# Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms

Jonathan Stewart[a] Michael Schweinberger[b] Michal Bojanowski[c]

Martina Morris[d]

**Abstract**

Multilevel network data provide two important benefits for ERG modeling. First, they facilitate estimation of the decay parameters in geometrically weighted terms for degree and triad distributions. Estimating decay parameters from a single network is challenging, so in practice they are typically fixed rather than estimated. Multilevel network data overcome that challenge by leveraging replication. Second, such data make it possible to assess out-of-sample performance using traditional cross-validation techniques. We demonstrate these benefits by using a multilevel network sample of classroom networks from Poland. We show that estimating the decay parameters improves in-sample performance of the model and that the out-of-sample performance of our best model is strong, suggesting that our findings can be generalized to the population of interest.

**Keywords:** *multilevel network; social network; $p^\star$-model; exponential-family random graph model; curved exponential-family random graph model.*

[a] Department of Statistics, Rice University, 6100 Main St, Houston, TX 77005, USA; email address: jrs6@rice.edu. Jonathan Stewart was partially supported by NSF awards DMS-1513644 and DMS-1812119 from the National Science Foundation.

[b] Corresponding author; Department of Statistics, Rice University, 6100 Main St, Houston, TX 77005, USA; email address: michael.schweinberger@rice.edu. Michael Schweinberger was partially supported by NSF awards DMS-1513644 and DMS-1812119 from the National Science Foundation.

[c] Department of Quantitative Methods & Information Technology, Kozminski University, 57/59 Jagiellonska St, 03-301 Warsaw, Poland; email address: mbojanowski@kozminski.edu.pl.

[d] Department of Sociology, Department of Statistics, University of Washington, Seattle, WA 98195-4322, USA; email address: morrism@u.washington.edu.

# 1 Introduction

Exponential-family random graph models (ERGMs) or $p^\star$-models (Wasserman & Pattison, 1996) have become one of the dominant statistical methods for analyzing social networks (Wasserman & Faust, 1994; Kolaczyk, 2009), as evidenced by a growing body of research articles, books (Lusher *et al.*, 2013; Harris, 2013), and software.[1] When properly specified, ERGMs can be used to investigate a wide range of network processes, both dyadic independent (e.g., degree heterogeneity and homophily by nodal attributes) and dyadic dependent (e.g., cyclical and transitive triadic closure).

Triadic processes, in particular, have long been of interest in social network analysis (Heider, 1946; Cartwright & Harary, 1956; Wasserman & Faust, 1994). Early studies used methods from mathematical graph theory to examine the relative frequencies of triadic configurations (see, e.g., the so-called triad census of Holland & Leinhardt, 1970). That work led to some of the foundational theories of social network analysis: that regularities in triadic configurations at the micro-level cumulate up to signature patterns at the macro-level, such as clustering and polarization (Rapoport, 1963). So when the first statistical models with dyadic dependence induced by counts of triadic configurations were proposed – the Markov random graphs of Frank & Strauss (1986) – it was natural that applied research focused on model specifications that used counts of triadic configurations to explain the clustering observed in empirical networks. But those specifications turned out to be problematic. It took some time to understand why, and to appreciate how (and how not) to represent dyadic dependence induced by triadic processes in statistical models. Strauss (1986) first observed that dyadic dependence induced by 2-star and triangle counts in combination with strong homogeneity assumptions (Frank & Strauss, 1986) leads to near-degenerate models, placing most probability mass on networks with almost no edges or almost all possible edges (Jonasson, 1999; Handcock, 2003; Schweinberger, 2011; Butts, 2011; Chatterjee & Diaconis, 2013).

That work led eventually to a better understanding of why the simple homogenous Markov specifications do not behave as expected, and to the development of more appropriate, parsimonious specifications of dyadic dependence in ERGMs. The most widely used of the new specifications are curved terms such as alternating $k$-stars and $k$-triangles (Snijders *et al.*, 2006) or, equivalently, geometrically weighted degrees and triads (Hunter & Handcock, 2006; Hunter, 2007; Hunter *et al.*, 2008).

[1] The publicly available software for ERGMs includes 19 `R` packages found at `CRAN.R-project.org` (in alphabetical order, `Bergm`, `blkergm`, `btergm`, `dnr`, `EpiModel`, `ergm`, `ergm.count`, `ergm.ego`, `ergm.graphlets`, `ergm.rank`, `ergm.userterms`, `ergmharris`, `fergm`, `GERGM`, `gwdegree`, `hergm`, `statnetWeb`, `tergm`, `xergm`), and the program `pnet` (Wang *et al.*, 2006).

## 1.1 Curved ERGMs with geometrically weighted terms

The promise of curved ERGMs with geometrically weighted terms was first demonstrated in the papers of Snijders *et al.* (2006), Hunter & Handcock (2006), Hunter (2007), and Hunter *et al.* (2008). Expressed in terms of sequences of degree and shared partner counts, curved terms are weighted sums of those sequences, where the weights decrease geometrically, as governed by a decay parameter. The homogenous Markov random graph terms based on the $k$-star and triangle counts in Frank & Strauss (1986) imply that each additional $k$-star and triangle configuration has the same influence on the log odds of the conditional probability of an edge. By contrast, the geometrically weighted terms imply declining marginal influence, where the rate of decline is controlled by the decay parameter. This term is then multiplied by the usual coefficient, which in this context is often called the "base parameter." Geometrically weighted terms give rise to curved exponential families of distributions in the sense of Efron (1975), therefore such ERGMs are known as curved ERGMs (Hunter & Handcock, 2006; Hunter, 2007). A growing body of applied research has demonstrated the usefulness of these curved ERGMs (see, e.g., Lusher *et al.*, 2013; Harris, 2013, and references therein). That said, some statistical challenges have emerged.

## 1.2 Statistical inference for curved ERGMs

While geometrically weighted terms are attractive on scientific grounds and better behaved in practice, estimating the decay parameters of these terms from a single network by maximum likelihood methods (Hunter & Handcock, 2006) or Bayesian methods (Koskinen, 2004; Caimo & Friel, 2011; Everitt, 2012; Bomiriya *et al.*, 2016) has proven to be difficult.

The seminal paper of Snijders *et al.* (2006), which introduced alternating $k$-star and $k$-triangle terms and a version of the geometrically weighted degree term, applied a curved ERGM to the Lazega law firm advice network (Lazega, 2001). Snijders *et al.* did not estimate the decay parameters, but fixed them at values found by trial and error. Hunter & Handcock (2006) introduced Monte Carlo maximum likelihood methods to estimate decay parameters and were able to estimate the decay parameters of some geometrically weighted model terms using the same law firm advice network, but conditioned on the observed number of edges (as did Obando & De Vico Fallani, 2017). We were only able to find four published papers that estimated decay parameters of geometrically weighted model terms without conditioning on the observed number of edges (Hunter, 2007; Koskinen *et al.*, 2010; Suesse, 2012; Almquist & Bagozzi, 2015). Three of them used the same network, the Lazega law firm advice network (Hunter, 2007; Koskinen *et al.*, 2010; Suesse, 2012).

Both of the heuristic approaches to using curved ERGMs in practice – fixing the decay parameters at values found by trial and error or conditioning on the observed number of edges – are undesirable. Fixing decay parameters at values other than the maximum likelihood estimates (MLEs) will change the estimates for all of the other model parameters, and can negatively affect both the in-sample and the out-of-sample performance of the model. Conditioning on the number of edges in the observed network also imposes a steep cost, as it limits statistical inference to networks with the same number of edges.

One reason that the estimation of the decay parameter is so challenging is that geometrically weighted terms are nonlinear functions of the product of the base and decay parameters (Hunter, 2007). As such, these two parameters are "mixed up," and difficult to estimate. In theory, estimation of both parameters is possible: well-specified models are identifiable and sensitive to changes in all parameters as long as the base parameters are not zero and the network contains at least four nodes. However, even well-specified models are less sensitive to changes in decay parameters when the base parameters are small or the decay parameters are large. As a consequence, a network may not contain much information about decay parameters (in the statistical sense of Fisher information), making it challenging to estimate them.

## 1.3 Multilevel network data facilitate statistical inference for curved ERGMs

The increasing availability of multilevel network data (e.g., Wang *et al.*, 2013; Zappa & Lomi, 2015; Lomi *et al.*, 2016; Slaughter & Koehly, 2016; Hollway & Koskinen, 2016; Lazega & Snijders, 2016; Hollway *et al.*, 2017) provides new opportunities to strengthen statistical inference for curved ERGMs. Multilevel network data come in many forms. Snijders (2016) presents a representative sample of the diverse forms that multilevel network structure can assume. Among the multitude of multilevel network structures, two basic forms of multilevel networks can be distinguished: multiple networks (e.g., multiple school networks) and multilevel networks with ties within and between two sets of nodes (e.g., a set of students and a set of school classes in a school). We consider here a simple example that combines both flavors of multilevel networks: we have multiple school networks and, within each school, we have students (level-1 units) nested in school classes (level-2 units), with ties among students within and between school classes—although in the multilevel network we will use the between-class ties are unobserved by the data collection design. Such data can be used to strengthen statistical inference for curved ERGMs in at least three ways.

First, multilevel networks help estimate decay parameters of geometrically weighted terms by providing replication. In the running example, if we assume that the network in each school class is generated by a curved ERGM with a size-adjusted parameterization (Krivitsky *et al.*, 2011; Krivitsky & Kolaczyk, 2015), then the sample of networks comprises replications from the same data-generating process. The replication provides additional information (in the statistical sense of Fisher information) that improves estimation of all of the parameters in a model. Recent advances in the statistical theory of ERGMs have shown that the MLEs of parameters, including the decay parameters of geometrically weighted terms, exist and are close to the data-generating values of the parameters with high probability, provided a large multilevel network consists of many networks of similar sizes (Schweinberger & Stewart, 2018). In practice, estimation from multilevel networks can reduce standard errors of maximum likelihood estimators and the posterior uncertainty in Bayesian approaches to ERGMs (Koskinen, 2004; Caimo & Friel, 2011; Everitt, 2012; Bomiriya *et al.*, 2016).

Second, multilevel networks can have computational advantages. This is especially true in our running example, where the edges within school classes do not depend on edges outside of school classes. In this case, the probability mass function of a multilevel network factorizes into class-dependent probability mass functions. The factorization implies that the within- and between-class contributions to the likelihood function can be computed separately, which allows them to be performed in parallel on multi-core computers or computing clusters.

Third, multilevel networks make it possible to assess the out-of-sample performance of ERGMs via cross-validation: the replicates can be split into two subsets, a training subset used to estimate the model, and a held-out subset used to assess the out-of-sample performance of the estimated model. It is worth noting that the assessment of out-of-sample performance serves a different purpose than the traditional assessment of goodness-of-fit (Hunter *et al.*, 2008). Goodness-of-fit checks assess in-sample performance: how well an estimated model reproduces other features of the same observed network that was used to estimate the model. By contrast, cross-validation assesses out-of-sample performance: how well the estimated model predicts features of networks that were not used to estimate the model. As a consequence, cross-validation helps strengthen the basis for sample-to-population inference.

## 1.4 Purpose of our paper

We demonstrate the advantages outlined in Section 1.3 by estimating a set of curved ERGMs from a multilevel network consisting of 304 third-grade school classes with 6,594 students, sampled from a population with 309,285 third-grade students in Poland (Dolata, 2014). Our primary focus is on geometrically weighted triadic closure terms for directed networks (Butts, 2008; Robins *et al.*, 2009). We compare the results from a model that fixes the decay parameter at two values (0 and .25) commonly used in practice (e.g., Hunter *et al.*, 2008; Goodreau *et al.*, 2009; Hunter *et al.*, 2012), to the results from the same model when the decay parameter is estimated. In addition, we explore four other alternative specifications of directed geometrically weighted triadic closure terms, capturing different forms of cyclical and transitive closure (Wasserman & Faust, 1994). All of the models use size-adjusted parameterizations for the density and reciprocity terms (Krivitsky *et al.*, 2011; Krivitsky & Kolaczyk, 2015; Butts &

Almquist, 2015). We assess the performance of the models in three ways: convergence properties, in-sample performance (goodness-of-fit), and out-of-sample performance (cross-validation).

Our findings show that the convergence properties of all curved ERGMs are excellent, and that the in-sample performance of curved ERGMs is superior when decay parameters are estimated rather than fixed. In addition, the best-fitting curved ERGM shows strong out-of-sample performance, which suggests that our findings can be generalized to the population interest.

A software implementation of the proposed models and methods is available in the `R` package `hergm` (Schweinberger & Luna, 2018), which depends on `R` package `ergm` (Hunter *et al.*, 2008). The package supports parallel computing on multi-processor computers and computing clusters.

## 1.5 Comparison with existing approaches

There is a growing body of research articles and books concerned with multilevel network data, models, and methods (e.g., Wang *et al.*, 2013; Zappa & Lomi, 2015; Lomi *et al.*, 2016; Slaughter & Koehly, 2016; Hollway & Koskinen, 2016; Lazega & Snijders, 2016; Hollway *et al.*, 2017). For the type of multilevel network considered here, existing approaches include

- pooling the network data and estimating a common model, without adjusting for network size (e.g., Kalish & Luria, 2013). That assumes that the coefficients are the same for all networks and ignores the potential impact of network size.

- estimating a model from each network separately (e.g., Hunter *et al.*, 2008; Goodreau *et al.*, 2009). That allows coefficients to vary from network to network, but does not pool information across networks to facilitate the estimation of the decay parameters of curved ERGMs. While the separate estimates can be combined into a single estimate by using meta-analysis (Lubbers, 2003; Lubbers & Snijders, 2007), estimating decay parameters from each network separately does not pool information across networks and is challenging for the reasons discussed above (Section 1.2).

- Bayesian approaches (e.g., Schweinberger & Handcock, 2015; Slaughter & Koehly, 2016) that assume the coefficients are random variables with common mean and variance. While flexible, existing Bayesian methods are associated with high computational costs.

None of these existing approaches have dealt with the problem of missing data.

By contrast, we

- pool the network data and estimate a common model, adjusting for network size by using methods proposed by Krivitsky *et al.* (2011) and Krivitsky & Kolaczyk (2015): that is, we assume that coefficients are functions of size-invariant parameters and size-dependent offsets.

- exploit the strength of the pooled network data to estimate the decay parameters of curved ERGMs, and increase the precision of other estimators, while keeping the model parsimonious and computations feasible for networks with thousands of nodes.

- distinguish between the process that generates the population network and the process that determines which network data are observed (Schweinberger *et al.*, 2017).

- incorporate modern missing-data methods for statistical network analysis, assuming that missing responses are ignorable as defined by Handcock & Gile (2010) and Koskinen *et al.* (2010).

- use out-of-sample prediction assessment to assess sample-to-population inference.

- provide a careful substantive interpretation of the key coefficients in these curved ERGMs.

To compare our work to the only four papers that estimated decay parameters without conditioning on the observed number of edges (Hunter, 2007; Koskinen *et al.*, 2010; Suesse, 2012; Almquist & Bagozzi, 2015), we note that all of them focus on a single triadic closure term (GW-ESP) for undirected networks, are based on a single network without sampled or missing data, one network with 36 nodes (Hunter, 2007; Koskinen *et al.*, 2010; Suesse, 2012) and the other network with 143 nodes (Almquist & Bagozzi, 2015). By contrast, we estimate the decay parameters of five triadic closure terms for directed networks on a sample of networks with missing data, and use out-of-sample performance assessment to justify inference to the population of third-grade class networks in Poland.

The remainder of our paper is structured as follows. We describe the population network of interest and the sampled network data in Section 2. A population network model is introduced in Section 3 and likelihood-based inference for the population network model is discussed in Section 4. We present the results in Section 5.

## 2 Population network and sampled network data

The data we use are sampled multilevel network data collected by the Polish Institute for Educational Research[2] as a part of the study "Quality and Efficiency of Education and Institutionalization of Research Facilities" (Dolata & Rycielski, 2014).

The population consists of all third-grade classes in 8,924 Polish primary schools during academic year 2010/2011. A total of 309,285 students attended third grade that year. A two-stage sampling design was used to generate a sample of school classes from the population. In the first stage, a stratified sample of 176 schools was generated, with strata defined by city size and the number of third-grade school classes. More details on the stratified cluster sampling design can be found in Maluchnik & Modzelewski (2014). In the second stage, 306 third-grade school classes were sampled from the 176 schools. If the school had one or two third-grade school classes, all were included. If the school had three or more third-grade school classes, two were selected by simple random sampling without replacement.

The study sought to interview all 6,607 students in the sampled school classes by in-class surveys, however parental consent was required for students to participate (Maluchnik & Modzelewski, 2014). Interview data were collected from 5,625 students (85%). The data from the remaining students are missing due to a combination of missing parental consent, absence on the day of the survey, and inadmissible or garbled responses. Participating students could still nominate students who did not participate, so the data set contains information on more students than participants. We removed the two smallest classes with 6 and 7 students because of the small sizes. The resulting data set used in this analysis is based on 5,612 interviews from 304 sampled school classes and provides information on 6,594 students.

Figure 1 shows the distribution of the sizes of the 304 sampled school classes and the percentages of students with missing data in each class. Class sizes range from 11 to 33, with a median of 22. Missingness ranges from 0 to 45%, with a median of 13%. There are 44 school classes (14%) without missing data.

The network data consist of directed edges from student $i$ to student $j$, where a directed edge indicates that student $i$ expressed interest in playing with student $j$. The name generator was: "Name people from your class that you would most like to play with" (translated from Polish). Nominations were restricted to other students in the same school class, so the data do not contain observations of between-class edges. In addition to the network data, two nodal attributes were collected from school records: the sex of students and the International Socio-Economic Index (ISEI) of parents. Due to high levels of missingness, we do not use the ISEI of parents in our analyses.

The observed outdegree distribution is shown in Figure 2 and reveals a notable spike at 5. While there was no upper bound on the number of nominations allowed,

[2] Instytut Badań Edukacyjnych, www.ibe.edu.pl.

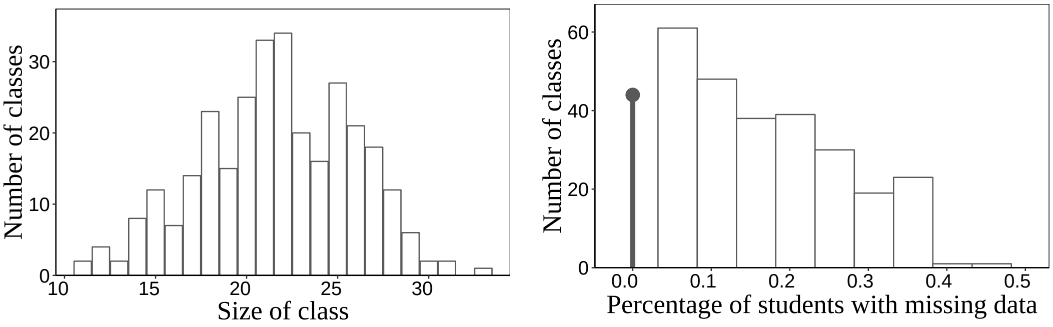


Figure 1: Left: Size distribution of sizes of the 304 school classes. Right: Distribution of the percentage of students with missing data in each school. The vertical bar at 0 shows the 44 school classes without missing data.

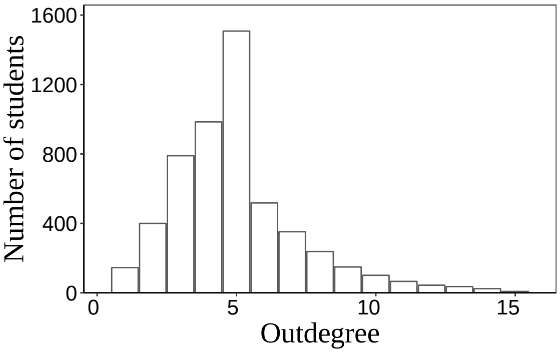


Figure 2: Observed outdegrees of students in the 304 school classes.

the questionnaire provided 5 lines for nominating playmates. It seems likely that some students interpreted the 5 lines as a limit on the number of nominations, while others did not. This has implications for modeling outdegrees, which we discuss in Section 3.2.

The mean outdegree and indegree of male students, computed from the 44 classes without missing data, are 4.61 and 4.83, respectively; for female students, the mean outdegree and indegree are 5.28 and 5.04, respectively. Table 1 shows the distribution of nominations by female and male students, based on the 44 classes without missing data.

# 3 Population network model

The true population of interest consists of all students in third-grade school classes in Poland. In this population, there may be edges both within and between school

| | | Receiver | | Total | |
|---|---|---|---|---|---|
| | | Male | Female | Ties | Students |
| Sender | Male | 1782 | 333 | 2115 | 459 |
| | Female | 437 | 1921 | 2358 | 447 |
| Total | Ties | 2219 | 2254 | 4473 | |
| | Students | 459 | 447 | | 906 |

Table 1: Distribution of nominations by female and male students. The counts are the total number of edges in each category across the 44 school classes without missing data.

classes, and both may be of scientific interest. The modeling framework we present here is capable of modeling both within- and between-class edges, provided data on both are available. To clarify which assumptions our model makes and under which conditions our model-based conclusions hold, we specify the general form here. When we turn to our application, the lack of data on between-class edges will constrain the model specification to a more limited form.

Let $X_{i,j} = 1$ if student $i$ expressed interest in playing with student $j$ and let $X_{i,j} = 0$ otherwise, and denote by $\mathcal{A}_k$ the set of all students in school class $k = 1, \ldots, K$. We denote the within-class networks by $\mathbf{X}_k = (X_{i,j})_{i \in \mathcal{A}_k,\, j \in \mathcal{A}_k}$, the between-class networks by $\mathbf{X}_{k,l} = (X_{i,j})_{i \in \mathcal{A}_k,\, j \in \mathcal{A}_l}$ $(k \neq l)$, and the population network by $\mathbf{X} = (\mathbf{X}_{k,l})_{k,l}^{K}$.

We assume that the population network $\mathbf{X}$ was generated by a random graph model with a probability mass function of the form

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) \quad = \quad \left[\prod_{k=1}^{K} \mathbb{P}(\mathbf{X}_k = \mathbf{x}_k)\right] \, \mathbb{P}(\mathbf{X}_{k,l} = \mathbf{x}_{k,l},\; k \neq l = 1, \ldots, K).$$

The population network model therefore makes two fundamental model assumptions:

- The within-class edges of students can depend on other edges among students in the same school class, but do not depend on edges to students outside of the school class.

- The between-class edges of students can depend on other between-class edges, but do not depend on within-class edges.

While the lack of data on between-class edges means that we cannot learn the probability law governing between-class edges (unless we make the unrealistic assumption that within- and between-class edges are governed by the same probability law), we can use our model to learn the probability law governing the within-class networks of the population network. In particular, we can use our model to examine whether playing preferences in the population of third-grade students in Poland show evidence of reciprocity, heterogeneity and homophily by sex, and triadic closure of different types (Wasserman & Faust, 1994).

## 3.1 Model specification

We focus here on the specification of within-class models, since we do not have data on between-class edges.

We assume that the within-class models are ERGMs with probability mass functions of the form

$$\mathbb{P}_\theta(\mathbf{X}_k = \mathbf{x}_k) \;\; = \;\; \exp\left(\sum_{i=1}^{p} \eta_{k,i}(\theta)\, s_{k,i}(\mathbf{x}_k) - \psi_k(\theta)\right), \quad k = 1, \dots, K,$$

where $s_{k,i} : \mathbb{X}_k \mapsto \mathbb{R}$ are network features of within-class network $\mathbf{x}_k \in \mathbb{X}_k$ and $\eta_{k,i} : \Theta \mapsto \mathbb{R}$ are the weights of the network features, called the natural parameters of the exponential family. The natural parameters $\eta_{k,i} : \Theta \mapsto \mathbb{R}$ may depend on the sizes of school classes and may be non-linear functions of a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^q$, which is the case in curved ERGMs with geometrically weighted terms. The function $\psi_k(\theta)$ ensures that the probability mass function $\mathbb{P}_\theta(\mathbf{X}_k = \mathbf{x}_k)$ sums to 1.

We start with a description of size-adjusted parameterizations for edges and mutual edges in Section 3.2 and discuss modeling outdegrees in Section 3.3. We then turn to the model terms of primary interest: heterogeneity and homophily by sex terms in Section 3.4 and triadic closure terms in Section 3.5, based on curved ERGMs with geometrically weighted terms. A graphical summary of all model terms is shown in Figures 4 and 5 below.

## 3.2 Size-adjusted parameterizations

The sizes of the sampled school classes described in Section 2 range from 11 to 33. If network density changes with network size, this has implications for model specification. The issue is related to density-dependence versus frequency-dependence in the ecology literature (e.g., DeBenedictis, 1977), and sparse versus dense graphs in mathematical graph limit theory (e.g., Chatterjee & Diaconis, 2013).

Consider an undirected Bernoulli random graph, which is equivalent to an ERGM with the number of edges as sufficient statistic and natural parameter $\eta(\theta) = \theta$. Here, $\theta$ is the log odds of the probability of an edge. Holding $\theta$ constant as the network size increases preserves the probability of an edge – i.e., the expected network density – but increases the expected degrees of nodes by a factor proportional to the change in network size. Thus, increasing network size by a factor of 10 would result in nodes having, on average, 10 times more edges. That is equivalent to the density-dependence

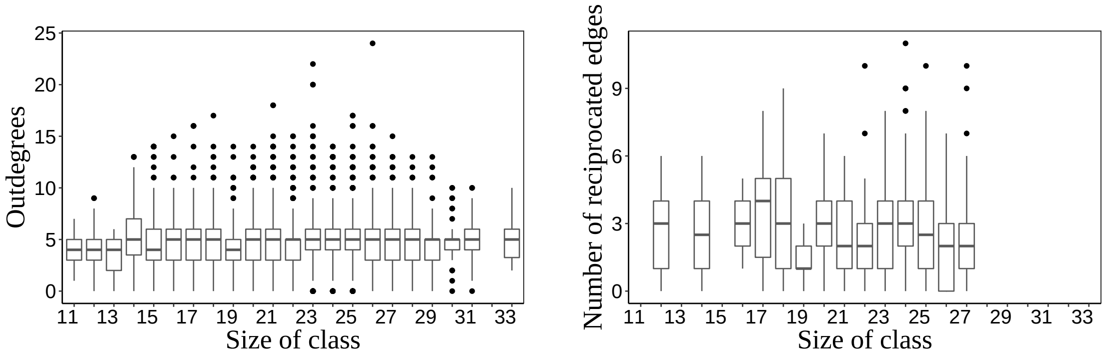


Figure 3: Left: Boxplots of the observed outdegrees of students in the 304 school classes. Right: Boxplots of the observed number of reciprocated edges in the 44 school classes without missing data.

assumption in the ecology literature, and the dense-graph regime in graph limit theory (Lovász, 2012).

Constant expected network density may be a reasonable assumption for the growth process in some non-social networks, and some of the mathematical-statistical work on ERGMs makes this assumption (e.g., Chatterjee & Diaconis, 2013). In the social science literature, however, it has long been recognized that constant network density is an unrealistic assumption for most social networks (Mayhew & Levinger, 1976). People do not have infinite resources for engaging with others and it is therefore more credible that, as the network size increases, the expected degrees of nodes are either constant or bounded above (Krivitsky *et al.*, 2011; Krivitsky & Kolaczyk, 2015; Butts & Almquist, 2015). That is equivalent to the frequency-dependence assumption in the ecology literature and the sparse-graph regime in graph limit theory (Lovász, 2012).

As shown in Figure 3, our data are consistent with the assumption that the expected degrees are either constant or bounded above: the median observed outdegree lies between 4 and 5 for sampled school classes of all sizes. That may partly reflect the fact that the questionnaire, while not limiting nominations, provided 5 lines, as discussed in Section 2. However, the outdegrees of the students who made more than 5 nominations do not appear to increase with network size either, suggesting that the expected degrees of all students are network size-invariant.

There is a small but growing body of work focused on developing size-invariant parameterizations for ERGMs (Krivitsky *et al.*, 2011; Krivitsky & Kolaczyk, 2015; Butts & Almquist, 2015). The assumption that the expected mean degree, rather than the expected network density, should be size-invariant leads to a per capita scaling adjustment, where the expected number of edges scales linearly, rather than quadratically, with the number of nodes. As proposed in Krivitsky *et al.* (2011), ERGMs can achieve size-invariance of expected mean degree by including a size-dependent offset. In the undirected Bernoulli random graph model, for example, the size-adjusted specification includes a size-dependent offset of $-\log|\mathcal{A}|$, where $|\mathcal{A}|$ denotes the number of nodes in $\mathcal{A}$:

$$\eta_1(\theta) \quad = \quad \theta_1 - \log|\mathcal{A}|. \tag{1}$$

Here, $\theta_1 \in \mathbb{R}$ is a size-invariant parameter that does not depend on the size of $\mathcal{A}$. Krivitsky *et al.* (2011) showed that for Bernoulli random graphs with parameterizations of the form (1), the expected mean degree is constant in the limit as the number of nodes increases without bound, and that the size-invariant parameter $\exp(\theta_1)$ can be interpreted as the limiting expected mean degree. This simple interpretation of $\exp(\theta_1)$ in terms of expected mean degree will change once other terms are added to the model, but the size-invariance of the expected mean degree will still be preserved. Krivitsky *et al.* (2011) showed that the size-dependent offset $-\log|\mathcal{A}|$ provides per capita scaling for all dyadic independence terms, including degree heterogeneity and homophily by nodal attributes.

In directed networks, a natural hypothesis is that a constant fraction of edges will be reciprocated. This implies the number of mutual edges will scale with the number of edges rather than the number of possible edges, and the expected number of reciprocated edges per student should not increase with network size. Again, our data are consistent with this invariance assumption. Figure 3 shows the observed number of mutual edges in the 44 school classes without missing data does not increase with class size.

If a mutual edge term with a size-invariant natural parameter is added to a model to capture the reciprocity effect, along with an edge term with a size-dependent natural parameter of the form (1), then the penalty imposed by the size-dependent offset $-\log|\mathcal{A}|$ implies that the reciprocity effect vanishes in the limit as the number of nodes increases without bound (Krivitsky & Kolaczyk, 2015). To prevent this, Krivitsky & Kolaczyk proposed to adjust the natural parameter of the mutual edge term by adding the size-dependent offset $\log|\mathcal{A}|$ in order to cancel the penalty:

$$\eta_2(\theta) \quad = \quad \theta_2 + \log\,|\mathcal{A}|,$$

where $\theta_2 \in \mathbb{R}$ is the size-invariant reciprocity parameter. A model with size-adjusted edge and mutual edge terms implies that the log odds of the conditional probability of $X_{i,j} = 1$ given the rest of the network $\mathbf{X}_{-(i,j)} = \mathbf{x}_{-(i,j)}$ has the form:

$$\log \frac{\mathbb{P}(X_{i,j} = 1 \mid \mathbf{X}_{-(i,j)} = \mathbf{x}_{-(i,j)})}{\mathbb{P}(X_{i,j} = 0 \mid \mathbf{X}_{-(i,j)} = \mathbf{x}_{-(i,j)})} \quad = \quad \begin{cases} \theta_1 - \log|\mathcal{A}| & \text{if } X_{j,i} = 0 \\ \theta_1 + \theta_2 & \text{if } X_{j,i} = 1, \end{cases}$$

where $\mathbf{X}_{-(i,j)}$ refers to the network $\mathbf{X}$ excluding $X_{i,j}$.

We use such size-adjusted edge and mutual edge terms in our network population model, with $\log |\mathcal{A}_k|$ for each of the school classes $\mathcal{A}_k$. Note that we are here not interested in the asymptotic properties of size-adjusted parameterizations – such as the asymptotic mean degree – and that the asymptotic properties can change when dyadic dependence terms are added to the model. We are concerned with small school classes with 11 to 33 students, so asymptotic properties based on school classes with infinite numbers of students are neither interesting nor relevant. We use size-adjusted parameterizations to allow school classes of different sizes to have different edge and mutual edge coefficients.

## 3.3 Outdegree terms

We noted in Section 2 that the observed outdegree distribution shows a sharp spike at 5, which is likely to be an artifact of the questionnaire design. The spike is not captured by conventional approaches to modeling outdegrees: the traditional edge count term produces a Poisson-like distribution without a spike, and a geometrically weighted outdegree term does not reproduce the observed distribution either. We explored both approaches and found that neither of them captures the outdegree distribution. We therefore model the outdegrees by using outdegree terms of the form $\theta_{2+l} \sum_{i=1}^{|\mathcal{A}_k|} \mathbb{1}(\sum_{j \in \mathcal{A}_k \,:\, j \neq i} x_{i,j} = l)$ for outdegrees $l = 1, \ldots, 6$. These terms ensure that the model reproduces, on average, the observed outdegrees 1 through 6, as confirmed by the goodness-of-fit assessment in Appendix D. Note that a model with outdegree 5 term but without the other outdegree terms would be more parsimonious and would capture the spike at outdegree 5, but we found that the resulting model fails to capture the rest of the outdegree distribution. We therefore include outdegree $1, \ldots, 6$ terms. The tail of the outdegree distribution is determined by the other model terms, and looks Poisson in our application.

Last, but not least, it is worth noting that the number of nodes with outdegree $k$ should not be confused with the number of $k$-out-stars, $k = 1, \ldots, 6$: e.g., the number of nodes with outdegree 2 is a number between 0 and $n$, whereas the number of 2-out-stars is a number between 0 and $n \binom{n-1}{2} \approx n^3/2$. The number of 2-out-stars can be much larger than the number of edges, which is at most $n\,(n-1) \approx n^2$. As a consequence, 2-out-star terms can overwhelm edge terms, leading to near-degenerate models that concentrate probability mass on networks with almost no 2-out-stars or almost all possible 2-out-stars (Handcock, 2003; Schweinberger, 2011; Butts, 2011; Chatterjee & Diaconis, 2013). By contrast, the outdegree terms we use cannot overwhelm edge terms, making them well-behaved alternatives.

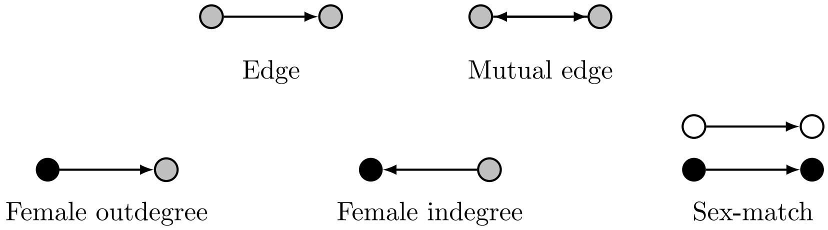


Figure 4: Graphical representations of the network features that are used as sufficient statistics in all models. Circles represent students, directed lines with one direction represent directed edges, and directed lines with two directions represent mutual edges. Black-colored circles represent female students, white-colored circles represent male students, and gray circles represent either female or male students.

## 3.4 Nodal attribute terms

We assess the influence of students' sex on degree heterogeneity and homophily with the following terms:

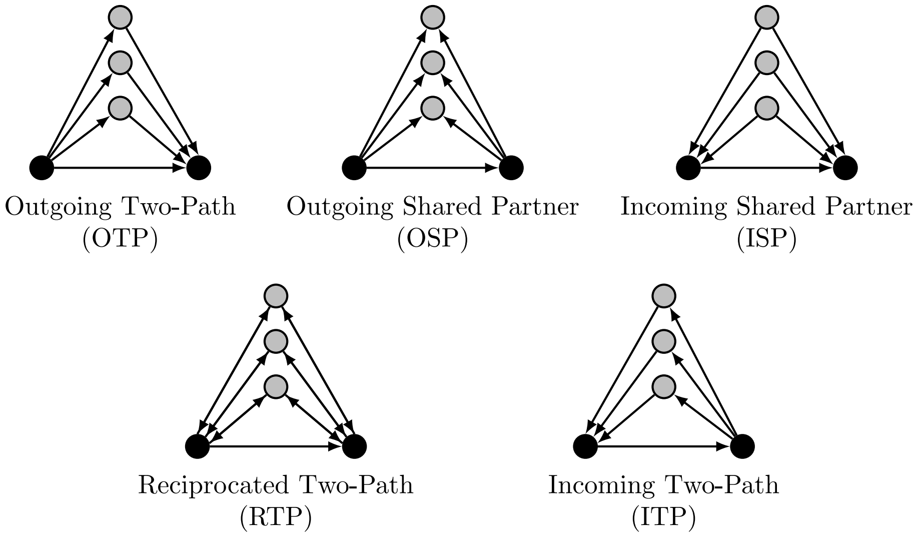
- A sex-specific outdegree term (female outdegree) of the form $\theta_9 \sum_{i \in \mathcal{A}_k,\, j \in \mathcal{A}_k} x_{i,j}\; c_i$.
- A sex-specific indegree term (female indegree) of the form $\theta_{10} \sum_{i \in \mathcal{A}_k,\, j \in \mathcal{A}_k} x_{i,j}\; c_j$.
- A sex homophily term (sex-match) of the form $\theta_{11} \sum_{i \in \mathcal{A}_k,\, j \in \mathcal{A}_k} x_{i,j}\; \mathbb{1}(c_i = c_j)$.

Here, $c_i$ is an indicator that is 1 if student $i$ is female and is 0 otherwise, and $\mathbb{1}(c_i = c_j)$ is an indicator that is 1 if the sex of students $i$ and $j$ matches and is 0 otherwise.

Note that we do not include indegree terms (other than the female indegree term), because the model without indegree term is more parsimonious and the in-sample and out-of-sample performance of models without indegree terms turns out to be excellent, as shown in Sections 5.4 and 5.5.

## 3.5 Triadic closure terms

To capture triadic closure in social networks, we use geometrically weighted (GW) terms based on counts of the following configurations (Butts, 2008; Robins *et al.*, 2009): outgoing two-path (OTP), outgoing shared partner (OSP), incoming shared partner (ISP), reciprocated two-path (RTP), and incoming two-path (ITP). We follow here the naming convention of Butts (2008); the same configurations with different names are used in Robins *et al.* (2009) using alternating $k$-triangle parameterizations. Graphical representations of these configurations are provided in Figure 5.



Outgoing Two-Path (OTP)

Outgoing Shared Partner (OSP)

Incoming Shared Partner (ISP)

Reciprocated Two-Path (RTP)

Incoming Two-Path (ITP)

Figure 5: Graphical representations of the triadic closure configurations used to construct GW statistics of the form $\sum_{i \in \mathcal{A}_k \neq j \in \mathcal{A}_k} x_{i,j} \;\; \mathbb{1}\left(\mathrm{T}_{\text{type}}(i,j) = m\right)$ $(m = 1, \ldots, |\mathcal{A}_k| - 2)$. The plots show black-colored pairs of nodes with $m = 3$ configurations of the specified type, where directed lines with one direction represent directed edges and directed lines with two directions represent mutual edges.

These five configurations capture different forms of cyclical and transitive closure in social networks. Their relative frequencies play an important role in the global structure of social networks, because transitive triads are the basic building blocks of hierarchical structure, while cyclical triads produce more egalitarian systems (Chase, 1980).

The first three, OTP, OSP, and ISP, capture purely transitive closure. All are based on the 030$T$ configuration in the triad census of Holland & Leinhardt (1970); each closes one of the three legs of that triad, but represents a distinct social process. Closing the OTP leg is the classic “a friend of my friend is my friend” effect; the OSP leg means that if we both nominate the same person as a friend, then one of us will nominate the other as a friend; and the ISP leg means that if the same person nominates both of us as a friend, then one of us will nominate the other as a friend. By contrast, ITP captures purely cyclical closure and RTP captures both forms of closure. In addition, the RTP term captures reciprocity, and may hence be useful for studying the interaction of reciprocity with cyclical and transitive closure in the ERGM framework, as Block (2015) did in the stochastic actor-oriented modeling framework (Snijders, 2001).

The GW terms for these triadic closure configurations are based on sufficient statistics that count the number of pairs of nodes with $m$ configurations of the specified type, within each school class $\mathcal{A}_k$ $(k = 1, \ldots, K)$:

$$s_{k,11+m}(\mathbf{x}_k) \;=\; \sum_{i \in \mathcal{A}_k \,\neq\, j \in \mathcal{A}_k} x_{i,j}\; \mathbb{1}\left(\mathrm{T}_{\text{type}}(i,j) \;=\; m\right), \quad m = 1, \ldots, |\mathcal{A}_k| - 2,$$

where $\mathrm{T}_{\text{type}}(i,j)$ counts the number of configurations of the specified type and $\mathbb{1}(\mathrm{T}_{\text{type}}(i,j) \;=\; m)$ is an indicator function, which is 1 if students $i$ and $j$ have $m$ configurations of the specified type in school class $\mathcal{A}_k$ and is 0 otherwise.

For each type of GW term, the natural parameters are given by

$$\eta_{k,11+m}(\theta) \;=\; \theta_{12}\, \exp(\alpha) \left[1 - (1 - \exp(-\alpha))^m\right], \quad m = 1, \ldots, |\mathcal{A}_k| - 2,$$

where $\theta_{12}$ is called the base parameter and $\alpha > 0$ is called the decay parameter. The motivation for these parameterizations is explained in the seminal papers of Snijders *et al.* (2006), Hunter & Handcock (2006), and Hunter (2007). As explained in Section 1.2, GW terms with $\theta_{12} > 0$ and $\alpha > 0$ ensure the value of each additional configuration of this type is positive but declining. We demonstrate that in Section 5.3.3 below.

An interesting special case of the GW-OTP arises when $\alpha = 0$. The term then reduces to a simpler form, called a transitive edge term, with sufficient statistic

$$s_{k,12}(\mathbf{x}_k) \;=\; \sum_{i \in \mathcal{A}_k \,\neq\, j \in \mathcal{A}_k} x_{i,j} \max_{h \in \mathcal{A}_k,\, h \,\neq\, i,j} x_{i,h}\, x_{h,j}$$

and natural parameter

$$\eta_{k,12}(\theta) \;=\; \theta_{12}.$$

Transitive edge terms differ from the triangle terms of Frank & Strauss (1986) by counting only the first triangle in which two nodes are involved. They are less prone to degeneracy and have turned out to be useful for capturing transitive closure in practice (e.g., Snijders *et al.*, 2010; Krivitsky, 2012; Hunter *et al.*, 2012). The assumption that $\alpha = 0$ is quite strong, however, and provides a useful comparison for the model where the decay parameter $\alpha$ is unrestricted, so we include it in Section 5.

# 4 Likelihood-based inference for population network models

To infer the probability law governing the within-class networks of the population network, we use likelihood-based inference.

To state the likelihood, let $\mathcal{S} \subseteq \{1, \ldots, K\}$ be the set of indices of the sampled school classes and let $u_{i,j} = 1$ if $x_{i,j}$ is unobserved and $u_{i,j} = 0$ if $x_{i,j}$ is observed. Note that $u_{i,j} = 1$ can occur in any of the following situations:

1. Students $i$ and $j$ were members of different school classes, and therefore $x_{i,j}$ is unobserved by the sampling design.
2. Students $i$ and $j$ were in the same school class, but the school class was not sampled.
3. Students $i$ and $j$ were in the same school class and the school class was sampled, but the response of student $i$ was not observed due to missing parental consent or an inadmissible response by student $i$.

More details on the sampling design and the missing data can be found in Section 2.

The likelihood is thus proportional to

$$\begin{aligned} L(\theta) &\propto \sum_{\substack{x_{i,j} \in \{0,1\} \\ \textit{for all } (i,j) \textit{ with } u_{i,j} = 1}} \left[\prod_{k=1}^{K} \mathbb{P}_\theta(\mathbf{X}_k = \mathbf{x}_k)\right] \mathbb{P}(\mathbf{X}_{k,l} = \mathbf{x}_{k,l},\ k \neq l = 1, \ldots, K) \\ &= \sum_{\substack{x_{i,j} \in \{0,1\} \\ \textit{for all } (i,j) \textit{ with } u_{i,j} = 1}} \prod_{k \in \mathcal{S}} \mathbb{P}_\theta(\mathbf{X}_k = \mathbf{x}_k), \end{aligned}$$

where the summation is over all values of $x_{i,j} \in \{0,1\}$ for all pairs of students $(i,j)$ for which $x_{i,j}$ is unobserved. It is worth noting that the between-class probability mass function is eliminated by summation over all possible values of the unobserved between-class edges and that the functional form of the between-class probability mass function is immaterial as long as it is sums to 1.

To derive the likelihood, we have assumed that the missing responses are ignorable for the purpose of likelihood-based inference for the population network model, as explained by Handcock & Gile (2010) and Koskinen *et al.* (2010). In other words, we have assumed that the missing responses due to missing parental consent and inadmissible responses by students do not depend on the unobserved edges.

Monte Carlo maximization of likelihoods of the form $L(\theta)$ given sampled and missing network data are described by Handcock & Gile (2010). We use an implementation of these Monte Carlo maximization methods in `R` package `hergm`.

# 5 Results

Using the Polish school multilevel network described in Section 2, we demonstrate that multilevel networks help estimate the decay parameters of curved ERGMs and provide new opportunities for assessing the out-of-sample performance of ERGMs via cross-validation. We first review all model specifications (Section 5.1) and then assess whether the Monte Carlo maximum likelihood procedure for estimating the parameters of all models converged (Section 5.2). We then interpret the estimates of all parameters and all models (Section 5.3). And finally we turn to model assessment, reviewing the in-sample performance of each model (Section 5.4) and the out-of-sample performance of the best-fitting model (Section 5.5).

## 5.1 Model specifications

We consider nine model specifications, all of which contain the same edge, mutual edge, outdegree, heterogeneity and homophily by sex terms as described in Section 3, but differ in the type of GW terms:

- Models 1–4 focus on GW-OTP (which is the default type for the `dgwesp` term in `R` packages `ergm` and `hergm`):
  - Model 1 is fit without the GW-OTP term, which is equivalent to fixing both the base parameter and the decay parameter at 0.
  - Model 2 leaves the base parameter unrestricted but fixes the decay parameter at 0, which is equivalent to an ERGM with a transitive edge term, as discussed in Section 3.5.
  - Model 3 leaves the base parameter unrestricted but fixes the decay parameter at .25, a value that was used in some of the early papers (Hunter *et al.*, 2008; Goodreau *et al.*, 2009), and has been adopted by others.
  - Model 4 leaves both the base parameter and the decay parameter unrestricted.
- Models 5–8 have GW terms of types OSP, ISP, RTP and ITP respectively, and leave both the base parameter and the decay parameter unrestricted.
- Model 9 has GW terms of types OTP and ITP along with a geometrically weighted indegree term, called GW-Indegree, and leaves the base and decay parameters of all three GW terms unrestricted.

Note that Models 1–9 have size-adjusted edge and mutual edge coefficients, but the other coefficients do not have size-adjustments. These simple size-adjustments suffice here, because the size of school classes are similar: the median class size is 22, and 246 of the 314 classes have 22 ±5 students. Indeed, we show in Sections 5.4 and 5.5 that these models have excellent in-sample performance and out-of-sample performance, which suggests that these simple size-adjustments suffice. A less parsimonious model does not seem worth it—for the data set we use. However, it goes without saying that for other data sets more sophisticated size-adjustments may be needed, based on either size-dependent offsets or size-dependent covariates, as discussed in Section 6.

We estimated the unrestricted parameters of Models 1–9 using the Monte Carlo maximum likelihood methods described in Section 4.

## 5.2 Convergence

To assess whether the Monte Carlo maximum likelihood procedure for estimating the parameters of Models 1–9 converged, we used trace plots of the sufficient statistics of the model, as is common practice (Hunter & Handcock, 2006; Hunter *et al.*, 2008; Hunter *et al.*, 2008). All trace plots show excellent convergence, so for brevity we present just the trace plots for Model 4 in Appendix B. The trace plots for other models may be obtained from the authors upon request.

The resulting estimates of parameters and the assessment of in-sample and out-of-sample performance are discussed in Sections 5.3, 5.4, and 5.5, respectively.

## 5.3 Estimates

The estimates of the unrestricted parameters of Models 1–4, Models 5–8, and Model 9 reported by the Monte Carlo maximum likelihood procedure are shown in Tables 2, 3, and 4, respectively. The standard errors of the estimates are based on the inverse Fisher information matrix (Hunter & Handcock, 2006).

We provide below a careful interpretation of the parameter estimates of all models. We believe that interpreting models is important: models that cannot be interpreted are black boxes, and black boxes do not advance scientific knowledge. Curved ERGMs with GW terms are complex models, and many papers using them interpret them only broadly, e.g., by stating that GW-OTP captures transitivity. There are some good introductions to interpreting GW terms for undirected networks in the seminal paper of Snijders *et al.* (2006) and in Hunter (2007), but those papers do not have (a) directed network data; (b) sampled data; (c) missing data; and (d) size-adjustments for multiple networks of different sizes. To advance proper use of curved ERGMs with GW terms, it is imperative to help users understand how these complex models can be interpreted, in particular in the presence of sampled and missing data, and size-adjustments.

To interpret the individual and joint impact of the parameter estimates, we use the log odds of the conditional probability that a student $i$ nominates another student $j$ as a playmate along with log odds ratios or differences in log odds (based on changes of sufficient statistics, i.e., change statistics). Log odds and log odds ratios are widely used in logistic regression and categorical data analysis (Agresti, 2002) and have long been used in the ERGM literature for interpretive purposes (e.g., Snijders *et al.*, 2006; Hunter & Handcock, 2006; Krivitsky, 2012). Both of these metrics focus on how the effects in the model influence the presence or absence of a single tie. Both condition on the rest of the network and assume that all other tie variables are fixed. Differences in the conditional log odds ratios emphasize how the odds of a single tie change if the tie does versus does not create one or more of the configurations of interest. A useful benchmark for conditional log odds ratios is the value zero. This implies the two configurations compared lead to the same conditional probability of a tie.

The log odds of the conditional probability that a student $i$ nominates another student $j$ as a playmate given the rest of the network $\mathbf{X}_{-(i,j)} = \mathbf{x}_{-(i,j)}$ and the sex indicators $c_i$ and $c_j$ of students $i$ and $j$ is defined as follows:

$$\operatorname{logit}(\mathbb{P}(X_{i,j} = 1 \mid \mathbf{X}_{-(i,j)} = \mathbf{x}_{-(i,j)},\ c_i,\ c_j)) = \log \frac{\mathbb{P}(X_{i,j} = 1 \mid \mathbf{X}_{-(i,j)} = \mathbf{x}_{-(i,j)},\ c_i,\ c_j)}{\mathbb{P}(X_{i,j} = 0 \mid \mathbf{X}_{-(i,j)} = \mathbf{x}_{-(i,j)},\ c_i,\ c_j)}.$$

Note that the conditional probability of the tie between students $i$ and $j$ is conditional on the rest of the network, that is, everything else in the network is considered fixed. For each of Models 1–8, the conditional log odds is given by

$$\begin{aligned}
& \underbrace{(\theta_1 - \log|\mathcal{A}_k|) \ + \ (\theta_2 + \log|\mathcal{A}_k|)\, x_{j,i} \ + \ \ldots}_{\textit{effects of edge, mutual edge, outdegree}} \\
+ \ & \underbrace{\theta_9\, c_i + \theta_{10}\, c_j + \theta_{11}\, \mathbb{1}(c_i = c_j)}_{\textit{effects of sex}} \\
+ \ & \underbrace{\sum_{m=1}^{|\mathcal{A}_k|-2} \left[\eta_{k,11+m}(\theta)\ s_{k,11+m}(\mathbf{x}_{-(i,j)},\ x_{i,j} = 1) - \eta_{k,11+m}(\theta)\ s_{k,11+m}(\mathbf{x}_{-(i,j)},\ x_{i,j} = 0)\right]}_{\textit{effects of triadic closure}},
\end{aligned}$$

where the dots refer to the effect of the outdegree of student $i$. The one exception is Model 9, which has three GW terms instead of one, so the log odds contains three differences in GW terms rather than one difference. Here, we have assumed that students $i$ and $j$ belong to the same school class, denoted by $\mathcal{A}_k$.

We interpret these effects one by one, with the exception of the effect of outdegrees (which are fit to match the artifact produced by the questionnaire design, see Section 3.3). As a running example, we use Model 4, the model with the GW-OTP and unrestricted base and decay parameter. Models 5–8, which only differ in the GW term that captures the effect of triadic closure, are compared in Section 5.3.3. Model 9 is discussed at the end of Section 5.3.

### 5.3.1 Edges and reciprocity effects

Interpreting the sign and magnitude of the edge and mutual edge coefficients is different when using size adjustments, so those coefficients need to be interpreted with care. The size-adjusted effects for edges and mutuals, as a function of class size, are shown in Figure 6.

|  | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
|  | No GW | GW-OTP(0) | GW-OTP(.25) | GW-OTP(free) |
| $\theta_1$ Edges | .152 (.015) $\ast\ast\ast$ | $-$.720 (.020) $\ast\ast\ast$ | $-$1.001 (.019) $\ast\ast\ast$ | $-$.706 (.016) $\ast\ast\ast$ |
| where $\eta_{k,1}(\theta) = \theta_1 - \log\|\mathcal{A}_k\|$ is the edge coefficient of $\mathcal{A}_k$ $(k = 1, \ldots, K)$ |  |  |  |  |
| $\theta_2$ Mutual | $-$1.501 (.021) $\ast\ast\ast$ | $-$1.703 (.022) $\ast\ast\ast$ | $-$1.900 (.023) $\ast\ast\ast$ | $-$1.992 (.023) $\ast\ast\ast$ |
| where $\eta_{k,2}(\theta) = \theta_2 + \log\|\mathcal{A}_k\|$ is the mutual edge coefficient of $\mathcal{A}_k$ $(k = 1, \ldots, K)$ |  |  |  |  |
| Female: |  |  |  |  |
| $\theta_9$ Outdegree | .244 (.018) $\ast\ast\ast$ | .228 (.016) $\ast\ast\ast$ | .211 (.016) $\ast\ast\ast$ | .206 (.016) $\ast\ast\ast$ |
| $\theta_{10}$ Indegree | $-$.077 (.018) $\ast\ast\ast$ | $-$.046 (.016) $\ast\ast$ | $-$.067 (.015) $\ast\ast\ast$ | $-$.098 (.013) $\ast\ast\ast$ |
| $\theta_{11}$ Sex-match | 1.599 (.016) $\ast\ast\ast$ | 1.231 (.014) $\ast\ast\ast$ | 1.032 (.012) $\ast\ast\ast$ | .900 (.011) $\ast\ast\ast$ |
| GW: |  |  |  |  |
| $\theta_{12}$ Base | 0 (fixed) | 1.055 (.018) $\ast\ast\ast$ | 1.237 (.016) $\ast\ast\ast$ | .713 (.012) $\ast\ast\ast$ |
| $\alpha$ Decay | 0 (fixed) | 0 (fixed) | .25 (fixed) | .913 (.014) $\ast\ast\ast$ |

Table 2: Monte Carlo maximum likelihood estimates and standard errors of all parameters, with the exception of outdegree parameters, which can be found in Appendix C. Significance at levels .1, .05, and .001 is indicated by $\ast$, $\ast\ast$, and $\ast\ast\ast$, respectively. A graphical representation of GW-OTP is shown in Figure 5.

First, note that the interpretation of the size-invariant edge parameter $\theta_1$ is more complicated than in the simple "edges-only" Bernoulli model discussed in Krivitsky *et al.* (2011) and Section 3.2. In the simple Bernoulli model, $\exp(\theta_1)$ is the limiting expected mean degree. In ERGMs with additional terms that interpretation no longer holds, because the limiting expected mean degree will reflect the impact of these additional terms. However, one can still interpret the size-adjusted coefficients, $\theta_1 - \log|\mathcal{A}_k|$, in terms of their effect on the conditional log odds of a tie. To do so, note that the sizes of the school classes range from 11 to 33 and the estimates of the size-invariant edge parameter $\theta_1$ range from $-1.001$ to .290 in Models 1–8, so the size-adjusted edge coefficients satisfy $\eta_{k,1}(\theta) = \theta_1 - \log|\mathcal{A}_k| < -2.1$ for all models and all school classes $\mathcal{A}_k$. The strong and negative edge coefficients imply that the conditional odds of a tie is negative, unless the tie creates one or more network configurations with a strong and positive weight.

The mutual edge coefficients are likewise size-adjusted. While the estimates of the size-invariant mutual edge parameter $\theta_2$ are negative, almost all size-adjusted mutual edge coefficients $\eta_{k,2}(\theta) = \theta_2 + \log|\mathcal{A}_k|$ are positive (we address the one exception below). For example, the estimate of $\theta_2$ under Model 4 is $-1.992$, but the estimates of the size-adjusted mutual edge coefficients $\eta_{k,2}(\theta)$ are positive and range from .41 (class size 11) to 1.50 (class size 33). This suggests that reciprocity is a powerful force in these classroom networks: the change in the log odds of the conditional probability

|  | Model 5 | Model 6 | Model 7 | Model 8 |
|---|---|---|---|---|
|  | GW-OSP | GW-ISP | GW-RTP | GW-ITP |
| $\theta_1$ Edges | $-.524$ (.015) $\ast\ast\ast$ | $-.501$ (.015) $\ast\ast\ast$ | .290 (.013) $\ast\ast\ast$ | $-.070$ (.014) $\ast\ast\ast$ |
| where $\eta_{k,1}(\theta) = \theta_1 - \log\lvert\mathcal{A}_k\rvert$ is the edge coefficient of $\mathcal{A}_k$ $(k = 1, \ldots, K)$ |  |  |  |  |
| $\theta_2$ Mutual | $-1.834$ (.023) $\ast\ast\ast$ | $-1.829$ (.023) $\ast\ast\ast$ | $-2.661$ (.031) $\ast\ast\ast$ | $-1.449$ (.022) $\ast\ast\ast$ |
| where $\eta_{k,2}(\theta) = \theta_2 + \log\lvert\mathcal{A}_k\rvert$ is the mutual edge coefficient of $\mathcal{A}_k$ $(k = 1, \ldots, K)$ |  |  |  |  |
| Female: |  |  |  |  |
| $\theta_9$ Outdegree | .253 (.017) $\ast\ast\ast$ | .199 (.016) $\ast\ast\ast$ | .207 (.017) $\ast\ast\ast$ | .251 (.019) $\ast\ast\ast$ |
| $\theta_{10}$ Indegree | $-.127$ (.014) $\ast\ast\ast$ | $-.112$ (.015) $\ast\ast\ast$ | $-.110$ (.015) $\ast\ast\ast$ | $-.146$ (.018) $\ast\ast\ast$ |
| $\theta_{11}$ Sex-match | .961 (.011) $\ast\ast\ast$ | .954 (.012) $\ast\ast\ast$ | 1.214 (.016) $\ast\ast\ast$ | 1.255 (.015) $\ast\ast\ast$ |
| GW: |  |  |  |  |
| $\theta_{12}$ Base | .522 (.009) $\ast\ast\ast$ | .471 (.008) $\ast\ast\ast$ | .435 (.010) $\ast\ast\ast$ | .134 (.005) $\ast\ast\ast$ |
| $\alpha$ Decay | 1.097 (.016) $\ast\ast\ast$ | 1.226 (.015) $\ast\ast\ast$ | .685 (.022) $\ast\ast\ast$ | 2.105 (.068) $\ast\ast\ast$ |

Table 3: Monte Carlo maximum likelihood estimates and standard errors of all parameters, with the exception of outdegree parameters, which can be found in Appendix C. Significance at levels .1, .05, and .001 is indicated by $\ast$, $\ast\ast$, and $\ast\ast\ast$, respectively. The size adjustments $-\log\lvert\mathcal{A}_k\rvert$ range from $-3.5$ to $-2.4$. Graphical representations of GW terms of types OSP, ISP, RTP, and ITP can be found in Figure 5.

that a student $i$ nominates another student $j$ as a playmate when the nomination is reciprocated is

$$\begin{aligned} & \operatorname{logit}(\mathbb{P}(X_{i,j} = 1 \mid X_{j,i} = 1,\ \mathbf{X}_{-(i,j),-(j,i)} = \mathbf{x}_{-(i,j),-(j,i)},\ c_i,\ c_j)) \\ - \quad & \operatorname{logit}(\mathbb{P}(X_{i,j} = 1 \mid X_{j,i} = 0,\ \mathbf{X}_{-(i,j),-(j,i)} = \mathbf{x}_{-(i,j),-(j,i)},\ c_i,\ c_j)) \\ = \quad & -1.992 + \log\lvert\mathcal{A}_k\rvert + \ldots, \end{aligned}$$

where $\mathbf{X}_{-(i,j),-(j,i)}$ denotes the network $\mathbf{X}$ excluding $X_{i,j}$ and $X_{j,i}$ and the dots refer to the effect of student $j$'s outdegree. The size-adjusted coefficients range from .41 (class size 11) to 1.50 (class size 33); so, the conditional odds are multiplied by $\exp(.41) = 1.51$ to $\exp(1.50) = 4.48$ when nominations are reciprocated rather than unreciprocated.

There is one exception to the general rule of a positive size-adjusted mutual effect: Model 7. This model has two reciprocity effects: the baseline reciprocity $\eta_{k,2}(\theta) = \theta_2 + \log\lvert\mathcal{A}_k\rvert$ and the reciprocity-triad effect in the form of GW-RTP. The baseline reciprocity estimate is $-2.661 + \log\lvert\mathcal{A}_k\rvert$, which ranges from $-.263$ (class size 11) to .836 (class size 33). It is small but negative for classes with 11–14 students, and

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Base terms: | | | | GW-OTP: | | | |
| $\theta_1$ Edges | $-1.042$ | (.017) | $* * *$ | $\theta_{12}$ Base | .891 | (.010) | $* * *$ |
| $\theta_2$ Mutual | $-1.483$ | (.028) | $* * *$ | $\alpha_1$ Decay | 1.311 | (.020) | $* * *$ |
| Female: | | | | GW-ITP: | | | |
| $\theta_9$ Outdegree | .094 | (.011) | $* * *$ | $\theta_{13}$ Base | $-.273$ | (.017) | $* * *$ |
| $\theta_{10}$ Indegree | $-.013$ | (.010) | | $\alpha_2$ Decay | 1.896 | (.106) | $* * *$ |
| $\theta_{11}$ Sex-match | .838 | (.012) | $* * *$ | GW-Indegree: | | | |
| | | | | $\theta_{14}$ Base | .837 | (.015) | $* * *$ |
| | | | | $\alpha_3$ Decay | 1.077 | (.048) | $* * *$ |

Table 4: Monte Carlo maximum likelihood estimates and standard errors of all parameters of Model 9, with the exception of outdegree parameters. Significance at levels .1, .05, and .001 is indicated by $*$, $**$, and $* * *$, respectively. A graphical representation of GW-OTP and GW-ITP is shown in Figure 5.

positive for larger classes. The negative effect of the baseline reciprocity term in small school classes will be offset by the positive reciprocity-triad term if a tie creates one or more configurations of type RTP. So a tie that creates one of the mutual legs of the RTP configuration gets both the baseline mutual effect (which may be slightly negative) and the GW-RTP effect (which is larger and positive). Even in small classes this net effect will be positive, and the model suggests that, for small classes, reciprocity is more likely to occur in the context of an RTP configuration than by itself.

### 5.3.2 Sex effects

There is evidence for both moderate degree heterogeneity and strong homophily by sex.

Under all models, the estimate of the female outdegree parameter is small and positive, the estimate of the female indegree parameter is small and negative, and the estimate of the sex-match parameter is large and positive.

These three sex-related terms, along with the edge term, saturate the model for the sex-mixing matrix (see Table 1) in the sense that the counts in the sex-mixing matrix are completely determined by the number of edges and the sex-related sufficient statistics (female outdegree, female indegree, and sex-match). As a result, when there are no missing data, the MLE reproduces the observed sex-mixing matrix, because it matches the observed number of edges and sex-related sufficient statistics. When there are missing data – as in the Polish multilevel network – the MLE reproduces the sex-mixing matrix averaged over all possible realizations of the missing data, because it matches the conditional expectation of the number of edges and the sex-related

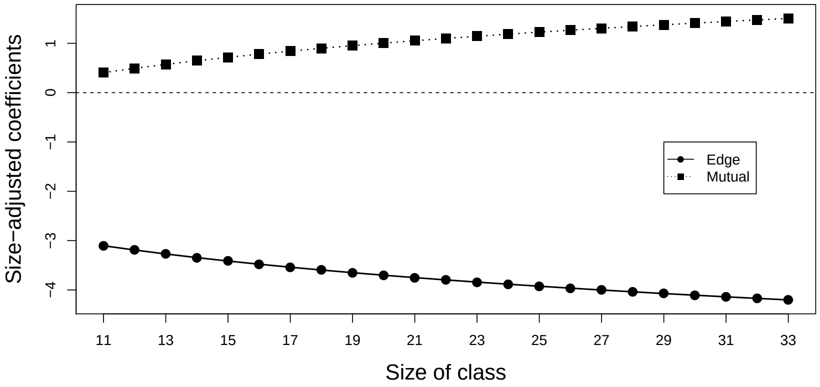


Figure 6: Size-adjusted edge and mutual edge coefficients based on the parameter estimates of Model 4.

sufficient statistics given the observed network data, as discussed in Appendix A. Note that the MLE does not reproduce the sex-mixing matrix in Table 1 based on the subset of 44 school classes without missing data. Instead, the MLE reproduces the sex-mixing matrix based on the whole set of 304 school classes, averaged over all possible realizations of the missing data. The Monte Carlo MLE, which we use as an approximation of the intractable MLE, does so approximately.

To interpret the coefficient values, consider Model 4 with estimates .206 (female outdegree), $-.098$ (female indegree), and .900 (sex-match). Note that the sex-match coefficient is the same for males and females, by construction. But this does not mean that an equal fraction of ties will be sex matched for both sexes; the level of homophily is determined by the net impact of all three sex-specific parameters.

The change in the log odds of the conditional probability that a female student $i$ nominates another student $j$ as a playmate when $j$ is female rather than male is

$$\begin{aligned} & \operatorname{logit}(\mathbb{P}(X_{i,j} = 1 \mid \mathbf{X}_{-(i,j)} = \mathbf{x}_{-(i,j)}, c_i = 1, c_j = 1)) \\ - & \operatorname{logit}(\mathbb{P}(X_{i,j} = 1 \mid \mathbf{X}_{-(i,j)} = \mathbf{x}_{-(i,j)}, c_i = 1, c_j = 0)) \\ = & (.206 - .098 + .900) - .206 \quad = \quad .802. \end{aligned}$$

The fact that the conditional log odds increases by .802 indicates that female students are more likely to choose another female than a male as a playmate. This is due to both the (negative) in- and (positive) out-degree differences for females, and the sex-match effect.

For males, we can calculate the analogous comparison. The change in the log odds of the conditional probability that a male student $i$ nominates another student $j$ as a playmate when $j$ is male rather than female is

$$\begin{aligned}
&\operatorname{logit}(\mathbb{P}(X_{i,j}=1 \mid \mathbf{X}_{-(i,j)}=\mathbf{x}_{-(i,j)},\ c_i=0,\ c_j=0)) \\
-&\operatorname{logit}(\mathbb{P}(X_{i,j}=1 \mid \mathbf{X}_{-(i,j)}=\mathbf{x}_{-(i,j)},\ c_i=0,\ c_j=1)) \ = \ .900-(-.098) \ = \ .998.
\end{aligned}$$

The conditional log odds increases by .998, so male students tend to choose male playmates over female playmates. Here, the net effect is determined by the marginal negative indegree effect for females and the sex-match effect. Note that this relative homophily effect is somewhat stronger for males than for females: compared to females, males are relatively more likely to choose a sex matched playmate.

Finally, the conditional log odds of a tie between two females versus between two males is given by

$$\begin{aligned}
&\operatorname{logit}(\mathbb{P}(X_{i,j}=1 \mid \mathbf{X}_{-(i,j)}=\mathbf{x}_{-(i,j)},\ c_i=1,\ c_j=1)) \\
-&\operatorname{logit}(\mathbb{P}(X_{i,j}=1 \mid \mathbf{X}_{-(i,j)}=\mathbf{x}_{-(i,j)},\ c_i=0,\ c_j=0)) \\
= &\ (.206-.098+.900)-(.900) \ = \ .108.
\end{aligned}$$

The conditional log odds increases by .108, so female-female ties are more common than male-male ties. All three sex-specific effects are combining to generate this net effect.

What is interesting here is that males are relatively more likely to choose sex matched playmates than females, but female-female ties are still more common than male-male ties. This illustrates some of the subtleties in interpreting parameters for even the simpler dyadic-independent terms in ERGMs. This is not an ERGM-specific issue; all generalized linear models (GLMs) (McCullagh & Nelder, 1983) for counts have this property. GLMs decompose the observed patterns in cross-tabulated counts into marginal and interaction effects (here, degree heterogeneity by sex and sex-match, respectively). The resulting parameters can be combined in different ways to highlight specific effects (similar to contrasts in ANOVA). The direct homophily effect in our models, represented by $\theta_{11}$, is the same for both males and females, by construction. But the effect of sex on mixing between males and females is also influenced by the sex-linked degree heterogeneity: females are less likely to be nominated (by both sexes) and more likely to nominate others (of both sexes). The net result is higher rates of female sex matched ties, but greater relative propensities for sex-match among males than females.

### 5.3.3 Triadic closure effects

We turn finally to the effect of triadic closure, first comparing Model 1 without triadic closure to Models 2–4 with triadic closure captured by GW-OTP, and then comparing Models 4–8 with GW terms of types OTP, OSP, ISP, RTP and ITP.

Models 1–4 impose a sequence of restrictions on the base and decay parameter of the GW term. Model 1 excludes the GW term, which is equivalent to assuming that both the base and decay parameter of this term are 0, and there is no propensity for triadic closure. Models 2–4 include the GW-OTP term, with different restrictions on the decay parameter, but all 3 models show a strong and significant base parameter, which suggests Model 1 is misspecified. Comparing the estimates in Model 1 to the corresponding estimates in Models 2–4 shows a moderate to large impact of this misspecification on all of the other estimates. For example, the estimate of the sex-match parameter decreases from 1.599 (Model 1) to .900 (Model 4), a reduction of more than 40%. A similar decrease can be seen in the mutual edge parameter. The decrease in the estimate of the sex-match parameter with the inclusion of GW-OTP indicates that triadic closure accounts for some of the homophily by sex, as found in previous studies of school friendship networks (e.g., Lubbers, 2003; Goodreau *et al.*, 2009).

Models 2 and 3 fix the decay parameter at two values repeatedly used in the literature (e.g., Hunter *et al.*, 2008; Goodreau *et al.*, 2009), while Model 4 leaves it free to be estimated. A key finding is that the estimate of the decay parameter, .913 (Model 4), is significantly greater than 0, and more than 3 times greater than the other fixed value of .25 (Model 3). That value was chosen by trial and error in the original papers, based on qualitatively optimizing the goodness-of-fit to the Add Health school friendship networks (Hunter *et al.*, 2008; Goodreau *et al.*, 2009). Our results suggest this decay value does not generalize to all networks, or even to all school friendship networks. Fixing the decay parameter at a value other than the MLE results again has a moderate to large impact on the estimates of all other parameters in the model. As shown in Section 5.4 below, the differences between these model specifications have a considerable impact on goodness-of-fit.

To interpret the estimates of the base and decay parameter of GW-OTP in Model 4, recall that the effect of triadic closure on the log odds of the conditional probability that student $i$ nominates student $j$ as a playmate is

$$\sum_{m=1}^{|\mathcal{A}_k|-2} \left[\eta_{k,11+m}(\theta)\; s_{k,11+m}(\mathbf{x}_{-(i,j)},\; x_{i,j} = 1) - \eta_{k,11+m}(\theta)\; s_{k,11+m}(\mathbf{x}_{-(i,j)},\; x_{i,j} = 0)\right],$$

where

$$\eta_{k,11+m}(\theta) \;=\; \theta_{12}\, \exp(\alpha) \left[1 - (1 - \exp(-\alpha))^m\right], \quad m = 1, \ldots, |\mathcal{A}_k| - 2.$$

If the edge $X_{i,j} = 1$ increases the number of OTP shared playmates of $(i, j)$ from 0 to 1 relative to the network with $X_{i,j} = 0$, assuming the rest of the network is the same, then the contribution of GW-OTP to the log odds of the conditional probability of the edge is

$$\eta_{k,11+1}(\theta) - 0 \;=\; \theta_{12}\, \exp(\alpha) \left[1 - (1 - \exp(-\alpha))\right] \;=\; \theta_{12}.$$

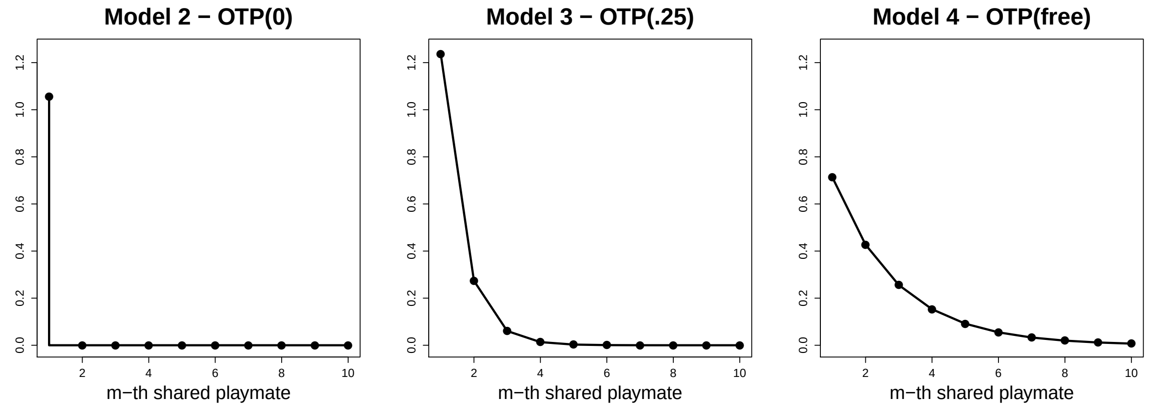


Figure 7: Estimated added value of additional shared playmates of type OTP under models 2–4, as explained in the text. The added value of the first shared playmate is $\theta_{12}$, while the added value of $m$-th shared playmate is $\theta_{12}\,(1-\exp(-\alpha))^{m-1}$ ($m = 2,\ldots,|\mathcal{A}_k|-2$). To make the plots, we used the estimates of $\theta_{12}$ and $\alpha$ shown in Table 2.

If the edge $X_{i,j} = 1$ increases the number of OTP shared playmates of $(i,j)$ from 1 to 2, then the contribution of GW-OTP to the log odds of the conditional probability of the edge is

$$\eta_{k,11+2}(\theta) - \eta_{k,11+1}(\theta) \;=\; \theta_{12}\,(1-\exp(-\alpha)).$$

If $\alpha > 0$, then $(1-\exp(-\alpha)) < 1$, so it acts as penalty on $\theta_{12}$, reducing the value of the second shared playmate. The smaller the value of $\alpha$, the larger this penalty becomes. When $\alpha = 0$ – the transitive tie specification in Model 2 – the penalty zeros out the value of the second OTP shared playmate.

In general, if the edge $X_{i,j} = 1$ increases the number of shared playmates of $(i,j)$ from $m-1$ to $m$ relative to the network with $X_{i,j} = 0$, assuming the rest of the network is the same, then the log odds of the conditional probability of the edge increases by

$$\eta_{k,11+m}(\theta) - \eta_{k,11+m-1}(\theta) \;=\; \theta_{12}\,(1-\exp(-\alpha))^{m-1}, \quad m = 2,\ldots,|\mathcal{A}_k|-2.$$

If $\theta_{12} > 0$ and $\alpha > 0$, then $\theta_{12}\,(1-\exp(-\alpha))^{m-1}$ decreases geometrically as $m$ increases. In other words, the added value of the $m$-th shared playmate decreases at a geometric rate, controlled by the decay parameter $\alpha$:

$$\underbrace{\theta_{12}}_{\textit{added value } m=1} \;>\; \underbrace{\theta_{12}\,(1-\exp(-\alpha))}_{\textit{added value } m=2} \;>\; \underbrace{\theta_{12}\,(1-\exp(-\alpha))^2}_{\textit{added value } m=3} \;>\; \ldots$$

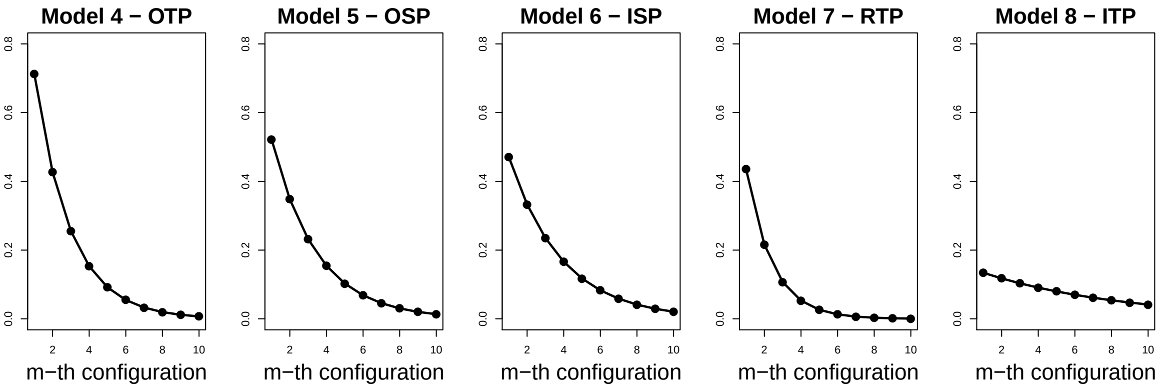


Figure 8: Models 4–8: Added value of additional configurations of type OTP (Model 4), OSP (Model 5), ISP (model 6), RTP (Model 7), and ITP (Model 8), as explained in the text. The added value of the first configuration of the specified type is $\theta_{12}$, while the added value of $m$-th configuration is $\theta_{12}\,(1-\exp(-\alpha))^{m-1}$ ($m = 2, \ldots, |\mathcal{A}_k| - 2$). To make the plots, we used the estimates of $\theta_{12}$ and $\alpha$ shown in Tables 2 and 3.

And, as in the case of $m = 2$, when $\alpha = 0$ the penalty zeros out contributions for all shared partners beyond the first.

A graphical representation of the predicted added value of additional shared playmates for Models 2–4 is shown in Figure 7, using the estimates of the base $\theta_{12}$ and decay parameters $\alpha$ under Models 2–4 from each model. The decay parameter values rise from 0 in Model 2 to .913 in Model 4, and the impact is clearly visible, lowering the penalty on the value of additional shared partners, and increasing predicted density in the right tail of the distribution. Under Model 2, the added value of the first shared playmate is 1.055, while the added value of all subsequent shared playmates is 0 ($m = 2 \ldots, |\mathcal{A}_k| - 2$). Under Models 3 and 4, the added value of the first shared playmate is 1.237 and .713, respectively, and the added value of the $m$-th shared playmate is $1.237 \times .221^{m-1}$ and $.713 \times .599^{m-1}$, respectively ($m = 2, \ldots, |\mathcal{A}_k| - 2$). The added value of additional shared playmates is always positive, but it decreases at a geometric rate, and the rate of decrease is slower when the value of the decay parameter is higher. The rate of geometric decay is high enough to ensure that the added value of the fifth shared playmate is less than .1 in all cases.

In terms of the impact on the odds of a tie, the postive effect of adding the first shared partner is still not enough to outweigh the large negative estimated edge coefficients $\eta_{k,1}(\theta) = \theta_1 - \log|\mathcal{A}_k|$, which are less than $-3$ under Models 2–4. So the log odds of a tie are still negative if the tie only adds a shared playmate, but they can become positive if that tie has other benefits such as reciprocity or homophily by sex.

Turning to Models 5–8, we find that the base parameter estimates of all of the the

GW terms are positive and significant according to Table 3, and the decay parameters are also large and positive. While there is a positive tendency toward each type of triadic closure, there are substantial differences in the specific base and decay parameter estimates, and the joint effect of these differences can be seen in Figure 8, which plots the added value each model assigns to additional configurations. Note that the models are displayed in order by type of closure: the three transitive closure specifications (Models 4-6, GW-OTP, OSP and ISP), followed by GW-RTP, which represents both transitive and cyclical closure, and finally the cyclical closure specification GW-ITP in Model 8.

A clear distinction can be seen in Figure 8 between the added value assigned by transitive versus purely cyclical (Model 8: ITP) specifications, and this follows directly from the parameter estimates. In Model 8, the base parameter estimate is much smaller that in any other model, and this will reduce the overall value of these cyclic triads, relative to the transitive triads. However, the decay parameter estimate is much larger that in the other models, and this reduces the rate at which the added value of additional configurations declines. The joint effect is the lower, flatter distribution of added value we see in the last panel of Figure 8. In the hierarchical world of children, it is not surprising that egalitarian cyclic triads have lower value than transitive hierarchical triads. The difference in the base parameter estimates between Models 4 and 8 – $\exp(.713 - .134) = 1.78$ – implies a nearly 80% increase in the odds of a tie if it forms the first triad of type OTP, compared to a triad of type ITP. But the decay parameter estimate for the GW-ITP term is a surprisingly large 2.105 – almost an order of magnitude larger than the commonly used fixed estimate of .25. While this increases the value of multiple cyclic triads formed by a single tie, the low overall value keeps the net impact in line with the transitive triads.

By contrast, the transitive GW terms of types OSP, and ISP and the combined transitive and cyclic term of type RTP in Models 5–7 display a pattern more similar to the transitive GW-OTP in Model 4. Recall that OTP, OSP, and ISP all lead to the same transitive triad $030T$ in the triad census of Holland & Leinhardt (1970), but each closes one of the three legs of that triad. Comparing the base parameter estimates GW terms of types OTP, OSP, and ISP suggests that the GW-OTP has the strongest initial triadic closure effect, increasing the relative odds of a tie by about 20 to 30% ($\exp(.713 - .522)$ to $\exp(.713 - .435)$). OTP is the classic "a friend of my friend is my friend" dynamic. By contrast, OSP suggests that pairs of playmates nominate the same shared playmates, whereas ISP suggests that pairs of playmates are nominated by the same shared playmates. Both of these latter social forces make sense, but the stronger effects for OTP may explain why only it has a special cultural phrase.

Models can include multiple GW terms. An example is Model 9, which contains three GW terms: GW-OTP, GW-ITP, and GW-Indegree. The estimates and standard errors of all parameters of Model 9, including the base and decay parameters

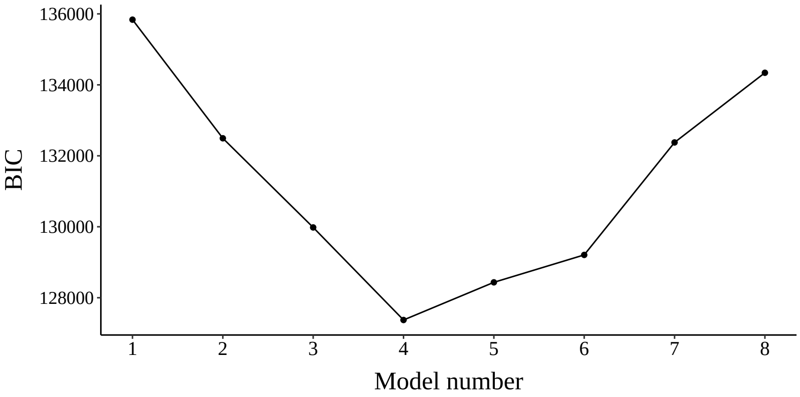


Figure 9: BIC of Models 1–8. The BIC of Model 9, not plotted, is 239,581.

of all three GW terms, can be found in Table 4. We do not attempt to interpret them here, although these estimates can be interpreted by using conditional log odds and log odds ratios as explained above. While Model 9 demonstrates that the base and decay parameters of multiple GW terms can be estimated, we caution that the interpretation of models with multiple GW terms is more complicated, and possible correlations among GW terms may raise multicollinearity issues (as in ordinary regression with correlated predictors).

Last, but not least, we turn to the question of which GW terms to use. GW terms can be selected based on AIC or BIC (see, e.g., Hunter *et al.*, 2008). The BIC of Models 1–9 is shown Figure 9. It is notable that the BIC of Model 4 with unrestricted decay parameter is much lower than the BIC of Models 1, 2, and 3 with restricted decay parameter, underscoring once again the importance of estimating, rather than fixing, decay parameters. Among the models with GW terms of types OTP, OSP, ISP, RTP and ITP, the models capturing transitive closure (Models 4, 5, and 6) clearly outperform the models capturing cyclical closure (Models 7 and 8) in terms of BIC, while Model 9 with three GW terms is heavily penalized by the BIC. The BIC hence agrees with the informal observation made above: it is transitive closure, rather than cyclical closure, that drives network formation in the Polish multilevel network.

### 5.3.4 Standard errors

In addition to facilitating the estimation of decay parameters, the standard errors of the decay parameter estimates in Tables 2, 3, and 4 demonstrate that multilevel networks, by providing replication, help reduce the uncertainty about the decay parameter estimates.

The standard errors of the decay parameter estimates for the directed, transitive GW terms of types OTP, OSP, and ISP range from .014 (GW-OTP in Model 4) to .020 (GW-OTP in Model 9). As noted before, we know of only four other published papers that estimated the decay parameters in a curved ERGM, Hunter (2007), Koskinen *et al.* (2010), Suesse (2012), and Almquist & Bagozzi (2015). These are not strictly comparable studies as all of them are based on undirected networks, which were slightly larger than our largest network (Hunter, Koskinen *et al.*, and Suesse: 36; Almquist & Bagozzi: 143; here: 11 to 33). Still, the comparison is suggestive, as the standard errors reported for the GW-ESP decay parameter estimates in their models are .109 (Hunter, 2007), .151 (Suesse, 2012), and .099 and .706 (Almquist & Bagozzi, 2015) – roughly an order of magnitude higher than ours. Note that Koskinen *et al.* (2010) follow a Bayesian approach and do not report standard errors, but summaries of the posterior suggest that the posterior standard deviation may be as large as the standard errors reported by Hunter (2007) and Suesse (2012) for the same network, the Lazega law firm advice network.

## 5.4 In-sample performance: goodness-of-fit

The traditional approach to evaluating the goodness-of-fit (GOF) of ERGMs is to assess how well the model predicts observed network features that were not included in the model (Hunter *et al.*, 2008). This is done by comparing the statistics from the observed network to statistics from networks simulated from the model. Because the comparison relies on the same network that was used to estimate the model, this is an assessment of the in-sample performance of ERGMs. The purpose of this type of assessment is to evaluate the generative performance of the fitted model: to determine whether a parsimonious set of terms that capture local, micro-level effects are able to reproduce the overall macro-level structural signatures in the network. As always with statistical assessments, bad performance allows hypotheses to be rejected, and good performance is not a form of proof, but in this case simply implies that the data are consistent with the hypothesized generative model.

The presence of missing data complicates GOF comparisons, because we want to compare model-based predictions to the 304 sampled school classes, but 260 of them have missing data. We could compare model-based predictions of subgraph statistics to fully observed subgraphs: e.g., we could compare model-based predictions of the number of mutual edges to the number of pairs of students for which both edges are observed. Such comparisons have at least two disadvantages, however. First, we would make the implicit assumption that the pairs of students for which one edge is present while the other one is missing or for which both edges are missing do not reciprocate edges. Second, it would reduce the number of pairs of students on which the comparison is based. The issue exists for both dyadic statistics (e.g., mutual edges) and triadic statistics (e.g., transitive edges), but it tends to be worse when the statistic involves more edges.

To avoid these two disadvantages, we compare model-based predictions of statistics to the conditional expectation of those statistics given the observed data. In other

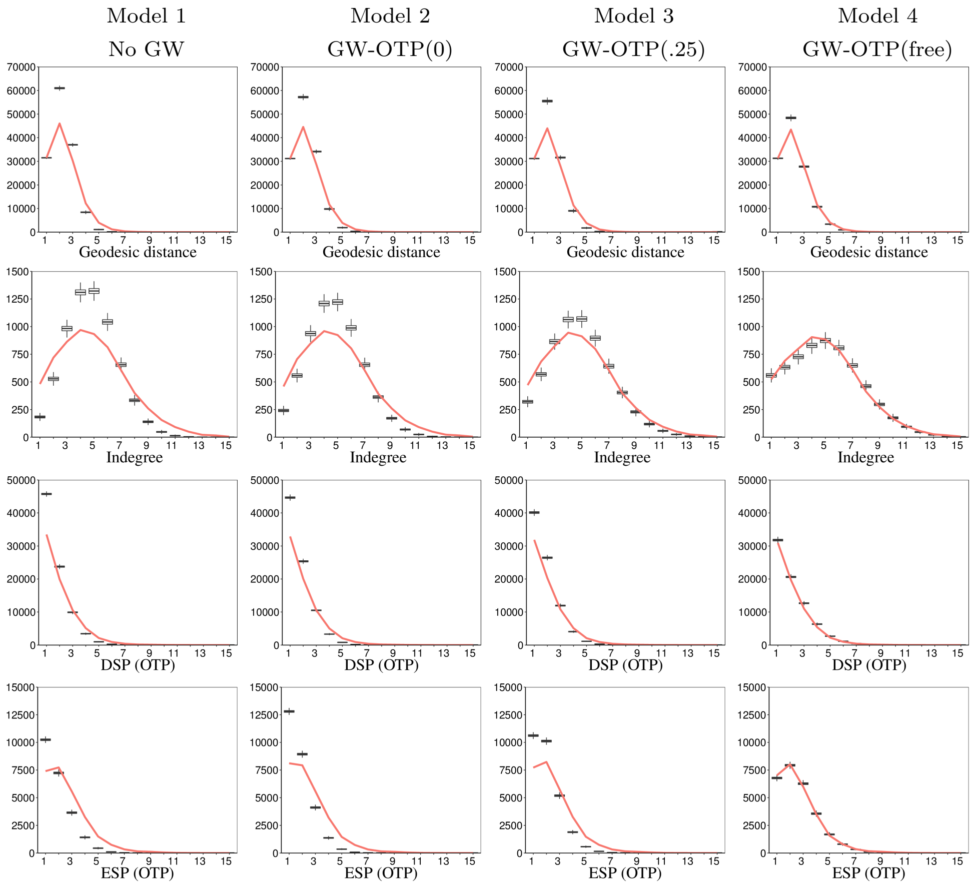
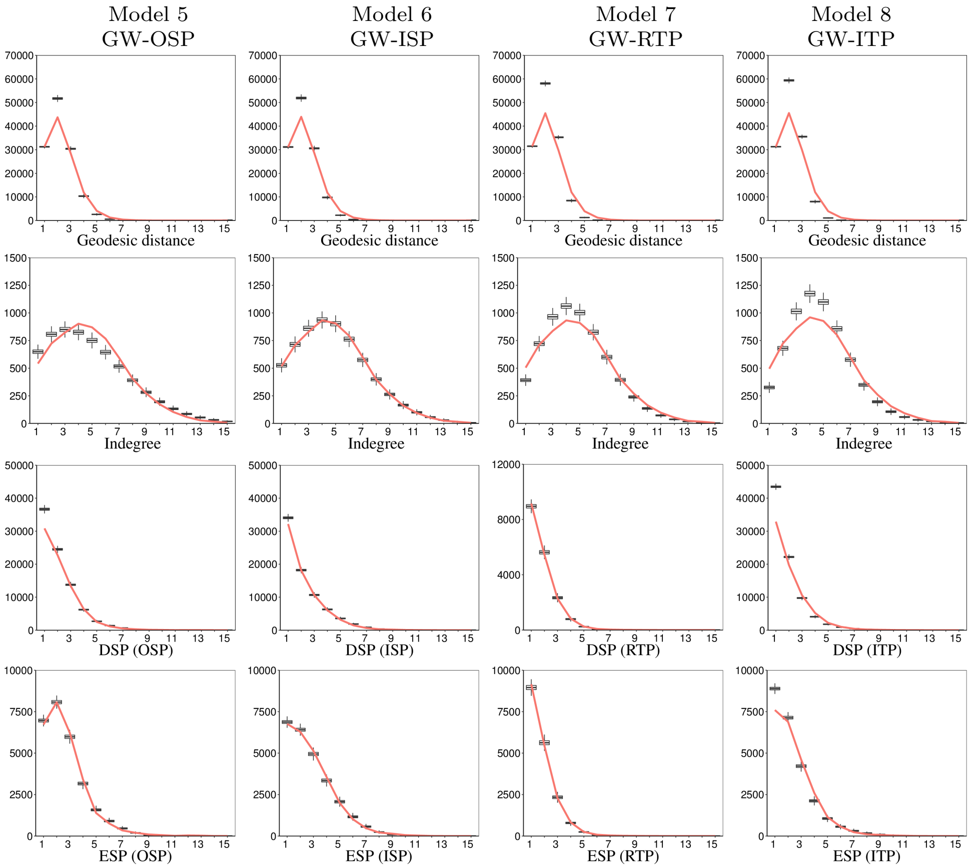


Figure 10: In-sample performance of Models 1–4. The red curves indicate the conditional expectations of the statistics given the observed data. The in-sample performance of Models 1–4 in terms of outdegrees is assessed in Appendix D.

words, we compare model-based predictions to weighted averages of those statistics, averaging over all possible values of the missing data, with the weights given by the conditional distribution of missing data given observed data. The conditional expectations of statistics cannot be calculated analytically, but it is possible to approximate them by Markov chain Monte Carlo sample averages of those statistics based on simulations of networks from the conditional distribution of missing data given the observed data.

Figures 10 and 11 compare the GOF of Models 1–8, using the statistics proposed by Hunter *et al.* (2008): distributions of geodesic distances, indegrees, the number of dyads (unconnected or connected) with $m$ shared partners (DSP), and the number of connected dyads with $m$ shared partners (ESP). For each model, we use the directed versions of the DSP and ESP statistics that match the type of the GW term in the model. Additional GOF plots for other types of DSP and ESP statistics are shown in



Model 5
GW-OSP

Model 6
GW-ISP

Model 7
GW-RTP

Model 8
GW-ITP

Figure 11: In-sample performance of Models 5–8. The red curves indicate the conditional expectations of the statistics given the observed data. The in-sample performance of Models 5–8 in terms of outdegrees is assessed in Appendix D.

Appendix E, and GOF plots for the outdegrees can be found in Appendix D. All GOF plots are based on 10,000 simulated networks generated from the estimated models. Given the missing data, we compare the statistics of the simulated networks to the conditional expectation of the statistics given the observed data. The conditional expectations of statistics are model-dependent and can therefore vary from model to model, but the variation is small, as can be seen in Figures 10 and 11.

The GOF performance of Models 1–4 different quite a bit, reflecting the impact of estimating, rather than fixing, the decay parameter of the GW-OTP. Model 1, which does not have a GW term, is unable to match any of the GOF statistics of the observed network data. The models with the GW-OTP (Models 2–4) do progressively better, as the fixed decay parameter value gets closer to the MLE. Model 4 – which estimates the decay parameter – shows superior GOF performance across the board. It provides a very good fit to the indegree distribution, especially when compared to

Model 1, without requiring a specialized term like a geometrically weighted degree. This is a classic example of how a macro-level network signature, like the indegree distribution, may be consistent with a generative process rooted in a very different dynamic, like triad closure. Model 4 also matches the full DSP and ESP distributions almost perfectly, using a single parsimonious curved ESP term with two parameters. The fact that this model also fits the DSP distribution indicates that an additional DSP term is not required.

The performance of Models 5–8 with GW terms of types OSP, ISP, RTP and ITP is shown in Figure 11. It is evident that the models with transitive triad terms (OTP, OSP, and ISP) outperform the models with cyclical triad terms (RTP and ITP). These results reinforce the findings from Section 5.3 that cyclical closure fails to capture the micro-level patterns that lead to hierarchical structure in the nomination of playmates.

In summary, reciprocity, attribute homophily and triadic closure are important micro-level determinants in the nomination of playmates. Curved ERGMs with GW terms are parsimonous models that can accurately capture the observed triadic closure along with other aggregate network patterns. Multilevel network data make it possible to estimate the MLEs of the triadic closure decay parameters, which results in better goodness of fit.

## 5.5 Out-of-sample performance: cross-validation

The final advantage of multilevel networks we will demonstrate here is that such data make it possible assess the out-of-sample performance of ERGMs using the traditional statistical principle of cross-validation. We can divide the 304 school classes into two subsets, use one as a training subset to estimate the model, and the other as a held-out subset to assess the predictive power of the estimated model. For convenience, we use the GOF statistics from Section 5.4 to assess the predictive power of ERGMs, but in principle any network statistics could be used.

We assess the out-of-sample performance of models by generating 100 model-based predictions as follows:

- Step 1: Stratify the 304 school classes by size and sample without replacement 50% of the school classes from each stratum (rounded up) to create a training data set for estimating models and a held-out data set for model-based predictions.

- Step 2: Estimate models based on the training data set.

- Step 3: Compare model-based predictions of the GOF statistics to the observed GOF statistics for the held-out data set.

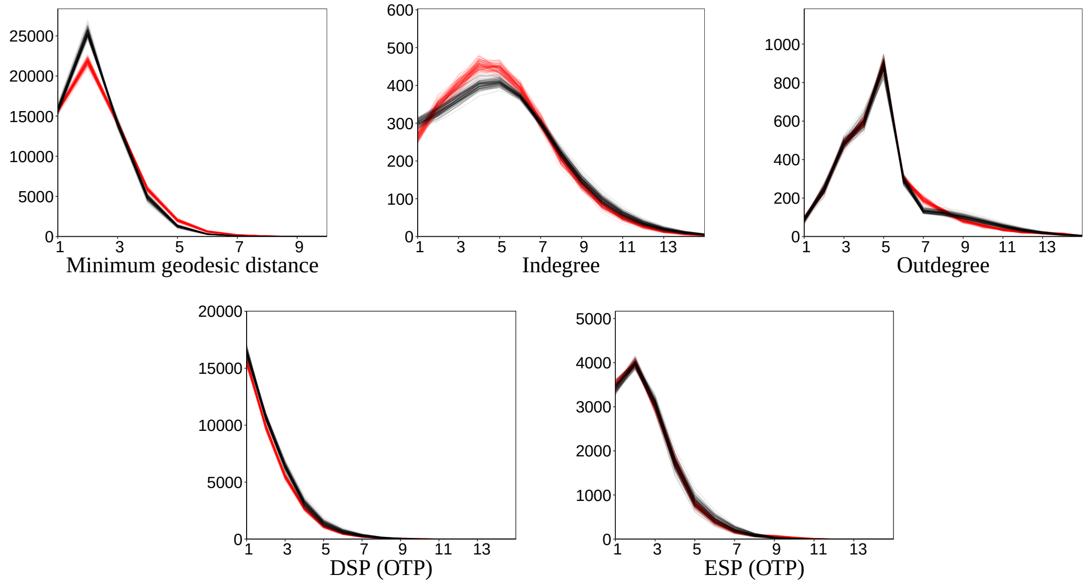


Figure 12: Out-of-sample predictions based on Model 4 with GW-OTP and estimated decay parameter. Each black curve represents one of the 100 out-of-sample predictions while each red curve represents one of the 100 out-of-sample observations.

Some remarks are in order. Step 1 uses stratified random sampling based on class size to facilitate the comparison of network statistics. The reason is that network statistics depend on class size and stratifying by class size helps compare network statistics across multiple random splits of the 304 school classes. The 50-50% split implies that the training data set is small while the held-out data set is large, relative to conventional cross-validation procedures with more observations in the training data set than the held-out data set. The small training data set makes the estimation more challenging as there is less information about the parameters of interest (in the statistical sense of Fisher information), but it has the advantage of reducing computing time. Step 2 is the most time-consuming step of the procedure, because it requires estimating curved ERGMs from 100 different training subsets. Even when parallel computing on multi-core computers or computing clusters is used, estimating curved ERGMs from 100 different training subsets can take days or weeks (depending on how the parallel computing is implemented and how much computing power is available). Step 3 generates out-of-sample predictions for the held-out data sets by using the estimates of the size-invariant parameters $\theta_1, \dots, \alpha$ obtained in Step 2 and the size-dependent offsets $\log |\mathcal{A}_k|$ based on the sizes of the school classes $\mathcal{A}_k$ in the held-out data set. For each held-out data set, 10,000 model-based predictions are generated, averaged, and compared to the observed held-out data set.

To demonstrate the cross-validation approach we use Model 4 (GW-OTP with estimated decay parameter), because the in-sample performance of Model 4 is the best of all of the models. Assessing the out-of-sample performance of other models is possible but time-consuming.

Figure 12 shows the results of the out-of-sample predictions based on Model 4. Overall, the out-of-sample predictions seem to be close to the observed network data. The strong out-of-sample performance suggests that our findings can be generalized to the population of third-grade classes in Poland.

# 6 Discussion

We have demonstrated that multilevel network data facilitate the estimation of curved ERGMs with GW terms, without fixing the decay parameters or conditioning on the observed number of edges. The MLE of the decay parameter for the traditional GW-OTP term was significantly different than the fixed values commonly used in practice. When we fixed the decay parameters at these values, we found this also affected all of the other parameter estimates in the model, in some cases quite substantially. The model with the estimated decay parameter had much better in-sample performance characteristics, and showed remarkable goodness of fit across the board. We also estimated the decay parameters of four additional GW triad specifications for directed networks. To the best of our knowledge, estimates of those decay parameters have never been published before.

The multilevel network data improved statistical inference in other ways also, reducing the uncertainty in our parameter estimates, and allowing us to conduct a traditional cross-validation analysis, to complement the traditional in-sample goodness of fit assessments used for ERGMs. When used with recently developed size-invariant parameterizations, the multilevel analytic framework provides a robust basis for curved ERGM estimation and performance assessment.

Substantively, our results suggest that the nomination of playmates among third-grade school children in Poland is driven by reciprocity, heterogeneity and homophily by sex, and transitive closure. These results agree by and large with the results obtained by others (e.g., Lubbers, 2003; Lubbers & Snijders, 2007; Goodreau *et al.*, 2009), though we have more confidence in the triadic effects now that we have the MLEs for the decay parameters. As we demonstrated in Section 5.3, fixing the decay parameter of GW terms at values far from the MLE affects all other parameter estimates, and can lead to incorrect inferences. In our application the value of additional shared partners was 2-3 times greater than the fixed levels suggested, and the effects of both reciprocity and homophily by sex fell by 30-40% once the decay term was properly estimated.

An important direction of future research is the development of more sophisticated size-adjustments for curved ERGMs. We have used here curved ERGMs with a simple form of size-adjusted parameterization based on Krivitsky *et al.* (2011) and Krivitsky & Kolaczyk (2015) and have demonstrated that both the in-sample and out-of-sample performance of the resulting models is excellent. While encouraging, it is worth remembering that the sizes of the school classes in our application range from 11 to 33, different but similar. If the sizes of school classes were more dissimilar, the simple size-adjusted parameterization we used may not be appropriate. However, more sophisticated size-adjusted parameterizations for curved ERGM terms are possible. In particular, Krivitsky & Kolaczyk (2015) developed a size-adjusted parameterization for the transitive edge term for undirected networks, which is equivalent to the undirected, transitive GW term with decay parameter fixed at 0. It would be interesting to investigate size-adjusted parameterizations for GW terms with unrestricted decay parameters, although the fact that GW terms are nonlinear functions of products of base and decay parameters requires a careful analysis of size-adjustments, which is beyond the scope of our paper. Last, but not least, an interesting idea would be to use network size as a covariate (Slaughter & Koehly, 2016). However, we do not expect a substantial improvement in in-sample and out-of-sample performance, because our simple size-adjusted parameterization shows strong in-sample and out-of-sample performance, in particular for triadic effects.

We provide a software implementation of the proposed models and methods in the form of `R` package `hergm`, which supports parallel computing on multi-processor computers and computing clusters. In the near future, we intend to split `R` package hergm into two `R` packages:

- `mlergm:` ERGMs with known block structure (multilevel ERGMs with nodes belonging to known blocks, with ties within and between blocks).
- `hergm:` ERGMs with unknown block structure (hierarchical ERGMs with nodes belonging to unknown blocks, with ties within and between blocks).

Both of them will be released to `https://cran.r-project.org` (R Core Team, 2018). The code we used here will be included in `mlergm`.

# References

Agresti, Alan. (2002). *Categorical data analysis.* 2 edn. Hoboken: John Wiley & Sons.

Almquist, Z. W., & Bagozzi, B. E. (2015). Using radical environmentalist texts to uncover network structure and network features. *Sociological Methods & Research*, 1–56.

Block, P. (2015). Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks*, **40**, 163–173.

Bomiriya, R. P., Bansal, S., & Hunter, D. R. (2016). Modeling homophily in ergms for bipartite networks. *International Conference on Robust Statistics 2016*.

Brown, L. (1986). *Fundamentals of statistical exponential families: With applications in statistical decision theory*. Hayworth, CA, USA: Institute of Mathematical Statistics.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, **38**, 155–200.

Butts, C. T. (2011). Bernoulli graph bounds for general random graph models. *Sociological Methodology*, **41**, 299–345.

Butts, C. T., & Almquist, Z. W. (2015). A flexible parameterization for baseline mean degree in multiple-network ERGMs. *Journal of Mathematical Sociology*, **39**, 163–167.

Caimo, A., & Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, **33**, 41–55.

Cartwright, D., & Harary, F. (1956). Structural balance: a generalization of Heider's theory. *Psychological Review*, **63**, 277–293.

Chase, I. (1980). Social process and hierarchy formation in small groups: a comparative perspective. *American Sociological Review*, **45**, 905–924.

Chatterjee, S., & Diaconis, P. (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics*, **41**, 2428–2461.

DeBenedictis, P. A. (1977). The meaning and measurement of frequency-dependent competition. *Ecology*, **58**, 158–166.

Dempster, A. P., Laird, N. M., & Rubin, R. B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Dolata, Roman (ed). (2014). *Czy szkoła ma znaczenie? Zróżnicowanie wyników nauczania po pierwszym etapie edukacyjnym oraz jego pozaszkolne i szkolne uwarunkowania*. Vol. 1. Warsaw: Instytut Badań Edukacyjnych.

Dolata, Roman, & Rycielski, Piotr. (2014). Wprowadzenie: założenia i cele badania szkolnych uwarunkowań efektywności kształcenia SUEK. *In:* Dolata (2014).

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, **3**(6), 1189–1242.

Efron, B. (1977). Discussion on the paper by Professor Dempster et al.. *Journal of the Royal Statistical Society, Series B*, **39**, 29.

Everitt, R. G. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *Journal of Computational and Graphical Statistics*, **21**, 940–960.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, **22**, 700–725.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, **81**, 832–842.

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.

Handcock, M. S. (2003). Statistical models for social networks: Inference and degeneracy. *Pages 1–12 of:* Breiger, Ronald, Carley, K., & Pattison, P. (eds), *Dynamic social network modeling and analysis: Workshop summary and papers*. Washington, D.C.: National Academies Press.

Handcock, M. S., & Gile, K. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics*, **4**, 5–25.

Harris, J. K. (2013). *An introduction to exponential random graph modeling*. Thousand Oaks, California: Sage.

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology*, **21**, 107–112.

Holland, P. W., & Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology*, **76**, 492–513.

Hollway, J., & Koskinen, J. (2016). Multilevel embeddedness: The case of the global fisheries governance complex. *Social Networks*, **44**, 281–294.

Hollway, J., Lomi, A., Pallotti, F., & Stadtfeld, C. (2017). Multilevel social spaces: The network dynamics of organizational fields. *Network Science*, **5**, 187–212.

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social Networks*, **29**, 216–230.

Hunter, D. R., & Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, **15**, 565–583.

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. *Journal of Statistical Software*, **24**, 1–29.

Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, **103**, 248–258.

Hunter, D. R., Krivitsky, P. N., & Schweinberger, M. (2012). Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, **21**, 856–882.

Jonasson, J. (1999). The random triangle model. *Journal of Applied Probability*, **36**, 852–876.

Kalish, Y., & Luria, G. (2013). Brain, brawn, or optimism? Structure and correlates of emergent military leadership. *In:* Lusher *et al.* (2013).

Kolaczyk, E. D. (2009). *Statistical analysis of network data: Methods and models.* New York: Springer-Verlag.

Koskinen, J. (2004). *Essays on Bayesian inference for social networks.* Ph.D. thesis, Stockholm University, Dept. of Statistics, Sweden.

Koskinen, J. H., Robins, G. L., & Pattison, P. E. (2010). Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Statistical Methodology*, **7**, 366–384.

Krivitsky, P. N. (2012). Exponential-family models for valued networks. *Electronic Journal of Statistics*, **6**, 1100–1128.

Krivitsky, P. N., & Kolaczyk, E. D. (2015). On the question of effective sample size in network modeling: An asymptotic inquiry. *Statistical Science*, **30**, 184–198.

Krivitsky, P. N., Handcock, M. S., & Morris, M. (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, **8**, 319–339.

Lazega, E. (2001). *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership.* Oxford, UK: Oxford University Press.

Lazega, E., & Snijders, T. A. B. (eds). (2016). *Multilevel network analysis for the social sciences.* Switzerland: Springer-Verlag.

Lomi, A., Robins, G., & Tranmer, M. (2016). Introduction to multilevel social networks. *Social Networks*, 266–268.

Lovász, L. (2012). *Large networks and graph limits*. Providence: American Mathematical Society.

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, Miranda J, & Snijders, T. A. B. (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*. Cambridge, UK: Cambridge University Press.

Maluchnik, Michał, & Modzelewski, Michał. (2014). Próba badawcza i proces zbierania danych. *In:* Dolata (2014).

Mayhew, B. H., & Levinger, R. L. (1976). Size and density of interaction in human aggregates. *American Journal of Sociology*, **82**, 86–110.

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. London: Chapman & Hall.

Obando, C., & De Vico Fallani, F. (2017). A statistical model for brain networks inferred from large-scale electrophysiological signals. *Journal of the Royal Society Interface*, 1–10.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rapoport, A. (1963). *Mathematical models of social interaction*. New York: John Wiley & Sons. In: R. A. Galanter and R. R. Lace and E. Bush *Handbook of Mathematical Psychology V. 2*.

Robins, G. L., Pattison, P. E., & Wang, P. (2009). Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. *Social Networks*, **31**, 105–117.

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, **106**(496), 1361–1370.

Schweinberger, M., & Handcock, M. S. (2015). Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society, Series B*, **77**, 647–676.

Schweinberger, M., & Luna, P. (2018). HERGM: Hierarchical exponential-family random graph models. *Journal of Statistical Software*, **85**, 1–39.

Schweinberger, M., & Stewart, J. (2018). Finite-graph concentration and consistency results for random graphs with complex topological structures. Available at https://arxiv.org/abs/1702.01812.

Schweinberger, M., Krivitsky, P. N., & Butts, C. T. (2017). Foundations of finite-, super-, and infinite-population random graph inference. Available at https://arxiv.org/abs/1707.04800.

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. B. (2001). The statistical evaluation of social network dynamics. *Pages 361–395 of:* Sobel, M.E., & Becker, M.P. (eds), *Sociological Methodology*. Boston and London: Basil Blackwell.

Snijders, T. A. B. (2016). The multiple flavours of multilevel issues for networks. *Pages 15–46 of: Multilevel network analysis for the social sciences*. Springer.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, **36**, 99–153.

Snijders, T. A. B., Steglich, C. E. G, & van de Bunt, G. (2010). Introduction to actor-based models for network dynamics. *Social Networks*, **32**, 44–60.

Strauss, D. (1986). On a general class of models for interaction. *SIAM Review*, **28**, 513–527.

Suesse, T. (2012). Marginalized exponential random graph models. *Journal of Computational and Graphical Statistics*, **21**(4), 883–900.

Wang, P., Robins, G., & Pattison, P. (2006). *PNet. Program for the simulation and estimation of Exponential Random Graph (p*) Models.* Melbourne School of Psychological Sciences, University of Melbourne.

Wang, P., Robins, G., Pattison, P., & Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks*, **35**, 96–115.

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

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regression for social networks: I. An introduction to Markov graphs and $p^*$. *Psychometrika*, **61**, 401–425.

Zappa, P., & Lomi, A. (2015). The analysis of multilevel networks in organizations: models and empirical tests. *Organizational Research Methods*, **18**, 542–569.

# A Maximum likelihood estimation of curved ERGMs with missing data

One of the most appealing properties of MLEs is that, in the simplest case when ERGMs do not contain curved ERGM terms and there are no missing data, MLEs match the expected and observed values of the sufficient statistics: e.g., the MLE of ERGMs with edge terms ensures that the expected number of edges equals the observed number of edges. However, we are dealing here with curved ERGMs with missing data, so the interpretation of MLEs is more complicated. We review here some important implications.

To do so, let $X_{obs}$ be the collection of all edge variables whose values are observed and $X_{mis}$ be the collection of all edge variables whose values are unobserved. By definition, the MLE maximizes the probability of the observed network data $x_{obs}$. Maximizing the probability of the observed network data $x_{obs}$ is equivalent to solving

$$\nabla_\theta \, \log p_\theta(x_{obs}) \quad = \quad 0,$$

which in turn is equivalent to solving

$$\begin{aligned} \nabla_\theta \, \log p_\theta(x_{obs}) \quad &= \quad \mathbb{E}_\theta \left[\nabla_\theta \, \log p_\theta(x_{obs}, X_{mis}) \mid X_{obs} = x_{obs}\right] \\ &= \quad (\nabla_\theta \, \eta(\theta))^\top \, \mathbb{E}_\theta \left[s(x_{obs}, X_{mis}) \mid X_{obs} = x_{obs}\right] \\ &- \quad \big((\nabla_\theta \, \eta(\theta))^\top \, \mathbb{E}_\theta \left[s(X_{obs}, X_{mis})\right]\big) \\ &= \quad 0, \end{aligned} \tag{2}$$

where the first line follows from a well-known missing-data identity dating back to Fisher (1925) and Dempster *et al.* (1977) (see the discussion of Efron, 1977), while the second line follows from exponential-family theory (Brown, 1986). Here, the expectation $\mathbb{E}_\theta \left[s(x_{obs}, X_{mis}) \mid X_{obs} = x_{obs}\right]$ is with respect to the conditional distribution of $X_{mis}$ given $X_{obs} = x_{obs}$, the expectation $\mathbb{E}_\theta \left[s(X_{obs}, X_{mis})\right]$ is with respect to the joint distribution of $X_{obs}$ and $X_{mis}$, and $(\nabla_\theta \, \eta(\theta))^\top$ is the matrix of partial derivatives of natural parameters $\eta_i(\theta)$ with respect to parameters $\theta_j$.

Equation (2) implies that the MLE $\hat{\theta}$ ensures that

$$(\nabla_\theta \, \eta(\theta)|_{\theta=\hat{\theta}})^\top \, \mathbb{E}_{\hat{\theta}} \left[s(X_{obs}, X_{mis})\right] \quad = \quad (\nabla_\theta \, \eta(\theta)|_{\theta=\hat{\theta}})^\top \, \mathbb{E}_{\hat{\theta}} \left[s(x_{obs}, X_{mis}) \mid X_{obs} = x_{obs}\right].$$

To discuss the implications of the maximum likelihood equation shown above, consider one of the sex-related sufficient statistics, the female outdegrees summed across all school classes. Denote the sum of female outdegrees by $s_i(x_{obs}, x_{mis})$ and its natural parameter by $\eta_i(\theta) = \theta_i$. The partial derivative of $\eta_i(\theta)$ with respect to $\theta_i$ is 1, whereas the partial derivative of $\eta_i(\theta)$ with respect to $\theta_j$ is 0 for all $j \neq i$. As a consequence, the MLE ensures that the unconditional and conditional expectation of female outdegrees match:

$$\mathbb{E}_{\hat{\theta}}\left[s_i(X_{obs}, X_{mis})\right] \;=\; \mathbb{E}_{\hat{\theta}}\left[s_i(x_{obs}, X_{mis}) \mid X_{obs} = x_{obs}\right].$$

When there were no missing data, the MLE matches the observed female outdegrees, $s_i(x_{obs})$:

$$\mathbb{E}_{\hat{\theta}}\left[s_i(X_{obs})\right] \;=\; s_i(x_{obs}). \tag{3}$$

Otherwise, when there are missing data, it matches the conditional expectation of female outdegrees given the observed network data, $\mathbb{E}_{\hat{\theta}}\left[s_i(x_{obs}, X_{mis}) \mid X_{obs} = x_{obs}\right]$:

$$\mathbb{E}_{\hat{\theta}}\left[s_i(X_{obs}, X_{mis})\right] \;=\; \mathbb{E}_{\hat{\theta}}\left[s_i(x_{obs}, X_{mis}) \mid X_{obs} = x_{obs}\right]. \tag{4}$$

Two remarks are in order.

First, the left-hand side of equations (3) and (4) is the same, but the right-hand side is not: when there are missing data, the sufficient statistic – the sum of female outdegrees across all school classes – cannot be computed, so it is replaced by a conditional expectation of the sufficient statistic given the observed network data. In other words, the sufficient statistic is averaged over all possible realizations of the missing data, where the possible realizations of the missing data are weighed by the conditional probabilities of the missing data given the observed network data.

Second, the female outdegrees are summed across all school classes, both school classes without missing data and school classes with missing data, which has subtle implications: the MLE matches the conditional expectation of the sum of female outdegrees summed across all school classes, but there is no guarantee that it matches the observed female outdegrees of school classes without missing data. To match the observed female outdegrees of school classes without missing data, class-specific female outdegree parameters would be needed. While it is possible to include class-specific female outdegree parameters, the resulting models would have a large number of parameters and would not be parsimonious, which would increase computational costs (e.g., computing time) as well as statistical costs (e.g., standard errors).

By the same argument, the MLE matches the conditional expectation of the number of edges, mutual edges, female outdegrees, female indegrees, sex-matched edges, and the number of students with outdegrees $1, \ldots, 6$. The GW terms are more complicated: the MLE matches weighted sums of conditional expectations of the number of configurations of the specified type. The weights are given by partial derivatives, and most of the partial derivatives are neither 0 nor 1, because the natural parameters of GW terms are nonlinear functions of products of parameters, and all natural parameters of GW terms depend on the same two parameters (the base and decay parameter).

Last, but not least, it is worth noting that we use Monte-Carlo based approximations of MLEs (as explained in Section 4), because it is infeasible to compute exact MLEs. However, the arguments concerning the behavior of MLEs we presented above also shed some light on the behavior of approximate MLEs, such as Monte Carlo MLEs.

# B Convergence

To assess whether the Monte Carlo maximum likelihood procedure converged, we used trace plots of the sufficient statistics of the model, as is common practice (Hunter & Handcock, 2006; Hunter *et al.*, 2008; Hunter *et al.*, 2008). We present trace plots of the sufficient statistics of Model 4 in Figure 13. The trace plots for all other models may be obtained from the authors upon request. None of these trace plots shows signs of non-convergence. Trace plots of the other models are not shown, but those trace plots do not show signs of non-convergence either.

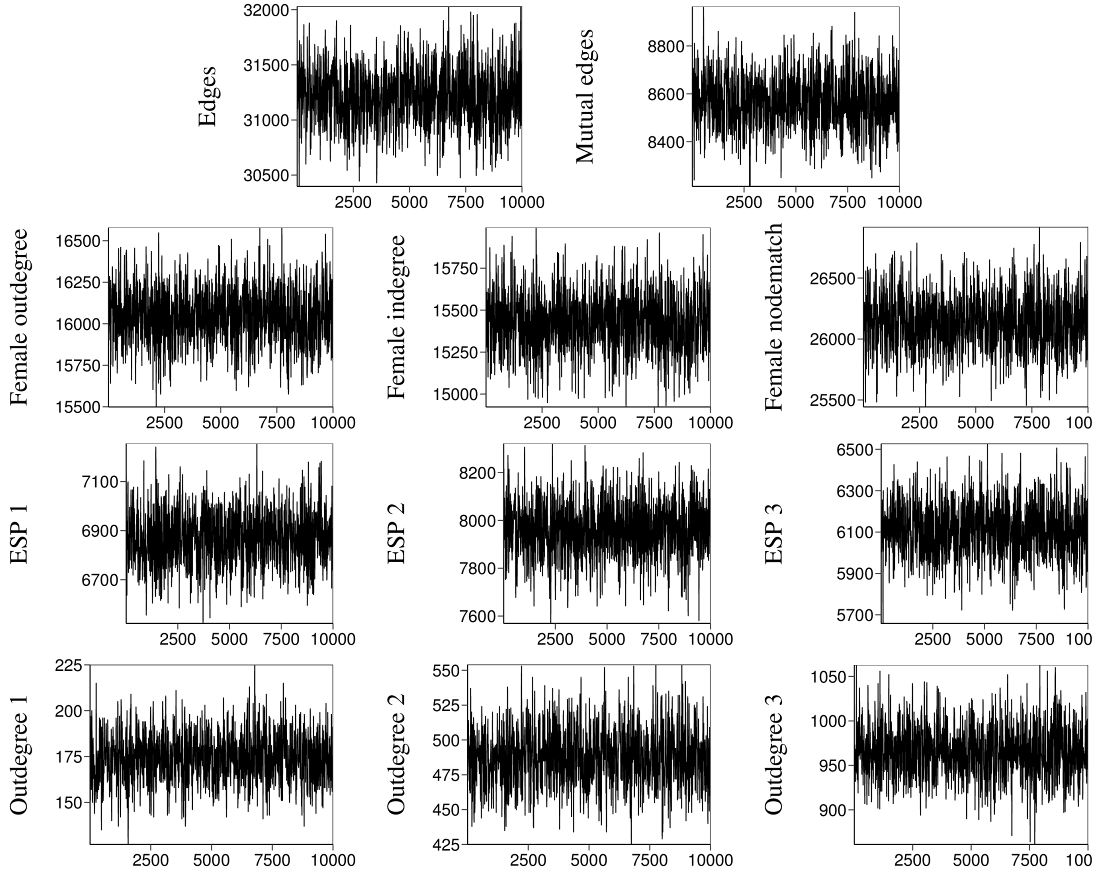


Figure 13: Model 4: Trace plots of sufficient statistics of Model 4 with GW of type OTP and estimated decay parameter. ESP1, ESP2, and ESP3 refer to the number of pairs of students with 1, 2, and 3 edgewise shared partners of type OTP, respectively.

|  | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
|  | No GW | GW-OTP(0): | GW-OTP(.25): | GW-OTP(free): |
| $\theta_3$ Outdegree 1 | $-.930$ (.074) $***$ | $-.116$ (.074) | .006 (.075) | $-.587$ (.075) $***$ |
| $\theta_4$ Outdegree 2 | $-.855$ (.046) $***$ | $-.022$ (.048) | .503 (.049) $***$ | .593 (.051) $***$ |
| $\theta_5$ Outdegree 3 | $-.665$ (.034) $***$ | .137 (.038) $***$ | .765 (.040) $***$ | 1.362 (.044) $***$ |
| $\theta_6$ Outdegree 4 | $-.603$ (.032) $***$ | .094 (.037) $**$ | .731 (.039) $***$ | 1.565 (.043) $***$ |
| $\theta_7$ Outdegree 5 | $-.067$ (.029) | .487 (.034) $***$ | 1.044 (.036) $***$ | 1.920 (.040) $***$ |
| $\theta_8$ Outdegree 6 | $-.797$ (.046) $***$ | $-.409$ (.047) $***$ | .027 (.049) | .790 (.051) $***$ |

Table 5: Monte Carlo maximum likelihood estimates, including standard errors, of all outdegree parameters of Models 1–4. Monte Carlo maximum likelihood estimates of all other parameters can be found in Table 2. Significance at levels .1, .05, and .001 is indicated by $*$, $**$, and $***$, respectively. A graphical representation of GW-OTP is shown in Figure 5.

# C Outdegree estimates

Tables 5 and 6 show Monte Carlo maximum likelihood estimates, including standard errors, of the outdegree parameters of Models 1–4 and Models 5–8, respectively.

# D In-sample performance of Models 1–8 in terms of outdegrees

We present plots for assessing the in-sample performance of Models 1–8 in terms of outdegrees in Figure 14.

# E In-sample performance of Models 5–8 in terms of DSP and ESP

Figures 16 and 15 show the in-sample performance of Models 5–8 with GW terms of types OSP, ISP, RTP and ITP in terms of DSP and ESP statistics of types OSP, ISP, RTP and ITP.

|  | Model 5 | Model 6 | Model 7 | Model 8 |
|---|---|---|---|---|
|  | GW-OSP | GW-ISP | GW-RTP | GW-ITP |
| $\theta_3$ Outdegree 1 | $-.768$ $(.075)$ $***$ | $-1.196$ $(.076)$ $***$ | $-1.239$ $(.076)$ $***$ | $-1.344$ $(.075)$ $***$ |
| $\theta_4$ Outdegree 2 | $.270$ $(.050)$ $***$ | $-.101$ $(.053)$ $*$ | $-.633$ $(.047)$ $***$ | $-.919$ $(.051)$ $***$ |
| $\theta_5$ Outdegree 3 | $1.010$ $(.044)$ $***$ | $.807$ $(.048)$ $***$ | $-.045$ $(.039)$ | $-.451$ $(.042)$ $***$ |
| $\theta_6$ Outdegree 4 | $1.248$ $(.043)$ $***$ | $1.267$ $(.049)$ $***$ | $.246$ $(.038)$ $***$ | $-2.180$ $(.041)$ $***$ |
| $\theta_7$ Outdegree 5 | $1.654$ $(.040)$ $***$ | $1.885$ $(.046)$ $***$ | $.832$ $(.035)$ $***$ | $.411$ $(.037)$ $***$ |
| $\theta_8$ Outdegree 6 | $.577$ $(.053)$ $***$ | $.956$ $(.055)$ $***$ | $-.015$ $(.049)$ | $-.321$ $(.049)$ $***$ |

Table 6: Monte Carlo maximum likelihood estimates, including standard errors, of all outdegree parameters of Models 5–8. Monte Carlo maximum likelihood estimates of all other parameters can be found in Table 3. Significance at levels .1, .05, and .001 is indicated by $*$, $**$, and $***$, respectively. Graphical representations of GW terms of types OSP, ISP, RTP, and ITP are shown in Figure 5.

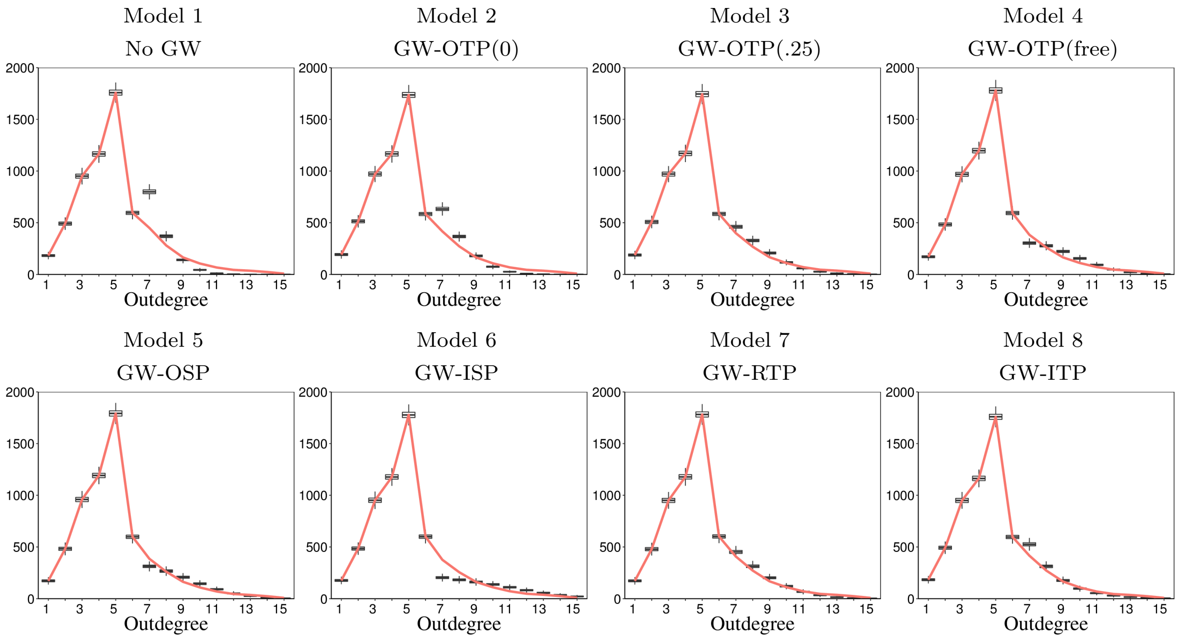


Figure 14: In-sample performance of Models 1–8 in terms of outdegrees. The red curves indicate the conditional expectations of the outdegrees given the observed data.

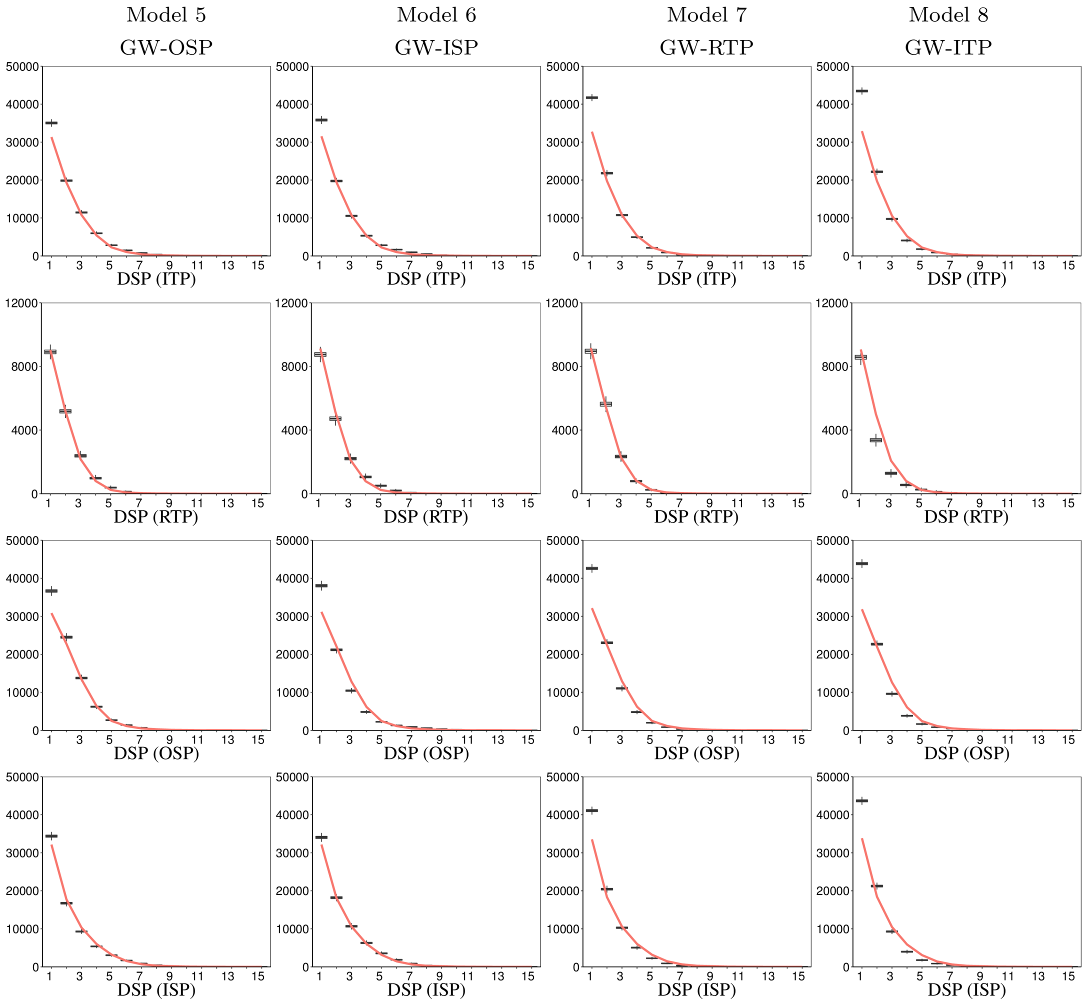


Figure 15: In-sample performance of Models 5–8 with GW terms of types OSP, ISP, RTP, and ITP in terms of DSP statistics of types OSP, ISP, RTP, and ITP. The red curves indicate the conditional expectations of the DSP statistics given the observed data.

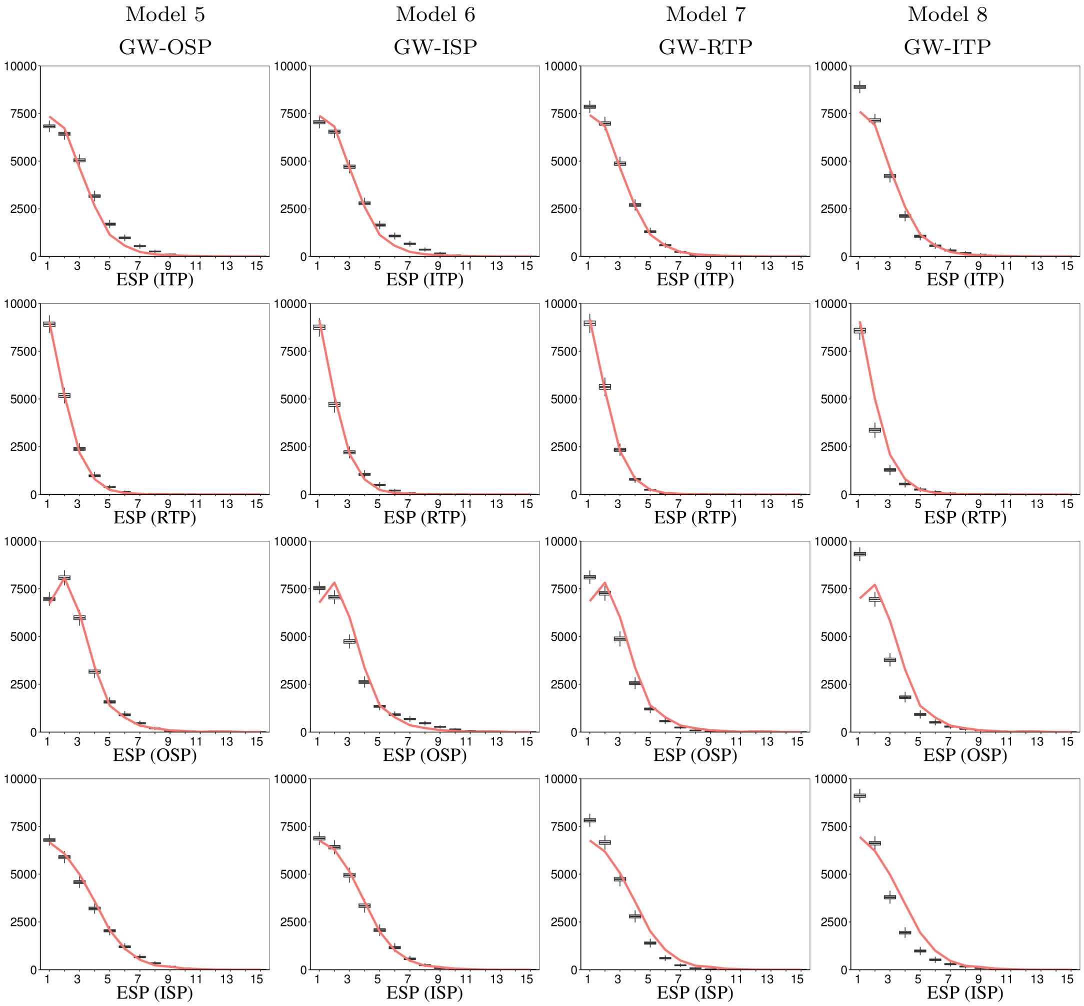


Figure 16: In-sample performance of Models 5–8 with GW terms of types OSP, ISP, RTP, and ITP in terms of ESP statistics of types OSP, ISP, RTP, and ITP. The red curves indicate the conditional expectations of the ESP statistics given the observed data.