Generalizations of Edge Overlap to Weighted and Directed Networks

Heather Mattie* and Jukka-Pekka Onnela

Biostatistics Department, Harvard T.H. Chan School of Public Health, 651 Huntington Ave., Boston, MA 02115

E-mail: hemattie@hsph.harvard.edu

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              | $k_i^{\text{in}}$ | In a directed network, the number of incoming edges to a node $i$          |
| out-degree             | $k_i^{\text{out}}$ | 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 random graph | $G(n, p)$ | A random graph of $n$ nodes and edges generated by connecting a pair of nodes with some 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. |
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. 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. 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. Studies have also leveraged network properties to target highly connected individuals in public health interventions. The structure of connections in contact networks have also been shown to affect statistical power in cluster randomized trials. 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.

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. 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.

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. Specifically,
several microscopic mechanisms are known to give rise to microscopic, mesoscopic, and macroscopic structure. For example, triadic closure, the process of getting to know a friend of a friend, can generate network communities. 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.

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. 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”.

The clustering coefficient has subsequently been extend to weighted and directed networks. 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. Most social networks are more clustered than corresponding random networks. 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 paper is organized as follows. In Section 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 3 and discuss our conclusions in Section 4. Supplementary material is contained in Appendices A, B, C and D.

2. Methods

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

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

(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. 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. (2) as the proportion of total weight associated with ties to common friends nodes $i$ and $j$ share, and denote it $o_{ij}^W$:

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

(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. (3) as the proportion of paths of length two between two connected individuals, and denote it $o^{D}_{ij}$:

$$
o^{D}_{ij} = \frac{\sum_{k=1}^{n}(A_{ik}A_{kj} + A_{jk}A_{ki})}{\min(k_{i}^{in}, k_{j}^{out}) + \min(k_{j}^{in}, k_{i}^{out}) - 1}.
$$

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 $\min(k_{i}^{in}, k_{j}^{out}) + \min(k_{j}^{in}, k_{i}^{out}) > 1$.

### 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.\textsuperscript{28} 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.\textsuperscript{29} 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 $p$.\textsuperscript{28, 30, 31} 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.
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. 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. (4).

$$O_{ij} = \frac{N_{ij}}{(K_i - 1) + (K_j - 1) - N_{ij}}$$ (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. This results in Eqs. (5) and (6) (for details, see Appendix A.1.).

\[
\mathbb{E}[O_{ij}] = \frac{p}{2-p} \tag{5}
\]

\[
\text{Var}(O_{ij}) = \frac{np^2}{(2np - 2 - np^2)^2} + \frac{n^2p^4(2np - 2 + np^2)}{(2np - 2 - np^2)^4}. \tag{6}
\]

Our second approach incorporates results from 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. (7) and (8) (for details, see Appendix B.1.):

\[
\mathbb{E}[O_{ij}] = \frac{p}{2-p} \tag{7}
\]

\[
\text{Var}(O_{ij}) = \frac{np^2}{(2np - 2 - np^2)^2}. \tag{8}
\]
Note that the expressions for the mean, Eqs. (5) and (7), are equivalent, but the expressions for the variance, Eqs. (6) and (8) differ, with the expression for Eq. (8) corresponding to the first term of Eq. (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 $W_{ij} \sim \text{geometric}(1-p)$. 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 $S_i \sim \text{negative binomial}(n-1, 1-p)$.

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. (9):

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

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. (10) and (11) (for details, see Appendix A.2.):
\[
\mathbb{E}[O_{ij}^W] = p
\]  \hspace{1cm} (10)

\[
\text{Var}(O_{ij}^W) = \frac{p + 1}{n}.
\]  \hspace{1cm} (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. (12) and (13). Again, the expressions for the mean are equivalent for both approaches, and the variance expressions are quite similar (for details, see Appendix B.2.):

\[
\mathbb{E}[O_{ij}^W] = p
\]  \hspace{1cm} (12)

\[
\text{Var}(O_{ij}^W) = \frac{np^2(p + 2)}{2(np - 1)^2}.
\]  \hspace{1cm} (13)

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 \(K_{in}^i\) and \(K_{out}^i\). Due to the definition given above, 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. (14).
\[ O_{ij}^D = \frac{\sum_{k=1}^{n}(A_{ik}A_{kj} + A_{jk}A_{ki})}{\min(K_i^{in}, K_j^{out}) + \min(K_j^{in}, K_i^{out}) - 1}. \] (14)

The mean and variance of the denominator can now be calculated and used to derive the expressions in Eqs. (15) and (16) (for details, see Appendix A.3.):

\[ E[O_{ij}^D] = \frac{np^2}{e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} (np)^j \right]^2 - 0.5} \] (15)

\[ \text{Var}(O_{ij}^D) = \frac{2n^2 p^4}{(2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} (np)^j \right]^2 - 1)^2} + \frac{32n^3 p^5 e^{np}}{e^{np-1} \left[ 1 - \frac{np}{e^{np-1}} \right]} \] (16)

The second approach again replaces the denominator with its expectation, and the mean and variance derivations result in the expressions in Eqs. (17) and (18) (for details, see Appendix B.3.). 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. (18) is equivalent to the first term of the variance resulting from the first approach in Eq. (16).

\[ E[O_{ij}^D] = \frac{np^2}{e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} (np)^j \right]^2 - 0.5} \] (17)

\[ \text{Var}(O_{ij}^D) = \frac{2n^2 p^4}{(2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} (np)^j \right]^2 - 1)^2} \] (18)

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 2 displays the simulation results and accuracy of our approximations. The top row contains the results for the mean unweighted overlap (Figure 2a), mean weighted overlap (Figure 2b) and mean directed overlap (Figure 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 2d), weighted overlap (Figure 2e) and directed overlap (Figure 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 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 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 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.

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 3). 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
Figure 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 completely 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.

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
Figure 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.

were linguistically homogeneous but had variability in caste. Each village contained 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%).

Table 2: The types of social interactions recorded for individuals in each village.

| 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      |

We first calculated the average unweighted overlap for each type of social relationship (labeled 1-12, see Table 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 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 8 in Appendix C). 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 4 and 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 Appendix C.

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 6 shows the distributions of raw and standardized weighted overlap for F/F, M/M and M/F ties. See Appendix C for figures stratified by attribute availability, caste and age.

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 Appendix D. In Figure 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. 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. However, this trend could also be due to the significant difference in the average degree for males and females across the villages (Figure 21 in Appendix C). 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
Figure 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.

Figure 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.
Figure 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.

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

Figure 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 6b). Again, all values are significantly higher than expected for each type of tie, as we would expect from Figure 5 above.
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 approximations 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 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.

Acknowledgments

We thank Banerjee et. al. for making the India data set publicly available.

References

(1) Christakis, N. A.; Fowler, J. H. *N. Engl. J. Med.* 2007, 357, 370–379.

(2) Christakis, N. A.; Fowler, J. H. *N. Engl. J. Med.* 2008, 358, 2249–2258.

(3) Fowler, J. H.; Christakis, N. A. *BMJ* 2008, 337.

(4) Fowler, J. H.; Christakis, N. A. *J. Health Econ.* 2008, 27, 1400–1405.

(5) Goodreau, S.; Kitts, J.; Morris, M. *EPL* 2009, 87.

(6) Onnela, J.-P.; Landon, B.; Kahn, A.; Ahmed, D.; Verma, H.; O’Malley, A.; Bahl, S.; Sutter, R.; Christakis, N. *Social Science and Medicine* 2016.

(7) Banerjee, A.; Chandrasekhar, A.; Duflo, E.; Jackson, M. *Science* 2013, 341.

(8) Valente, T. *Models and Methods in Social Network Analysis* 2005, 98–116.

(9) VanderWeele, T. *Sociological Methods and Research* 2011, 40, 240–255.

(10) Kim, D.; Hwong, A.; Stafford, D.; Hughes, D.; O’Malley, A.; Fowler, J.; Christakis, N. *The Lancet* 2015, 386.

(11) Staples, P.; Ogburn, E.; Onnela, J.-P. *Scientific Reports* 2015, 5.
(12) Kim, D.; O’Malley, A.; Onnela, J.-P. The Social Geography of American Medicine. 2016.

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

(14) Landon, B.; Keating, N.; Barnett, M.; Onnela, J.-P.; Paul, S.; O’Malley, A.; Keegan, T.; Christakis, N. JAMA 2012, 308, 265–273.

(15) Sima, C.; Panageas, K.; Heller, G.; Schrag, D. Appl Health Econ Health Policy 2010, 8, 37–51.

(16) Fortunato, S. Physics 2010, 486, 75–174.

(17) Bianconi, G.; Darst, R.; Iacovacci, J.; Fortunato, S. Physics Review 2014, 90.

(18) Kumpula, J.; Onnela, J.-P.; Saramaki, J.; Kaski, K.; Kertesz, J. Physical Review Letters 2007, 99.

(19) Porter, M.; Onnela, J.-P.; Mucha, P. J. Notices of the AMS 2009, 56, 1082 – 1166.

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

(21) Watts, D.; Strogatz, S. Nature 1998, 393, 440–442.

(22) Saramaki, J.; M. Kivelä, J.-P. O.; Kaski, K.; Kertesz, J. Physical Review E 2007,

(23) Tore, O. Social Networks 2013, 35, 159–167.

(24) Reinert, G. Probability and Statistics for Network Analysis. University Lecture, 2012.

(25) Newman, M. Phys. Rev. E 2003, 68.

(26) Newman, M. Networks: An Introduction; Oxford University Press, 2010.
(27) Onnela, J.-P.; Saramaki, J.; Hyvonen, J.; Szabo, G.; Lazer, D.; Kaski, K.; Kertesz, J.; Barabasi, A.-L. *PNAS* 2007, 104, 7332–7336.

(28) Erdős, P.; Rényi, A. *Publicationes Mathematicae* 1959, 6, 290–297.

(29) Garlaschelli, D. *New Journal of Physics* 2009, 11.

(30) Erdős, P.; Rényi, A. *Publ. Math. Inst. Hung. Acad. Sci.* 1960, 5.

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

(32) Skellam, J. *Journal of the Royal Statistical Society* 1946, 109.

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

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

(35) Lin, K. Motif Counts, Clustering Coefficients and Vertex Degrees in Models of Random Networks. PhD dissertation, Oxford University, 2016.

(36) Hijmans, R. Global Administrative Areas: Boundaries Without Limits. 2009; Accessed: 2016-05-12.

(37) Mukerjee, P. Vizualyse. 2013; Accessed: 2016-05-12.

(38) Hwong, A.; Onnela, J.; Kim, D.; Stafford, D.; Hughes, D.; Christakis, N. Not Created Equal: Sex Differences in the Network-Based Diffusion of Public Health Interventions. Ph.D. thesis, Harvard T.H. Chan School of Public Health, 2016.

(39) Hwong, A.; Staples, P.; Onnela, J. Simulating Network-Based Public Health Interventions in Low-Resource Settings. Ph.D. thesis, Harvard T.H. Chan School of Public Health, 2016.
A. Approach 1 Mean and Variance Derivations

A.1. Original Erdős-Rényi Overlap

Edge overlap is considered a random variable with mean and variance (See Eq. (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.

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

(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.\(^{32}\) Table 3 summarizes these distributions.

Table 3: The distribution, mean and variance for each random variable included in Erdős-Rényi overlap.

| 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\((2np - 2 - np^2)\) | \( 2np - 2 - np^2 \) | \( 2np - 2 + np^2 \) |

Edge overlap is the ratio of two random variables and its mean and variance can be
approximated using a Taylor series expansion. 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

$$g(x) = g(\theta) + \sum_{i=1}^{k} g_i'(\theta)(x_i - \theta_i) + O(n^{-1})$$ \hspace{1cm} (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 (21) shows the Taylor series expansion for $g(N_{ij}, H_{ij})$ about the mean, $\theta = (\mathbb{E}(N_{ij}), \mathbb{E}(H_{ij}))$.

$$g(N_{ij}, H_{ij}) = g(\theta) + \sum_{i=1}^{2} g_i'(\theta)(x_i - \theta_i) + O(n^{-1})$$ \hspace{1cm} (21)

$$= g(\theta) + g_{N_{ij}}'(\theta)(N_{ij} - \mathbb{E}(N_{ij})) + g_{H_{ij}}'(\theta)(H_{ij} - \mathbb{E}(H_{ij})) + O(n^{-1})$$

$$= g(\theta) + g_{N_{ij}}'(\theta)(N_{ij} - \mathbb{E}(N_{ij})) + g_{H_{ij}}'(\theta)(H_{ij} - \mathbb{E}(H_{ij})) + O(n^{-1})$$

Using the above approximation, the expectation of the ratio, $\mathbb{E}[g(N_{ij}, H_{ij})]$, can be derived as in Eq. (22).
\[ E[g(N_{ij}, H_{ij})] = E[g(\theta)] + g_{N_{ij}}'(\theta)(N_{ij} - E(N_{ij})) \]  
\[ + g_{H_{ij}}'(\theta)(H_{ij} - E(H_{ij})) + O(n^{-1}) \]  
\[ = E[g(\theta)] + E[g_{N_{ij}}'(\theta)(N_{ij} - E(N_{ij}))] + E[g_{H_{ij}}'(\theta)(H_{ij} - E(H_{ij}))] \]  
\[ = E[g(\theta)] + g_{N_{ij}}'(\theta)E[N_{ij} - E(N_{ij})] + g_{H_{ij}}'(\theta)E[H_{ij} - E(H_{ij})] \]  
\[ = E[g(\theta)] + 0 + 0 \approx g(E(N_{ij}), E(H_{ij})) = \frac{E(N_{ij})}{E(H_{ij})} \]  
\[ = \frac{np^2}{2np - 2 - np^2} \approx \frac{p}{2 - p} \]

Using the definition of variance and the result that \( E[g(N_{ij}, H_{ij})] \approx g(\theta) \) from Eq. (22), the variance of \( g(N_{ij}, H_{ij}) \) can be first approximated by Eq. (23).

\[ \text{Var}(g(N_{ij}, H_{ij})) = E \left\{ [g(N_{ij}, H_{ij}) - E(g(N_{ij}, H_{ij}))]^2 \right\} \]  
\[ \approx E \left\{ [g(N_{ij}, H_{ij}) - g(\theta)]^2 \right\} \]
Using the first order Taylor expansion for \( g(N_{ij}, H_{ij}) \) from Eq. \([21]\), we have

\[
\text{Var}(g(N_{ij}, H_{ij})) \approx \mathbb{E}\{[g(\theta) + g_N'(\theta)(N_{ij} - \mathbb{E}(N_{ij}))(N_{ij} - \mathbb{E}(N_{ij}))\}^2 + g_H'(\theta)(H_{ij} - \mathbb{E}(H_{ij})) \}^2
\]

In this case, \( g(N_{ij}, H_{ij}) = N_{ij}H_{ij} \), \( g_N'(\theta) = \frac{1}{H_{ij}} \), \( g_H'(\theta) = -\frac{N_{ij}}{H_{ij}} \), and \( \theta = (\mathbb{E}(N_{ij}), \mathbb{E}(H_{ij})) \).

Placing these expressions into \([24]\) we have that

\[
\text{Var}(g(N_{ij}, H_{ij})) \approx \frac{\text{Var}(N_{ij})}{\mathbb{E}^2(H_{ij})} + \frac{\mathbb{E}^2(N_{ij})\text{Var}(H_{ij})}{\mathbb{E}^3(H_{ij})} - 2\frac{\text{Cov}(N_{ij}, H_{ij})\mathbb{E}(N_{ij})}{\mathbb{E}^3(H_{ij})}
\]

\[
= \frac{\mathbb{E}^2(N_{ij})}{\mathbb{E}^2(H_{ij})} \left[ \frac{\text{Var}(N_{ij})}{\mathbb{E}^2(N_{ij})} + \frac{\text{Var}(H_{ij})}{\mathbb{E}^2(H_{ij})} - 2\frac{\text{Cov}(N_{ij}, H_{ij})}{\mathbb{E}(N_{ij})\mathbb{E}(H_{ij})} \right]
\]

\[
= \frac{np^2}{(2np - 2 - np^2)^2} + \frac{n^2p^4(2np - 2 + np^2)}{(2np - 2 - np^2)^4} - 2\frac{np^2\text{Cov}(N_{ij}, H_{ij})}{(2np - 2 - np^2)^3}
\]

Note that \( \text{Cov}(N_{ij}, H_{ij}) > 0 \) since \( N_{ij} \perp H_{ij} \). 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 (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

\[
g(N_{ij}, H_{ij}) = g(\theta) + g_{N_{ij}}'(\theta)(N_{ij} - \theta N_{ij}) + g_{H_{ij}}'(\theta)(H_{ij} - \theta H_{ij}) + \frac{1}{2}g_{N_{ij}N_{ij}}''(\theta)(N_{ij} - \theta N_{ij})^2 + \frac{1}{2}g_{H_{ij}H_{ij}}''(\theta)(H_{ij} - \theta H_{ij})^2 + O(n^{-1})
\]

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

\[
\mathbb{E}[g(H_{ij}, N_{ij})] = \mathbb{E}[g(\theta) + g_{N_{ij}}'(\theta)(N_{ij} - E(N_{ij})) + g_{H_{ij}}'(\theta)(H_{ij} - E(H_{ij})) + \frac{1}{2}g_{N_{ij}N_{ij}}''(\theta)(N_{ij} - E(N_{ij}))^2 + \frac{1}{2}g_{H_{ij}H_{ij}}''(\theta)(H_{ij} - E(H_{ij}))^2 + g_{N_{ij}H_{ij}}''(\theta)(N_{ij} - E(N_{ij}))(H_{ij} - E(H_{ij})) + O(n^{-1})]
\]

For \(g(N_{ij}, H_{ij}) = \frac{N_{ij}}{H_{ij}}, g''_{N_{ij}N_{ij}} = 0, g''_{N_{ij}H_{ij}} = \frac{1}{H_{ij}}, g''_{H_{ij}H_{ij}} = \frac{2N_{ij}}{H_{ij}^2}\). Plugging these expressions into (27) results in Eq. (28).
\[ \mathbb{E}[g(N_{ij}, H_{ij})] = \frac{\mathbb{E}(N_{ij})}{\mathbb{E}(H_{ij})} + \frac{\text{Var}(H_{ij})\mathbb{E}(N_{ij})}{\mathbb{E}^3(H_{ij})} - \frac{\text{Cov}(N_{ij}, H_{ij})}{\mathbb{E}^2(H_{ij})} \]  

(28)

\[ = \frac{np^2}{2np - 2 - np^2} + \frac{(2np - 2 + np^2)(np^2)}{(2np - 2 - np^2)^3} - \frac{\text{Cov}(N_{ij}, H_{ij})}{(2np - 2 - np^2)^2} \]

\[ = \frac{p}{2 - p} + \frac{(2np - 2 + np^2)(np^2)}{(2np - 2 - np^2)^3} - \frac{\text{Cov}(N_{ij}, H_{ij})}{(2np - 2 - np^2)^2}. \]

Again, \( \text{Cov}(N_{ij}, H_{ij}) > 0 \) 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 (28) in the approximation of the mean.

### A.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. (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.

\[ O^W_{ij} = \frac{\sum_{k=1}^{N_{ij}} (w_{ik} + w_{jk})}{s_i + s_j - 2w_{ij}} \Rightarrow O^W_{ij} = \frac{\sum_{k=1}^{N_{ij}} (W_{ik} + W_{jk})}{S_i + S_j - 2W_{ij}} = \frac{V_{ij}}{M_{ij}} \]  

(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 $W_{ij} \sim \text{geometric}(1 - p)$. 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 $S_i \sim \text{negative binomial}(n - 1, 1 - p)$.

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. (30)) and variance (Eq. (31)) of $V_{ij}$.

$$ E[V_{ij}] = E[E[V_{ij} | N_{ij}]] = E\left[\frac{2N_{ij}}{(1 - p)}\right] \tag{30} $$

$$ = \frac{2}{(1 - p)} E[N_{ij}] = \frac{2np^2}{(1 - p)} $$

$$ \text{Var}(V_{ij}) = E[\text{Var}(V_{ij} | N_{ij})] + \text{Var}(E[V_{ij} | N_{ij}]) \tag{31} $$

$$ = E\left[\frac{2pN_{ij}}{(1 - p)^2}\right] + \text{Var}\left(\frac{2N_{ij}}{(1 - p)}\right) $$

$$ = \left[\frac{2p}{(1 - p)^2}\right] E[N_{ij}] + \left[\frac{2}{(1 - p)}\right]^2 \text{Var}(N_{ij}) $$

$$ = \frac{2np^2(p + 2)}{(1 - p)^2} $$
The distribution of $M_{ij}$ is more convoluted. In fact, it is unknown, and its mean and variance must be calculated directly (Eqs. (32) and (33)). Table 4 summarizes all of these distributions.

\[
\mathbb{E}[M_{ij}] = \mathbb{E}[S_i] + \mathbb{E}[S_j] - \mathbb{E}[2W_{ij}] \\
= \frac{(n-1)p}{(1-p)} + \frac{(n-1)p}{(1-p)} - \frac{2}{(1-p)} \approx \frac{2np - 2}{(1-p)}
\]

\[
\text{Var}(M_{ij}) = \text{Var}(S_i) + \text{Var}(S_j) + \text{Var}(2W_{ij})
\]

\[
= \frac{(n-1)p}{(1-p)^2} + \frac{(n-1)p}{(1-p)^2} - \frac{4p}{(1-p)^2} = \frac{2np}{(1-p)^2}
\]

Table 4: The distribution, mean and variance for each random variable included in weighted Erdős-Rényi overlap.

| 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                    | $\frac{2np-2}{(1-p)}$ | $\frac{2np}{(1-p)^2}$ |

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

\[
\mathbb{E}[g(V_{ij}, M_{ij})] \approx g(\mathbb{E}(V_{ij}), \mathbb{E}(M_{ij})) = \frac{\mathbb{E}(V_{ij})}{\mathbb{E}(M_{ij})} = \frac{np^2}{np - 1} \approx p
\]
\[ \text{Var}(g(V_{ij}, M_{ij})) \approx \frac{\mathbb{E}^2(V_{ij})}{\mathbb{E}^2(M_{ij})} \left[ \frac{\text{Var}(V_{ij})}{\mathbb{E}^2(V_{ij})} + \frac{\text{Var}(M_{ij})}{\mathbb{E}^2(M_{ij})} - \frac{2 \text{Cov}(V_{ij}, M_{ij})}{\mathbb{E}(V_{ij})\mathbb{E}(M_{ij})} \right] \]  

\[ = p^2 \left[ \frac{p + 2}{2np^2} + \frac{1}{2np} - \frac{(1 - p)^2 \text{Cov}(V_{ij}, M_{ij})}{2np^2(np - 1)} \right] = p + \frac{1}{n}. \]

Note that \( \text{Cov}(V_{ij}, M_{ij}) > 0 \) since \( V_{ij} \perp M_{ij} \). 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

\[ \mathbb{E}[g(V_{ij}, M_{ij})] = \frac{\mathbb{E}(V_{ij})}{\mathbb{E}(M_{ij})} + \frac{\text{Var}(M_{ij})\mathbb{E}(V_{ij})}{\mathbb{E}^3(M_{ij})} - \frac{\text{Cov}(V_{ij}, M_{ij})}{\mathbb{E}^2(M_{ij})} \]

\[ \approx p + \frac{n^2p^3}{(np - 1)^3} - \frac{(1 - p)^2 \text{Cov}(V_{ij}, M_{ij})}{4(np - 1)^2}. \]

Again, \( \text{Cov}(V_{ij}, M_{ij}) > 0 \) since \( V_{ij} \perp M_{ij} \). 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. (36) in the approximation of the mean.

### A.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. (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 \( j \). \( K_i^{\text{in}} \) and \( K_i^{\text{out}} \) denote the in and out-degree distributions of node \( i \), respectively. Note that because \( K_i^{\text{in}} \) and \( K_i^{\text{out}} \) are identically dis-
tributed for each node $i$, $\min(k_{i}^{in}, k_{j}^{out}) = \min(k_{j}^{in}, k_{i}^{out})$, and w.l.o.g., we write their sum as $2\min(K_{j}^{in}, K_{i}^{out})$. 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.

\[
O_{ij}^D = \frac{\sum_{k=1}^{n}(A_{ik}A_{kj} + A_{jk}A_{ki})}{\min(k_{i}^{in}, k_{j}^{out}) + \min(k_{j}^{in}, k_{i}^{out})} - 1 \Rightarrow \frac{D_{ij}}{2\min(K_{j}^{in}, K_{i}^{out})} - 1 = \frac{D_{ij}}{C_{ij}} \quad (37)
\]

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 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. (38)) and an upper bound for the variance (Eq. (39)) can be derived. We denote the minimum of two random variables as $K_{(1)}$ and $K_{in}^{i}, K_{out}^{i}$ as simply $K_{i}$. Table 5 summarizes these random variables.
\[ \mathbb{E}[K_{(1)}] = \sum_{k=1}^{(n-1)} P(K_{(1)} \geq k) = \sum_{k=1}^{(n-1)} P(K_1 \geq k)^2 \] (38)

\[ = \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} P(K_i = j) \right]^2 = e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 \]

\[ \text{Var}(K_{(1)}) = 2\text{Var}(K_i) = \frac{2npe^np}{e^{np} - 1} \left[ 1 - \frac{np}{e^{np} - 1} \right] \] (39)

Table 5: The distribution, mean and variance for each random variable included in directed Erdős-Rényi overlap.

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

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

\[ \mathbb{E}[g(D_{ij}, C_{ij})] \approx g(\mathbb{E}(D_{ij}), \mathbb{E}(C_{ij})) \] (40)

\[ = \frac{\mathbb{E}(D_{ij})}{\mathbb{E}(C_{ij})} = \frac{np^2}{e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 0.5} \]
\[ \text{Var}(g(D_{ij}, C_{ij})) \approx \frac{\text{E}^{2}(D_{ij})}{\text{E}^{2}(C_{ij})} \left[ \text{Var}(D_{ij}) + \frac{\text{Var}(C_{ij})}{\text{E}^{2}(C_{ij})} - 2 \frac{\text{Cov}(D_{ij}, C_{ij})}{\text{E}(D_{ij})\text{E}(C_{ij})} \right] \] (41)

\[ = \frac{2n^2p^4}{(2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 1)^2} + \frac{32n^3p^5e^{np}}{e^{np-1}} \frac{1 - \frac{np}{e^{np-1}}}{(2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 1)^2} - \frac{4np^2\text{Cov}(D_{ij}, C_{ij})}{(2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 1)^3} \]

Note that \( \text{Cov}(D_{ij}, C_{ij}) > 0 \) since \( D_{ij} \perp C_{ij} \). 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

\[ \text{E}[g(D_{ij}, C_{ij})] = \frac{\text{E}(D_{ij})}{\text{E}(C_{ij})} + \frac{\text{Var}(C_{ij})\text{E}(D_{ij})}{\text{E}^3(C_{ij})} - \frac{\text{Cov}(D_{ij}, C_{ij})}{\text{E}^2(C_{ij})} \] (42)

\[ = \frac{np^2}{e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 0.5} + \frac{16n^2p^5e^{np}}{e^{np-1}} \frac{1 - \frac{np}{e^{np-1}}}{(2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 1)^3} - \frac{\text{Cov}(D_{ij}, C_{ij})}{(2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 1)^2} \]
Again, 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. (42) in the approximation of the mean.

B. Approach 2 Mean and Variance Derivations

B.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. (43).

$$O_{ij} = \frac{N_{ij}}{H_{ij}} \approx \frac{N_{ij}}{E[H_{ij}]} = \frac{1}{2np - 2 - np^2} N_{ij}$$  \hspace{1cm} (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. (44)) and variance (Eq. (45)) can be easily derived.

$$E[O_{ij}] = \frac{1}{2np - 2 - np^2} E[N_{ij}] = \frac{np^2}{2np - 2 - np^2} \approx \frac{p}{2 - p}$$  \hspace{1cm} (44)

$$\text{Var}(O_{ij}) = \frac{1}{(2np - 2 - np^2)^2} \text{Var}(N_{ij}) = \frac{np^2}{(2np - 2 - np^2)^2}$$  \hspace{1cm} (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.

**B.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 $O^W_{ij}$ as Eq. (46).

$$O^W_{ij} = \frac{V_{ij}}{M_{ij}} \approx \frac{V_{ij}}{\mathbb{E}[M_{ij}]} = \frac{(1 - p)}{2np - 2} V_{ij} \tag{46}$$

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. (47)) and variance (Eq. (48)) are now easily derived.

$$\mathbb{E}[O^W_{ij}] = \frac{(1 - p)}{2np - 2} \mathbb{E}[V_{ij}] = \frac{(1 - p)}{2np - 2} \left( \frac{2np^2}{1 - p} \right) \approx p \tag{47}$$

$$\text{Var}(O^W_{ij}) = \frac{(1 - p)^2}{(2np - 2)^2} \text{Var}(V_{ij})$$

$$= \frac{(1 - p)^2}{(2np - 2)^2} \left( \frac{2np^2(p + 2)}{(1 - p)^2} \right)$$

$$\approx \frac{np^2(p + 2)}{2(np - 1)^2}$$

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.
B.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. (49).

\[
O_{ij}^D = \frac{D_{ij}}{C_{ij}} \approx \frac{D_{ij}}{\mathbb{E}[C_{ij}]} = \frac{1}{2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 1} D_{ij} \tag{49}
\]

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. (50)) and variance (Eq. (51)) are now easily derived.

\[
\mathbb{E}[O_{ij}^D] = \frac{1}{2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 1} \mathbb{E}[D_{ij}] \tag{50}
\]

\[
= \frac{np^2}{e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 0.5}
\]

\[
\text{Var}(O_{ij}^D) = \frac{1}{(2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 1)^2} \text{Var}(D_{ij}) \tag{51}
\]

\[
= \frac{2np^2}{(2e^{-2np} \sum_{k=1}^{(n-1)} \left[ \sum_{j=k}^{(n-1)} \frac{(np)^j}{j!} \right]^2 - 1)^2}
\]

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.
C. 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 7 and 8 show the distributions of raw and standardized unweighted overlap for all edges in the network before stratification. Similarly, figure 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 10 and 11. Figure 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 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 13 and 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 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 17 and 18. Each age category contains edges connecting two nodes belonging to the same age category. Figure 16 shows the distributions of raw and standardized weighted overlap stratified by age.

Figure 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 2 above for full descriptions interaction types.

Figure 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 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 2 above for full descriptions interaction types.
Figure 9: Distribution of average weighted overlap (a) and standardized weighted overlap (b) for all villages.

Figure 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 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 12: Distribution of average weighted overlap (left) and standardized weighted overlap (right) stratified by the presence or absence of nodal attributes.
Figure 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 2 above for full descriptions interaction types.

Figure 14: Distribution of standardized unweighted overlap for each village for each type of social interaction stratified by caste.
Figure 15: Distribution of average weighted overlap (a) and standardized weighted overlap (b) stratified by caste.

Figure 16: Distribution of raw weighted overlap (a) and standardized weighted overlap (b) stratified by age.
Figure 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 2 above for full descriptions interaction types.
Figure 18: Distribution of standardized unweighted overlap for each village for each type of social interaction stratified by age.
Figure 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 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 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.
D. Additional Results

Figure 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 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 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 9b). This is expected given the values in Figure 8 for each type of social interaction are also significantly higher than expected, and that humans do not typically create friendships randomly.

Figure 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 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. 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 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 20 in Appendix C, 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 12a) follows, not surprisingly, the same pattern as its unweighted counterpart (Figure 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 12b), which is again expected from the unweighted distributions in Figure 11.
The median average unweighted overlap is quite similar for the OBC and ‘Other’ caste categories across all interaction types (Figure 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 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 15a). This pattern holds after standardization (Figure 15b), which is a departure from what was seen in the unweighted case (Figure 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 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 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 16a showcases the same pattern among the age categories.
for the average weighted overlap as seen in the unweighted plot in Figure 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 16b).