based data. *Appl Health Econ Health Policy*, 8:37–51.

Skellam, J. (1946). The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society*, 109.

Staples, P., Ogburn, E., and Onnela, J.-P. (2015). Incorporating contact network structure in cluster randomized trials. *Scientific Reports*, 5.

Stork, D. and Richards, W. (1992). Nonrespondents on communication network studies. *Groups Organization Management.*

Stuart, A. and Ord, K. (1998). *Kendall's Advanced Theory of Statistics: v.1.* Wiley-Blackwell.

Tore, O. (2013). Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35:159–167.

Valente, T. (2005). Network models and methods for studying the diffusion of innovations. *Models and Methods in Social Network Analysis*, pages 98–116.

van der Laan, M. J. (2007). Super learner. *Statistical Application in Genetics and Molecular Biology.*

VanderWeele, T. (2011). Sensitivity analysis for contagion effects in social networks. *Sociological Methods and Research*, 40:240–255.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications.* Cambridge University Press.

Watts, D. and Strogatz, S. (1998). Collective dynamics od 'small-world' networks. *Nature*, 393:440–442.

Wellman, B. and Wortley, S. (1990). Different strokes from different folks: Community ties and social support. *American Journal of Sociology*, 96(3):558–588.

Wiese, J., Min, J.-K., Hong, J., and Zimmerman, J. Assessing call and sms logs as an indication of tie strength.

Xiang, R., Neville, J., and Rogati, M. (2010). Predicting tie strength in a new medium.

Zamal, F. A., Liu, W., and Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. *AAAI Publications.*