Social Networks 76 (2024) 1–11

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Same but different☆
A comparison of estimation approaches for exponential random graph models for multiple networks

Petro Tolochko *, Hajo G. Boomgaarden

Department of Communication, University of Vienna, Austria

ARTICLE INFO

Keywords:
ERGM
Simulation study
Hierarchical modeling
Multiple network analysis

ABSTRACT

The Exponential Random Graph family of models (ERGM) is a powerful tool for social science research as it allows for the simultaneous modeling of endogenous network characteristics and exogenous variables such as gender, age, and socioeconomic status. However, a major limitation of ERGM is that it is mainly used for descriptive analysis of a single network. This paper examines two methods for estimating multiple networks: hierarchical and integrated. We contrast the two approaches, evaluate their accuracy and discuss the advantages and drawbacks of each. Furthermore, we make recommendations for future researchers on how to proceed with multiple network analysis depending on various factors such as the number of networks and the hierarchical structure of the data. This research is important as it highlights the need for the analysis of multiple networks in order to gain a more comprehensive understanding of social phenomena and the potential for new discoveries.

1. Dealing with multiple networks

The core of human interaction and the subject of virtually all social sciences are relations. Statistical social network analysis, specifically Exponential Random Graph Models (ERGM) are a family of methods that allow to quantitatively analyze this relational and interdependent structure, and are gaining popularity within the social sciences. Such models allow modeling the endogenous structural characteristics of a network simultaneously with the effect of exogenous variables (i.e., gender, age, socioeconomic status, etc.). Importantly, ERGM is not simply a class of statistical models, but also a set of theoretical assumptions about (social) network structures, hence allowing researchers to analyze network data in a robust, theoretically justified way. Within this framework, prior research across various disciplines of social science, for instance, addresses the issue of self-identification (Kruse and Kroneberg, 2019), diffusion of crime (Papachristos and Bastomski, 2018), cooperation and conflict between countries (Fritz et al., 2022), HIV prevalence (Krivitsky and Morris, 2017), voting and political participation networks legislative networks (Kirkland and Gross, 2014), political attitudes (Lazer et al., 2010), political communication networks (Song, 2015), migration communication flows (Verdery et al., 2018), or organizational structures (Pilny and Atouba, 2018), among many others. Given the centrality of networks for contemporary social and political phenomena, the relevance of social network analysis is only likely to grow in the discipline.

The ERGM framework, however, is generally limited to estimating the parameters of a single network. Indeed, possibly due to the high cost and difficulty in obtaining high-quality social network data, as well as the relative difficulty in the estimation of such models, the vast majority of studies involving ERGMs focus either on a specific single network, or a small set of independent networks that are compared/contrasted in the study (please refer to Lusher et al. (2013) for an overview of studies with ERGM applications). Nevertheless, such an approach allows researchers to thoroughly investigate only a single realization of a social network providing valuable information about the structure and social processes within a specific setting. But it inevitably leaves the question of whether the structural effects obtained from the model pertain only to this specific network, or whether these structures are actually generalizable to a wider set of networks in a relatively similar context. In their paper (Goodreau et al., 2009) stressed that even though the ERGM framework is theoretically justified and is often used in a hypothesis-testing scenario, the vast majority of applications remain mainly descriptive since they focus on a single network in isolation.

☆ Funding: This project has received funding from the European Union's HORIZON2020 Research & Innovation program under Grant Agreement no. 870612.
* Correspondence to: Kolingasse 15, 1170, Vienna, Austria.
E-mail address: petro.tolochko@univie.ac.at (P. Tolochko).

https://doi.org/10.1016/j.socnet.2023.05.003

Available online 22 May 2023

In recent years, however, there has been an increase in the number of studies that focus on multiple networks simultaneously to investigate broader structural phenomena. Some examples of multi-network studies include the analysis of school classes structure (Lubbers, 2003; Lubbers and Snijders, 2007), comparison of multiple political discussion networks (Goodreau et al., 2009; Eveland and Kleinman, 2013; Song, 2015; Minozzi et al., 2020), analysis of self-identification of children with immigrant background (Kruse and Kroneberg, 2019), or investigating organizational advice networks (Agneessens et al., 2022), among others. While the examples of studies that use a multi-network setup are not dominant in the social science literature, arguably because of the difficulties in obtaining and analyzing the data, they are becoming more commonplace.

In essence, this multilevel approach is characterized by following dependency assumptions: a set of N nodes (e.g., people), broken down into K separate blocks (e.g., classes) such that edges E within a block A_k are structurally dependent on each other (Lazega and Snijders, 2015; Schweinberger and Handcock, 2015, e.g.,). The nodes N are independent across K blocks (i.e., no structural possibility for nodes from different blocks to have a tie). Finally, the edges E have an "informational dependency" across blocks K in a sense that they are formed according to a similar dependency structure (e.g., reciprocity, transitivity, etc.). In other words, different blocks have a similar data generating processes behind the tie formation. Therefore, knowing something about the structural properties of block A_i may inform the generation of ties in block A_j.

Thus, the need for an analytic framework capable of dealing with such clustered structure, or a "population" (Lehmann and White, 2021) of networks, has become more evident. The purpose of this paper is to provide a general overview of the estimation possibilities for multiple networks, and investigate the differences and similarities between them in terms of their procedures and outcomes.

Several estimation methods have been proposed to deal with this structure. For example, Lubbers (2003) and Lubbers and Snijders (2007) were interested in investigating the variance in structural characteristics between 57 school classes. They proposed a "two-step regression" approach to analyze the estimates from multiple networks to determine their general significance. The formula they proposed was:

$$\hat{\theta}_m = \mu_\theta + U_m + E_m \tag{1}$$

with $\hat{\theta}_m$ being the estimated parameter for class m, μ_θ being the average coefficient, $U_m \sim Normal(0, \sigma^2_\theta)$ being the deviation of class m, and E_m is the estimation error. This approach is essentially a meta-analysis with a known level-1 variation (Goldstein, 2011). A very similar two-step method was proposed by Snijders and Baerveldt (2003). This methodology was further improved by Viechtbauer (2010) developing a framework for multilevel meta-analysis and An (2015) adjusting the meta-analysis application to social networks. Substantively, this method of estimation was used by Song (2015), in which multiple political discussion networks were analyzed. Slaughter and Koehly (2016) improved this approach by using Bayesian hierarchical models to pool network effects together. Similar approaches were implemented by Kruse and Kroneberg (2019), analyzing self-identification in German schools by incorporating school-specific variance across multiple ERGMs, as well as Minozzi et al. (2020), analyzing college political discussion networks by estimating a multilevel hierarchical Bayesian model to incorporate partial pooling of the estimates across 112 networks. Furthermore, another interesting extension of a hierarchical approach was proposed by Agneessens et al. (2022), where authors specified the dyad as a lower-level hierarchical structure and groups as higher-level structures.

The meta-analytic, "*hierarchical*" approach described above is a two-step method, in which individual network statistics are estimated in the first step and then coefficients are pooled relative to the grand mean. However, this is not the only possible way to deal with multiple networks. An alternative approach to multiple (multilevel) network estimation is the holistic estimation or "*integrated*" approach (e.g., Snijders, 2016), in which the entire population is estimated as a single network, but with structural zeros imposed in such a way that actors from subpopulations (i.e., individual networks) cannot interact with each other. Unlike the previously described paradigm, this is a one-step solution to the estimation problem. Since ERGM estimation is extremely computationally intensive (for n nodes, there are $2^{n(n-1)}$ possible directed networks, with complexity increasing exponentially as n grows), it is generally not well-suited for very large networks, with a few hundred nodes being the upper limit (Lusher et al., 2013). Therefore, the approach of simultaneously estimating several networks as a single large network instance was previously limited to a very small number of (sub-) networks (e.g., Kalish and Luria, 2013). Nevertheless, the increase in computational power and development of more efficient algorithms today allow for such estimation with the inclusion of more than a couple of network instances (subnetworks). The main assumption is that the supernetwork X is generated from a distribution $\mathbb{P}_{N,\theta}$, with the goal of estimating θ to model the tie formation mechanisms (e.g., Stewart and Schweinberger, 2018; Schweinberger and Handcock, 2015).

Within both of these approaches – a multilevel two-step vs. integrated supernetwork – networks can be estimated either with complete pooling estimates (i.e., all networks are assumed to have been generated by the same data generating process) or with no pooling (i.e., all networks are assumed to be generated by independent data generating processes). Partial pooling, however, is almost always a preferred method of dealing with clustered data (e.g., McElreath, 2020), since no-pooling does not take into account the clustering structure of the data and therefore is prone to underfitting, while complete pooling may drastically overfit the data. Both approaches can be extended to incorporate partial pooling (i.e., combining all available information from individual networks and the clusters these networks are in) — this can be done with a hierarchical Bayesian model for the two-step approach, or by modeling local and global dependence simultaneously, for the integrated approach (Schweinberger and Handcock, 2015); the R package `mlergm` (Stewart and Schweinberger, 2018), for example, is capable of modeling such networks. Partial pooling is also important when dealing with hierarchical or clustered data since the estimation can use information not only between individual networks but also between different clusters of networks.

There could be several reasons for choosing one estimation approach over another — practical, and, even more importantly, theoretical. In practice, the integrated approach may be significantly easier to estimate and diagnose, as described later in the paper. Theoretically, a researcher dealing with multiple networks would have to make an assumption about the data-generating process of these networks, specifically about the strength of the "informational dependency" discussed above, which refers to the similarity of the dependency structure across networks. If one assumes that the networks come from different data-generating processes, then the hierarchical approach should be used. On the other hand, by using the integrated approach, one implicitly assumes that networks come from a single distribution (e.g., Ripley et al., 2023).

While both procedures are generally thought to be significantly better than estimating ERGMs one-by-one (or even simply investigating a singular network, as has been the predominant modus operandi when dealing with ERGMs), they are a relatively recent addition to social network analysis. To our knowledge, they have not been systematically and empirically compared with one another, and no specific advantages or drawbacks of either one have been outlined.

The current paper proposes to systematically compare the aforementioned methods of multiple network estimation by using a relatively unique empirical dataset (Study 1), as well as simulation modeling (Study 2). In Study 1, we investigate whether the two approaches provide similar or divergent results and whether they can lead to conflicting interpretations.

Given the nature of the empirical data (described in more detail in the next section), namely that the data come from three different countries, schools, and classes, it is implausible to assume that all of these networks were generated by the same principle. Therefore, Study 1 aims to determine the extent to which these different estimation scenarios converge in the results despite the violated assumption. However, in less evident scenarios (e.g., when hierarchical structure is not explicitly specified), it may be very difficult to know whether the networks come from the same data-generating process and can be accurately estimated with the same model specification. This raises the probability of making the wrong assumption. In the case of conflicting results, a researcher in an empirical setting would potentially have no recourse in determining which of the estimation methods is more accurate.

Study 2 employs a Monte-Carlo simulation design to systematically compare the two approaches and various scenarios based on a "ground truth" that we determine beforehand. The main question is how well these different approaches can recover the "true" parameters. It is important to understand how the estimation behaves when the assumption that all networks come from the same data-generating process is violated. This is particularly relevant because in a real-world scenario, it could be difficult or not always possible to determine whether every network in the sample comes from the same distribution (one way to check is by looking at network descriptives and comparing the variable distributions (e.g., Ripley et al., 2023), but this still remains an assumption that the researchers have to make). Therefore, it is crucial to determine to what extent these different estimation approaches lead to similar conclusions under various data-generating conditions.

The paper further discusses the steps taken for the estimation of models for both of the paradigms, the difficulties, and specificities of each individual approach, and finally, makes recommendations for future researchers on how to proceed with multiple network analysis depending on various factors such as the number of networks, hierarchical structure of the data, estimation time, etc.

2. Estimation of empirical networks (Study 1)

2.1. Data

The data were collected in 2021, as part of an ongoing international research project [Omitted for Review]. The data contains information on 121 networks (school classes), nested in 15 different schools, across three different countries (Germany, Italy, and Portugal). Importantly, this dataset is quite unique because it has several hierarchical levels (classes, schools, and countries), allowing us to compare how partial pooling of estimates works on multiple clustering levels (unlike only on the network level as has been done previously). Furthermore, the dataset offers a collection of sociocentric networks (in contrast to egocentric networks), networks that have a known boundary and with data on all nodes and in the population of interest.

The data were collected across middle and high schools — age range from 9 to 20 (M_{age} = 14.82; SD_{age} = 1.38). The network data were collected by asking the students to name up to three "friends" within their class, furthermore, sociodemographic variables (age, gender, socioeconomic status, etc.) were measured. The average number of nodes was M_N = 19.92; SD_N = 6.77. The average in-degree of the networks was $M_{in-degree}$ = 2.31; $SD_{in-degree}$ = .53. A visualization of a subset of networks is present in Fig. 1. In-degree distribution for the complete sample of networks is presented in Fig. 2.

2.2. Method

2.2.1. Software

The analyses for this project were conducted in R, with the `ergm` modeling toolkit package (Hunter et al., 2008). Next, the `mlergm` (Stewart and Schweinberger, 2018) was used for the integrated, supernetwork estimation. Finally, `Stan` statistical modeling language (Carpenter et al., 2017) and the `brms` package (Bürkner, 2017) were used for the Bayesian multilevel modeling.

2.2.2. Parametrization and goodness-of-fit

Exponential Random Graph Models are quite difficult to fit. Sometimes ERGM parameters fit to the empirical data do not allow the algorithm to converge because of the nature of real-world graphs (e.g., Lusher et al., 2013). Therefore, one of the difficulties for the multinetwork analysis is estimating the "master" model (e.g., Minozzi et al., 2020), i.e., a model with equal specifications across all networks to make comparisons possible. Therefore, the first step is to find the set of specifications (structural network statistics) that works for all empirical networks in the data. This is an iterative process — once the parametrization that is applicable to all networks has been found, the goodness of fit is obtained for all networks. This process is repeated until the goodness of fit is acceptable, and the simulation procedures can recreate the empirical network structure with the chosen parametrization. There are numerous ways to parametrize an ERGM model, (please refer to Lusher et al. (2013) or Hunter et al. (2008) for a list of possible network statistics), therefore this is a non-trivial task for a large collection of networks.

In this specific instance, this problem is compounded by the fact that we need the same parametrization across two modeling strategies – hierarchical and integrated – potentially resulting in less-than-ideal parametrization. We estimate the models with the following parameters: density (θ_0), reciprocity (θ_1), geometrically weighted edgewise shared partnership (gwesp; Hunter, 2007; θ_2), and geometrically weighted in-degree (gwidegree; Hunter, 2007; θ_3). It is important to note that this parametrization was chosen not because of any theoretical underpinnings, but rather because it is a standard parametrization that often provides reasonable goodness of fit. Furthermore, nodal covariates, such as gender are also highly relevant for estimating especially school classes. We have opted *not* to include it in this investigation, however, due to the large proportion of missingness at nodal covariate level (roughly 90% of all networks has at least one node with missing data).[1] Nevertheless, three out of 121 networks could not be estimated with this parametrization, leaving 118 networks in total. The networks that failed to converge were also removed from the integrated approach, leaving the set of networks for the hierarchical and integrated approaches exactly the same.

To assess the goodness-of-fit for this analysis, we calculated the average p-value for the simulated networks based on the chosen parametrization. For the hierarchical approach (i.e., separate networks), the average p-value was 0.89, while for the integrated approach, the average p-value was 0.80. Higher p-values indicate that the simulated networks are coming from the same probability distribution as the empirical networks. Although the fit for the hierarchical estimation was slightly better than for the integrated estimation, the test statistic rarely exceeded 2, and the fit was deemed acceptable (Lusher et al., 2013).[2]

2.3. Modeling

As discussed above, we consider two general methods of modeling — hierarchical (i.e., a hierarchical regression model) and integrated (estimation of the whole supernetwork). Furthermore, the hierarchical

[1] When the models are ran with imputed gender data, structural estimates stay virtually the same, indicating the robustness of current parametrization.

[2] Please note that for an actual empirical scenario, much more attention should be paid to parametrizing the models and assessing the goodness-of-fit. The parametrization should be theoretically sound and result in an acceptable goodness-of-fit.

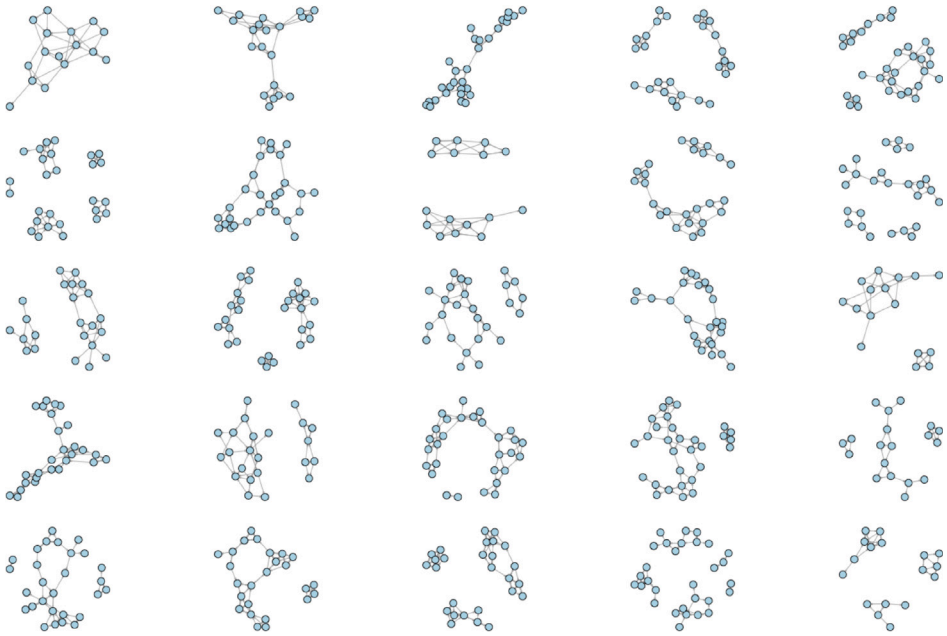

Fig. 1. Subset of 25 class networks.

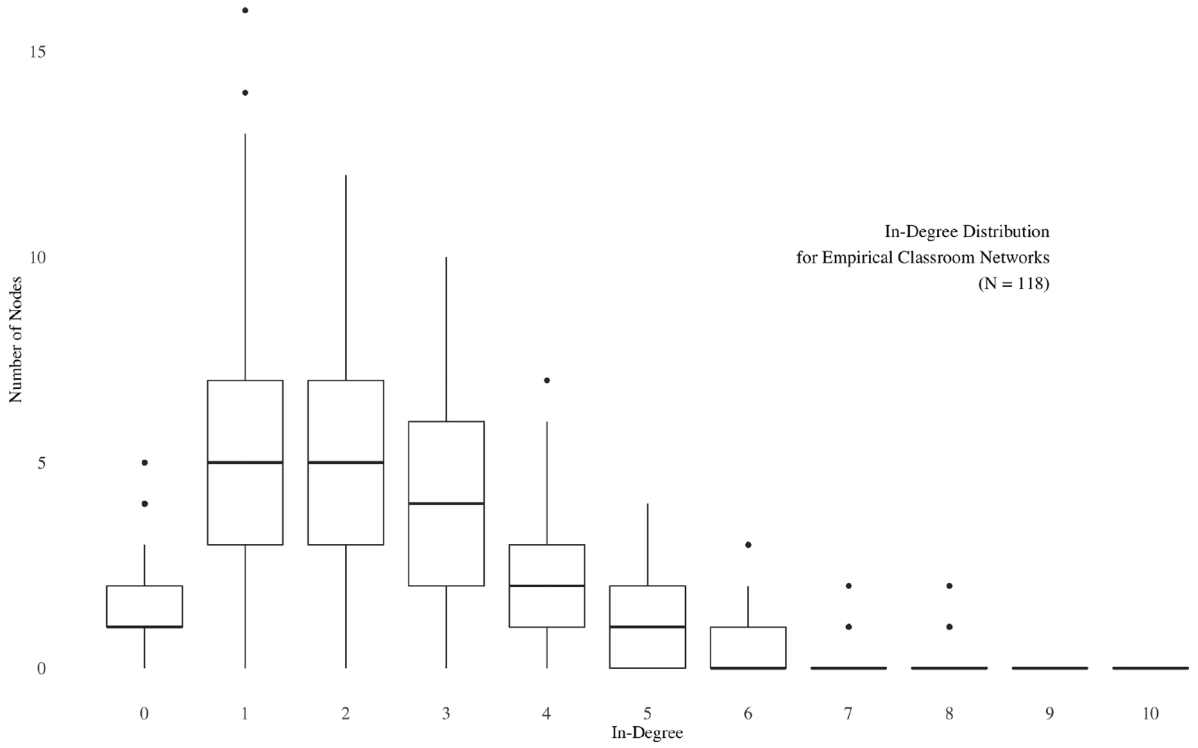

Fig. 2. In-degree distribution for 118 Networks. Note: The x-axis shows the number of nodes with in-degree x.

approach can accommodate additional grouping, therefore we break it down to a "simple hierarchical" method — with standard errors clustered on the network level, "country hierarchical" — with standard errors clustered on the network and country level, and "country-school hierarchical" — with clustering on network, country, and school levels.

2.3.1. Hierarchical estimation

This is a two-step method. The first step is to sequentially estimate the ERGM networks with the parametrization discussed above. Once all the networks were sequentially estimated, we move to the regression modeling. The basic hierarchical Bayesian model for meta-analysis (e.g., Röver, 2017) has a form of:

$$\begin{aligned} \hat{\theta}_k &\sim Normal(\theta_k, \sigma_k^2) \\ \theta_k &\sim Normal(\mu, \tau_k) \\ \mu &\sim Normal(0, 1) \\ \tau_k &\sim HalfCauchy(0, 1) \end{aligned} \tag{2}$$

where $\hat{\theta}_k$ is the estimated effect size from network k, σ_k^2 is the standard error of the estimate, θ_k is the true effect size that follows a Normal distribution with a mean of μ, the pooled effect size, and variance τ_k

signifying between network heterogeneity. μ and τ_k have a Normal and Half-Cauchy priors, respectively. This model can be extended to incorporate information on higher-level clustering and nesting:

$$\begin{aligned}
\hat{\theta}_k &\sim Normal(\theta_k, \sigma_k^2) \\
\theta_k &\sim Normal(\mu, \tau_k) \\
\mu &\sim Normal(\alpha_{country}, 1) \\
\alpha_{country} &\sim Normal(0, \tau_{country}) \\
\tau_k &\sim HalfCauchy(0, 1) \\
\tau_{country} &\sim HalfCauchy(0, 1)
\end{aligned} \tag{3}$$

for a single hierarchical level, or

$$\begin{aligned}
\hat{\theta}_k &\sim Normal(\theta_k, \sigma_k^2) \\
\theta_k &\sim Normal(\mu, \tau_k) \\
\mu &\sim Normal(\alpha_{country}, 1) \\
\alpha_{country} &\sim Normal(\alpha_{school}, \tau_{country}) \\
\alpha_{school} &\sim Normal(0, \tau_{school}) \\
\tau_k &\sim HalfCauchy(0, 1) \\
\tau_{country} &\sim HalfCauchy(0, 1) \\
\tau_{school} &\sim HalfCauchy(0, 1)
\end{aligned} \tag{4}$$

for two levels.

2.3.2. Integrated estimation

As mentioned previously, `mlergm` R package will be used for the integrated approach. The basic idea of the integrated approach is to merge all the networks in the sample together, imposing structural zeros between nodes from different networks (i.e., no possibility of a tie forming). The same parametrization is chosen as previously: density (θ_0), reciprocity (θ_1), geometrically weighted edgewise shared partnership (θ_2), and geometrically weighted in-degree (θ_3). The whole network is then estimated.

2.4. Results

Here we present the results obtained by the different methods of estimation (hierarchical with no country grouping, hierarchical with country grouping, hierarchical with country and school grouping, and integrated). The posterior modes and 95% Credibility Intervals (for the integrated approach, the 95% Confidence Interval is presented) are shown in Fig. 3.

As can be seen from Fig. 3, $\hat{\theta}$s for hierarchical approaches are very close to each other, albeit models that incorporate group variance have significantly wider CIs. It should be noted that the regularization (i.e., wide confidence intervals) in the hierarchical scenario is implicit, and comes from the pooling of the estimates to the grand mean. Nevertheless, explicit regularization can also be achieved by using either a highly informative prior or a shrinkage prior. For more information please refer to (e.g., Gelman et al., 2013; Piironen and Vehtari, 2017).

On the other hand, $\hat{\theta}$s obtained with the integrated approach is consistently different from the first two — either significantly lower (in the case of θ_0 and θ_2), or higher (in the case of θ_1 and θ_3). Moreover, in three out of four cases, the CIs of the integrated approach do not even overlap with the wide CIs of the hierarchical group model estimates.

Since all of these estimates are of the same sign, in the binary hypothesis testing scenario, one would reject the null hypothesis. While rejecting the null hypothesis based on statistical significance can be informative, it does not provide information about the size of the effect. In other words, simply knowing that a difference exists between two groups may not be sufficient to make informed decisions. For example, if we are interested in the difference in network density based on some external factor, we would not only like to know that the difference does not equal zero, but also the actual difference between the groups. The effect size could determine whether it has "practical significance", therefore allowing us to make decisions about the practicality of interventions and whether it is worth the financial investment. An effect size of a given magnitude may be considered significant in one context but not in another. Estimating the parameter and its uncertainty correctly is thus important for making informed decisions. Thus, simply looking at the statistical significance of an estimate is not enough to draw meaningful conclusions from the data (e.g., Nickerson, 2000; Gelman and Stern, 2006; McShane et al., 2019).

Given that the empirical networks come from three different countries, different schools (potentially in different socioeconomic areas), and the age distribution across the networks, the assumption that all of the networks come from the same distribution might not be plausible. Nevertheless, we can look at the estimated τ parameters from the most complex hierarchical model to understand whether the spread in the network and/or country/school estimates is significant or not. Fig. 4 shows that the network, country, (and school) variances are quite substantial, providing more evidence that the different networks in the sample are coming from different data-generating processes. Importantly, the group variances are actually different for different $\hat{\theta}$ parameters, indicating that including the group information into the model may be beneficial, especially if one intends to predict, for example, specific group effects.

2.4.1. Out-of-sample predictive accuracy

Bayesian hierarchical models incorporate priors, which can potentially have a regularizing effect on parameter estimates. This is particularly evident in the case where three out of four parameters are shrunk towards 0, as depicted in Fig. 3. If this regularizing effect is indeed present, then the model's out-of-sample predictive accuracy should theoretically improve, as the model would be less likely to overfit the sample data.

Next, we attempt to understand whether the hierarchical models are better at out-of-sample prediction. This would suggest that the differences in effect sizes could be attributed to regularization. Specifically, we employ a leave-one-out cross-validation method in which the models are trained on all networks except for one held-out network. We then use the models to make predictions at every tie level (i.e., predicting $P(Y_{ij} = 1)$ for the held-out network). Since we have rather sparse networks ($M_{density} = 0.13$), we further examine the area under the Precision–Recall Curve (AUPRC), which is a good method for performance evaluation for the classification of imbalanced binary events (e.g., Sofaer et al., 2019). AUPRC is a summary statistic for model performance based on the various classification thresholds of precision and recall. Higher values of AUPRC would indicate that the model is better at predicting yet-unseen data and is not overfitting the training data. Fig. 5 shows the box plots of AUPRC for each of the estimation methods, over 118 out-of-sample networks.

As can be seen from the figure, the AUPRC is virtually the same for all the estimation methods, with the integrated approach having a slightly lower median AUPRC (although statistically indistinguishable from the other models). It should be noted that all models have an acceptable predictive accuracy, with the average baseline being $\approx$13% (average proportion of dyads with an edge in a network, thus the baseline curve would be equivalent to randomly selecting 13% of dyads and assigning an edge between them). These findings suggest that neither the integrated approach nor the hierarchical approach is better at predicting yet-unseen data, and that the regularizing effect of the hierarchical models is not strong enough to improve the model's out-of-sample predictive accuracy.

3. Monte-Carlo simulation (Study 2)

The empirical results suggest that the different methods of estimation actually provide rather different outcomes, that might change the way the researcher interprets the results in an empirical setting. While there is no way to discover which of the methodologies actually comes

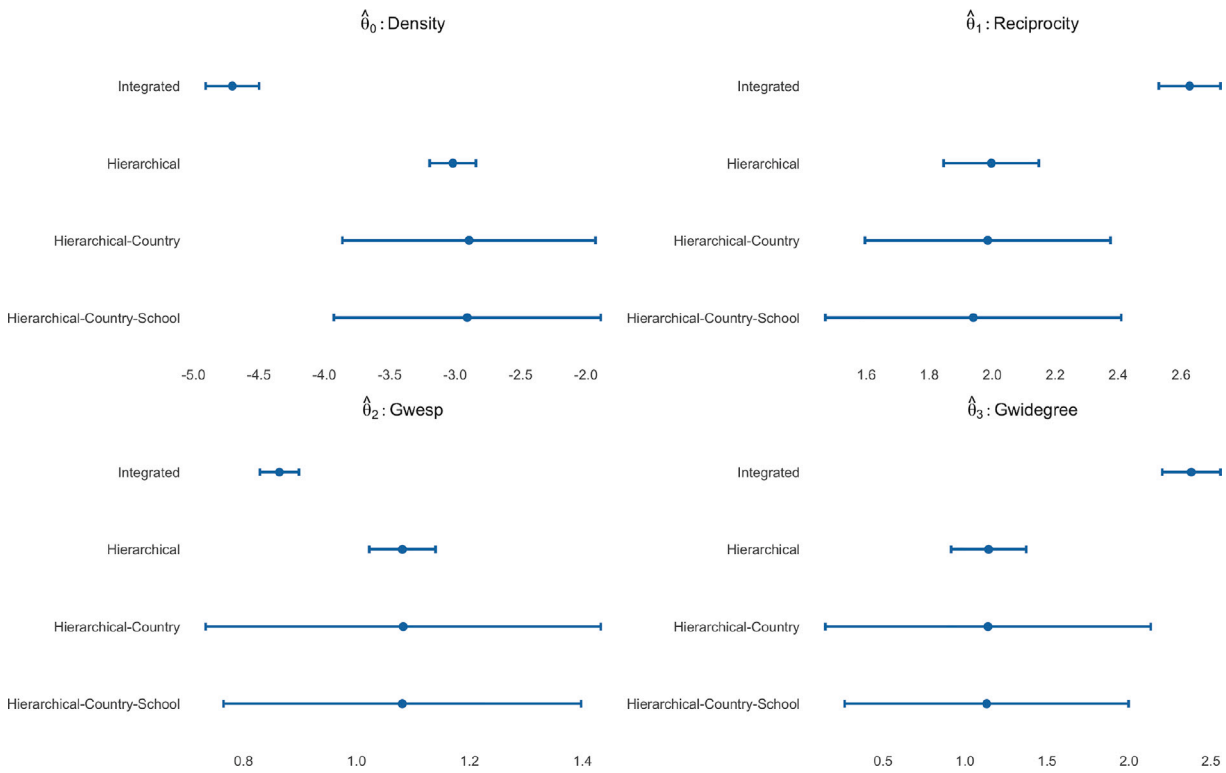

Fig. 3. Posterior modes and 95% CIs.

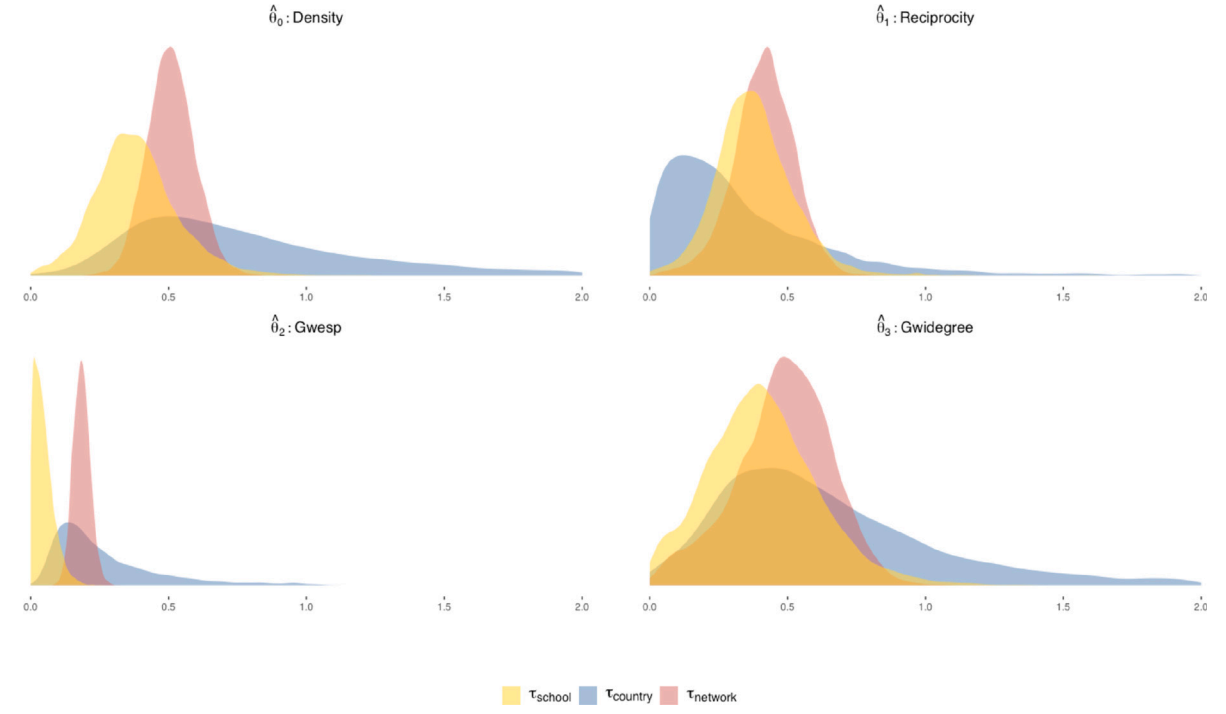

Fig. 4. Posterior distribution of variances.

closer to the "truth" with empirical data, one can devise a simulation set-up, in which the "truth" is actually known (i.e., parameters chosen by the researcher). This allows to consider which method comes closer to estimating the "true parameters". Simulation is crucial because in this scenario we know the "true" parameter vector θ — something that is impossible with empirical data. Thus, changing the "true" parameters according to the cluster in which the network is nested, would allow us to compare how each of the estimation procedures (hierarchical vs. integrated) is able to recover them.

3.0.1. *Modeling*

With a set of simulated models, with different "true" θ according to their cluster membership, we can now compare the two estimation approaches described earlier. As mentioned previously, `mlergm`

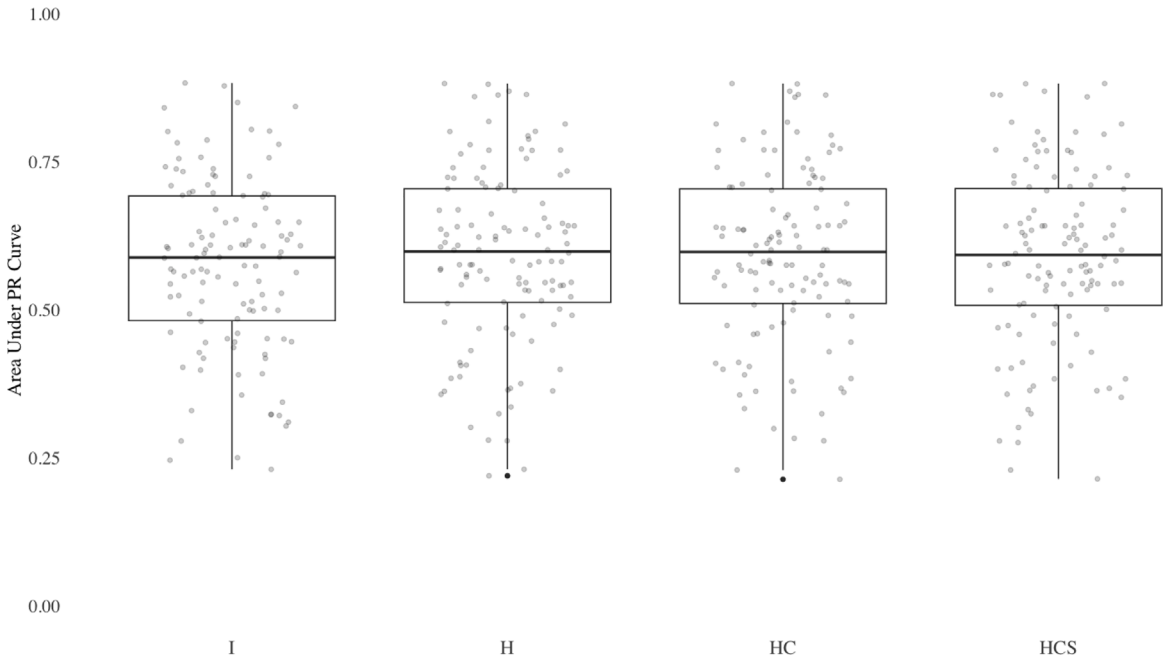

Fig. 5. AUPRC for different estimation methods. Note: I – "Integrated", H – "Hierarchical", HC – "Hierarchical with Country grouping", HCS – "Hierarchical with Country and School grouping".

R package will be used for the integrated approach and a Bayesian hierarchical model for the two-step method (Models (2) and (3)). The main question is how well are these methods able to recover the parameters under different set setups (i.e., the grand mean, individual cluster parameters, etc.).

3.1. Data generation

The exponential graph modeling framework allows simulating networks based on chosen network parameters (e.g., network size, the density of the network, number of triangles, etc.). We further refer to the input parameters as θ (the ground truth) and the output parameters as $\hat{\theta}$ (estimated parameters). In order to create the parameter vectors θ with a hierarchical structure (for reasons of computational tractability we only consider one level of hierarchical grouping for the simulation study), one could use the following model:

$$\begin{aligned} &\theta_{ij} \sim Normal(\mu_i + \delta_j, \sigma^2_{ij}) \\ &\sum_{j=1}^{J} \delta_j = 0 \end{aligned} \tag{5}$$

where θ_{ij} is the ith network coefficient for group j, μ is the "grand mean" of the coefficient, δ_j is the group effect for group j and σ^2_{ij} is the variance of the coefficient. However, due to the inherent simulation variance of the ERGM estimation, the decision was taken to simplify the model for the input parameters to:

$$\begin{aligned} &\theta_{ij} = \mu_i + \Delta_j \\ &\Delta_j \sim Normal(0, \delta_j) \end{aligned} \tag{6}$$

removing the variance of the coefficient at the current step, since it will be added during the MCMC simulation.

The parameter vector θ is thus created by only selecting μ and the group effects Δ_j (which is drawn from Normal distribution with δ_j, so that $\mathbb{E}\Delta_j = 0$, and thus the group effect is the same for all parameters in the parameter vector θ). Once θs are created, they are fed to the ERGM simulation algorithm (from the `ergm` package) to simulate random networks. To obtain $\hat{\theta}$s, the simulated networks are sequentially estimated using the same package for further hierarchical modeling, or combined and estimated with the `mlergm` package for the integrated approach. The μ vector was chosen to be the same throughout the simulation runs (thus only θ is changing from run to run).

Choosing theoretically sound parametrization is even less important here than in was in Study 1, since we are not interested in the interpretation of results. The only interest here lies in how well can we estimate the parameters that were used as ERGM simulation inputs. Therefore, a decision was taken simplify the parameter vector from Study 1, and use $\mu = -1, .5, -.5$. This reduction in the number of simulated/estimated parameters makes the simulation more manageable from a computational standpoint.[3] The μ_0 (*density/edges* parameter), μ_1 (*mutual* parameter), and μ_2 (*gwesp* parameter). Therefore, if $\Delta_1 = .2$, then, e.g., $\theta_{1,1} = -.8, \theta_{2,1} = .7, \theta_{3,1} = -.3$.

3.2. Monte-Carlo simulation parameters

We now turn to the parameters that guide the simulation runs. We consider the simulation conditions that might be typical of a study that employs multiple networks and is occupied with estimating the general networks parameters from the sample of networks (see Table 1): the number of groups per population of networks ($K = \{1, 3, 5\}$), the size of these groups ($S_k = \{10, 20, 30\}$), the number of nodes in a single network ($N_{nodes} = \{20, 30, 40\}$), and finally, the group effect size $\delta_j = \{.1, .3, .6\}$ (term from Eq. (5)). In total, this gives $3^4 = 81$ simulation conditions. Each of the simulation conditions is replicated 30 times with 3 parameters, therefore all simulation runs result in $N = 7290$ separate observations. Each run, Δ_j is resampled and θ is calculated. The input parameters for these conditions were selected based on "realistic" real world scenarios (e.g., a standard deviation of .2 on a log odds scale is realistic, while a standard deviation of 2 is less so). Another consideration was that of practicality — the total node count (i.e., from all networks) per simulation condition ranged from 200 nodes ($K = 1$, $S_k = 10$, $N_{nodes} = 20$) to 6000 nodes ($K = 5$, $S_k = 30$, $N_{nodes} = 40$), thus making multiple simulation runs of larger conditions computationally expensive even on a modern computer. For each simulation run, the absolute error (i.e., $|\hat{\theta} - \theta|$) is computed.

[3] Still, on a modern laptop with an M1 processor and 16 GB of RAM, the complete simulation took several weeks in total to run.

Table 1
Monte-Carlo simulation input parameters.

Simulation factors	Simulation parameters
K (Number of groups)	1, 3, 5
S_k (Group size)	10 (small), 20 (medium), 30 (large)
N_{nodes} (Nodes per network)	20 (small), 30 (medium), 40 (large)
δ_j (Group effect size)	.1 (small), .3 (medium), .6 (large)

3.3. Results

We first present the absolute error (AE), defined as $\Delta\theta = |\hat{\theta} - \theta|$. Fig. 6 presents the results of the simulation runs. Three columns represent θ_0, θ_1, and θ_2, respectively. Separate panels are broken down into the number of groups for the simulation run, while the three rows represent the other three simulation factors. "H", "HG", and "I" stand for "Hierarchical", "Hierarchical-Group", and "Integrated". Hierarchical-Group is omitted from the single group condition.

As can be seen, the absolute error is, unsurprisingly, very small for single group conditions across all models and simulation factors. The absolute error, nevertheless, begins to increase with the number of groups and the group effect size. This pattern is the most pronounced for the integrated estimation method since it appears that it struggles to estimate the between-group variability. The hierarchical models are less affected by the number of groups and the group effect size. The increase in group size and the number of nodes per network are allowing for a more precise estimation, but only for the hierarchical models; the integrated model is not affected by these factors.

We now move to modeling the MAE estimates based on different simulation factors. Table 2 reports the results of a Bayesian Gamma GLM with log-link[4] predicting the Mean Absolute Error (average out of 30 runs) as the outcome, with marginal means for every simulation factor and estimation method. All models included varying intercepts on the $\hat{\theta}$ parameters. $\hat{R}$ for all parameters did not exceed 1, indicating good convergence of models (Gelman and Rubin, 1992).

The table reports log-transformed predicted MAE values, therefore the higher the value, the higher the estimated MAE for that method. As can be seen from the table, the MAE estimates are consistently lowest for the hierarchical group models, for every simulation factor. Hierarchical models without the grouped random errors are also performing well although the difference between grouped and non-grouped models (as is the difference between all models) is statistically significant. The only scenario, where the integrated method performed on par with the hierarchical method, is the single group scenario. Moreover, we see an increase in estimation error as the number of groups gets higher and as the group effect increases. This difference is much more pronounced for the integrated method than for the other two. The increase in group size, as well as network size, helps all estimation methods to achieve better accuracy.

Judging from the results of the simulation study, the most important factors (in terms of their effect) for the estimation of the MAE are the number of groups and the group effect size. In other words, attempting to capture between- and within-group variability should be one of the priorities for the empirical researcher who is using multiple networks and ERGMs in their setup. It would seem that the clear winner here is the Hierarchical regression model with clustered random errors since it is the only model that is able to capture the between-group variability, and generally performs well. Nevertheless, there are several practical considerations involved that do not necessarily deal strictly with the accuracy of the estimate, but rather with the practicality of analysis.

[4] Gamma regression model provided the best posterior predictive distribution for the data (non-negative, right-skewed mean absolute error values).

4. General discussion

As a unique method for dealing with relational data, social network analysis provides opportunities for researchers to investigate phenomena that otherwise would be off limits with a different methodology. Social network analysis is already a mature methodology and a theoretical paradigm. Nevertheless, even the approaches that purport to be robust hypothesis testing methods – such as the Exponential Random Graph Models – in essence analyze only a single draw from the potential distribution of networks. In order to make generalizable claims, one needs to analyze a sample of networks representative of the larger distribution. There are several difficulties associated with such an approach. Two of the major ones are the effort needed to obtain network data and the question as to how to analyze multiple network data. This paper deals with the second problem.

The results from both studies provide novel insights into the estimation procedures of multiple ERGMs. Study 1 showed that the Hierarchical and the integrated approaches to network estimation do not converge on the same results on the large set of empirical networks. A simple hierarchical model and the hierarchical model with additional group (country and school) information did produce similar posterior modes, however, the models with added group information had significantly wider CIs, implying that the groups (country and school) add additional uncertainty to the estimation procedure. The integrated approach produced estimates that differed significantly from the hierarchical estimation. Moreover, the integrated approach provides very narrow CIs, indicating the high certainty of the estimate.

As mentioned previously, in order to make a well-informed choice regarding the estimation method, the researcher would need to make certain assumptions about the data-generating processes that the networks come from. If the data are believed to come from a single distribution, then an integrated approach might be best suited. Otherwise, a hierarchical approach should theoretically account for the differences in the data clusters. While given the empirical example provided in the current study this choice is arguably relatively straightforward — several countries, several schools within these countries as well different students' ages across networks make the assumption of the same data-generating process less probable, a researcher in an empirical setting could potentially be hard-pressed to actually choose which of these estimation options is correct given a less obvious scenario than presented here. One simple recommendation to choose one method over another would be if out-of-sample performance. If one method performs better at predicting at this evaluation strategy, this would imply that it is not overfitting the sample an is better at anticipating the "true" data-generating process. In our empirical example, however, we found no difference in the out-of-sample predictive performance of different methods. This might have been either due to idiosyncrasies of our specific sample, or a general trend of the methods and, therefore, warrants further investigation. Thus, a simulation scenario allows us to investigate how well different estimation methods approach the "real" data-generating process. And, importantly, to understand how the violation of the aforementioned assumption reflects on the obtained results.

Using a Monte-Carlo simulation setup (Study 2), we tested under which conditions which estimation method produces optimal results. The simulation study investigated a number of factors that are most likely conditions one would encounter when doing empirical research with multiple social networks: different number of groups, different group sizes, different sizes of networks (number of nodes), and finally different group effect sizes. In total, the simulation had 7290 separate observations, with an absolute error of the estimation computed for each observation.

The estimation method based on a hierarchical regression model (with errors clustered on the group level) was the most accurate, with the smallest mean absolute error across all conditions. The hierarchical model with no grouping came in a close second. The integrated model

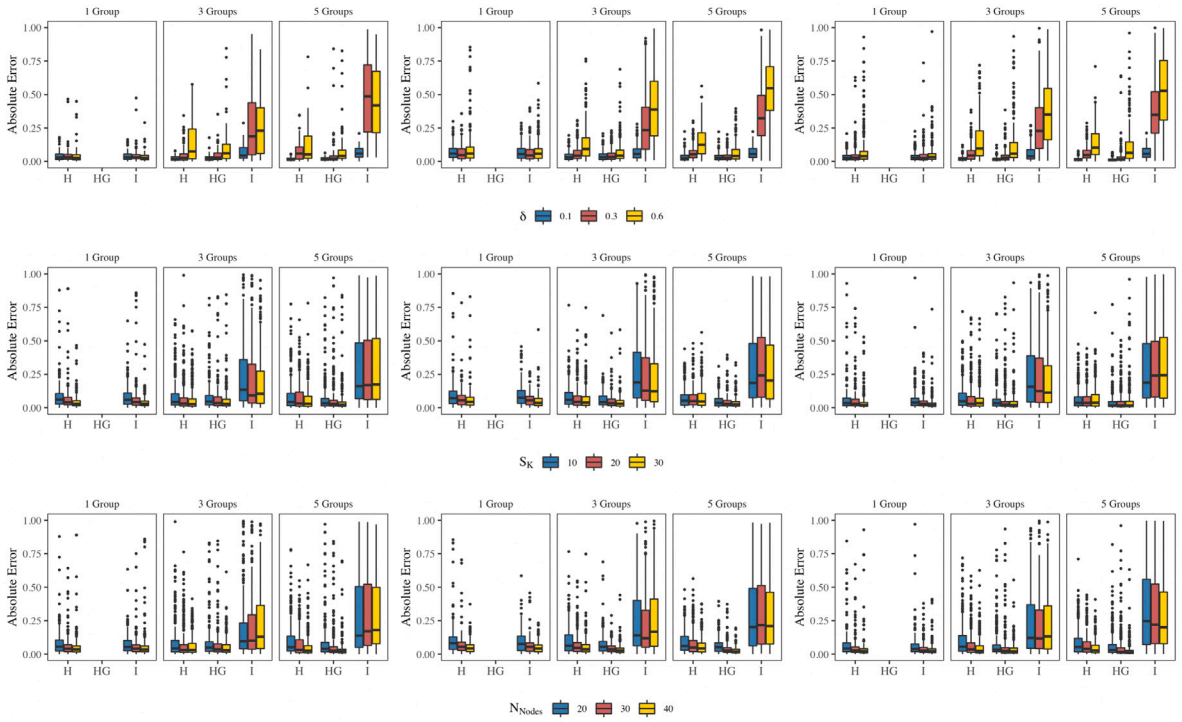

Fig. 6. Simulation results.

Table 2
Gamma models predicting MAE.

Simulation factors	Hierarchical (No group)	Hierarchical (Group)	Integrated
K **(Number of groups)**			
1	−2.69[−2.74; −2.65]	–	−2.60[−2.65; −2.56]
3	−2.51[−2.56; −2.46]	−2.74[−2.79; −2.70]	−1.13[−1.17; −1.08]
5	−2.75[−2.79; −2.70]	−2.54[−2.59; −2.50]	−0.89[−0.93; −0.84]
S_k **(Group size)**			
10	−2.43[−2.48; −2.39]	−2.62[−2.67; −2.56]	−1.27[−1.32; −1.23]
20	−2.57[−2.61; −2.52]	−2.80[−2.85; −2.74]	−1.29[−1.34; −1.24]
30	−2.67[−2.72; −2.62]	−2.83[−2.89; −2.77]	−1.36[−1.42; −1.31]
N_{nodes} **(Nodes per network)**			
20	−2.30[−2.35; −2.25]	−2.54[−2.60; −2.48]	−1.20[−1.25; −1.15]
30	−2.64[−2.69; −2.59]	−2.78[−2.84; −2.73]	−1.35[−1.39; −1.30]
40	−2.78[−2.83; −2.73]	−2.95[−3.01; −2.89]	−1.39[−1.44; −1.34]
δ_j **(Group effect size)**			
.1	−3.35[−3.39; −3.32]	−3.65[−3.70; −3.60]	−2.79[−2.83; −2.76]
.3	−2.70[−2.74; −2.66]	−3.11[−3.16; −3.06]	−1.26[−1.30; −1.21]
.6	−2.03[−2.07; −1.98]	−2.09[−2.16; −2.04]	−0.76[−0.81; −0.71]

Note: The values presented are log-transformed predicted MAE and 95% Credibility Intervals. Each model across one experimental factor (i.e., one row) is statistically distinguishable from other models. The difference was calculated by subtracting the posterior distribution of models from each other, if the 95% CI did not cross 0, then the model parameters were deemed "statistically distinguishable". The hierarchical Group model is missing from $K = 1$ factor because there are no groups to cluster the random intercepts on.

was the least accurate, with the largest mean absolute error across all conditions. When it comes to the simulation factors, the number of groups and the group effect size are the two most important factors influencing the absolute error. And, predictably, increasing the number of networks and the size of networks will decrease the absolute error.

4.1. Recommendations

At face value, it would seem like the hierarchical models the recommended analysis methods when dealing with multiple networks, given the lowest estimation error. Nevertheless, several practical considerations are not immediately apparent from the results that may impact the choice of method. We here discuss the pros and cons of each approach and provide some guidance in which situations one might want to use which analysis.

As mentioned above, hierarchical models with varying intercepts based on groups provide the most accurate estimates, closely followed by hierarchical models with varying intercepts only on the network level. One obvious drawback is the fact that a researcher who has a set of networks at their disposal, does not always possess the information on the group affiliation of these networks. In this scenario, the model with only network-level intercepts is a much more fitting choice. The integrated approach does comparatively well only in a scenario in which all networks come from the same population, and drops in

accuracy quite drastically when the number of groups increases (and even more so if the group effect size increases). However, within the integrated approach, the network model is significantly easier to fit. This comes from mainly two factors: (1) network specification and (2) goodness-of-fit. As noted previously, ERGMs are very sensitive to the specification chosen, and quite often empirical networks simply cannot be estimated with a chosen specification (e.g., Lubbers, 2003; Lusher et al., 2013). Thus, for the integrated approach, researchers should be able to find a theoretically driven specification easier since they need to specify only a single model. The same rationale goes for the goodness-of-fit – researchers should be able to find a correct specification and arrive at an acceptable goodness-of-fit for a single model.

The reverse is true for the hierarchical approach (regardless of the inclusion of groups, etc.), for which researchers should find a correct specification that suits every single network in the dataset. Furthermore, with increasing numbers of networks in the set, the probability that several of the empirical networks cannot be estimated also increases. Therefore, the choice is whether to drop networks that cannot be estimated, or to restart the estimation procedure with a different specification. The first option is obviously suboptimal because of the loss of data, while the second option is likely to take a significant amount of time.[5] Finally, when all network models do converge, the next step is to calculate goodness-of-fit statistics for every single model. Once again, the chosen specification that works for estimating all the models in the set, might not produce acceptable goodness-of-fit statistics on some empirical networks, creating the same choice of dropping the networks, or restarting the whole procedure with a different specification.

The following recommendations for analyzing multiple networks with ERGMs are given. If the data is known to come from different distributions and there is information on the group affiliation of each individual network – the hierarchical grouped approach is the optimal choice. By contrast, hierarchical models without the group information also provide fairly accurate results and should be the preferred choice in a study setup, where grouping information is not available (which is probably the vast majority of scenarios). Finally, the integrated approach should be chosen only in those situations, where the researcher is fairly confident that the data comes from the same distribution (no significant group effects). The ease of estimation of the integrated models wins over the marginal increase in accuracy provided by the hierarchical models for one group scenario.

4.2. Limitations and conclusion

Several limitations of the current study should be noted. The first major limitation refers to Study 1. The set of empirical networks comes from the same project, and the data were obtained in the same period. In order to increase the robustness of the design, one would ideally use several sets of empirical networks to better understand the details of multi-network estimation. However, due to the aforementioned difficulties with obtaining such data, this study had to resort to a single case study. Moreover, future researchers interested in this topic should investigate whether network data with a temporal component can be analyzed in the same way.

Second, this study does not provide in-depth recommendations on the goodness-of-fit for a hierarchical estimation scenario. While not crucial for the simulation set-up discussed above, this step is critical for estimating ERGMs with empirical data. To the best of our knowledge, so far there exists no "integrated" method of goodness-of-fit evaluation for multiple networks apart from evaluating them one-by-one (e.g., Minozzi et al., 2020). This is straightforward with a relatively few networks, but quickly becomes very difficult as the number of networks grows. While outside the scope of the current study, future research should investigate possible ways of evaluating the goodness-of-fit that incorporate multiple networks.

Third, the Monte-Carlo simulation study allows us to investigate phenomena that are not possible with empirical data, simply because we *know* the truth beforehand. However, the simulation setup in Study 2 was somewhat limited in the number of simulated variables by the computational expense of the simulation. Even though we believe that the current factors were chosen as the most common variables that may come up in a multi-network study, there might be other scenarios that are not covered by the current simulation. For example, currently, we used a rather simple parametrization of networks (i.e., density, reciprocity, and gwesp). Future research should explore other variants of parametrization and how they affect the accuracy of different methods. Additionally, the aforementioned temporal aspect of the network data is not implemented in the design, although it could have an effect.

Regardless of these limitations, we believe this study provides a valuable contribution to understanding multiple ERGM estimation procedures. We showed that different estimation methods often arrive at different results, and have designed a rigorous simulation study to test the current methods, and show in which situations which methods perform the best. Importantly, this study implicitly shows that very often results based on a single network ERGM estimation may not be as generalizable as one might like. We do not advocate for ERGM methodology to be used *only* in a multi-network set-up. There are many instances of social network analysis in which researchers might be interested in a specific network and do not care about the general pattern of effects. However, when we think about a theory-testing paradigm, we often think of "the more data the better" adage but are quick to generalize findings based on a single network. It might be very difficult to obtain high-quality network data for every single research question, thus we do not adopt a dogmatic view that all ERGM research should inevitably incorporate multiple networks. We do hope, however, that the current study does push researchers to pay closer attention both to the limitations of single network designs and to the broader implications of generalizability of results. Only through the combination of careful design and applicability of the data we could attempt to minimize the error in our research.

[5] During the course of the simulation, multiple runs had to be restarted several times simply because some of the simulated networks in a single run could not be estimated with the chosen specification.

Funding

This project has received funding from the European Union's HORIZON2020 Research & Innovation program under Grant Agreement no. 870612.

References

Agneessens, F., Trincado-Munoz, F.J., Koskinen, J., 2022. Network formation in organizational settings: Exploring the importance of local social processes and team-level contextual variables in small groups using Bayesian hierarchical ergms. Social Networks.

An, W., 2015. Multilevel meta network analysis with application to studying network dynamics of network interventions. Social Networks 43, 48–56.

Bürkner, P.-C., 2017. brms: An r package for Bayesian multilevel models using stan. J. Stat. Softw. 80, 1–28.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. J. Stat. Softw. 76 (1), 1–32.

Eveland, W.P., Kleinman, S.B., 2013. Comparing general and political discussion networks within voluntary organizations using social network analysis. Polit. Behav. 35 (1), 65–87.

Fritz, C., Mehrl, M., Thurner, P.W., kauermann, G., 2022. Exponential random graph models for dynamic signed networks: An application to international relations.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. CRC Press.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statist. Sci. 457–472.

Gelman, A., Stern, H., 2006. The difference between significant and not significant is not itself statistically significant. Amer. Statist. 60 (4), 328–331.
Goldstein, H., 2011. Multilevel Statistical Models, Vol. 922. John Wiley & Sons.
Goodreau, S.M., Kitts, J.A., Morris, M., 2009. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. Demography 46 (1), 103–125.
Hunter, D.R., 2007. Curved exponential family models for social networks. Social Networks 29 (2), 216–230.
Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M., 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. J. Stat. Softw. 24 (3), nihpa54860.
Kalish, Y., Luria, G., 2013. Brain, brawn, or optimism? Structure and correlates of emergent military leadership. In: Exponential Random Graph Models for Social Networks. pp. 226–236.
Kirkland, J.H., Gross, J.H., 2014. Measurement and theory in legislative networks: The evolving topology of congressional collaboration. Social Networks 36, 97–109, Special Issue on Political Networks.
Krivitsky, P.N., 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. Ann. Appl. Stat. 11 (1), 427.
Kruse, H., Kroneberg, C., 2019. More than a sorting machine: Ethnic boundary making in a stratified school system. Am. J. Sociol. 125 (2), 431–484.
Lazega, E., Snijders, T.A., 2015. Multilevel Network Analysis for the Social Sciences: Theory, Methods and Applications, Vol. 12. Springer.
Lazer, D., Rubineau, B., Chetkovich, C., Katz, N., Neblo, M., 2010. The coevolution of networks and political attitudes. Polit. Commun. 27 (3), 248–274.
Lehmann, B., White, S., 2021. Bayesian exponential random graph models for populations of networks. arXiv preprint arXiv:2104.05110.
Lubbers, M.J., 2003. Group composition and network structure in school classes: a multilevel application of the p* model. Social Networks 25 (4), 309–332.
Lubbers, M.J., Snijders, T.A., 2007. A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. Social Networks 29 (4), 489–507.
Lusher, D., Koskinen, J., Robins, G., 2013. Exponential RandOm Graph Models for Social Networks: Theory, Methods, and Applications, Vol. 35. Cambridge University Press.
McElreath, R., 2020. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. CRC Press.
McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2019. Abandon statistical significance. Amer. Statist. 73 (sup1), 235–245.
Minozzi, W., Song, H., Lazer, D.M., Neblo, M.A., Ognyanova, K., 2020. The incidental pundit: Who talks politics with whom, and why? Am. J. Polit. Sci. 64 (1), 135–151.
Nickerson, R.S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol. Methods 5 (2), 241.
Papachristos, A.V., Bastomski, S., 2018. Connected in crime: The enduring effect of neighborhood networks on the spatial patterning of violence. Am. J. Sociol. 124 (2), 517–568.
Piironen, J., Vehtari, A., 2017. Sparsity information and regularization in the horseshoe and other shrinkage priors.
Pilny, A., Atouba, Y., 2018. Modeling valued organizational communication networks using exponential random graph models. Manage. Commun. Q. 32 (2), 250–264.
Ripley, R.M., Snijders, T.A., Boda, Z., Vörös, A., Preciado, P., 2023. Manual for Rsiena, Vol. 1. University of Oxford, Department of Statistics, Nuffield College. University of Groningen: Department of Sociology, p. 2023.
Röver, C., 2017. Bayesian random-effects meta-analysis using the bayesmeta r package. arXiv preprint arXiv:1711.08683.
Schweinberger, M., Handcock, M.S., 2015. Local dependence in random graph models: characterization, properties and statistical inference. J. Amer. Statist. Assoc. 77 (3), 647.
Slaughter, A.J., Koehly, L.M., 2016. Multilevel models for social networks: Hierarchical Bayesian approaches to exponential random graph modeling. Social Networks 44, 334–345.
Snijders, T.A., 2016. The multiple flavours of multilevel issues for networks. In: Multilevel Network Analysis for the Social Sciences. Springer, pp. 15–46.
Snijders, T.A., Baerveldt, C., 2003. A multilevel network study of the effects of delinquent behavior on friendship evolution. J. Math. Sociol. 27 (2–3), 123–151.
Sofaer, H.R., Hoeting, J.A., Jarnevich, C.S., 2019. The area under the precision–recall curve as a performance metric for rare binary events. Methods Ecol. Evol. 10 (4), 565–577.
Song, H., 2015. Uncovering the structural underpinnings of political discussion networks: Evidence from an exponential random graph model. J. Commun. 65 (1), 146–169.
Stewart, J., Schweinberger, M., 2018. mlergm: Multilevel exponential-family random graph models. R package version 0.1.
Verdery, A.M., Mouw, T., Edelblute, H., Chavez, S., 2018. Communication flows and the durability of a transnational social field. Social Networks 53, 57–71, The missing link: Social network analysis in migration and transnationalism.
Viechtbauer, W., 2010. Conducting meta-analyses in r with the metafor package. J. Stat. Softw. 36 (3), 1–48.