Contents

1 Introduction

When analyzing dynamical processes in complex environments such as social settings, it is often required to take into account the topological structure that supports their evolution. For systems composed of interacting people, this translates to determining the structure of a social network. The gold standard for this purpose is provided by the so-called *sociometric studies*, which involve surveying all or almost all individuals of the community of interest and gathering their complete web of relationships. This approach results in a global network, but executing a census can easily turn out to be impractical or too costly: moreover, trying to collect the full relational data on a population, even if restricted, can lead to several difficulties, such as boundary specification problems and the greater need to account for missing data ([Newman, 2018], [Perry et al., 2018]). An alternative to this sampling design is given by *egocentric sampling*, which collects information about many non-overlapping personal networks, each centered on an individual in the population ([Perry et al., 2018]). This kind of data is cheaper to obtain, and thus it is much more frequently collected, sometimes even in surveys that are not primarily focused on social networks and other social structures ([Newman, 2018]). Potentially, this means that there is already an extensive amount of network data available for the analysis of researchers. However, ego-centered networks are incomplete in the information they provide, and, more specifically, they are not suited to capture the global structure of a network. A complete topological structure is nonetheless a prerequisite for the analysis of a dynamical process, whether it is an epidemic, the changing of opinions, or any other social phenomenon.

Before detailed social network structure data became abundant thanks to the spread of information technologies, social contagion processes typically relied on the assumption of the homogeneous mixing of the population: agents of the same type are interacting with other agents completely at random and are indistinguishable from each other. This approximation allows to simplify the process evolution to a deterministic and tractable system of differential equations ([Pastor-Satorras et al., 2015]). More recently, this approach has been superseded by the simulation of more realistic social interaction patterns, leading to more accurate models ([Pastor-Satorras et al., 2015], [Eubank et al., 2004], [Longini et al., 2005]), but often relying on census data. In principle, conventional theory-based random graph models could be used to generate different types of network structures for the investigation of socio-dynamical systems, but in practice they fall short in providing realistic topologies. They can serve as tools for evaluating the role of different graph

properties on the system evolution, but they are not predisposed to incorporate empirical network data, and cannot be trusted on their prediction of process behavior; they have often been designed not for fitting data, but to test network theories and capture basic formation mechanisms: as an example, the Barabási-Albert model, though occasionally used as a model for some empirical networks, has been shown unable to describe many characteristics of real-world systems ([Barabási, 2016]).

Relatively little research has been conducted on social contagion processes in settings where comprehensive sociometric data are unavailable. This led scholars to consider ego networks as a viable alternative. Although egocentric data do not reveal the complete network structure or the exact relation of nodes with the broader system, they may still be able to capture relevant aspect of the structure on which the dynamics under study runs. Therefore, examining whether and how much ego-centric network information can improve the analysis of spreading processes is a very relevant topic for applied complex systems scientists. In particular, in this thesis, I will compare egocentric-data-driven models of structural dependencies against widely used, simple, and data-agnostic methods to assess the potential advantages.

In summary, in this work we discuss the reconstruction of global networks on the basis of egocentric-sampling data, aiming to inform the dynamics of spreading processes in social environments. To this end, we present and apply different models on egocentric networks obtained for the study of opinion dynamics concerning Covid-19 and vaccination programs. The data presents two types of relationships, and the methods applied for the reconstruction of the complete social network are analyzed, evaluated and compared. This kind of data is of particular interest due to its generality: it contains multiple types of personal relationships, contacts of participants are not identified, and the structure of the local networks is only obtained through generic proxies.

For egocentric data, and especially of the kind just described, there is no consolidated rigorous inferential framework, and estimation methods have seen little development ([Krivitsky and Morris, 2017]). Nonetheless, the family of exponential random graph models emerged as the most common choice in this context. Following the existing literature, I develop a data-driven reconstruction approach that leverages minimal information about the systems' units and general measures of their relations with their social connections. I compare this method with simple but successful network generation algorithms, measuring

differences in global metrics and the downstream impact of the different structures on the trajectory of some spreading processes. The results of this analysis aim at understanding the utility of incorporating this type of data into the analysis of dynamical systems. We deem the literature on the discussed research question to be currently limited, which motivates us to provide some empirical results that may inform related future works in the field of network reconstruction.

2 Background

2.1 Network theory and models

In this first paragraph, we briefly summarize the core ideas building up the backbone of network science. We also present the field-specific terminology that will be used in the rest of this work to present the details behind the approach and to measure the goodness of the approach.

Among the real-world systems that are considered of interest across many different disciplines, most are probably complex, meaning that their components interact in non-trivial ways. Thus, the global behavior these systems exhibit cannot be reduced to the individual agents, and cannot be modeled without taking into account the relationships between elements. A useful tool to capture the interactions of a system is the *network*, or graph, which is in essence a mathematical structure composed of a set of objects, referred to as *nodes*, or vertices, representing the agents of the modeled system, and a set of *edges*, or ties, identifying the pairwise relationships between agents ([Newman, 2018]). Formally, we denote a network as the pair (N, Y), where $N = \{1, ..., n\}$ is the set of nodes, and Y is a real $n \times n$ matrix with Y_{ij} encoding the relationship between node i and node j. This object is commonly referred to as *adjacency matrix* ([Jackson, 2010]).

Different types of networks can be defined. A network can be undirected, if its edges have no direction, or directed, if a link from i to j does not necessarily imply a link from j to i. If an edge connects a node with itself, it is called a *self-edge*, and if two or more edges connect the same nodes in the same direction, they are called *multi-edges*. A graph that does not allow self-edges and multi-edges is called *simple*; otherwise, it is called a *multigraph*. The network structure thus defined simplifies the system of interest, modeling the pattern of components' interactions, but discarding other kinds of information. To have the graph capture additional aspects of the relevant phenomenon, nodes, and edges are usually given labels and attributes. In particular, a network is *weighted* if its edges do

not merely represent the absence or presence of connections, but instead are characterized by some non-binary value.

For a graph $G = (N, Y)$ that is undirected, simple and unweighted, we have $Y_{ij} \in \{0, 1\}$ for all $\{i, j\} \in N^2$, with 0/1 coding for tie absence/presence. Also, $Y_{ii} = 0$ for all $i \in N$, and, being undirected, Y is symmetric. These are the type of networks which we focus on in this study.

Let E denote the set of edges of network $G = (N, Y)$, that is, $E = \{(i, j) \in N^2 : Y_{ij} > 0\}$. A *walk* in a network G between nodes i and j is a sequence of edges $v_1v_2, v_2v_3..., v_{K-1}v_K$, with $v_1 = i$ and $v_K = j$. If each node in the sequence $v_1, v_2, ..., v_K$ is distinct, then we call the sequence a *path* from i to j. A *cycle* is a walk with distinct nodes except for the terminal ones, which are the same, $v_1 = v_K = i$. The *distance* between two nodes is the length of the shortest path between them; "length" typically refers to the number of edges of the path, or, if applicable, to the sum of edge weights. A network is said to be *connected* if any two nodes are joined by at least one path. We call the *components* of a network its distinct maximal connected subgraphs. A *tree* is a connected network with no cycles. A *complete network* is one where an edge exists between any pair of nodes, i.e. $Y_{ij} > 0$ for all $i, j \in N^2, i \neq j$. A *clique* is a maximal complete subgraph of a network ([Jackson, 2010]).

The set of nodes which are linked to some node i is called the *neighborhood* of i. Its cardinality, i.e. the number of edges connected to i for simple graphs, is the *degree* of i, denoted k_i. This concept is generalized to weighted networks with the *strength*, or weighted degree, s_i, summing the weights of links incident to i. The *degree distribution* p_k provides the relative frequency, or probability, of finding a node of degree-k in the network.

Degrees are also measures of centrality, meaning they quantify the importance of a node in a network. Together with degrees, there are various others centrality metrics (see [Newman, 2018]), among which:

1. *Closeness centrality*, the inverse of the mean distance of a node from all other nodes:

$$x_i = \frac{n}{\sum_j d_{ij}}.$$

2. *Betweenness centrality*, which measures how well situated a node is in terms of the

paths that it lies on:

$$x_i = \frac{1}{n^2} \sum_{s,t \in N} \frac{n_{st}^i}{g_{st}},$$

where g_{st} is the number of shortest paths from node s to node t, and n_{st}^i is the number of shortest paths from s to t that pass through node i;

3. *Eigenvector centrality*, which is a self-referential metric, as the node's centrality is proportional to the centrality of its neighbors:

$$x_i = \kappa^{-1} \sum_j Y_{ij} x_j,$$

where κ^{-1} is a proportionality constant; the measure can be viewed as an extension of degree centrality where a node's importance is affected not only by the number of its neighbors, but also by their prominence.

To gain insight into the centrality and compactness of an overall graph, useful metrics are the average and standard deviation of node-centrality indexes, as well as the centralization measures introduced by [Freeman, 1979]. The centralization measure of a network for some centrality measure x_i is defined as:

$$x(G) = \sum_{i \in N} \left| \max_{i \in N}(x_i) - x_i \right|,$$

and it quantifies the range or variability of the individual centrality indices.

Let m be the cardinality of the set of edges E. The average degree of a network can be written as:

$$\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{2m}{n}.$$

The maximum number of edges of $G = (N, Y)$ is equal to $\binom{n}{2}$. We call the fraction of present to maximum edges the *density* of G:

$$\rho = \frac{m}{\binom{n}{2}} = \frac{\langle k \rangle}{n-1}.$$

A network is said to be *sparse* if $\rho \to 0$ as $n \to \infty$; by the above expression, sparsity is equivalent to the average degree growing in a sublinear fashion with the number of nodes.

The above measure ρ can be adapted to measure the local density around a node, that is,

the density of the subgraph induced by its neighborhood:

$$C_i = \frac{2m_i}{k_i(k_i - 1)},$$

where m_i is the number of links between neighbors of i. This is also referred to as the *local clustering coefficient*. To capture the clustering of the whole network, one can either use the *average clustering coefficient*:

$$\langle C \rangle = \frac{1}{n} \sum_{i=1}^{n} C_i,$$

or sum C_i across all nodes i, thereby obtaining the *global clustering coefficient* C_Δ, which can be written as:

$$C_\Delta = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}},$$

where a triangle is a completely connected triad of nodes, and a *connected triple* refers to an ordered set of three nodes uvw such that u connects to v and v connects to w. The average and global clustering coefficients, though similar measures, can exhibit discrepancies, as the first one gives more weight to low-degree nodes.

Up until now, we have seen the graph G as fixed. However, we often consider G to be a random variable taking values on a space of graphs $\mathcal{G}$, and refer to it as *random graph*. The uses of random graphs include studying the mechanisms originating empirical network structures, testing the significance of some graph property, and obtaining estimators of network measures. We will assume that all graphs in $\mathcal{G}$ have the same number of nodes. In this framework, by *random graph model*, we mean a collection of probability distributions $P_\theta(G)$ over $\mathcal{G}$, indexed by a set of parameters $\theta \in \Theta$ ([Kolaczyk, 2009]). Below, some relevant classes of models which can be employed in network reconstruction are outlined.

2.1.1 Conditionally uniform models

Conditionally uniform models assume G to be uniform over $\mathcal{G}$, with the space $\mathcal{G}$ being restricted to graphs satisfying a set of properties. Typically, the researcher fixes some statistics she wishes to control for and tests some other properties of interest using the random graph distribution as the null hypothesis ([Snijders, 2011]). The simplest model in this family is the *Erdős–Rényi (ER) model*([Erdös and Rényi, 1959]), which fixes just the number of edges in the graph. The ER random graph, denoted with $G(n, m)$, is then

uniformly distributed over all graphs with n nodes and m edges, and can be sampled by choosing m node pairs uniformly at random from all dyads and linking them. More often, however, by Erdős–Rényi random graph model we refer to a slight variant of this, denoted $G(n, p)$, where, instead of fixing the number of edges, we fix the probability of any edge p. We sample from this model simply by creating an edge between each node pair with probability p. The degree distribution for $G(n, p)$ then follows the binomial distribution:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}. \tag{1}$$

The average number of edges if $\langle m \rangle = \binom{n}{2} p$ and the average degree is $\langle k \rangle = (n-1)p$. Also, global clustering is $C = \langle k \rangle / (n-1) = p$. Since most real networks are sparse, we have that $C \xrightarrow{n \to \infty} 0$, which differs from most empirical networks. Also, the binomial distribution (1) can be approximated by the Poisson distribution:

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}, \tag{2}$$

which makes the model more analytically tractable.

The ER random network exhibits topologically different regimes depending on the average degree value, $\langle k \rangle$ ([Barabási, 2016]). If $\langle k \rangle \leq 1$, the network has few links and the relative size of the largest component tends to 0 as n increases. When $\langle k \rangle > 1$, the largest component contains a non-zero fraction of the nodes, and we call it the *giant component*. For $\langle k \rangle > \ln(n)$, the graph is connected.

2.1.2 Network growth models

Other modeling methods induce $P_\theta(G)$ implicitly through a recurrent generative mechanism. This approach allows for the specification of the mechanism by which a network structure emerges and changes, thereby giving rise to so-called *network growth models*. A famous member of this family is the *Barabási-Albert (BA) model* ([Barabási and Albert, 1999]). The BA model generates networks through the following algorithm: starting with n_0 nodes with an arbitrary number of links among them, at each time step a new node is added to the network; then $m \leq n_0$ edges are created between the new node and m other nodes from the existing network. There, the probability that an added edge will be connected to an existing node i of degree k_i is given by $k_i / (\sum_j k_j)$. Through the combination of just two simple mechanisms, i.e. growth and preferential attachment, the BA model can

generate networks with more realistic properties compared to the ER model. In particular, it generates *scale-free networks*, that is, graphs whose degree distribution follows a power law distribution:

$$p_k \sim k^{-\gamma}. \tag{3}$$

The motivation behind the term *scale-free* lies in the limiting behavior of the system as $n \to \infty$. The degree standard deviation σ_k diverges, provided that γ is between 2 and 3, which is typically the case in empirical data ([Barabási, 2016]). Thus, this leads to arbitrarily large nodes, which means that the graph can be considered to be without any internal scale, meaning that there is no single typical degree value that represents most of the nodes in the network, as there is always a non-negligible probability of finding nodes with degree orders of magnitude larger than the average. Conversely, if k is Poisson-distributed as in (2), then the standard deviation is well defined and finite: $\sigma_k = \langle k \rangle^{1/2}$. In this case, the average degree, $\langle k \rangle$ serves as a characteristic scale for the system ([Barabási, 2016]).

Compared to ER random graphs, scale-free distributions display fat tails, which means that the graph has a larger number of low- and high-degree nodes, whereas there is a lower probability to observe $k \approx \langle k \rangle$. The presence of hubs has also the effect of shrinking distances in the network. However, the topology of these graphs is sensitive to the value of the degree exponent γ, and becomes indistinguishable from the $G(n, p)$ model for $\gamma > 3$.

2.1.3 Models of structural dependencies

Another type of random graph models are the so-called *models of structural dependencies*. This family of models aims to capture the structural dependencies between edge variables (i.e. the random elements of the adjacency matrix), and possibly other endogenous or exogenous factors. Conditionally uniform models and network growth models are useful for many purposes, like hypothesis testing and graph generation, but less so as models to be used in real world application studies, meaning as statistical models to be fit on empirical data ([Kolaczyk, 2009]). Widely used classes of models that have been designed with the intent to estimate the structural pattern of empirical networks are, for instance, *exponential random graph models* (ERGMs) and *stochastic block models* (SBMs).

ERGMS constitute an exponential family of distributions, as they assume the logarithm of graph probability to be a linear combination of parameters and statistics; this allows to

leverage the wealth of known properties and developed methods for this type of models. As we cover with more detail in section 2.7, this framework models networks on the basis of counts of some configurations (i.e. small, local subgraphs) of interest. Though ERGMs can be used for simulating networks, they are primarily employed on empirical data, for the investigation of potentially different hypotheses on a network formation process and structure ([Snijders, 2011], [Lusher et al., 2012]).

SBMs are another family of models that can be used to both generate and fit networks to data. They assume each node to belong to one[1] of two or more different groups, and have the probability of each tie being dependent on the group assignments of the end nodes. SBMs are popular both for their analytical tractability and for the numerous extensions and specifications that have been developed for different purposes. When using them for simulation, they can produce a wide range of network structures, including conventional community, core-periphery, and hierarchical structures ([Karrer and Newman, 2011]). They are also widely employed for fitting an empirical network with the aim of recovering its block structure, a task referred to as *a posteriori blockmodeling* ([Snijders and Nowicki, 1997]). We describe this framework in greater detail in section 2.6.

2.2 Egocentric network sampling

Ideally, a researcher examining the structure and properties of a social network would collect the full data on all individuals, their links, and their characteristics; however, this is seldom practically possible. A more realistic approach is sampling a subset of actors from the entire population and using the data to infer a network's global properties. This can be done through a number of network sampling schemes (see [Smith, 2012] for a summary). A strategy consists of selecting a random sample of nodes, together with all ties occurring among them; this is called a *subgraph sample*, and it has the considerable disadvantage of providing little information for most empirical networks, which are typically very sparse and whose induced subgraphs have therefore low density. This issue is fixed in *snowball sampling* schemes, in which, starting from a set of individuals, researchers interview their connections, and then the connections of their connections, and so on. In practice, this process can easily turn out to be quite demanding. By contrast, another procedure called *ego network sampling*, proves to be both informative and feasible: it consists of surveying a random sample of the population about their characteristics, their relationships, and

[1]Nodes can be allowed to belong to multiple groups, an approach referred to as *soft-clustering*([Lee and Wilkinson, 2019]).

possibly the ties among their social contacts. Its final result is a set of independent local networks, each centered on an individual, named *egocentric networks*, or ego-centered networks. The central individual is termed *ego*, and the nodes in its ego-network are called *alters*.

Egocentric sampling schemes can be designed so that alters are uniquely identified, meaning that we know any intersection between the sampled graphs, and they can vary in the information provided on the alter-alter ties. If alters cannot be uniquely identified, and connections between alters are not explicitly and directly collected, we call this egocentric sampling design "minimal". The minimal design reduces the burden of data collection, and therefore is more widespread. However, it also limits inference power, especially as it does not contain information on network transitivity, meaning the tendency of nodes sharing a neighbor to be connected ([Krivitsky and Morris, 2017]). In this work, we focus on egocentric data containing connections between alters, possibly through proxies, like reported measures of density. However, we keep the assumption that alters are not directly identified, which is not only true for the empirical data we are going to use, but more in general is typical in population-representative topic-specific surveys; indeed, these are designed to investigate specific research questions, and often deem superfluous or impractical the extraction of detailed network information.

[Handcock and Gile, 2010] outlined a likelihood-based inferential framework for network samples, which includes egocentric designs. This approach assumes a stochastic graph Y, and relies on using partial (egocentric) data on its realization y to estimate a model of the process generating the underlying complete network; for this reason, it is a *model-based* approach. An alternative to this is the *design-based* framework, which removes the distinction between Y and y, considering the underlying population as a set of fixed parameters to estimate. Both methods are discussed in the following sections.

2.3 Egocentric network inference

2.3.1 Model-based inference

In their model-based framework, [Handcock and Gile, 2010] assume sampling of specific edges in the global network, hence their work only suits designs in which alters are identified (or which at least allow for approximating the identities, for instance when an informative enough set of attributes is obtained). However, the distinction between

schemes with and without identification can be negligible if the sample size is much smaller than the population size, and even more for some network topologies; in some cases of non-identification then, possible intersections between sampled nets could be ignored.

The foundation of the framework of [Handcock and Gile, 2010] describes the distribution of sampled units (ties) obtained under an egocentric sampling design. We only consider networks on a fixed set of nodes $N = \{1, ..., n\}$, hence we will equivalently refer to Y both as an adjacency matrix and as a graph. Let Y be a random network taking values in $\mathcal{Y}$, the set of all possible undirected networks on N. The realized empirical network is y. In order to estimate the network-generating process for Y, suppose we conduct egocentric sampling on y. As a result, S is the $n \times 1$ vector such that $S_i = 1$ if node i was sampled, and $S_i = 0$ otherwise. Binary (symmetric) $n \times n$ matrix D is defined such that $D_{ij} = 1$ if Y_{ij} has been sampled, and $D_{ij} = 0$ otherwise. We assume that each actor is selected independently from the others, with probability ψ, and we collect all the edges incident to the sampled nodes. It follows that $D_{ij} = 1 \iff (S_i = 1 \vee S_j = 1)$ and D can be written as a deterministic function of S. Indeed, we have $D_{ij} = S_j + S_i - S_j S_i$, or equivalently $D = \mathbf{1}S^T + S\mathbf{1}^T - SS^T$. From this, we can write the probability distribution of D, called the *sampling design*, as follows:

$$P(D = d|Y, \psi) = \psi^{\sum_i s_i}(1 - \psi)^{n - \sum_i s_i}, \tag{4}$$

where $d = \mathbf{1}s^T + s\mathbf{1}^T - ss^T$, $s \in \{0, 1\}^n$.

From [4], we see that, for egocentric sampling, $P(D = d|Y = y, \psi) = P(D = d|\psi)\ \forall y \in \mathcal{Y}$, as indeed information collected by survey participants is not used to direct the sampling of new egos. Sampling designs exhibiting this property are termed *conventional*, and are a special case of *adaptive designs*, for which unobserved relationships are "missing at random", given the observed data.

Having determined the egocentric sampling design $P(D)$, we proceed to discuss how to conduct inference based on the data. Let Y^{obs} denote the restriction of Y_{ij} to indices $ij : D_{ij} = 1$, that is, the observed portion of the complete network. The unobserved part of Y is denoted Y^{mis} and defined analogously as the restriction of Y_{ij} on $ij : D_{ij} = 0$. For the realization of these processes, we set the following notation: $y = y^{\text{obs}} + y^{\text{mis}}$. We define the set $\mathcal{Y}(y^{\text{obs}})$ as the set of possible unobserved portions induced by y^{obs}, that is, $\mathcal{Y}(y^{\text{obs}}) = \{x : x + y^{\text{obs}} \in \mathcal{Y}\}$.

Assume the random graph Y follows a distribution $\mathbf{Y}$ parameterized by $\theta \in \Theta$ and taking values on the set of undirected networks of n nodes $\mathcal{Y}$, $Y \sim \mathbf{Y}(\theta)$. We collect a subset of the network, y^{obs}, through a selection of the edges d obtained by an egocentric sampling design with parameter ψ. Inference on θ is based on the likelihood $\mathcal{L}(\theta, \psi | d, y^{\text{obs}}) = P(Y^{\text{obs}} = y^{\text{obs}}, D = d | \theta, \psi)$, which can be complex; however, [Rubin, 1976] has shown that, whenever the sampling design is adaptive and the sampling parameters ψ are distinct from θ, it is equivalent to conducting inference with the *face-value likelihood* of θ given $Y^{\text{obs}} = y^{\text{obs}}$:

$$\mathcal{L}(\theta | y^{\text{obs}}) \propto \sum_{x \in \mathcal{Y}(y^{\text{obs}})} P(Y = y^{\text{obs}} + x | \theta). \tag{5}$$

Hence, for the special case of egocentric sampling, we can show that the two likelihoods are proportional:

$$\begin{aligned} \mathcal{L}(\theta, \psi | d, y^{\text{obs}}) &= \sum_{x \in \mathcal{Y}(y^{\text{obs}})} P(D = d | Y = y^{\text{obs}} + x, \theta, \psi) P(Y = y^{\text{obs}} + x | \theta, \psi) \\ &= \sum_{x \in \mathcal{Y}(y^{\text{obs}})} P(D = d | \psi) P(Y = y^{\text{obs}} + x | \theta) \\ &= P(D = d | \psi) \sum_{x \in \mathcal{Y}(y^{\text{obs}})} P(Y = y^{\text{obs}} + x | \theta) \\ &= \mathcal{L}(\psi | d) \mathcal{L}(\theta | y^{\text{obs}}), \end{aligned} \tag{6}$$

and so estimation can ignore the nuisance parameter ψ and the information provided by $D = d$.

2.3.2 Design-based inference

As previously stated, in the design-based framework our population network y is fixed, and so the only source of randomness comes from the sampling design. While in the likelihood-based method of the previous section we conducted inference on the parameters θ, under this alternative approach the unobserved characteristics of the population substitute θ and become the target of estimation. While it can certainly appear less flexible than a model-based framework as it removes randomness from the underlying network, this approach has two important advantages: it can incorporate survey weights in addressing survey selection biases, and it does not require egos to identify alters.

Another difference with the method in Section [2.3.1] is that this one does not require a model for the data, though a model may still be used to guide design-based inference

([Handcock and Gile, 2010]). For instance, [Krivitsky and Morris, 2017] specifies an exponential random graph model (see Section [2.7]) and conducts inference using design-based estimators for its sufficient statistics.

Before considering potential modeling approaches for our data, we briefly discuss the type of local information that can be obtained from egocentric networks, as it constitutes a crucial component in the specification of our models.

2.4 Egocentric network analysis

Egocentric network analysis (ENA) is the branch of network science studying egocentric networks, and so it is concerned with the subsets of some population graphs obtained by focusing on single nodes and their neighbors. In this work, we focus on simple undirected ego networks of order n_i. Keeping notation consistent with that for global networks, we use k_i to denote the number of nodes excluding the ego, $n_i - 1$. The number of edges is given by $k_i + m_i$, where m_i is the number of alter-alter links.

The main goals of ENA are explaining and predicting attributes of the ego from its network, and explaining the structure of the ego network itself. Through analysis of egocentric graphs we can extract a series of measures describing the social circle of subjects. Many of these metrics summarize either composition or structure of the network ([Perry et al., 2018]).

Composition deals with the distribution of the attributes of nodes, and their potential interactions. The simplest composition metrics are each alter attribute's absolute and relative frequencies. If we want to focus on the uniformity of some attributes in the network, we need a heterogeneity measure, for example, Blau's heterogeneity index

$$H = 1 - \sum_k p_k^2,$$

where p_k is the proportion of alters with some categorical attribute having value k. Another diversity measure is given by the *Shannon diversity index*, or Shannon's entropy, which is calculated as:

$$H = -\sum_k p_k \ln(p_k).$$

For continuous attributes, diversity can be measured by standard deviation.

When, instead of looking at alter-alter similarity, we analyze ego-alter similarity, we deal

with measures of homophily. For categorical attributes, a measure of this is the "proportion homophilous", meaning the proportion of alters sharing the ego's attribute value, or the equivalent E-I index

$$EI = \frac{\text{External} - \text{Internal}}{\text{External} + \text{Internal}},$$

where Internal is the number of alters similar to the ego, and External is the number of the different ones. For continuous attributes, homophily/heterophily can just be computed through the average distance ([Perry et al., 2018]).

Structure focuses on the networks' ties, and especially on links between the alters. The structural metrics of egocentric networks coincide with the local structural metrics for the whole networks discussed in section 2.1. For example, density: $2m_i/(k_i(k_i - 1))$, also referred to as the ego-level clustering coefficient. A useful extension of density is the fragmentation index, which, instead of counting the proportion of alter pairs that are directly linked, counts the proportion that are connected by paths of any length; this, however, together with other structural measures like ego's betweenness and alters' centrality, are less relevant for our purposes, as the structure of our considered ego-networks relies on categorical proxies for within-alter ties.

Local features of the original network are relatively straightforward to study through ego-networks, as long as we can consider the set of egos to be representative of the population. It is a more involved task to make inferences on global features of the underlying complete graph, for example in order to investigate the dynamics of spreading processes. Indeed, egocentric sampling provides us with a series of small local regions of the networks, with these regions being typically disjoint ([Newman, 2018]), either because the sample size is too small or, as is often the case because the alters are not identified. Hence, in order to estimate the complete graph, previous research has mostly adopted a simulation approach, by generating global graphs on the basis of the local information from egos. This can be achieved through a set of different models, which are outlined in the next sections.

2.5 Configuration model

The *configuration model* belongs to the family of conditionally uniform models. It is a random graph model assigning equal positive probability to any graph with a pre-specified number of nodes $n = |N|$ and a degree sequence, and probability zero to every other

network. Let $\{k_1, k_2, ..., k_n\}$ be a sequence of degrees, either fixed or sampled from a degree distribution p_k, then a sample from the configuration model ensemble can be obtained by assigning to each node $i \in N$ k_i half-edges, and iteratively connecting pairs of stubs at random. The resulting network will have $m = \frac{1}{2}\sum_{i=1}^{n} k_i$ edges, among which possibly some self-edges and multiedges. The expected number of edges linking any two different nodes i and j is $k_i k_j/(2m-1)$. As m increases, for constant k_i, k_j, we have that the probability of i and j being connected coincides with their expected number of connections, and we write

$$p_{ij} \approx \frac{k_i k_j}{2m-1} \approx \frac{k_i k_j}{2m} \quad \text{for } m \text{ large enough.} \tag{7}$$

For self-edges, the probability is:

$$p_{ii} \approx \frac{k_i(k_i-1)}{4m} \quad \text{for } m \text{ large enough.} \tag{8}$$

From this, we can show that the expected number of multiedges and self-edges is constant in n, assuming $\langle k^2 \rangle$ is constant; hence, as n increases, their fraction becomes negligible and the degree distribution converges to the target distribution.

Let p_k be the fixed degree distribution in the configuration model. If we choose an arbitrary node in the network with a positive degree and follow one of its edges, then the number of other links of the resulting neighbor follows the *excess degree distribution*, given by:

$$q_k = \frac{(k+1)p_{k+1}}{\langle k \rangle}. \tag{9}$$

The global clustering coefficient, measuring the average local density, can then be obtained (see [Newman, 2018] for a detailed description):

$$C = \sum_{k_i \geq 0} \sum_{k_j \geq 0} q_{k_i} q_{k_j} \frac{k_i k_j}{2m} = \frac{1}{n} \frac{(\langle k^2 \rangle - k)^2}{\langle k^3 \rangle}. \tag{10}$$

Therefore, if we assume that the moments of the degree distribution are constant, $C \xrightarrow{n \to \infty} 0$, and so we expect to find few or no alter-alter ties in our ego-networks, for large enough populations. Furthermore, this is just an implication of the following general property of the configuration model: as m increases, if the average excess degree, $E_q(k) = \sum_{k \geq 0} k q_k = (\langle k^2 \rangle - \langle k \rangle)/\langle k \rangle$ is constant, the probability of networks that are locally tree-like approaches 1. Thus, in the configuration model, we expect that, for large values of n, generated

networks have no loops.

These properties of the configuration model, i.e. tree-like substructures and the vanishing of clustering, make it particularly constrained for empirical social network modeling. In fact, typically, such networks exhibit a diverse and complex range of characteristics. Conversely, the model is more realistic when considering degree distributions, which can be arbitrarily chosen, and path lengths, as graphs exhibit the small-world effect.

Due to its shortcomings, the base configuration model has been generalized to incorporate additional structural features. For example, [Ball et al., 2012] extends the model to include also the specification of clustering and degree correlation. Their approach to induce non-vanishing clustering was to sample a series of group sizes, and then partition the nodes among the resulting clusters; then, all vertices belonging to the same groups were connected to each other, thereby giving rise to a network consisting of interconnected households. Previously, other works had investigated models with tunable clustering and the implications of local density on spreading processes (e.g. [Britton et al., 2007]), also in the form of a households-partitioned population, as found in [Ball et al., 2010]. Degree correlation, which is positive in empirical social networks ([Newman, 2018]), was incorporated in [Ball et al., 2012] by having a fraction of each node's neighbors be chosen among members of the same (or opposite) quantile of the degree distribution (excluding within-households edges). Additionally, to make the configuration model better able to capture important information for the modeling of spreading processes, [Britton et al., 2011] extended it to weighted networks, with integer-valued weights whose distribution is allowed to depend on the nodes' degrees. We also mention that, in a variant reminiscent of the difference between $G(n, m)$ and $G(n, p)$, called Chung-Lu model ([Chung and Lu, 2002]), instead of fixing the sequence of degrees, one chooses the sequence of expected degrees c_i, so that the edge probability of two different vertices is given by $p_{ij} = c_i c_j / 2m$, with m being the expected number of edges, and $\langle k_i \rangle \sim \text{Poisson}(c_i)$. Overall, despite all the extension that have been developed, the configuration model remains in general an inflexible and unutilized method for network reconstruction, and adapting it to any particular dataset may prove nonviable.

2.6 Stochastic block models

Another class of methods for generating and fitting networks is given by *stochastic block models* (SBMs). SBMs assume the nodes to be partitioned into communities and assign a

tie probability depending on the group of each node. This makes this family of models particularly suited to model networks exhibiting community structures.

Consider an undirected graph $G = (N, Y)$ with $n = |N|$ nodes partitioned into K groups, and a binary adjacency matrix, Y. The group assignments are collected in the $n \times 1$ vector g, where g_i denotes the community to which vertex i belongs, and it can be either fixed or random. In the base SBM, the number of edges between each dyad (i, j) has a Poisson distribution with mean $\psi_{g_i g_j}$. If we denote as $\omega_{g_i g_j}$ the expected value of Y_{ij}, then it follows that $\psi_{g_i g_j} = \omega_{g_i g_j}$ if $i \neq j$, and $\psi_{g_i g_i} = \frac{1}{2}\omega_{g_i g_i}$ for all $i \in N$. Let ω be the matrix of parameters ω_{rs}. Thus, for a given matrix ω and vector g, the graph distribution of the simple SBM is given by:

$$P(Y|\omega, g) = \prod_{i<j} \left[\frac{\omega_{g_i g_j}^{Y_{ij}}}{Y_{ij}!} e^{-\omega_{g_i g_j}} \right] \prod_i \left[\frac{(\frac{1}{2}\omega_{g_i g_i})^{Y_{ii}/2}}{(Y_{ii}/2)!} e^{-\frac{1}{2}\omega_{g_i g_i}} \right]. \tag{11}$$

See [Lee and Wilkinson, 2019] for further details.

The main drawback of this model is that the degrees are Poisson-distributed, thereby lacking the degree diversity observed in real social networks. This makes the generated graphs unrealistic, and can potentially misguide community detection. This issue can be easily solved by extending the model in [11] to incorporate degree heterogeneity as proposed by [Karrer and Newman, 2011]. This variant, known as *degree-corrected stochastic block model*, introduces a set of parameters θ_i controlling the expected degrees of nodes; the distribution in [11] is consequently modified by replacing $\omega_{g_i g_j}$ with $\theta_i \theta_j \omega_{g_i g_j}$, the new expected value of Y_{ij} . Then we have

$$P(Y|\omega, \theta, g) = \prod_{i<j} \left[\frac{(\theta_i \theta_j \omega_{g_i g_j})^{Y_{ij}}}{Y_{ij}!} e^{-\theta_i \theta_j \omega_{g_i g_j}} \right] \prod_i \left[\frac{(\frac{1}{2}\theta_i^2 \omega_{g_i g_i})^{Y_{ii}/2}}{(Y_{ii}/2)!} e^{-\frac{1}{2}\theta_i^2 \omega_{g_i g_i}} \right], \tag{12}$$

where we impose the constraint

$$\sum_i \theta_i \delta_{g_i, r} = 1 \tag{13}$$

to specify the model.

Compared to other methods, stochastic block models certainly stand out if modeling community structure is of primary importance. Also, they can be estimated when data is missing, making them suited to ego networks, for both undirected and directed ties. As

already stated, one is often interested in reconstructing systems of polytomous interactions; in this respect, a range of extensions of SBMs to weighted graphs have been developed: for instance, [Yang et al., 2011] model discrete tie strength of each dyad (i, j) with a geometric distribution with success probability $\omega_{g_i g_j}$, while in [DuBois et al., 2013] $i - j$, interactions follow a Poisson process with intensity $\exp(\omega_{g_i g_j})$. However, in studies of network reconstruction from egocentric data, it is considerably less employed relative to the ERGM methods we discuss next, and the same is true in the tools developed for performing inference on ego-networks.

2.7 Exponential random graph models

Another class of methods for generating and fitting networks is given *exponential random graph models* (ERGMs), which model the distribution over graphs with a fixed number of vertices, by describing the probability of each network as a function of its local structural patterns and other covariates. The local structures whose presence the graph probability is dependent on are termed *configurations*. Simple ties, reciprocated ties, stars and triangles are examples of configurations, whose scope can thus be at or above the level of a single dyad.

Then, the general form of ERGMs for the distribution of the random graph Y is:

$$P(Y = y|\theta) = \frac{\exp(\theta^T z(y))}{K(\theta)}, \tag{14}$$

where $z(y)$ is a vector whose elements are counts of some network configuration in y, θ is a real-valued vector of parameters such that θ_i governs the individual contribution of the statistic $z_i(y)$ to the probability of graph y, and $K(\theta)$ is the normalizing constant ensuring [14] is a probability distribution.

The configurations included in the model are decided on the basis of dependence assumptions on the network ties (see [Lusher et al., 2012]). ERGMS were first introduced with the work of [Frank and Strauss, 1986], with the proposal of the *Markov random graph model*, which assumes Markov dependence: the probability of a tie, conditional on the rest of the graph, depends on incident ties, whereas dyads not sharing nodes are assumed independent. This makes the probability of Y dependent on the number of edges, stars and triangles it contains.

While the Markov model has seen wide use by researchers, subsequent research ([Snijders,

2002], [Handcock, 2003]) has shown that, for specifications based on most empirical social networks, it induces a degenerate graph distribution, leading to problems in estimation and fitting to realistic data. This consideration led [Snijders et al., 2006] to develop the *social circuit model*, which assumes, together with Markov dependence, that two disjoint ties Y_{ij}, Y_{hm} are dependent, conditionally on the rest of the graph, if edges ih and jm exist. The model allows many new statistics, and in particular introduces the k-triangle (i.e. two connected nodes that are also linked to k other nodes), and k-independent 2-paths (i.e. two nodes that are also linked to k other nodes). While these configurations make the ERGM more suited to model triangles, they require the estimation of too many parameters, which led [Snijders et al., 2006] to propose the following statistics: the geometrically weighted degree statistics (GWD) $\sum_r w_r(\alpha) D_r(Y)$, the geometrically weighted edgewise shared-partner statistic (GWESP) $\sum_r w_r(\alpha) T_r(Y)$, and the geometrically weighted dyadic shared-partner statistic (GWDSP) $\sum_r w_r(\alpha) P_r(Y)$, where D_r is the number of degree-r nodes in Y, T_r is the number of r-triangles, P_r is the number of r-independent 2-paths, and $w_r(\alpha)$ is a function of the damping parameter α, which reduces the effect of D_r, T_r and P_r for higher r as α increases. The damping parameter can be either estimated or set fixed, with larger values reducing degeneracy problems ([Lusher et al., 2012], [Snijders, 2011]). The GWESP statistics is particularly interesting as it is our preferred way to include transitivity in ERGMs in this work. The number T_r of r-triangles in its formula can be defined as the number of linked node pairs with exactly r edgewise shared partners, meaning they share other r connected actors.

The model in [14] can be extended to include the effects of node-level and tie-level attributes in the following form:

$$P(Y = y|\theta, X = x) = \frac{\exp(\theta^T z(y, x))}{K(\theta, x)}, \tag{15}$$

where X stores individual or dyadic covariates, and the vector of statistics z can now include configurations involving network-attribute interactions.

The introduction of node attributes can control for effects such as homophily in network formation, while dyadic covariates may be for example another network on the same set of actors. Differently from structural statistics $z(y)$, which are endogenous in the model, these additional predictors are instead exogenous and treated as fixed.

The framework is amenable for estimation when data is missing, and for egocentric

networks, which makes it appropriate for our task. Evidently, when fitting an ERGM on egocentric data, one can only include statistics that can be extracted from ego-centered networks. ERGMs can be used to model both directed and undirected networks. However, they are typically applied to binary data, with edge weights being either ignored or used as covariates ([Lusher et al., 2012]). If dichotomizing a weighted graph in order to apply ERGMs leads, in general, to a loss of information, this is even more relevant for our purposes, as the accurate prediction of a dynamical process is typically dependent on a measure of tie strength. [Krivitsky, 2012] generalize ERGMs to valued networks where ties are counts, that is, Y_{ij} takes values on $\mathbb{N}_0$ for all i, j. Since the sample space of the random adjacency matrix Y, $\mathcal{Y}$, is not finite as in the binary case, the normalizing constant in [14] may be infinite. Hence, the model requires the specification of a function $h : \mathcal{Y} \to [0, \infty)$, called *reference measure*, which determines the support and the basic shape of the ERGM distribution. An ERGM with reference measure h assigns probability to a graph y equal to [14] multiplied by $h(y)$.

The general nature and versatility of ERGMs together with their interpretability and existing literature connecting them to network reconstruction make them the preferred tool to translate egocentric data into a global system, thereby informing the dynamics of social spreading processes. Thus, we briefly summarize the methods used to fit ERGMs in the following paragraph, and move on to applying them to empirical data in section [3].

2.8 Estimation, simulation and evaluation of ERGMs

The goal in fitting an ERGM is centering the distribution of statistics $z(Y)$ over the observed network statistics $z(y^{\text{obs}})$, that is, solving the moment equation $\mathbb{E}_\theta(z(Y)) = z(y^{\text{obs}})$ for θ. For ERGMs, as all exponential family distributions, the value of θ solving this equation coincides with the maximum likelihood estimate of θ ([Lusher et al., 2012]).

The loglikelihood function for the model in [14] is

$$l(\theta) = \theta^T z(y^{\text{obs}}) - \log K(\theta). \tag{16}$$

Since $\log K(\theta)$ involves a summation over the space $\mathcal{Y}$, it cannot be evaluated explicitly, and must be approximated numerically. We outline one of the methods applied to this task, derived from the work of [Geyer and Thompson, 1992], and called *Markov chain Monte Carlo maximum likelihood estimation* (MCMC-MLE).

First, we notice that maximizing [16] is equivalent to maximization of the log-likelihood ratio

$$l(\theta) - l(\theta_0) = (\theta - \theta_0)^T z(y^{\text{obs}}) - \log\left(\frac{K(\theta)}{K(\theta_0)}\right), \tag{17}$$

for an arbitrary θ_0. One can show that

$$\frac{K(\theta)}{K(\theta_0)} = \mathbb{E}_{\theta_0}\left[\exp\left\{(\theta - \theta_0)^T z(Y)\right\}\right], \tag{18}$$

and so, by the law of large numbers, we can approximate $\log\left(\frac{K(\theta)}{K(\theta_0)}\right)$ by first generating a Markov chain Monte Carlo sample $Y_1, ..., Y_m$ from model [14] under $\theta = \theta_0$, and computing

$$\log\left(\frac{1}{m}\sum_{i=1}^{m}\exp\left\{(\theta - \theta_0)^T z(Y_i)\right\}\right). \tag{19}$$

As $m \to \infty$, approximation [19] converges to the log of [18], and so the maximum of the approximation of the log-likelihood ratio converges to the maximum of [17], that is, the maximum likelihood estimator of θ ([Kolaczyk, 2009]).

To maximize [16], after simulating the m samples from the ERGM given θ_0, we update our estimate for θ through the Fisher scoring algorithm ([Longford, 1987]), and the resulting value takes the place of θ_0 in [17] in the next iteration.

To draw a random sample $Y_1, ..., Y_m$ from the ERGM [14] given θ_0 and $\mathcal{Y}$, we start from some graph on $\mathcal{Y}$ and employ a Metropolis-Hastings algorithm: at each step t, we propose a new graph y^* by performing a toggle on the current graph y^{t-1}, that is changing the state (y_{ij}^{t-1}) of the system for a single random dyad ij. Then, the proposed change in the graph is accepted in probability: we set $y^t = y^*$ with probability

$$\min\left\{1, \frac{\mathbb{P}(Y = y^*|\theta_0)q(y^{t-1}|y^*)}{\mathbb{P}(Y = y^{t-1}|\theta_0)q(y^*|y^{t-1})}\right\}, \tag{20}$$

where $q(y_1|y_2)$ is the proposal distribution, denoting the probability of proposing y_1 given that the current state is y_2. After the burnin period, the MCMC converges and graphs are sampled from the target distribution; however, the sequence of graphs is autocorrelated, and we tackle this issue by discard a number of iterations between sample points, typically less than what was necessary for burnin.

While, in principle, the log-likelihood approximation can be made arbitrarily accurate for

any choice of θ_0 if m is large enough, in practice, if θ_0 is too distant from the MLE, the approximation in [19] will fail ([Hunter et al., 2008]). One approach for choosing θ_0 is by setting it equal to the *maximum pseudo-likelihood estimator* (MPLE) of θ. The MPLE is obtained by maximizing the pseudo-log-likelihood:

$$\sum_{i,j} \log P(Y_{ij} = 1 | \theta, Y_{-ij} = y_{-ij}). \tag{21}$$

This method was developed by [Besag, 1974] in the context of spacial analysis and could be used in place of the computationally intensive MCMC-MLE algorithm to approximate the MLE with success if the dependencies of ties are weak ([Kolaczyk, 2009], [Hunter et al., 2008]). Since we are interested in fitting more realistic social network models, the MPLE will be used just as the θ_0 parameter of the MCMC algorithm.

Once we have fitted an ERGM, we are interested in determining if the model is a good representation of our data. Indeed, while it is true that, under the MLE of θ, the ERGM network statistics $z(Y)$ are centered on $z(y^{\text{obs}})$, this could be achieved by a multimodal probability distribution with low mass on networks resembling the data. If this is the case, networks generated from the model cannot be considered good reconstructions from the egocentric sample. To evaluate goodness-of-fit of a fitted ERGM, we generate from it a series of networks as we did during estimation, and compare the distribution of network statistics, not necessarily included in $z(\cdot)$, over these simulated graphs with those in the data (see section [3.3] for our implementation).

3 Methods

3.1 Data

The empirical data we are going to use has been collected from a survey organized by Bocconi University to understand what personal and collective factors influence people's decisions regarding their health and the health of their family, and in particular about prevention of infectious diseases ([Offeddu et al., 2025]). The participants provided abundant data about their generalities and personal relationships; here we are going to focus on just a window of the most relevant features and on a subset of $n = 12122$ subjects across 4 different countries (Germany, France, Italy, UK).

For each ego, the number of alters has been collected for two non-overlapping types of social relationships: being friends and being coworkers. Less precise data has been obtained on

alter-alter ties, in the form of a categorical measure of density within coworkers and within friends. Moreover, no information has been collected on possible friend-coworker ties. For each ego, numerous features have been collected, comprising: generalities (age, sex), socio-economic status (SES) variables (like education, employment, household income), family size, and political orientation; fewer attributes have been obtained for alters.

Our main interest relies in reconstructing the global network of coworker relationships and friendships through the best possible extrapolation from egocentric data. In order to make the data more amenable to this task, it has been preprocessed through a number of steps. First, we cleaned the data by filling or, when necessary, removing incomplete records, and discarding unreliable samples (for instance, impossibly high numbers of friends). The categorical proxy for alter-alter ties has been translated into a graph structure by assigning links at random to the dyads of each ego network, according to the reported density of ties; specifically, the proxy has been mapped to a fraction of alter-alter dyads taking possible values: 0, 0.1, 0.25, 0.5, 0.75, 0.9, 1. Next, we have dealt with the fact that egos have the tendency to approximate their number of connections to multiples of 5. This leads to a spiky and irregular degree distribution in the data for both friends and coworkers counts, which may then turn out problematic in model estimation. For both relationship types, first we simplified the distribution by: (1) mapping the all alter counts greater than 2 to the nearest multiple of 5 (e.g. counts from 3 to 7 were reassigned as 5, counts 8 to 12 as 10, and so on); at this point, the frequencies of nodes of degree k, with k a multiple of 5, is actually the sum of the frequencies of nodes with self-reported degree $k-2$, $k-1$, k, $k+1$, $k+2$; (2) we split the frequencies of nodes of degree k uniformly to $k-2$, $k-1$, k, $k+1$, $k+2$. This way, we produce a stepwise shape, where the frequencies of degrees 0, 1 and 2 are unchanged, while the frequencies of subsequent degrees are constant in groups of 5. To capture more information from the initial raw distribution, we: (1) smooth the simplified frequencies by fitting a Gamma model to the stepwise degree distribution[2]; reassign the degree k for each node, with k multiple of 5, to a value from $k-2$, $k-1$, k, $k+1$, $k+2$, with probability given by the estimated Gamma. In the end, we recover a degree distribution which seemingly behaves like the original one, without the spikes due to participants' reporting behavior. Finally, we only considered alters whose age is above 18, since no minors are present among the egos; hence we aim at reconstructing networks

[2]The choice of the Gamma distribution has been deemed appropriate by inspection of the form of the empirical degree distributions, and especially since it can handle 0-degree nodes, which are present in the data.

without minors.

In the end, the data we get to work with consists of a set of undirected, unweighted egocentric networks split across 4 countries. We are going to consider a country at a time, and model separately coworkers from friends. Egos are characterized by the following attributes: sex, age group, education, income, political orientation, number of cohabitants, number of non-cohabitants family members. Alters only feature the age-group attribute.

3.2 Specifying and fitting ERGMs for network reconstruction

To fit a model that allows a good network reconstruction, we consider a number of different ERGM specifications. Informed by the exploratory analysis in the following section, we model coworkers and friends separately; also, our parameters will include effects for all attributes, and terms for differential homophily. For capturing the degree distribution and transitivity of the data, we add terms for the geometrically-weighted degree and the geometrically-weighted edgewise shared-partner, respectively. However, these statistics carry the issue of deciding how to specify their damping parameter. As the decay α tends to 0, the GWD simply reduces to the number of non-isolated nodes, whereas, as α increases, the statistics becomes less affected by high-degree vertices. Similarly, as the damping for the GWESP goes to, the statistics becomes simply the number of triads; hence, too low decay values lead to specifications increasingly similar to the Markov model and so subject to degeneracy problems. For our task, we test, for each relationship type, 8 different ERGM specification varying in the modeling of degrees and transitivity. Five of these models have the decays of GWD and GWESP fixed, respectively, at: (0.25, 0.25), (0.5, 0.5), (0.8, 0.8), (1, 0.2), (1, 0.5); two models have only one decay fixed: GWD at 1 and GWESP at 0.5; the last specification considers both parameters free and to be estimated.

Each model has been estimated through the Geyer-Thompson MCMC-MLE algorithm described in Section [2.8], runned for 50 iterations, and constrained by fixing the maximum degree equal to the maximum observed degree, which simplifies the fitting for the free decay models. For each Markov chain, we sample 2000 points, at an interval of 2000 steps, after a burning phase of 2000 points.

We also perform additional estimations without sample space constraints, with smaller values for the MCMC parameters. Fitting has been performed through the **statnet** suite of R packages for network science ([Pavel N. Krivitsky et al.,]).

3.3 Goodness of fit and global metrics

To evaluate and compare fitted ERGMs we need to measure how well they captured the property of the network data. To do so, we simulate 100 graphs from the model of interest, compute and average relevant statistics over this sample, and compare the result with the observed features of the data. However, since we fit on egocentric samples, non-local metrics like path-lengths cannot be used for this purpose; instead, our main goodness-of-fit measures involve the degree distribution and edgewise shared-partner distribution. The Kullback–Leibler divergence is used as a measure of discrepancy between the true and approximated distributions.

After selecting the best ERGM specification for coworkers and the best one for friends, these are used to reconstruct a global network for each relationship type. The approximated population networks are then compared to benchmark reconstructions obtained through a Poisson random graph on the empirical density, and a scale-free network generated by the Barabási-Albert process adding a node with $m = \langle k \rangle / 2$ new links at each step. The key features on which the reconstruction methods are contrasted are the distribution of degrees, the average distance and the distribution of path lengths, and clustering.

3.4 Dynamical processes

In the last part of our work, we test the spreading of two processes on the reconstructed networks after discarding the few nodes not present in the largest component. The first one is a SEIR model, which is typically used for modeling influenza-like illnesses (see [Keeling and Rohani, 2008], [Pastor-Satorras et al., 2015]). The SEIR framework assumes that, at all points in time, every node is assigned one of 4 possible states, or compartments, denoted "S", "E", "I", "R" standing for "susceptible", "exposed", "infectious", and "recovered", respectively. In the context of disease transmission, "S" actors are those that have not come into contact with the disease; "E" actors have been infected but the disease is still in an incubation phase, meaning they cannot spread the illness; "I" actors have the disease and can pass it to others; "R" actors are those who have recovered, or died, from the disease, and are thus removed from the system. In the SEIR model, nodes transition from one state to the other in the following fashion:

$$\mathrm{S} \to \mathrm{E} \to \mathrm{I} \to \mathrm{R}$$

The process ends when all nodes are in the "R" state, i.e. have been removed; we set the starting point of the system to be a fully susceptible population apart from one exposed node chosen at random.

The state transitions are driven by the following rules. First, every infectious node in the network can infect any susceptible node to which it is connected, making it transition to the "exposed" state; the waiting time for a transmission across any susceptible-infectious edge is distributed as an exponential random variable with rate β, which is called the *transmission rate*. A node in the "E" state transitions to the "I" state with waiting time distributed as a gamma random variable with shape parameter k_E and scale parameter θ_E. Similarly, a node in the "I" state transitions to the "R" state with waiting time distributed as a gamma random variable with shape parameter k_I and scale parameter θ_I. This implies that the average incubation period is $k_E\theta_E$ and the average infectious period is $k_I\theta_I$; we call the inverses of these periods the incubation rate σ and the recovery rate γ, respectively.

For both the friends and coworkers networks, we simulate the epidemic diffusion according to the SEIR model for 45 different combination of transmission, incubation, and recovery rates, using values aligned with empirical observations. The parameters of the Gamma distributions are chosen so as to set the desired average incubation/recovery periods, with a variance of the periods set at 0.5 for all simulations. The parameter combinations are shown in Table [1]. Simulation has been performed with the help of the **epinet** R package for analyzing epidemics ([Groendyke and Welch, 2018]).

Parameter	**Values**
Transmission Rate (β)	0.2, 0.3, 0.4, 0.5, 0.6
Incubation Rate (σ)	1/3, 1/4, 1/5
Recovery Rate (γ)	1/5, 1/6, 1/7
Total Simulations	45

Table 1: Parameter combinations used for the SEIR simulations.

The second model is utilized to describe the dynamics of opinions in a population, and is referred to as the *Friedkin-Johnsen model*, first introduced in [Friedkin and Johnsen, 1990]. This model assumes that each node i in the network has both an external opinion $p_i(t)$ and an internal opinion, which is constant and equal to the initial value of the external opinion $p_i(0)$. The opinions are modeled as continuous states, and $p_i(t) \in [0, 1]$ at all

discrete times t. Opinions evolve through time according to:

$$p_i(t+1) = \frac{(1-\beta)p_i(0) + \beta \sum_{j \in N(i)} p_j(t)}{1 - \beta + \beta k_i}, \tag{22}$$

where $\beta \in [0, 1]$ weighs the effect of social contagion, and $p(0) \in [-1, 1]$ is the vector of internal opinions/personal biases. Since our networks are unweighted, the opinion of neighbors have equal weight $1/k_i$ in influencing nodes' states.

For each considered network topology, we run the above dynamical process through 20 choices of parameter β, going from 0.05 to 1 with steps of length 0.05, and 9 choices for the distribution of internal opinions. In particular, $p(0) \sim \text{Beta}(\alpha_1, \alpha_2)$, with α_1 and α_2 taking 3 possible values, with different combinations leading to different mean and polarization of starting/internal opinions. See Table [2] for the parameter combinations. The process is stopped once the opinions states converge.

Parameter	**Values**
Influence Effect (β)	$\{0.05, 0.10, \ldots, 1.00\}$
Int. Op. Shape 1	0.1, 0.5, 1
Int. Op. Shape 2	0.1, 0.5, 1
Total Simulations	180

Table 2: Parameter combinations used for the Friedkin-Johnsen simulations.

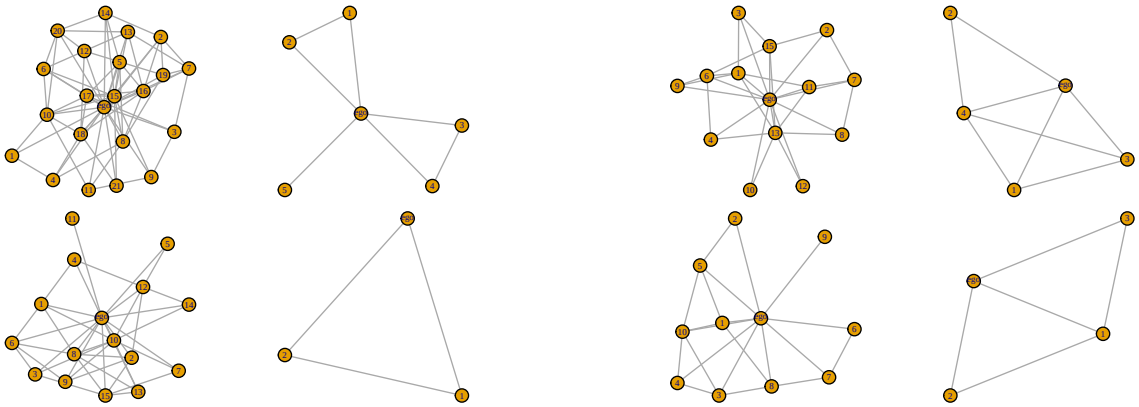
4 Results

4.1 Exploratory analysis

As a first step in our analysis, we explore the structural properties of our data. This has two main objectives: (1) provide general insight on the problem at hand and potentially inform model specification; (2) collect the properties which will act as a benchmark against which our models will be evaluated.

Our data appears as shown in Figure 1, where we plot a sample of 4 egocentric networks for the Italian portion of our data, including the ego.

In the upper panels of Figure 2, we show the degree distribution for egos across all countries, both for coworkers and friends. We see that the data corroborate the intuition of accounting separately for the 2 relationship types, as we observe different structural properties already from degree frequencies. In particular, the friends degree distribution seems more heterogenous, with a larger number of low-degree nodes and with lower mass



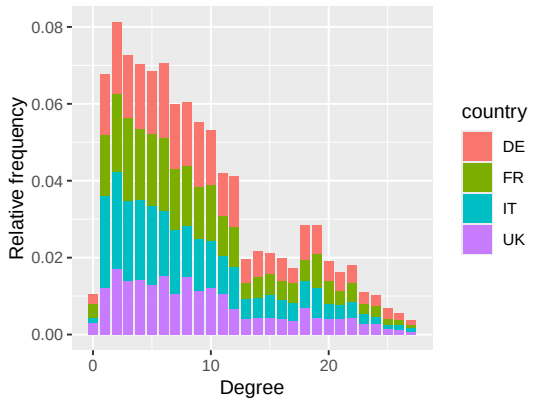
(a) Egocentric coworkers networks (b) Egocentric friends networks

Figure 1: Egocentric networks for 4 random nodes, isolating the coworkers subgraph and the friends subgraph.

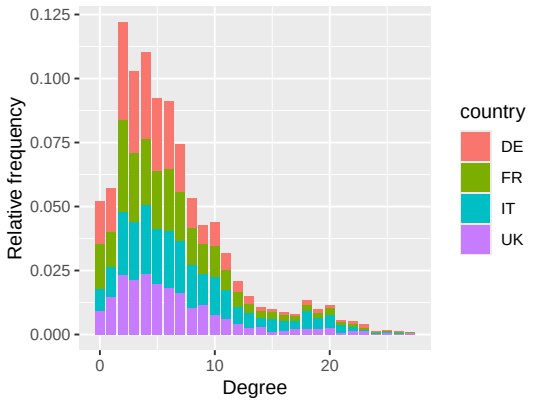
on average values. This may potentially be explained by differing structural properties due to the friends network emerging from a process subject to preferential attachment, in contrast with the coworkers network forming under different forces. This already shows the contribution of egocentric data in informing graph modeling by detecting cases where theoretical social network growth models may be inappropriate. In the lower panels, the degree distributions of Italian egos and coworkers are decomposed by the age attribute. We see that age seems to be on average non-informative on the number of friends, while there is possibly some effect going in the coworkers network. Dependencies like these are what we hope to observe again in the final reconstructed model in the greatest possible amount.

Another basic structural property which can be inspected is density. In Figure [3], the upper panels plot the distribution of edge-wise shared partners (ESP); as before, we observe uniformity across countries, but this time the relationship type seems to be less important. Still, among friends, ESP values are relatively more concentrated on the low end, indicating that in the friends network the tendency for transitivity is lower, or possibly that the decay parameter for the GWESP statistics is higher. The lower panels of the figure graph the distribution of the categorical density variable for Italian friends ego networks, decomposed by sex and by education level; the data shows a similar density across genders, whereas the education level may affect the connectedness of local networks.

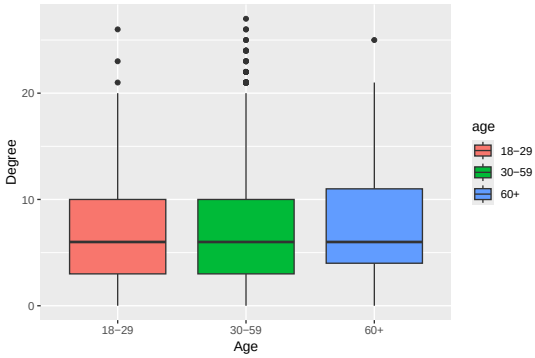
When studying social spreading processes, it is typical that age is an important factor in shaping the system's dynamics and so we are particularly interested in capturing its



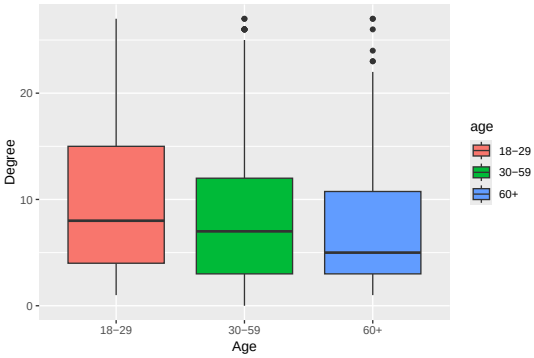
(a) Degree distribution of egos for coworkers



(b) Degree distribution of egos for friends



(c) Degree distribution of egos for Italian friends by age group

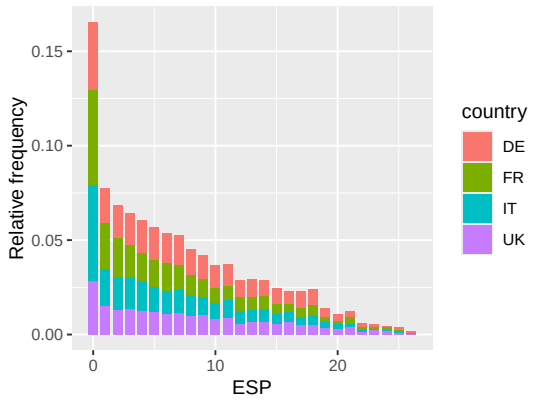


(d) Degree distribution of egos for Italian coworkers by age group

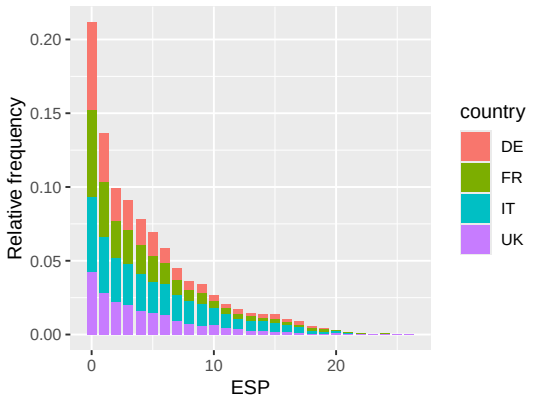
Figure 2: In panels a and b, the plots show the degree distribution across countries for the networks of coworkers and friends respectively. Panels c and d decompose by age the degree frequencies of friends and coworkers, respectively.

relationship with network structure, possibly beyond the mere effect on the average degree we observed before. To determine whether our models will achieve this result, we inspect the age attribute under the lens of the composition measures described in Section [2.4]. We start by inspecting the variability of this attribute: we plot the distribution of the Shannon entropy for the ego networks in Figure [4]. The egos are pretty uniform across countries in the diversity of age groups reported, both for coworkers and for friends. The density plots show that there is a considerable number of values with zero entropy, i.e. no variability in the age groups of alters; we see that networks which are completely homogeneous in this attribute are more common for the friends relationship. The boxplots at the bottom of each graph plot the Shannon diversity distribution for just nonzero diversity values: in this case, the average entropy is similar across relationship types and countries, while the variability of entropy is typically higher for coworkers networks.

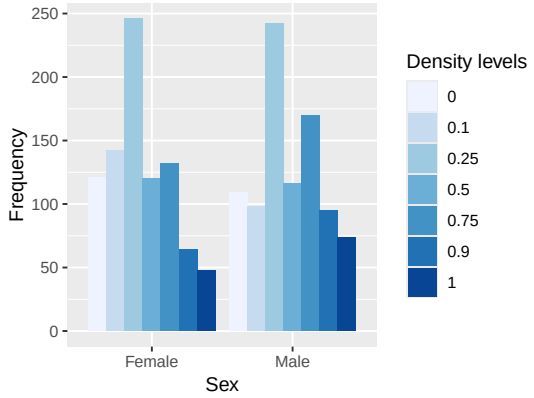
In Tables [3] and [4] we plot the mixing matrices for age-group across all countries,



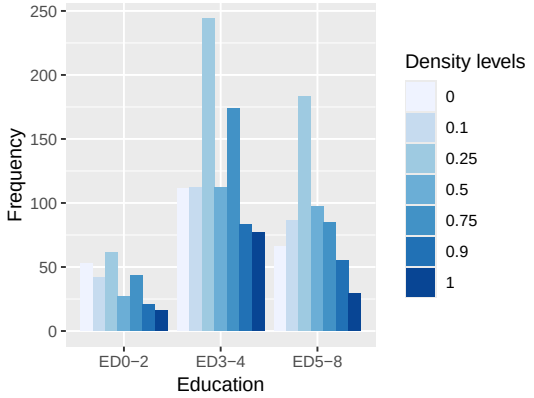
(a) ESP distribution of egos for coworkers



(b) ESP distribution of egos for friends



(c) Density distribution of egos for Italian friends by sex



(d) Density distribution of egos for Italian friends by education

Figure 3: In panels a and b, the plots show the distribution of edgewise shared partners across countries for the networks of coworkers and friends respectively. Focusing on just the Italian friends network ESP frequencies, panels c and d display the distribution of ego network densities by sex and education level respectively.

for coworkers and friendships respectively. Mixing matrices describes the number of contacts/interactions among subgroups of a population. In our case, rows indicate the age-group of egos, and columns the age-group of alters; the value in a cell is the proportion of alters of age-group corresponding to the column name, linked to egos of age-group corresponding to the row name, out of the total number of alters for each relationship-country pair. We see that the proportion of egos in each age-group is similar across countries, with the slight exception of the UK where the number of middle-aged participants is relatively lower. We also observe homophily effects at play, seemingly uniform among countries; for example, around half of the coworkers of egos aged 18-29 have age 18-29, despite this age accounting for just around 10% of egos, and this proportion becomes higher for the friendship networks. Similar effects are found also for the other two age categories.

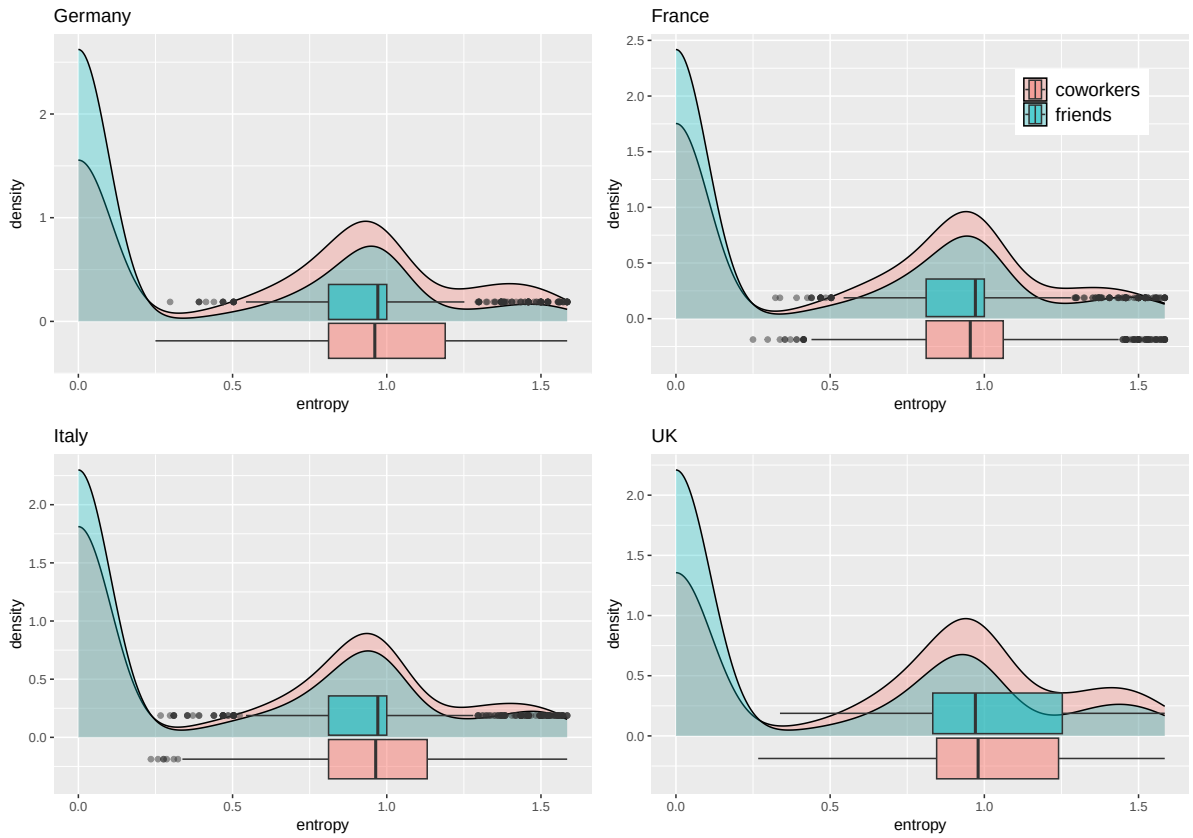


Figure 4: Distribution of Shannon entropy for age across countries and relationship types. The density plots considers all samples, whereas the boxplots exclude zero values.

To better explore age homophily/heterophily, in Figure [5], the six panels plot the distribution of the E-I index for the age attribute for Italian egos across relationship types, sex, and age groups. For both friends and coworkers, we observe a strong tendency for middle-aged individuals to be connected to other people of similar age, with homogeneity across sexes. Among older individuals, women tend to have a balanced E-I index, whereas old men have a strong tendency to connect to younger people. For egos aged 18 to 29, there is uniformity across sexes, but striking differences between coworkers and friends: the first show a less concentrated index with a prevalence for heterophily, whereas the latter exhibit a very strong pull towards age homophily, as would be expected. How much of these effects is a consequence of the prevalence of age categories, and how much it is due to preferences, remains unclear for now. Overall, we deduce we should model age homophily, also accounting for differences across age categories, while the effect of sex seems less obvious.

Table 3: Mixing matrix for Coworkers

	Germany			France		
	18-29	30-59	60+	18-29	30-59	60+
18-29	0.09	0.10	0.02	0.10	0.11	0.01
30-59	0.13	0.49	0.06	0.16	0.51	0.05
60+	0.02	0.08	0.02	0.01	0.05	0.01
	Italy			UK		
	18-29	30-59	60+	18-29	30-59	60+
18-29	0.09	0.10	0.02	0.14	0.11	0.02
30-59	0.12	0.55	0.08	0.14	0.35	0.05
60+	0.01	0.06	0.02	0.02	0.06	0.02

Mixing matrix for the age group of coworkers networks.

Table 4: Mixing matrix for Friendships

	Germany			France		
	18-29	30-59	60+	18-29	30-59	60+
18-29	0.15	0.05	0.01	0.15	0.04	0.01
30-59	0.10	0.51	0.06	0.12	0.53	0.08
60+	0.01	0.05	0.05	0.01	0.04	0.03
	Italy			UK		
	18-29	30-59	60+	18-29	30-59	60+
18-29	0.11	0.04	0.01	0.20	0.06	0.02
30-59	0.10	0.57	0.08	0.11	0.43	0.06
60+	0.01	0.05	0.04	0.01	0.05	0.05

Mixing matrix for the age group of friends networks.

4.2 Models specification and estimation

We fit a number of ERG models to both the friends network and the coworkers network, restricted to the Italian egos. Model estimation and selection for the two relationship types proceeds in an analogous manner. In light of the previous analysis, all considered specifications contain an homophily term for each individual age category; also, effects for the following attributes are included in each model: sex, age group, education level, income quantile, political opinion, number of cohabitants, number of non-cohabitant family members. Finally, the purely structural terms in the ERGMs are the number of edges, the GWD, and the GWESP terms. Models vary in the choice of the decay parameters for these last two statistics (see Section [3.2]). Fitting ERGMs can prove computationally intensive even with a small number of egos. For this reason we decide to speed up computations by first estimate ERGMs under the constraint of having maximum degree equal to the

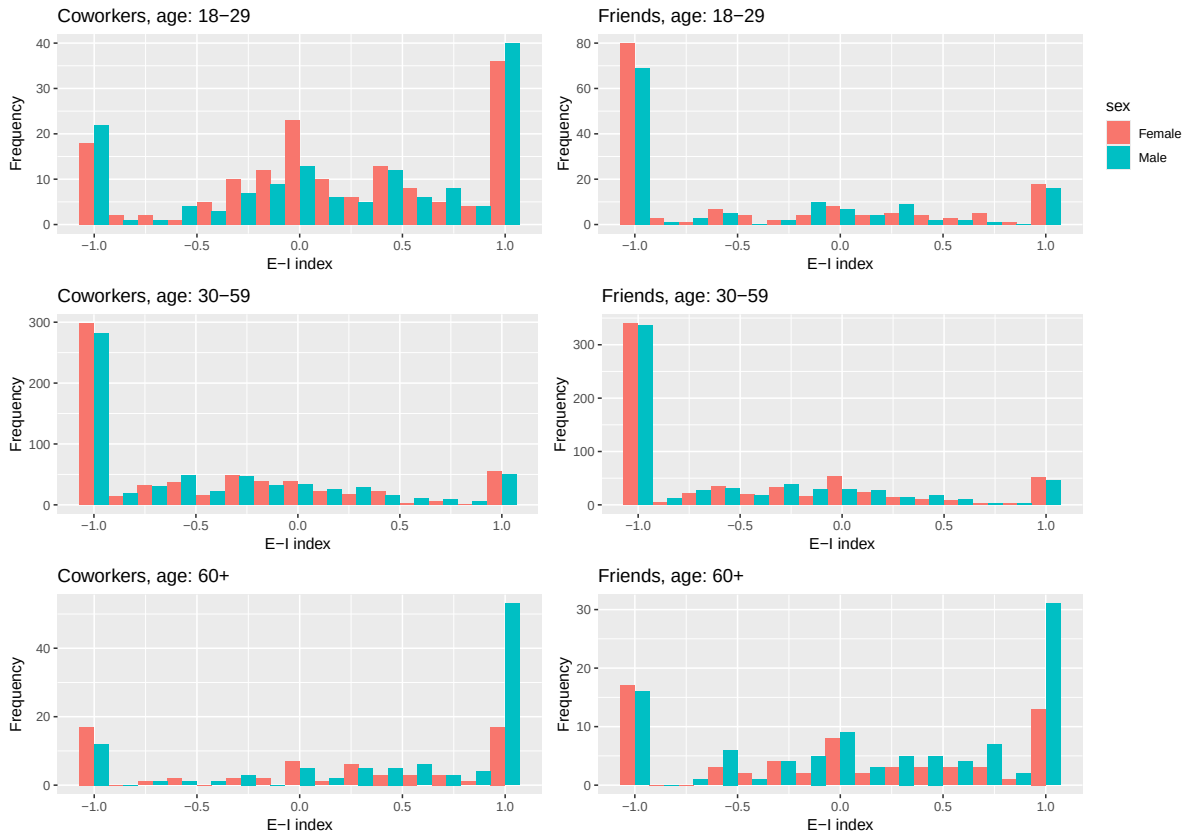


Figure 5: Distribution of the E-I index for the age attribute for Italian egos across relationship types, sex, and age groups.

observed maximum degree (which is 27 for both friends and coworkers). This restriction of the ERGM sample space allows us to evaluate more choices of decays, but produces suboptimal fits. We choose the best specifications on the basis of how much the ERGM statistics match the empirical ones, and of the KL divergence between the model and the empirical degree distributions. In a second phase, we improve the ability of the chosen models to capture transitivity by fitting them again without constraints on the sample space, and initializing the MCMC on the parameters found in the first fit.

4.2.1 Model estimation and selection: friends network

Among the eight specifications listed in Section [3.2], the friends network is better captured by the model with fixed decay 1 for GWD and fixed decay 0.8 for GWESP, since the degree distribution from networks simulated from this specification has the lowest Kullback–Leibler divergence from the empirical distribution, around 0.05.

The coefficients, their estimates and p-values are shown in Table [5][3]. The table shows that

[3]The baseline categories are: ”Female” for Gender, ”18-29 years old” for Age Group; ”Primary (ED0-2)”

age homophily effects are highly significant for each age category, and the only significant attribute is income. Statistics for education, political orientation, household and family size do not reach statistical significance.

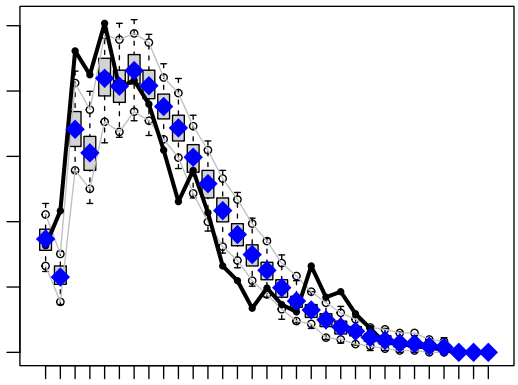
Variable	Estimate	Std. Error	p-value
Network Size Adjustment	-7.4827	0.0000	$< 1e-04^{***}$
Number of Edges	-0.5163	0.3444	0.1339
Degree Distribution (GWD, Decay = 1)	0.0581	0.2114	0.7833
Clustering (GWESP, Decay = 0.8)	0.6733	0.2976	0.0237^{*}
Male Gender Effect	-0.0276	0.0705	0.6957
Age 30-59 (vs. 18-29)	-0.1359	0.1611	0.3986
Age 60+ (vs. 18-29)	0.1575	0.2033	0.4385
Same Age Group: 18-29	1.1036	0.2889	0.0001^{***}
Same Age Group: 30-59	0.6174	0.1625	0.0001^{***}
Same Age Group: 60+	0.8319	0.2524	0.0010^{**}
Education: Secondary (ED3-4)	0.0876	0.0929	0.3459
Education: Higher (ED5-8)	0.0425	0.0886	0.6316
Income: Quintile 1 (Lowest)	-0.1767	0.1244	0.1553
Income: Quintile 2	-0.1341	0.0801	$0.0943^{\cdot}$
Income: Quintile 3	-0.1093	0.1201	0.3631
Income: Quintile 4	-0.0659	0.0950	0.4878
Income: Quintile 5 (Highest)	0.0084	0.1199	0.9440
Political Affiliation: EPP	-0.0090	0.1417	0.9496
Political Affiliation: GREENSEFA	-0.0479	0.1554	0.7578
Political Affiliation: Prefer not to answer	0.0008	0.0781	0.9915
Political Affiliation: NI	0.0560	0.3420	0.8700
Political Affiliation: Other	-0.1460	0.1960	0.4562
Political Affiliation: PfE	-0.0117	0.1181	0.9210
Political Affiliation: Renew Europe	-0.0748	0.1269	0.5558
Political Affiliation: S&D	-0.0097	0.0858	0.9099
Political Affiliation: The Left	-0.0293	0.0803	0.7152
Household Size (Cohabitants)	-0.0161	0.0210	0.4412
Non-Cohabitant Social Ties	-0.0008	0.0144	0.9542

Table 5: Monte Carlo Maximum Likelihood Results for the best fit on the Italian friends network under sample space constraints. Significance codes: '***' < 0.001, '**' < 0.01, '*' < 0.05, '·' < 0.1.

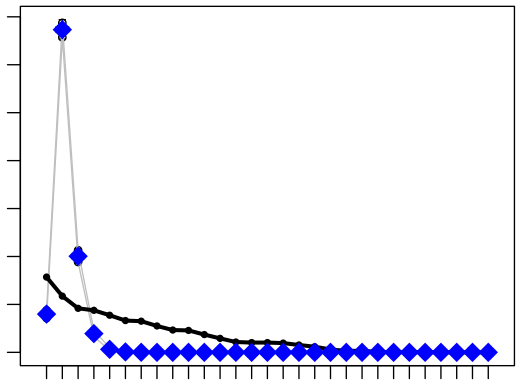
In the left panels of Figure [6], we compare the observed degree distribution with the one obtained by 100 samples from our chosen ERGM model; we see that indeed the two are very similar, by which we can confidently say that this feature is well captured by our model. However, also due to imposing the sample space constraint, transitivity is badly captured, as shown in the panels on the right: while the table of coefficients [5] shows a significant positive effect of GWESP, indicating higher transitivity, this effect comes short

for Education; "Prefer not to answer" for Income; "ECR" for Political Affiliation.

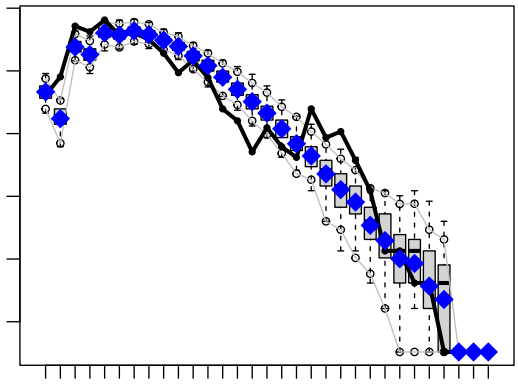
of reproducing the real extent of transitivity in the network.



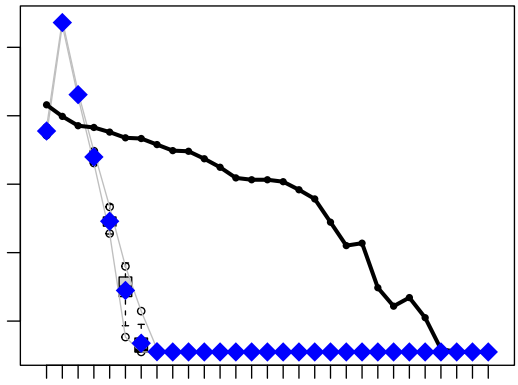
(a) Estimated and empirical degree distribution.



(b) Estimated and empirical ESP distribution.



(c) Log-odds of degrees.



(d) Log-odds of ESPs.

Figure 6: In panels a and b, the plots show the distribution of degrees and of edgewise shared partners for the network of Italian friends, respectively. The boxplots summarise the distribution of the degree/ESP frequencies in the simulated networks, with the blue sign indicating the average. The solid black line represent the empirical distribution. In panels b and c, we plot the lod-odds of degrees and ESPs, respectively, instead of probabilities.

To try getting the model to better approximate the ESP distribution, we fit the same ERGM, initialized with the parameter estimates of Table [5]. We run the algorithm for other 5 iterations without constraints. The resulting model improves over the base one in its capturing of the network, as the KL-divergence of the ESP distribution decreases from 3.52 to 2.74. Still, despite the improvement, we were not able to found a non-degenerate ERG model fully capable of representing the transitivity pattern in the data.

4.2.2 Model estimation and selection: coworkers network

Among the same eight specifications of Section [3.2], the coworkers network is better captured by the model with fixed decay 0.8 for GWD and fixed decay 0.8 for GWESP,

with Kullback–Leibler divergence from the empirical distribution around 0.16.

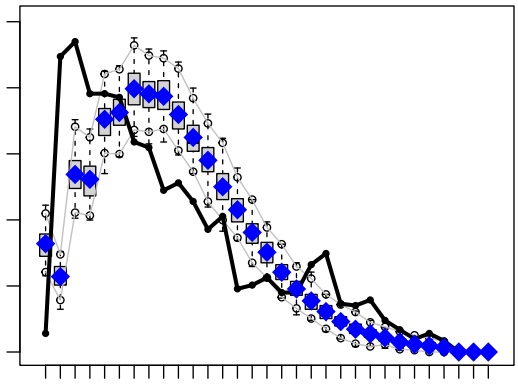
The coefficients, their estimates and p-values are shown in Table [6]. We see that also for coworkers, differential age-homophily effects are positive and significant. While we found statistical significance for income in the friends data, here the only significant attribute, apart from age, is education. Thus, according to these ERG models, we find evidence for the income level of a node affecting its number of friends, and the education level affecting its number of coworkers.

Variable	Estimate	Std. Error	p-value
Network Size Adjustment	-7.4827	0.0000	$< 1e-04^{***}$
Number of Edges	-0.4011	0.2838	0.1576
Degree Distribution (GWD, Decay = 0.8)	-0.2402	0.2003	0.2304
Clustering (GWESP, Decay = 0.8)	0.4303	0.2683	0.1088
Male Gender Effect	0.0523	0.0345	0.1300
Age 30-59 (vs. 18-29)	0.0880	0.1212	0.4678
Age 60+ (vs. 18-29)	0.1431	0.0800	$0.0737^{\cdot}$
Same Age Group: 18-29	0.9610	0.1886	$< 1e-04^{***}$
Same Age Group: 30-59	0.1104	0.1276	0.3872
Same Age Group: 60+	0.4321	0.2493	$0.0830^{(\cdot)}$
Education: Secondary (ED3-4)	0.0993	0.0442	0.0246^{*}
Education: Higher (ED5-8)	0.0796	0.0468	$0.0888^{(\cdot)}$
Income: Quintile 1 (Lowest)	-0.0895	0.0800	0.2634
Income: Quintile 2	-0.0733	0.0658	0.2651
Income: Quintile 3	-0.0869	0.0636	0.1721
Income: Quintile 4	-0.0520	0.0611	0.3943
Income: Quintile 5 (Highest)	-0.0115	0.0788	0.8836
Political Affiliation: EPP	-0.0446	0.0651	0.4936
Political Affiliation: GREENSEFA	-0.0481	0.1000	0.6306
Political Affiliation: Prefer not to answer	0.0245	0.0529	0.6440
Political Affiliation: NI	-0.1367	0.2253	0.5439
Political Affiliation: Other	-0.0774	0.1157	0.5037
Political Affiliation: PfE	0.0610	0.0756	0.4203
Political Affiliation: Renew Europe	-0.0955	0.0984	0.3318
Political Affiliation: S&D	0.0442	0.0551	0.4218
Political Affiliation: The Left	-0.0148	0.0515	0.7741
Household Size (Cohabitants)	-0.0039	0.0141	0.7833
Non-Cohabitant Social Ties	-0.0070	0.0081	0.3864

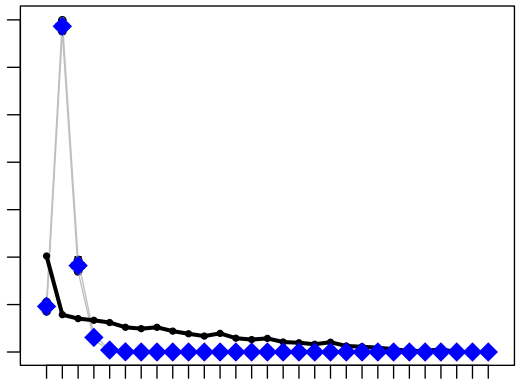
Table 6: Monte Carlo Maximum Likelihood Results for the best fit on the Italian coworkers network under sample space constraints. Significance codes: '***' < 0.001, '**' < 0.01, '*' < 0.05, '·' < 0.1.

As before, in the left panels of Figure [7], we compare the observed degree distribution with the one obtained by 100 samples from our chosen ERGM model; the two are somewhat

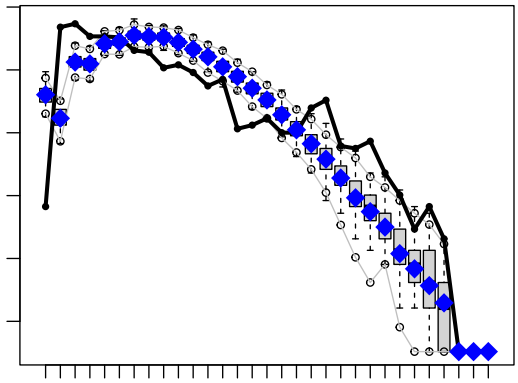
similar, but the distribution is captured less compared to the friends-network case. Again, while we find an almost-significant positive coefficient for GWESP, the right panels of Figure [7] display that the empirical network exhibits more transitivity than what is captured by the ERGM.



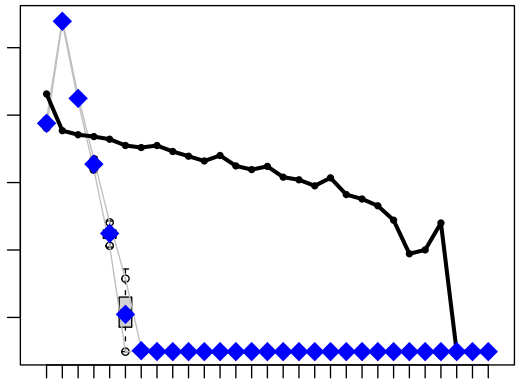
(a) Estimated and empirical degree distribution.



(b) Estimated and empirical ESP distribution.



(c) Log-odds of degrees.



(d) Log-odds of ESPs.

Figure 7: In panels a and b, the plots show the distribution of degrees and of edgewise shared partners for the network of Italian coworkers, respectively. The boxplots summarise the distribution of the degree/ESP frequencies in the simulated networks, with the blue sign indicating the average. The solid black line represent the empirical distribution. In panels b and c, we plot the lod-odds of degrees and ESPs, respectively, instead of probabilities.

To try getting the model to better approximate the ESP distribution, we fit the same ERGM, initialized with the parameter estimates of Table [6]. We run the algorithm for other 5 iterations without constraints. The resulting model improves over the base one in its capturing of the network, as the KL-divergence of the ESP distribution decreases from 4.17 to 3.89. Still, despite the improvement, we were not able to found a non-degenerate ERG model fully capable of representing the transitivity pattern in the data.

4.3 Models evaluation

In this section, we compare the networks reconstructed from the best ERGM models with two alternative methods: the Erdős–Rényi random graph and the Barabási-Albert model. The comparison involves evaluating both the global measures of the graphs and the evolution of a spreading process on these networks. For each network topology, 10 graphs of 10000 nodes are simulated, and average measures are analyzed.

4.3.1 Evaluation and comparison to benchmarks models: metrics

In Table [7] we show the main metrics for each network topology, obtained by averaging over the simulations. The ERGM-reconstructed coworkers network has an average degree of 8.45, very close to the ER value, and a degree standard deviation of 6.11, less than the scale-free deviation and more than the ER one . This, together with the distinct shape of the degree distribution as shown in the upper-left panel of Figure [8], already puts the ERGM network outside the reconstruction possibilities of the benchmark methods. In particular, we observe that the networks generated by the alternative models have drastically less very-low-degree nodes, whereas the ERGM is capable of capturing the presence of these actors, producing indeed a distribution more in line with the data. Similar results are obtained for the friends networks, shown in Table [8] and Figure [9].

The average distance of the coworkers ERGM network is 4.77, above that of the ER and the BA models, but similar enough to suggest that also the ERGM-graph exhibits the small-world effect. In the bottom-left panel of Figure [8], for each topology we plot the histogram of the average distance of nodes from all other nodes. We see indeed that the ERGM distances have the same magnitude of the simpler methods, which already are capable of generating realistic path lengths. However, the ERGM distance distribution is distinct, being more similar to the ER graph in the average, but more similar to the BA graph in variability, which may be possibly an important aspect in the approximation of the population graph, although it cannot be evaluated since distance information is not provided in an egocentric sampling design. Again, the friends network behaves similarly, and we can see in Figure [9] that the distribution of distances has a shape similar to the one emerging from the BA process, but scaled upwards.

The "Clustering" rows of Tables [7] and [8] reports the global and average clustering coefficients for the networks, with the ERGM being the only method exhibiting non-vanishing

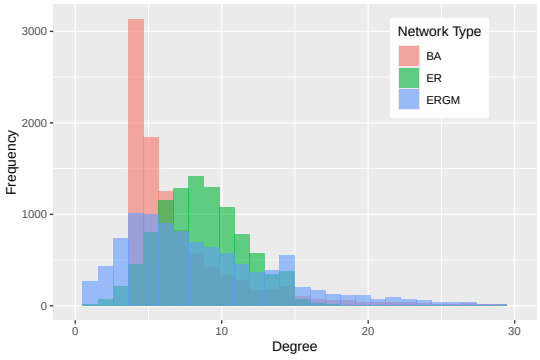
transitivity. This is particularly important since low transitivity is the among the most important unrealistic aspects of the ER and BA networks. Also, this good modeling result is obtained although the computational difficulties of fitting the model with the GWESP term have led to abandon the full capturing of the ESP distribution. In the upper-right panel of Figure [8], the difference from other methods is made even more evident by plotting the distribution of local clustering coefficients of the nodes. While the ERGM-network transitivities are spread between 0 and 1, the ER and BA networks have very similar distribution concentrated at 0.

The last six rows of Tables [7] and [8] display three centrality measures, i.e. closeness, betweenness, and eigenvalue centrality, through their average across nodes and their graph centralization measure. As for path lengths, the accuracy of these values cannot be assessed as they cannot be extracted by egocentric networks, but, still, finding them different between the ERGM-based graph and the benchmarks would provide further confirmation of the usefulness of adopting a flexible data-driven method for reconstruction. This is indeed the case for both relationship types: on the one hand, the centralization closeness and betweenness measures are intermediate for the ERGM-generated topology, meaning the closeness and betweenness individual indices have a smaller range then that of the BA network, and a greater range than that of the ER network; on the other hand, the ERGM graph exhibits a higher average betweenness centrality, suggesting there are more nodes acting as ”bridges” in the network, and a lower average closeness centrality, implying that traveling from a node to another in the ERGM network takes on average slightly longer when compared to the benchmarks.

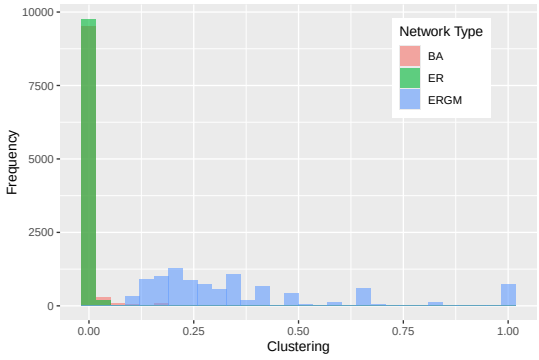
Finally, in the lower-right panel of Figure [8], we plot the degree distribution of the ERGM coworkers network decomposed by our most relevant attribute, the age group. The distribution obtained from the data is shown as black boxplots, whereas the reconstructed distribution as colored boxplots. We observe that the network indeed decently captures the age-related effect present in the data. For the friends data, Figure [9] shows the degree distribution decomposed by the income attribute, which our previous analysis has shown to be potentially informative. Again, we see that the ERGM network decently captures the degree dependence on this feature.

Metric	ER	BA	ERGM
Average Degree	8.38	8.00	8.45
Degree SD	2.90	9.14	6.11
Average Distance	4.57	4.02	4.70
Distance SD	0.19	0.29	0.39
Global Clustering	< 0.01	< 0.01	0.20
Average Clustering	< 0.01	< 0.01	0.36
Average Closeness Centrality	2.19e-05	2.50e-05	2.17e-05
Closeness Centralization	0.06	0.27	0.13
Average Betweenness Centrality	1.78e-04	1.51e-04	1.83e-04
Betweenness Centralization	0.002	0.08	0.01
Average Eigenvalue Centrality	0.28	0.13	0.14
Eigenvalue Centralization	0.72	0.99	0.99

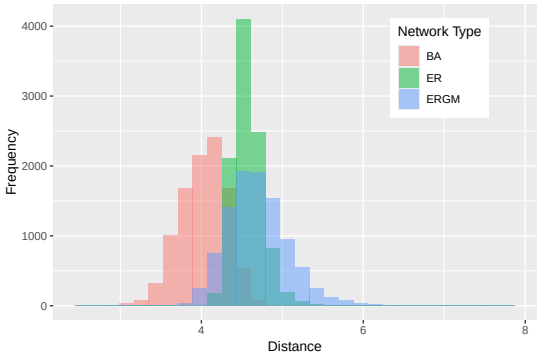
Table 7: Table of metrics of reconstructed coworkers networks.



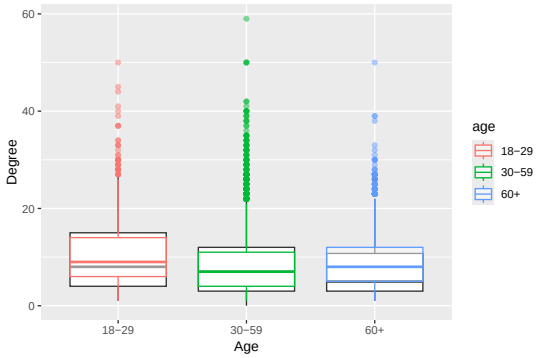
(a) Degree distribution by network topology.



(b) Local clustering coefficient distribution by network topology.



(c) Distance distribution by network topology.



(d) Degree distribution by age for the data and the ERGM.

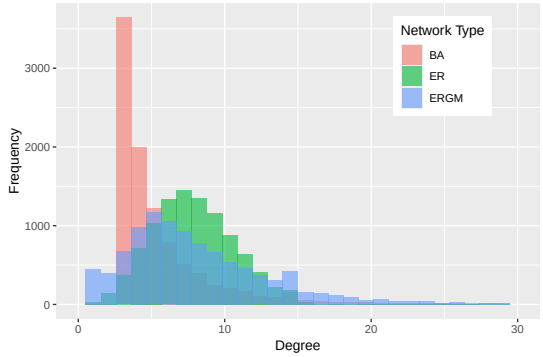
Figure 8: Panels [a], [b], and [c] show, for the reconstructed coworkers graphs, the distribution of degrees, clustering coefficients, and distances, respectively. In panel [d], the black boxplots represent the degree distribution of the data split by age; the colored boxplots plot the degree frequencies of the ERGM-reconstructed network.

4.3.2 Evaluation and comparison to benchmarks models: process simulation

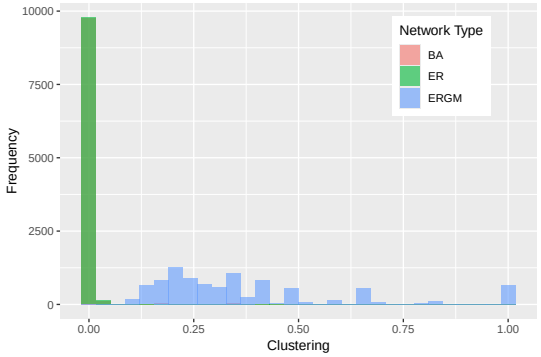
We conclude the Results section with the implementation of two simple spreading processes on the ERGM-based networks and the two theoretical benchmarks. In Figure [10], we plot

Metric	ER	BA	ERGM
Average Degree	7.61	6.00	7.72
Degree SD	2.77	6.86	5.55
Average Distance	4.77	4.48	4.92
Distance SD	0.22	0.36	0.41
Global Clustering	< 0.01	< 0.01	0.21
Average Clustering	< 0.01	< 0.01	0.37
Average Closeness Centrality	2.19e-05	2.50e-05	2.17e-05
Closeness Centralization	0.05	0.26	0.12
Average Betweenness Centrality	1.88e-04	1.74e-04	1.90e-04
Betweenness Centralization	0.002	0.12	0.01
Average Eigenvalue Centrality	0.32	0.01	0.13
Eigenvalue Centralization	0.68	0.99	0.99

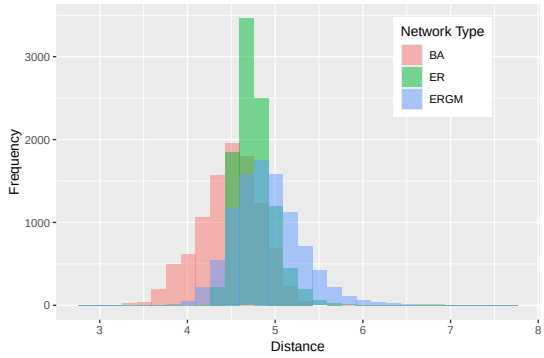
Table 8: Table of metrics of reconstructed friends networks.



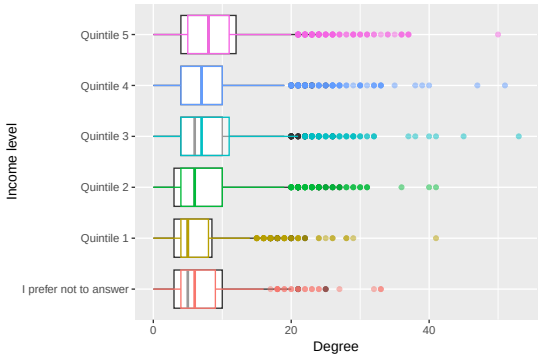
(a) Degree distribution by network topology.



(b) Local clustering coefficient distribution by network topology.



(c) Distance distribution by network topology.



(d) Degree distribution by income level for the data and the ERGM.

Figure 9: Panels [a], [b], and [c] show, for the reconstructed friends graphs, the distribution of degrees, clustering coefficients, and distances, respectively. In panel [d], the black boxplots represent the degree distribution of the data splitted by age; the colored boxplots plot the degree frequencies of the ERGM-reconstructed network.

the evolution through time of the SEIR compartments for the simulations with transmission rate $\beta = 0.4$ on the coworkers networks (the other transmission rate values provide very

similar results, and the same is true for friendship networks): each row corresponds to a disease state, and each column corresponds to a combination of incubation period and recovery period, displayed above the figure; in each plot, the evolution of the compartment is divided by network topology. What we observe invariably in all considered parameter combinations is that the ERGM-based net exhibits a lower epidemic peak with respect to the other graphs, meaning that the maximum number of infectious people reached during the epidemic is smaller. Since for all networks the final number of susceptible individuals is zero, this means that the ERGM-network topology slows down the spread of the disease across our chosen range of SEIR coefficients.

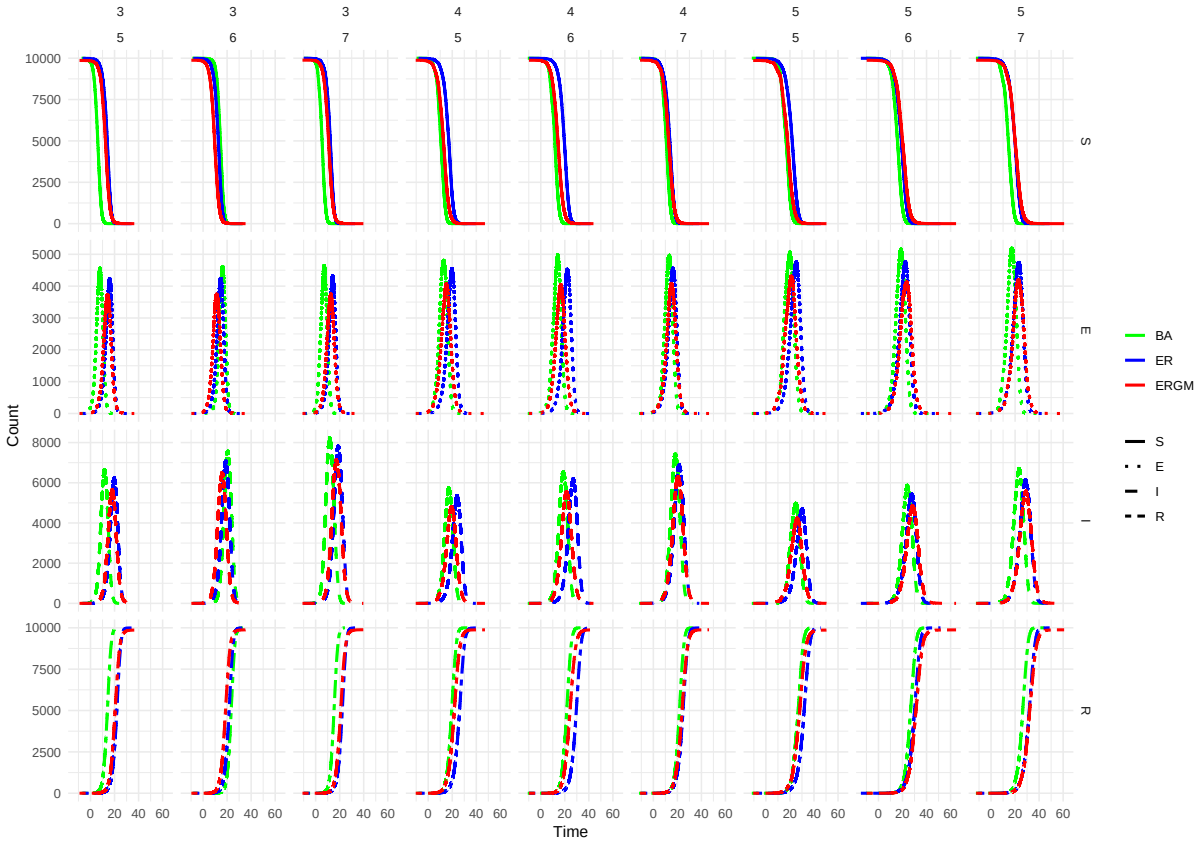


Figure 10: Evolution of the SEIR compartment for the simulations with transmission rate 0.4 on the coworkers networks. The rows and linestyle represent the disease state. The columns represent a combination of incubation period and recovery period, displayed above the figure. The colors refer to the three networks, with blue, green and red referring to the ER, BA and ERGM networks, respectively.

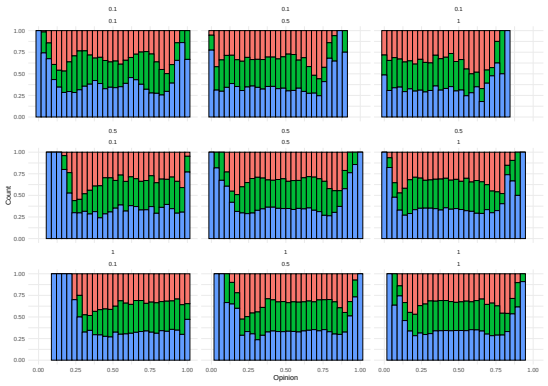
As for the Friedkin-Johnsen model, we found very similar results in the mean and variance of final opinions across the network topologies. However, there were, even for this very simple system, some differences in the dynamics on the three network topologies at almost all parameter combinations. In Figure [??], in the upper-left panel the x-axis represents

the opinions values between 0 and 1, and is split up into bins; through the colors, the graph shows what percentage all final opinions in each particular bin belongs to each of the three network topology cases. All the simulations considered in the plot have an influence parameter of 0.1; the rows and columns correspond to different combinations of shapes of the Beta distribution of initial opinions, displayed above each facet. The upper-right and lower-left panels are the same type of graph, but restricted to simulations with influence parameter of 0.5 and 0.9, respectively. Finally, the bottom-left panel displays the difference the range of final opinions, meaning the difference between their maximum and minimum, across all parameters combination, with the x-axis representing the influence effect, and the facets corresponding to the Beta distribution shape. These figures refer to coworkers networks, with outcomes for friendships being highly similar.

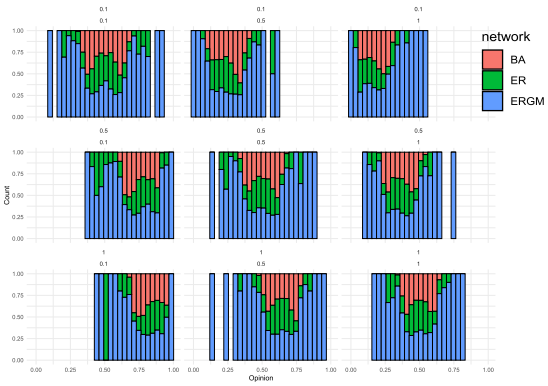
While we stated that most of the final opinions distribution is similar across network types, Figure [**??**] shows that the same is not true for the entire distribution. Indeed, the ERGM-based network tends to maintain some final opinions with more extreme values with respect to other topologies, and indeed this is shown in the bottom-left panel, where the range of the ERGM-network is always above that of both benchmarks.

5 Discussion

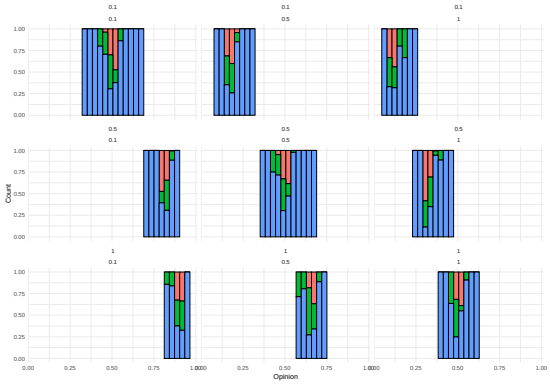
The aim of this study was to implement and assess ERGM-based methods for network reconstruction from empirical egocentric data, highlighting potential discrepancies with benchmark theoretical models regarding graph metrics and dynamical processes evolution. An exploratory analysis of the data has been conducted to inform model specification and model evaluation; due to the similarities in the features outlined by the data exploration, we opted for fitting model solely on the Italian egos. The analysis revealed a series of structural differences between friends and coworkers networks, among which a faster decay in the friendship network degree frequencies; this is also reflected in the best-found ERGM fits, which indeed exhibit, for the friendship network, a higher decay parameter for the weighted degrees, which would then lead to less high-degree nodes. As already stated, this suggests a difference in the underlying process generating the (global) graphs. For instance, a friendship network may emerge under the force of preferential attachment, which, as [Barabási and Albert, 1999] has shown, by itself can lead to a scale-free distribution more similar to the friends one, with a few hubs and many low degree nodes. In contrast, the coworking relationships are intuitively less likely to emerge from a similar process, and so a less heterogeneous degree distribution is what we would expect. Subsequently,



(a) Composition of final opinion values by network type, for simulations with influence parameter 0.1.



(b) Composition of final opinion values by network type, for simulations with influence parameter 0.5.



(c) Composition of final opinion values by network type, for simulations with influence parameter 0.9.

(d) Trend of the range of final opinions across parameter combinations.

Figure 11: In panels a, b and c, the plots are restricted to simulations with influence parameter 0.1, 0.5, and 0.9 respectively. Each facet corresponds to a combination of shapes of the Beta distribution of initial opinions, displayed above the facet. The x-axis represent the values of final opinions, and is split into bins. For each bin, the colors show the proportion of final opinions in that bin that belongs to each network topology. Panel d displays the difference between the maximum and minimum final opinions across all parameter combinations; its x-axis is the influence parameter β, while the facets correspond to the combinations of shapes of the distribution of initial opinions.

the distribution of degrees and local densities has been decomposed by several network attributes. Among these, some features did not display a clear effect on network structure, which has been further confirmed by the non-significance of ERGM coefficients; for instance, the sex attribute seems to be in general unrelated to the degree and local clustering of nodes. Other characteristics seemed to show an effect for one relationship type, and not for another; for instance, the number of coworkers seems to decrease as age increases, whereas a significant relationship cannot be observed for friends. Other attributes have displayed a correlation with degrees for both relationship types, but not necessarily for

the same reason. For instance, we find the number of friends and coworkers to be related to the income and education attributes, but, when fitting ERGMs, we observe that in the friends network only income is significant, and the same happens for the education attribute in the coworkers network. This suggests, on the one hand, that the effects of education on friends number and of income on coworker number can be explained by the other attributes; on the other hand, the number of friendships depends on the income level of the actor, as it is lower for people in low-income quantiles, and the number of coworkers depends on the education level, as it is lower for actors with the least formal education. Particular attention has been devoted to the age attribute as it is typically of great interest when interestend in reconstructing a network for age-dependent processes. Being the only feature obtained for alters, we could compute composition measures, which unveiled an overall general tendency for homophily. This was corroborated by the positive, and almost always significant, ERGM coefficients for the homophily of each age group, with the group of 18-29 years-old actors displaying the strongest tendency to make friends and coworkers of similar age.

The estimation of egocentric ERGMs can prove computationally heavy. For this reason, to fit models with different choices of GWD and GWESP decay parameters, we have restricted the sample space by putting a ceiling on the maximum degree. We have chosen the specifications minimizing the Kullback-Leibler divergence of the model degree distribution, obtained by averaging 100 simulations, from the empirical distribution; the obtained decays for GWD and GWESP were, respectively, $(1, 0.8)$ for friends, and $(0.8, 0.8)$ for coworkers. The selected models were then run for additional iterations without constraints, to improve the capturing of transitivity patterns. In the end, we have observed that the found models were indeed able to properly reproduce the degree distribution and the observed attribute effects found in the data, while they were less accurate in matching transitivity. We then used the fitted models for the networks of friends and coworkers to generate global networks of 10000 nodes, and compared the key measures of these graphs to those obtained by alternative networks generated either through the BA model, or as Poisson random graphs. We found relevant structural differences from the benchmark methods, which shows how much egocentric data is capable of potentially improving and better informing network reconstruction and systems simulation. In the context of our empirical data, we have observed a degree distribution which, while not differing in its average, stands out for an intermediate standard deviation, a considerably larger mass on very-low-degree

nodes, and less medium-degree actors with respect to the benchmarks. Despite the estimation challenges in capturing transitivity, the ERGM-reconstructed networks also exhibit considerable, non-vanishing clustering, which is instead not present in ER and BA networks, while typical in empirical social data; this structural difference is even more evident when inspecting the distribution of local clustering coefficients. The ability of reconstructing networks with various structural and compositional metrics, exploiting and approximating known empirical information, makes the use of egocentric data and ERGMs a valuable way for researchers to derive global network measures and simulate dynamical processes in a context where data is costly, scarce or incomplete. In our analysis, we observe that the ERGM-based method proves flexible in reproducing patterns of degrees, transitivity and attribute effects, without losing the small-world property of random networks, as made evident by the low average distance. Still, path lengths constitute another point of contrast from the theoretical models, as they are on average slightly greater and possess a higher variability. Moreover, we have observed differences in the average of some relevant centrality metrics, as well as in their corresponding centralization measures. Even though, as already stated, this element of the reconstructed graphs cannot be evaluated against the data, it still shows how much more flexible and informative a data-informed generated network can be, as the distribution of centrality measures is relevant in many types of spreading processes.

Finally, we simulated two simple dynamical processes on the three network topologies. We observed that even for these basic cases, the ERGM-based network exhibits differences in the evolution of systems' dynamics. In particular, the SEIR simulations consistently display a slower epidemic, with a lower epidemic peak for the ERGM graph, while the Friedkin-Johnsen simulations show that some more extreme opinions survive in the data-driven network, in contrast to the benchmarks whose final opinion homogeneity is much stronger. Interestingly, it is possible that, among all the possible reasons for these findings, these two results share, at least in part, a common explanation. Possibly, both slower epidemic spread and higher opinion heterogeneity may be due to some community structure in the graph, which we know is indeed present as a consequence of significant age-homophily effects.

6 Conclusion

In light of the findings of this study, we confirm the potential of ego-centered data, even with categorical proxies for within-alter ties, to inform the analysis of dynamical systems

on graph topologies. The chosen ERGM-based global network reconstruction method proved successful, in that the resulting network captured well the main features of the data, and showed distinctive patterns also in several global metrics, as well as in the dynamics of some spreading processes. Moreover, these results were obtained despite simplifications in the fitting of ERGMs, further highlighting the value of adopting a flexible, data-informed method for network reconstruction even in the presence of estimation challenges.

However, fitting network models on egocentric data is still an underexplored area in network science: no unified statistical framework exists, and, due to the computational hurdles in estimation, the few studies conducting ego-based graph reconstructions consider at most thousands of egos. Future research should then focus on developing new models for the task, as well as on improving the efficiency in ERGM fitting, possibly through the focus on improved Markov Chain Monte Carlo methods.

Advancing the integration of social systems and network processes requires empirical studies to better investigate social structures, their properties and functionalities. The potential of efficiently gathered ego-network data offers a promising avenue for inferring population-level structures, making it possible to more accurately describe complex social systems. This analysis aims to bridge the gap between network scientists, social scientists and public policy practitioners by highlighting the benefits of incorporating complex social relationships into their work. We compare various ego-based network reconstruction methodologies and advocate for their use to enhance the precision of modeling and simulations.

References

[Ball et al., 2012] Ball, F., Britton, T., and Sirl, D. (2012). A network with tunable clustering, degree correlation and degree distribution, and an epidemic thereon.

[Ball et al., 2010] Ball, F., Sirl, D., and Trapman, J. (2010). Analysis of a stochastic sir epidemic on a random network incorporating household structure. *Mathematical Biosciences*, 224(2):53–73.

[Barabási, 2016] Barabási, A. (2016). *Network Science*. Cambridge University Press.

[Barabási and Albert, 1999] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[Besag, 1974] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):pp. 192–236.

[Britton et al., 2007] Britton, T., Deijfen, M., Lagerås, A. N., and Lindholm, M. (2007). Epidemics on random graphs with tunable clustering.

[Britton et al., 2011] Britton, T., Deijfen, M., and Liljeros, F. (2011). A weighted configuration model and inhomogeneous epidemics. *Journal of Statistical Physics*, 145(5):1368–1384.

[Chung and Lu, 2002] Chung, F. and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences of the United States of America*, 99(25):15879–15882.

[DuBois et al., 2013] DuBois, C., Butts, C. T., and Smyth, P. (2013). Stochastic blockmodeling of relational event dynamics. In *International Conference on Artificial Intelligence and Statistics*.

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

[Eubank et al., 2004] Eubank, S., Guclu, H., Kumar, S., Marathe, M., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modeling disease outbreaks in realistic urban social networks. *Nature*, 429:180–4.

[Frank and Strauss, 1986] Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.

[Freeman, 1979] Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239.

[Friedkin and Johnsen, 1990] Friedkin, N. and Johnsen, E. (1990). Social influence and opinions. *Journal of Mathematical Sociology*, 15:193–206.

[Geyer and Thompson, 1992] Geyer, C. J. and Thompson, E. A. (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):657–683.

[Groendyke and Welch, 2018] Groendyke, C. and Welch, D. (2018). epinet: An r package to analyze epidemics spread across contact networks. *Journal of Statistical Software*, 83(11):1–22.

[Handcock, 2003] Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks. *Center for Statistics and the Social Sciences*.

[Handcock and Gile, 2010] Handcock, M. S. and Gile, K. J. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1).

[Hunter et al., 2008] Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29.

[Jackson, 2010] Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

[Karrer and Newman, 2011] Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1).

[Keeling and Rohani, 2008] Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

[Kolaczyk, 2009] Kolaczyk, E. (2009). Statistical analysis of network data: Methods and models. *Springer Series In Statistics*.

[Krivitsky, 2012] Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic Journal of Statistics*, 6(none):1100 – 1128.

[Krivitsky and Morris, 2017] Krivitsky, P. N. and Morris, M. (2017). Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in hiv prevalence in the us. *The Annals of Applied Statistics*, 11(1):427 – 455.

[Lee and Wilkinson, 2019] Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1).

[Longford, 1987] Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4):817–827.

[Longini et al., 2005] Longini, I., Nizam, A., Xu, S., Ungchusak, K., Hanshaoworakul, W., Cummings, D., and Halloran, M. (2005). Containing pandemic influenza at the source. *Science (New York, N.Y.)*, 309:1083–7.

[Lusher et al., 2012] Lusher, D., Koskinen, J. H., Robins, G., and Granovetter, M. S. (2012). *Exponential random graph models for social networks: theories, methods and applications*.

[Newman, 2018] Newman, M. (2018). *Networks*. Oxford university press.

[Offeddu et al., 2025] Offeddu, V., Colosi, E., Leone, L., Lucchini, L., Balsamo, D., Bonacina, F., Chiavenna, C., Colizza, V., Karsai, M., Koltai, J., Moreno, Y., Zagheni, E., Cucciniello, M., and Melegaro, A. (2025). Epidemic modelling with human behaviour: An interdisciplinary framework for the collection of empirical behavioural data across countries. In preparation; Preprint available at Open Science Framework.

[Pastor-Satorras et al., 2015] Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979.

[Pavel N. Krivitsky et al.,] Pavel N. Krivitsky, M. S. H., Hunter, D. R., Butts, C. T., Bojanowski, M., Klumb, C., Goodreau, S. M., and Morris, M. Statnet: Tools for the statistical modeling of network data.

[Perry et al., 2018] Perry, B. L., Pescosolido, B. A., and Borgatti, S. P. (2018). *Egocentric Network Analysis: Foundations, Methods, and Models*. Cambridge University Press.

[Rubin, 1976] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

[Smith, 2012] Smith, J. A. (2012). Macrostructure from microstructure: Generating whole systems from ego networks. *Sociological Methodology*, 42(1):155–205. PMID: 25339783.

[Snijders, 2011] Snijders, T. (2011). *Statistical Models for Social Networks*, pages 131–153. Annual Review of Sociology. Annual Reviews.

[Snijders and Nowicki, 1997] Snijders, T. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14:75–100.

[Snijders et al., 2006] Snijders, T., Pattison, P., Robins, G., and Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153.

[Snijders, 2002] Snijders, T. A. B. (2002). Markov chain monte carlo estimation of exponential random graph models. *J. Soc. Struct.*, 3.

[Yang et al., 2011] Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks–a bayesian approach. *Mach. Learn.*, 82(2):157–189.