

Research Master's programme Sociology and Social Research
Utrecht University, the Netherlands

MSc Thesis Suzanne Veerle Ekhart (6244386)
Applying ERGMs in analysis of large ego-network data
Accuracy and vulnerability for alter selection and alter sample size
May 2022

Supervisors:
Dr. R. Corten

Second grader:
Prof. Dr. F. van Tubergen

Preferred journal of publication: Social Networks
Word count: 9262

Abstract

In Social Network Analysis, one of the ways of gathering data of large networks is the use of surveys where individuals are asked about their connections. This type of data-collection does not always lead to a fully recorded, complete network. However, research questions can be about general properties of this complete network. These questions may be answered by using an extension of the already often used Exponential Random Graph Models (ERGM), suitable to analyze sampled ego-centered network data.

In this research, the accuracy of the ego-centered ERGM, as well as its vulnerability for certain biases that egos may have are tested using a large scale complete socio-centric data set from an online social network with approximately 10,4 million users. Biases that are included are biases that follow from egos nominating alters non-randomly and the maximum number of alters an ego can have. A model about gender homophily is used to discover the accuracy and vulnerability.

The findings suggest a low coverage and a high bias in the estimations of the ego-centered ERGMs, and little differences when changing the maximum number of alters per ego occur. However, the biases in alter selection seem to barely influence the results.

Contents

1	Introduction	2
2	Exponential Random Graph Models (ERGMs)	4
2.1	ERGMs and analysing sampled ego-network data	5
2.2	Utilization of ego-centered ERGMs	5
3	Earlier model evaluation	6
3.1	Ego-networks without alter-alter relationships	6
3.2	Ego-networks with alter-alter relationships	8
3.3	Measuring accuracy for large social network data	8
4	Data	9
5	Methods	10
5.1	Ego sampling and alter selection	10
5.1.1	No alter sampling	11
5.1.2	Baseline	11
5.1.3	Degree based	11
5.1.4	Embeddedness based	12
5.2	Model specification: gender homophily ERGM	13
5.3	Ground Truth	13
6	Results	14
6.1	Descriptive results	14
6.2	Results: gender homophily ERGM	16
7	Discussion	19
8	References	21
	Appendix I	23
	Appendix II	24

1 Introduction

Network Analysis is getting more and more important and popular in all fields of science, such as the study of Campana (2016) about criminal networks, the study of Kuperman and Abramson (2001) about the spread of infection disease and the study from Shao (2010) about human heartbeat dynamics. Collecting complete socio-centric data (from now complete network data) of a network can be hard; a researcher needs to define the boundaries of the network which are often not clear, and for large networks it can be very time consuming and expensive to record the connections of every individual in a network. Many of the complete networks are for example school classes, such as in the work of Knecht (2006). In school classes boundaries are clear and the complete network is small enough to collect data of all individuals. However, large complete network data is very rare. In social networks research, one of the ways of gathering data, especially within larger networks, is the use of ego-networks where individuals are asked about their connections. This type of data-collection does not always lead to a fully recorded, complete network. When questions about a very large network exist, it is almost impossible to gain data about the complete network and the choice for ego-network sampled data seems obvious. However, what if we don't know the complete network, but we do have questions about the network as a whole? This article aims to gather more information about Exponential Random Graph Models and their application on ego-networks. How can we learn things about the entire network using data that only contains incomplete ego-network data?

In earlier research, more descriptive analyses were used to gain insights about network properties. As such, in the work of Wellman (1979) and Dodds, Muhamad, and Watts (2003) a very descriptive analysis was used to analyze network data. In the last years, Social Network Analysis has shifted from describing network structures to identifying underlying micro-level mechanisms from which these structures follow (Lewis and Papachristos 2020). Most of the more conventional statistical models, such as regression analysis, come with some difficulties. Challenges can arise when looking at assumptions made when using the model but also with the loss of information in the analysis or the data collection (Krackhardt 1988; Newman 2002). When we want to estimate network properties from sampled network data, we require a proper statistical framework (Krivitsky, Morris, and Bojanowski 2019).

Here, Exponential Random Graph Models (ERGMs) become important. ERGMs are a type of model which can be used to analyze network data, and they are very accessible in terms of the used mechanisms, which can be on monadic, dyadic and higher-order scales (Lewis and Papachristos 2020). ERGMs and the use of exponential families are less problematic in terms of violation of assumptions compared to some more conventional methods such as regression analysis (Robins et al. 2007), where independence of the observations is assumed (Field 2013). Krivitsky and Morris (2017) focuses on ERGMs and provides extensions of the ERGM that can estimate models with the use of sampled ego-networks.

The extensions that make analysis of ego-centered data possible are especially interesting when answering

questions about large scale social networks where data collection of the complete network is not possible. Ego-centered ERGMs were previously evaluated using simulated, small-scale data, but evaluation using large-scale, real, complete data has not been performed yet. During ego-centered network data collection, egos are questioned about their relations and this type of data is, thus, in theory among others vulnerable for biases that result from the way egos select the alters they report and how much alters they can report. We want to discover how the accuracy of the method proposed by Krivitsky and Morris (2017) depends on differences in the way egos select their alters and the number of alters that is included in the data. In network data collection, participants are often asked to write down a number of their connections with a specific characteristic (such as friendship). But, often there is a maximum number of alters given that can be nominated by ego (Peter V. Marsden 2011). This type of data does not only capture a smaller part of the network, but also the type of data is different as every relationship or attribute in the data is provided by ego, and in the perception of the egos. In this research, we will research not only the influence of bias in the alter selection process, but also the influence of the maximum number of alters that is given during data collection on the accuracy of the ego-centered ERGMs.

Besides the sample size, the way egos select their alters can cause bias in the data and a bias in the results of the ERGM could follow. In order to use the ERGMs in further research, and in the analysis of ego-centered sampled data it is important to know how vulnerable the ERGM is on different alter selection methods from the egos. A researcher would be interested in whether particular designs to gather ego-network data result in highly accurate or inaccurate estimations of the overall network properties when using an ERGM for ego-centered data. The knowledge about whether sampling designs matter can also influence future data collection, as the researcher may influence on the way questions are asked in order to prevent certain types of bias in the data, in particular exert biases that influences the accuracy of the ego-centered ERGMs negatively. In order to test how the accuracy of the predictions of the model changes when different selection biases happen, we will estimate models with different ego samples and we will differ the selection method of the alters.

In this article I will go in depth on the appliance of the ERGM that was proposed by Krivitsky, Morris, and Bojanowski (2019). In order to test the functioning of the model as well as the constraints of the model I will apply the model to sampled ego-centered data from a large network that is fully captured in a data set (the population network) in order to look how accurate the predictions of the ERGM about network properties from the sampled ego-centered data are when applying them on a very large network. This research can extend our knowledge about the accuracy of the ERGM as we can extrapolate the ego-centered data to network properties that are known from the complete data, which is very useful when we want to learn something about very large social networks that cannot be analysed using a regular ERGM.

The rest of the article will continue as follows. First, some additional information about the ERGMs will be presented. There will be an overview on the ERGMs that can be applied on complete network data,

and this will be extended with information about the ego-centered ERGMs. Following with some examples from usage of the ego-centered ERGMs in empirical research. Next, an overview will be given of the model evaluations of the ego-centered ERGMs that have been carried out up to now. In section 4, the data will be presented following by the methods that are used in section 5. Then, the results of the analysis are presented and last the conclusion and discussion points are discussed.

2 Exponential Random Graph Models (ERGMs)

In recent research, the use of Exponential Family Random Graph Models (ERGMs) is becoming more popular. ERGMs are a class of statistical models that are able to overcome dependency issues by accounting for the presence and absence of network ties (Lusher, Koskinen, and Robins 2013). They are generalizations of the more known dyadic independence models such as the p1 model, which is an early example of this type of model (Robins et al. 2007). ERGMs are used to make causal claims and explain why an observed network is the way it is. They are also used to model network structures both in simulation studies and when modelling observed network data. (Lusher, Koskinen, and Robins 2013). The models can give new opportunities in exploring network effects and structures (Butts 2018).

Assuming that a network is being built from local patterns such as reciprocity, then connections are formed in response to other ties in their environment without independence of the ties. (Lusher, Koskinen, and Robins 2013) Ties are formed based on the presence (or absence) of other ties (Lusher, Koskinen, and Robins 2013; Robins et al. 2007) and similarity or dissimilarity in node level attributes which in itself results in interdependence of the presence and absence of ties in the network.

In an ERGM, the set of nodes is considered fixed and the set of possible networks and the probability that it occurs is represented by the probability distribution on this set of networks. (Robins et al. 2007) When we think of an ERGM very intuitively, we fit the model by maximizing the probability of observing the observed network given the model over all networks with the same number of edges that could have been observed and with that information choose our parameters (Cranmer et al. 2017). When specifying an ERGM, the researcher chooses a set of parameters of interest based on theoretical knowledge and fits this model to an observed social network. The parameters used to analyze the above mentioned network structures and processes are estimated based on this fit (Lusher, Koskinen, and Robins 2013). The model parameters can be based on theory driven approaches such as homophily or reciprocity (Robins et al. 2007).

A great advantage is the flexibility of the ERGM. It is possible to add monadic, dyadic, and higher-order structures to the model. (Lewis and Papachristos 2020) In its most general form, the maximum amount of parameters is $d - 1$, with d as the number of ties in the network. This would lead to a maximally saturated model where each tie has its own probability (Goodreau, Kitts, and Morris 2009). When a theory or model is too abstract, it is not as realistic as we would like it to be, and application to the real world can be hard.

But, when we decrease abstraction too much (as with the maximally saturated model), the model is not easy to understand (Raub, Buskens, and Van Assen 2011).

$$p(x) = \exp\{\theta^T z(x)\} / \Psi(\theta) \quad \text{with} \quad \Psi(\theta) = \sum \exp\{\theta^T z(x)\} \quad (1)$$

In an ERGM, the probability that a relation or tie exists is modelled as a linear function of predictors. This looks very much like the logistic regression but is in fact different. In logistic regression, the ties are only on one side of the regression equation and often limited to one predictor. (Field 2013) This is different from the ERGM, where the tie is present at both sides of the equation and often included in multiple predictors, which makes them recursively dependent. This ensures that the relation and interdependence of the ties is modeled more explicitly compared to a normal logistic regression. (Goodreau, Kitts, and Morris 2009)

2.1 ERGMs and analysing sampled ego-network data

As data collection of a complete network can be difficult and expensive, data can be observed by ego-network data collection where the connections to alters of an ego from a specific type are gathered, as well as all relations between those alters. In practice the information of the relation between the alters is perhaps not reliable (Crossley et al. 2015) due to the fact that the information is provided and biased by the perception of the ego (Peter V. Marsden 1990).

The conventional ERGMs are not suitable when analyzing sampled ego-network data. Thus, an extension of the ERGMs was developed by Krivitsky and Morris (2017). The inferential goal of the model was to fit ERGMs to unobserved networks based on samples of the complete network in the form of ego-networks. The gathered ego-network data can be analysed using the new types of ERGMs. The ERGM would allow us to recover possible complete networks from which the sample may have been drawn (Krivitsky, Morris, and Bojanowski 2019) which could give us information which we can use to analyze the network structures and processes empirically. Krivitsky and Morris (2017) present a method in which relationships between alters are not included. They later extended their method by including possibilities to extend a model with alter-alter relationships (Krivitsky, Morris, and Bojanowski 2019). With ERGMs to analyze ego-centered data we can gain knowledge about networks where only incomplete network data is gathered.

2.2 Utilization of ego-centered ERGMs

Up to now, ERGMs for ego-centered data is not often used. This is not very surprising as the method was published quite recently, and the extension with alter-alter relationships is at the time of writing not even published in a research journal. More general, ERGMs are already used in research in Social Sciences. The study of Khalilzadeh (2018) researched travelers and attitudes towards destinations for traveling. With

the use of semi-structured interviews, participants were asked to name four destinations where they would never choose to spend their vacation and some reasons why. A bipartite network with reasons and countries created using the information from the interviews was analysed using an ERGM.

The extension of the ERGMs as proposed by Krivitsky and Morris (2017) was, as far as I could find, used in two instances: a dissertation and a published article. The method where alter-alter relationships was included (Krivitsky, Morris, and Bojanowski 2019) is not yet used in a published article. The dissertation of Lee (2019) was about the trust relationship in the newsroom. The dissertation focuses on status-based homophily which they define as an individual tendency that similarity in social status breeds an informal connection on the work floor. To answer questions about this network homophily, the author used inferential network analysis with the ERGMs as primary method. In the research ego-centered data from journalists is used and analysis with an ERGMs is done with the use of the set of ERGM developed by Krivitsky and Morris (2017). The article from Hermans et al. (2017) explores the capacity to innovate the potential of multi-stakeholder platforms in Burundi, Rwanda and a part of the Democratic Republic of Congo. To do so, they apply ERGMs to look into the structural properties of the networks of the multi-stakeholder platforms. The network data they used consisted of ego-centered data and they did not know the size of the complete network. With the use of the set of ERGMs where we can analyze ego-centered data they overcome this problem.

3 Earlier model evaluation

The extension of the ERGMs that allows for analysis of ego-centered data is tested in the papers by Krivitsky et al. (2017; 2019). We will now provide an overview of the tests that were already done in order to determine the accuracy of the extension as well as how we will test the accuracy of the model when using large social network data in this paper.

3.1 Ego-networks without alter-alter relationships

In order to test the working of the model that was put forward in the published paper by Krivitsky and Morris (2017), a network with known ERGM parameters was simulated. From there, ego-centric samples were taken in order to create ego-centered data. The samples were taken using two different sampling designs. The first method was weighted and based on the way the National Health and Social Life survey was sampled. The second method was unweighted. Furthermore, they varied the sample sizes in order to test the effect of the samples size on the biases and the accuracy of the standard errors. As a result of the simulation study it was found that weighted sampling resulted in highly biased results for smaller sample size. The standard errors are found to be consistent for the sampling designs and more accurate when the sample size is higher.

The accuracy of the model is researched using simulated data. This can be problematic if we want to look into more complex, hidden social structures. Capturing these social structures in a simulated dataset can be hard and thus testing on real social data is desirable in order to test whether the ERGM behaves differently and whether the accuracy changes when using real social data.

Next to the model testing on simulated data, the ERGMs from Krivitsky and Morris (2017) were also evaluated with more substantive questions. These questions were about HIV prevalence in the US and were about race homophily in the population, possible differences in the propensity towards monogamy, and the impact of network features on overall network connectivity and network exposure by race and sex.

The data used in order to answer some substantive questions about HIV prevalence was from the National Health and Social Life Survey (NHSLs) from 1992. The data consisted of information on sexual partnerships of 3357 egos extended with some socio-demographic attributes of both the egos and the alters (the sexual partners of the egos). In total, 2555 alters were included in the data. In the ERGM they fit, they used homophily in sex and race as an edge covariate, the concurrency that was present (ego's having multiple relationships, and thus with a degree greater than one), and for modelling purposes they added a monogamy term which allows for group-specific propensities for monogamy.

In order to test the goodness of fit of the ERGM that was applied, they compared the observed degree distributions to realizations of the networks from the ERGMs.¹ They then reproduced complete networks from the fitted ERGMs and compared these networks with real epidemiological data in order to determine how accurate the estimated models are. Note that no full network was used to take samples from; the model predictions were compared to secondary data. The complete networks from the model predictions were quite well in line with the observed epidemiological data, which suggests accurate predictions of the ERGM.

Using the NHSLs data, it was possible to answer questions and look at the accuracy of the model using real social data of a very large social network. However, the network was not fully captured which made it impossible to compare the outcomes of the ERGM to the overall network properties that could have been known if the complete network was known. Instead of comparing to a full network, they used epidemiological data as a way to determine the accuracy of the outcomes of the ERGM. This gives an approximation of the accuracy, but is not a perfect method as differences can exist between the over-all epidemiological situation and the epidemiological situation for the group that was captured in the sample.

¹The authors chose the degree distribution, but it is possible to do it with every egocentric statistic that is not already in the model.

3.2 Ego-networks with alter-alter relationships

In the paper of Krivitsky, Morris, and Bojanowski (2019) an extension of the earlier ego-centered ERGMs was made, making it possible to include alter-alter relationships in the models. This extension was tested using simulated data from Faux Magnolia High. Real information of Faux Magnolia High was used to simulate a network with similar characteristics. They performed a parameter recovery simulation study in which they fit the ERGM to the complete network and then take 1000 independent egocentric samples with half of the network included. To each egocentric sample, they fit an egocentric ERGM having the same parameters as the complete network model they fit in the beginning. They now compared the distributions of the parameters to the ground truth, the parameters from the complete network. They could conclude that the results were quite accurate, the ego-centered data predicted over-all network properties well.

Krivitsky, Morris, and Bojanowski (2019) tested the model using simulated complete data and this made it possible to compare the model results to a complete network. Nevertheless, the network is very small: only about 1500 students are included in the data. Social networks that are of interest in sociology often cover a much larger network such as a city, neighborhood or online community. This makes it often not feasible to create a sampled ego-centered data set with about half of the whole community included. Thus, further tests on the method with larger data and a smaller proportion of the network captured in the data is needed in order to test how accurate the estimates of the ERGMs are when using larger networks and a smaller captured proportion. Also, the complete data was simulated and as apposed to real complete data which makes it vulnerable to errors in determining of the ground truth.

3.3 Measuring accuracy for large social network data

In this research we will provide a test of the accuracy of the model when analyzing a much larger network, where a smaller proportion of the full network is captured in the sample using the complete network of online social network Hyves. We will make use of real social data instead of simulating a data set in order to capture more hidden, complex social structures in the data. We will vary both the proportion of the network that is captured in the sampled data, and the sampling design used to create a sample of the complete network. In order to evaluate the ego-centered ERGM, we will fit one ERGM to different ego-samples of the data with different alters selected based on a limitation of alters per ego and a selecting strategy. Doing this, we simulate a survey data-collection where egos are sampled from the population and nominate a limited number of alters to write down in the survey.

In Social Sciences, and thus in Sociology, the way friendship networks form is of great interest. In the research about friendship networks, there is attention to homophily in friendship networks. This tendency of people to have more contact with similar people compared to contact with dissimilar people is researched a lot and proof is found in over a hundred studies (McPherson, Smith-Lovin, and Cook 2001). In this study we will

also focus on homophily, more specifically: gender homophily. The tendency of people to form friendships with individuals with the same gender more than with individuals from the opposite gender. The ERGMs will try to estimate to what extent gender-homophily exists and this will be compared to the actual levels of homophily (ground truth) in the complete data. We want to discover how well the ego-centered ERGMs predict the actual levels of homophily in the data, AND to what extent the predictions vary when changing both the number of alters that is included and the way the alters are nominated by ego.

A distribution of the coefficients of gender-homophily that are found for the different ego-samples and alter selections. This distribution will be compared to the actual level of gender-homophily in the complete data, with all egos and alters included. To determine how accurate the ERGMs are, we will look at the distribution of the coefficients for gender homophily. We want the estimated levels of homophily to be both reliable and valid, so the estimated values need to estimate the levels of homophily we find in the real data and the spread of the estimations cannot be too wide. If predictions have a great variance, the results cannot be interpreted as consistent.

4 Data

The data that will be used in this study is the complete Hyves network, recorded in July 2010. Hyves is a social networking website similar to the nowadays popular service *Facebook*. Users are able to send a friendship request to other users which can be reciprocated. If both users agree to a friendship connection on the service, they are recorded as friends. After that, they can send each other messages or post messages on each others timelines. In the network data, the nodes represent the users with a Hyves profile at the time of recording and the edges represent the friendship connections between the users. The data was retrieved directly from Hyves and is anonymized by deleting usernames.

There are 10,4 million users included in the data and there are a number of attributes of the users known. The gender and age as well as the location of the users is recorded. The location of the user is based on a location that was filled in by the user and the level at which is location is captured differs between users. For example, some users added a city such as Utrecht and others a specific neighborhood of the city such as Wittevrouwen in Utrecht.

According to Corten (2012), the data consists of people with on average 82 friendship connections to other users with a standard deviation of 135,23. Users are on average 27 years old (which is younger than the general population in the Netherlands) and men and women are equally represented in the data. Most of the users have a location in the Netherlands (about 86 percent), other users are spread over many other countries. (Corten 2012)

An ego-centered ERGM cannot yet deal with missing values very well and thus imputation of missing data is done. For the variable age, the mean age in the data is imputed (1.483.386 missing values imputed). A

random gender is given for every respondent that has an unknown gender (1.055.605 missing values imputed). For the complete data, this resulted in a dataset with 51 percent female users. These imputation methods are chosen as many other imputation methods would have the side-effect of causing a bias in the data which can influence the outcomes of the ERGM and thus the results of the analysis. The ERGMs were fitted using the package `ergm.ego` in R, which is free and open source (Krivitsky et al. 2003-2020). The code that was used can be requested from the author.

5 Methods

In order to compare the results of the ERGMs with the actual gender homophily levels, the ground truth (the actual level of homophily) needs to be discovered by looking at the complete data set. The ground truth will be determined looking at the log of the odds ratio on a homophily table. This is further explained in section 5.3.

After looking at the complete data, ego-centered samples of the data will be taken in order to fit ego centered ERGMs that estimate levels of gender homophily. For this project, for every ERGM a random sample of 1000 egos is taken in order to create an ego centered network data set. After sampling 1000 egos, the alters are sampled. In data collection a network data set can be created using a survey where, for example, a name generator is used. Ego is asked to report a certain amount of alters and write down some characteristics of the alters. Ego needs to determine which alters are in its ego network. This selection can contain bias and in this project the alters that are selected are selected based on theoretical reasoning about how egos could have bias when selecting their alters in for example a survey with a name generator. The theoretical reasoning behind the selection methods is explained for every method separately in paragraph 5.1.

After taking a selection of alters in a few different ways (see paragraph 5.1), we will look how the estimated models fit the data and whether there are differences in biases between alter sampling strategies. We will investigate how well the ERGM predicts levels of homophily when different alter selection methods are used based on the biases and coverage of the outcomes.

5.1 Ego sampling and alter selection

In this project, the alter selection method is varied in order to look at the vulnerabilities of the ego-centered ERGM, for alter selection bias of egos in ego-centered data collection. The egos are sampled randomly for all the different data-samples, we will not vary the ego sampling method. For every alter selection method, 100 different ego samples with 100 egos will drawn in order to eliminate effects due to any specific sample. The alters are selected using two different strategies, a baseline with random selection of alters and only an ego sample where no selection of alters is made. As selection strategies random, based on degree and based

on the embeddedness are included. The maximum number of alters that is included in the alter sample is varied; a sample with one, two, three, four, five, ten and twenty alters is taken which results in 100 samples for the baseline strategy and 700 alter selections for each of the other strategies.

5.1.1 No alter sampling

We'll start by running a model with no modifications at all. Only egos will be sampled, and all alters will be included in the data, where all alters (yellow) of all egos (red) are selected. This is visualized in figure 1.a. As a result, there will be no selection bias in the alter selection. The data now contains all of the egos true relationships. This is starting to resemble ego-centered network data. There are egos (red nodes) in the sample, and all alters in the complete data are included (yellow). When ego-centered network data is collected, however, the number of alters an ego can select is limited, but all alters that exist are included in this sample.

5.1.2 Baseline

A random selection of alters will be taken as a baseline selection strategy. A random selection of N alters is made for each ego. This can be seen in figure 1.b where per ego, a limited amount of alters is selected. We employ random sampling as a baseline model since the ego-centered ERGM assumes no bias in alter selection. We now have ego-centered data that is equivalent to survey network data, but is free of bias in how egos select their alters. The number of alters is limited (in the example in figure 1, to five), and hence alters are dropped for egos with networks that are larger than the limit.

For all egos the same number of alters is selected, differences in the probability to be selected between the alters follows, as alters from an ego with a lower degree have a higher probability to be selected compared to alters from an ego with a higher degree, as a greater proportion of the alters from the ego is selected. In data collection using surveys egos are in general also questioned about a fixed number of alters, without deviation between egos with more or less friends. Only less alters then the maximum is possible, as egos can choose to nominate less alters then the maximum which is also possible in the alter selection that we make in this project. If an ego has less friends then the maximum number of friends that is selected, all alters are included (this can also be seen in the visualization in figure 1), if an ego has less then five alters all alters are included in the selection.

5.1.3 Degree based

The first theory-based alter selection strategy is one in which egos select alters based on popularity. The degree in the complete data is a proxy for the users popularity. People tend to select alters with higher status or popularity during data collection (Light and Moody 2020), resulting in an ego network with a bias

in the number of friends alters they have, since alters with a high degree will be over represented in the ego-centered data with a restricted number of alters.

For the simulation of this process, the alters with the highest degree in the complete network are sampled. First, the data is restructured so that we get an edgelist with the sampled egos in the first column and their alters in the second. Second, the data is grouped by ego, and the alters are ordered within the ego-groups according to their degree in the complete network. The degree is obtained from a list with attributes in which the number of dyads in which a user participates in the whole data, is recorded. Finally, the N alters with the highest degree are selected within every ego. In figure 1.c a subset of ego networks is shown with the degree in the complete network as label for the nodes. One can see that for every ego, the alters with the highest degree are selected (yellow nodes). If an ego has less alters then the number of alters that is sampled, all alters are included in the sample.

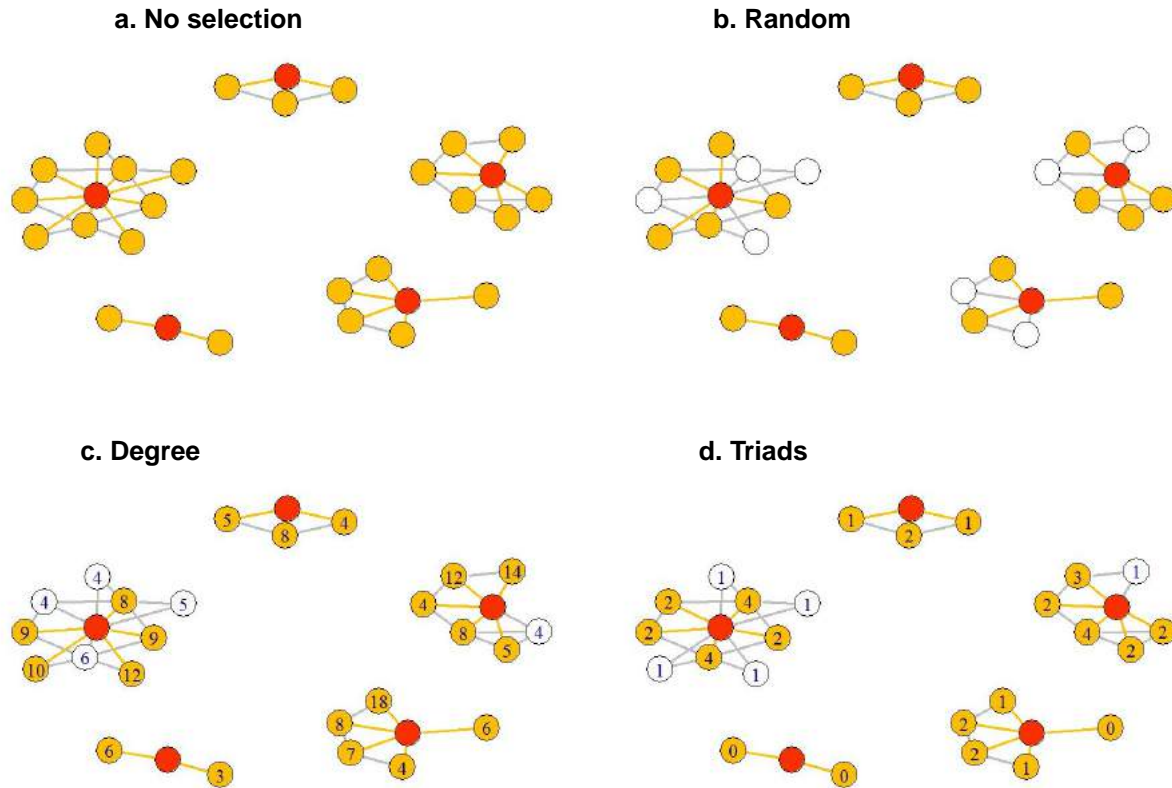


Figure 1: Selected alters (yellow) per ego (red), per selection method. ($N_{ego} = 5$, $N_{alter} = 5$)

5.1.4 Embeddedness based

In ego-network data collection, a name generator is often used and a not uncommon question that is asked is “with whom do you discuss important matters?” (Light and Moody 2020). Here, not only the interpretation

about what important matters are, but also the recall process of alters by the ego cause bias in the data. Evidence is found that when a respondent feels insecure, they may tend to report alters that are more embedded in their network. They would tend to report alters that are in a group with a high level of triadic closure instead of a person that is perhaps liked more but less connected to other friends (Smith, Menon, and Thompson 2012). A result of this process is that alters who have more friends in common with ego are more likely to end up in the sample.

In order to simulate this process, the alters with a higher number of closed triads with the ego are sampled. This is done by adding a column to the data with the number of closed triads ego and alter are part of in the complete data. Second, the data is restructured so that we get an edgelist with the sampled egos in the first column, their alters in the second and the number of closed triads they share in the third. Then, the ego sample is grouped by the ego and within the ego-groups the alters are sorted based on the number of closed triads ego and alter are part of. Last, the N alters with the highest number of shared closed triads are selected within every ego. In figure 1.d a network is displayed with blue edges when the number of closed triads ego and alter were part of was higher than one. Now we see that the sampled alters (yellow) are the ones with the highest number of shared, closed triads. If there are multiple nodes with the same number of shared, closed triads, the nodes that are sampled are randomly selected.

5.2 Model specification: gender homophily ERGM

In order to test the homophily in gender with the model, a nodematch term is added which looks at matching or non-matching gender. In the ERGM, no other control variables are added. This was done in order to get comparable results. When control variables are added, the effect of gender homophily would be influenced by the terms of the control variables and it would, thus, not be comparable to the ground truth anymore. An edges term that can be seen as the intercept for a network model, which looks at the log-odds of a tie to exist, is added.

5.3 Ground Truth

The results of the ERGM will be compared to the actual levels of gender homophily in the complete data. This will be done using the logarithm of the odds ratio calculated on the cross table about existence of ties and homophily/heterophily of the ties as described by Bojanowski and Corten (2014). The result is visualized in table 1 and the number of existing homophily and heterophily ties is filled in based on the edgelist of the Hyves data. When looking at the table, we observe that $6,96 * 10^{-6}$ percent of the possible heterophily ties and $8,78 * 10^{-6}$ percent of the possible homophily ties exist in the Hyves network. This indicates that the probability of finding an existing tie in the group with homophily ties is higher compared to the probability of finding an existing tie in the group with heterophily ties. The odds ratio that is calculated from the

homophily table 1 is equal to 1,26 and, thus, the log-odds is equal to $\ln(1,26) = 0,23$.

	Homophily	Heterophily
Tie exists	238.645.764	188.934.436
No tie	$\frac{(N_{male}^2 + N_{female}^2 - N)}{2} - 238.645.764$ $\approx 2,719 * 10^{13}$	$(N_{male} * N_{female}) - 188.934.436$ $\approx 2,716 * 10^{13}$

Table 1: Homophily table Hyves network

6 Results

6.1 Descriptive results

In order to explore to what extent the selection methods influence the descriptive statistics of the samples we made a plot visualizing the statistical kernel density of the average age of alters, the proportion of female alters, the average degree of alters and the average number of shared closed triads an alter has with its ego. The results are shown in figure 2. The colour represents the maximum number of alters included in the ego-networks and the rows divide the different alter selection methods.

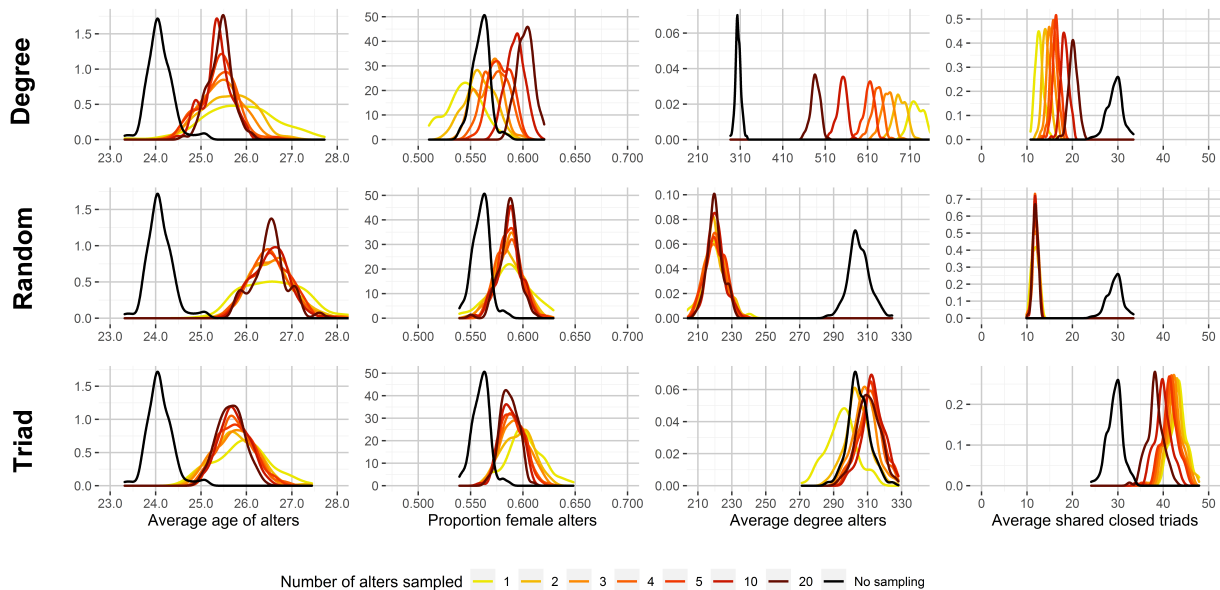


Figure 2: Descriptive statistics on the samples per alter selection method (*note: the x-axis for the average degree of alters differs between degree and random/triad.*).

For the average age of the alters we see that for all selection methods the average age is higher compared to the sample where all alters are preserved. The differences between the three methods are not very large

when looking at the average but we observe a higher variance when less alters are included, especially when selecting alters based on their degree. The higher average age for the alter selections compared to the ego sample with all alters included could be caused by the way outliers in age are spread over the various egos.

It could be the case that outliers in age (with a very low reported ² age) are over represented in ego networks from egos with a high degree. As we know, alters of high degree egos have a lower probability to end up in the selection. Thus, a single outlier in age in the sample with all alters has less influence on the average age in the selection.

When focusing on the proportion of female alters in the sample we can see that this is also higher for random and triad alter sampling compared to the sample with all alters included. But, when looking at alter selection based on the degree of the alter we see that the proportion of female alters in the sample scales with the maximum number of alters that is selected. When more alters are selected, the proportion of female alters is higher. A reason for this can be found in a difference in the average degree between female and male users. If female users on average have a lower degree, the proportion of female alters lowers when only the N alters with the highest degree are selected.

Moving to the average degree of the alters we observe a higher degree when using the degree as selection method. When only two alters per ego are selected, the average degree of the alters in the sample is almost 700 compared to 310 when all alters are included. This is a logical consequence of alter selection by degree. When only the two alters with the highest degree are selected, the average degree will be much higher than when all alters are preserved. When selecting randomly, the degree is much lower compared on alter selection based on triads and to the sample with all alters preserved, which have similar average degree of alters.

Last, we can observe an influence of shared closed triads as selection criterium on the average number of closed triads. When less alters are selected, the number of shared closed triads between the alters and their ego increases. This is theoretically understandable as we drop alters with lower number of shared closed triads and thus the average goes up. When randomly selecting the alters, the average number of shared closed triads is lower compared to the sample with all alters included. The degree selection method results in a lower average number of shared closed triads than when less alters are sampled. This indicates that users with a high number of shared closed triads, which are more embedded in the egos network, on average have a lower degree.

For all different alter selection methods and the different alter selection sizes the odds for gender homophily was determined and displayed in figure 3. The odds for a gender homophily tie can give in a descriptive way an idea what happens to the distribution of homophily and heterophily ties in the samples that result from the different alter selection methods. It is important to notice that these odds cannot be compared to

²The age in the data is self-reported by the users and thus not completely reliable.

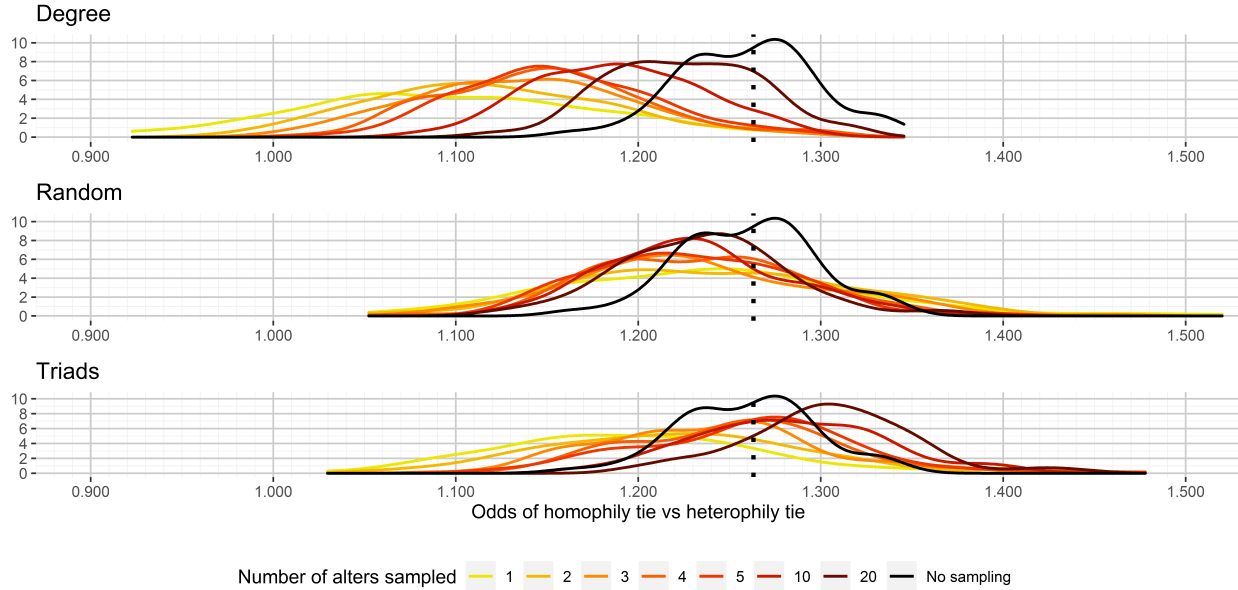


Figure 3: Density plots of the odds of homophily divided by the method and number of alters sampled.

the log-odds that is determined as the ground truth as this odds is not calculated based on both presence of ties and homophily/heterophily of ties, but only on the homophily/heterophily of ties as the presence of ties cannot be determined due to the type of network data (ego-centered).

We observe that the odds in the samples was about 1,25 when all alters were preserved, which implies that the number of homophily ties in the data was higher than the number of heterophily ties and gender homophily is likely. When alters were selected based on the degree the odds was found to be lower. The size of this bias was related to the number of alters that was selected. If fewer alters were selected, the odds became lower. Random selection of the alters resulted in similar results for all sizes of the selections and the odds was similar to the odds when no selection was made. Looking at the odds ratio for the selection based on triads, we find a higher odds when more alters are selected but differences are small. Also, it seems that when the number of alters increases, the domain of the density decreases which means that there is more variation between the different ego samples.

6.2 Results: gender homophily ERGM

The results of the ERGMs are shown in figure 4. The three alter selection methods are presented in three different plots. The first plot contains the coefficients with degree selection, the second with random selection and the last with selection based on the number of shared closed triads ego and alter have. The ground truth is added to the figures as a dotted, vertical line at 0,23. On the x-axis the coefficient about gender homophily is shown and on the y-axis, the statistical density is visualized.

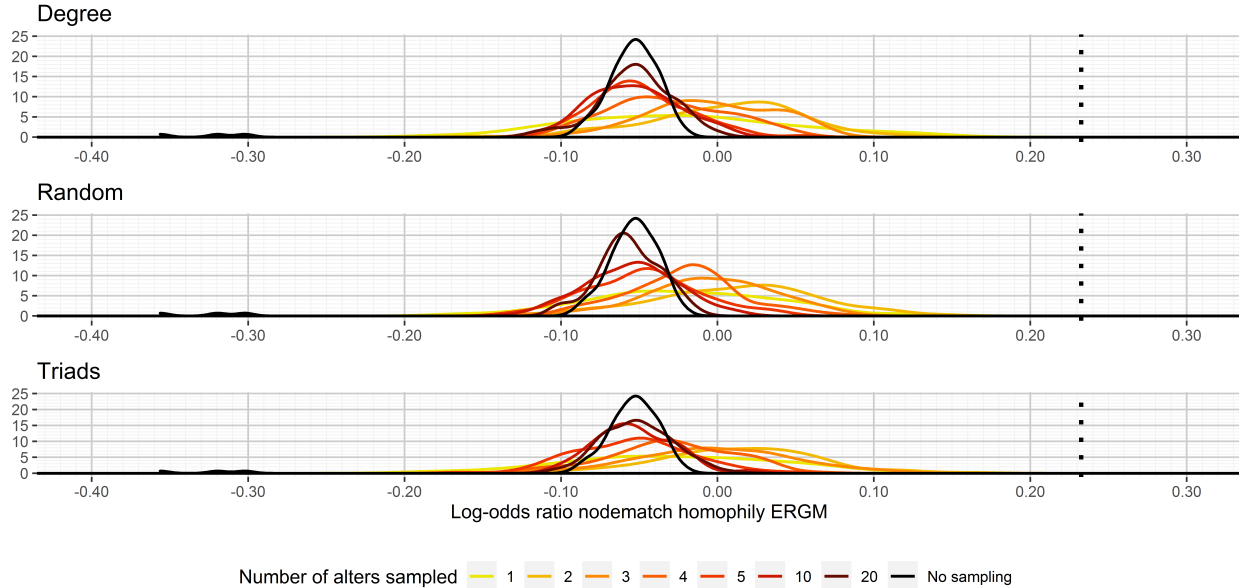


Figure 4: Density plot of the gender homophily coefficients that were fitted in the ERGM split by method and number of selected alters. The vertical line represent the log-odds of homophily in the complete data.

We observe that the three methods for alter selection have similar predictions when looking at gender homophily. All coefficients are around $-0,05$ which indicates that the log-odds of a within-group tie are $0,05$ smaller than an across-group tie. When looking at the differences between the different sizes of the alter selections (represented with different colors) we observe that the variance of the gender homophily term increases when a smaller number of alters per ego is selected. Also, the coefficient shifts up when less alters are added. When all alters are added, all 100 different ego-samples resulted in a negative homophily term, whereas some selections with a smaller number of alters sometimes resulted in positive homophily coefficients.

This could be caused by the way alters are selected and the probability alters end up in the ERGM data. It can be argued that users with a smaller number of friends have, on average, a stronger relationship with their friends in general. As alters of an ego with a low degree have a higher probability to end up in the selection (as mentioned in 5.1.2), the relative influence of stronger ties in the ERGM grows. We know from past research that strong ties are more likely to be homophily ties compared to weak ties (McPherson, Smith-Lovin, and Cook 2001), which would explain the higher levels of predicted homophily for smaller number of alters per ego.

The log-odds of the homophily ties in the complete network is equal to about $0,23$. As the result of the ERGM can be related to the log-odds directly we observe a great difference between the odds ratio and the predicted coefficients in the ERGM. The log-odds is about $0,23$ and the coefficients are almost all below $0,15$, which indicates a discrepancy in the measurements of gender homophily between the ground-truth

and the homophily terms in the ego-centered ERGMs. Looking at the bias, we observe that the difference between the mean coefficient and the ground truth lies for all alter selection methods between 0,22 and 0,29 (see appendix 2). This is confirmed when looking at the coverage, which is the proportion of ERGMs where the ground truth lies within the confidence interval around the coefficient of gender homophily. The coverage for all different alter selection methods was 0, which means that none of the ERGMs resulted in valid results. ³

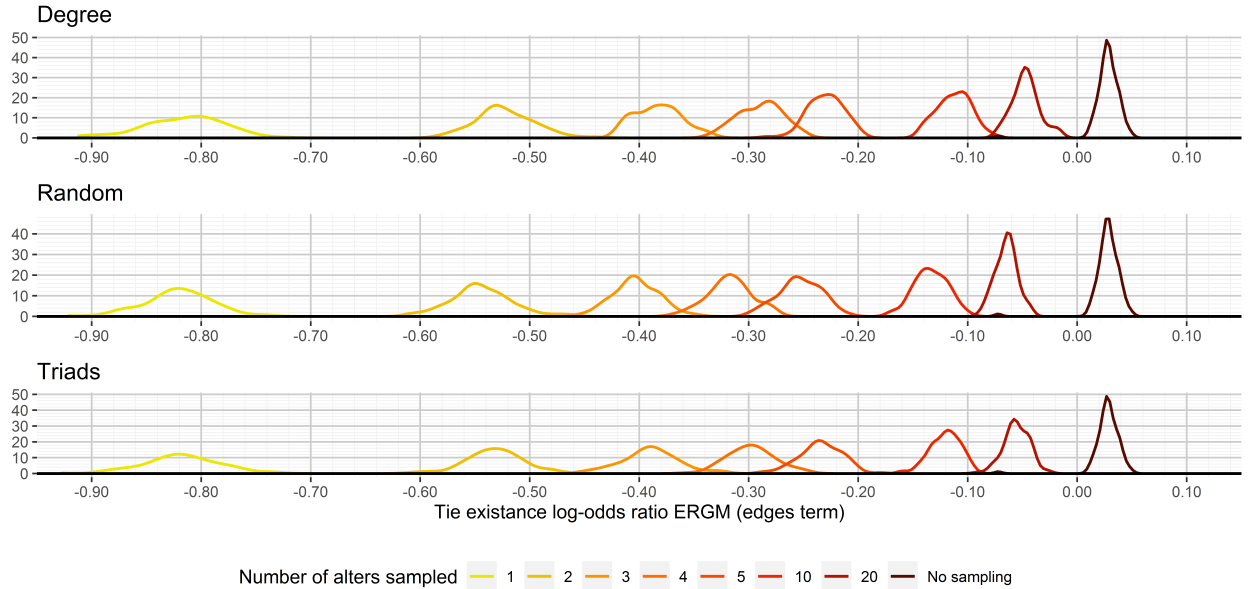


Figure 5: Density plot of the different edges coefficients that were fitted in the ERGM split by method and number of sampled alters.

When focusing on the edges term that was fitted in the model (visualized in figure 5), we can observe a large correlation between the number of alters that was selected and the edges term that was fitted by the ego-centered ERGM. The edges term in an ERGM can be interpreted as the overall density of the network. A smaller number of alters resulted in smaller predicted density, which indeed had to be expected given that less ties are included.

Last, the ego-centered ERGMs make a estimation of the population size: the pseudopopulation size. In figure 6, the pseudopopulation size is visualised. When the number of alters included increases, the pseudopopulation size increases as well. With more than two alters per ego included, the pseudopopulationsize for the random alter selections differ from the degree and triad alter selections. For those selections, the pseudopopulationsize is estimated lower compared to the estimates for degree and triads.

³A table with the coverage is not provided as it contains only zeroes, but the code with which it is calculated can be requested by the author.

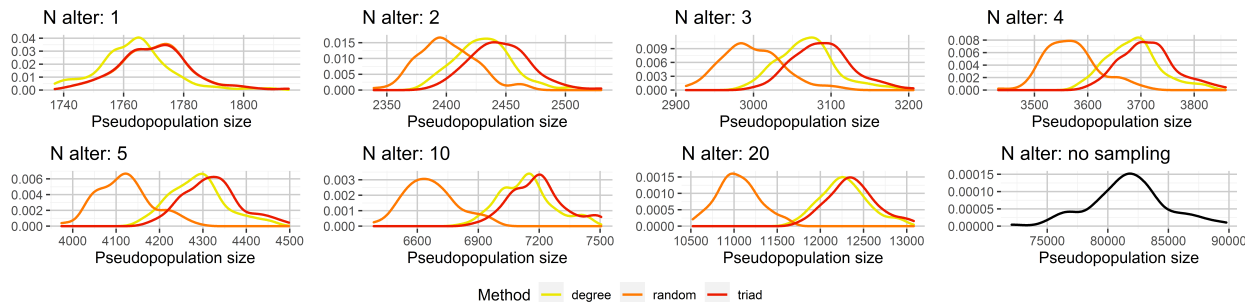


Figure 6: Estimations of the pseudopopulation size split into the number of alters included and the alter selection method used.

7 Discussion

We investigated the influence of bias an ego can have when nominating alters during network-data collection on the accuracy of ERGMs. Various alter selection methods (i.e. random, based on degree and based on embeddedness) were used to create ego samples with a specific maximum number of alters in order to create ego-centered network samples from the large social network Hyves. The size of the ego-centered networks was also varied by using different maximum numbers of alters. Finally, an ego centered ERGM about gender homophily was fitted on the sampled data and these results were compared to the ground-truth, which are the actual levels of gender homophily in the complete data.

The bias in alter selection was found to have minimal influence of the outcomes of the ERGMs when the number of alters was held constant. The predicted coefficients had greater spread when fewer alters were selected. Additionally, it was found that the number of alters selected had some influence on the outcomes of the ERGM. The levels of homophily were predicted higher when a smaller number of alters was included per ego. But, the significance of the differences can be questioned. Looking at the comparison of the ERGM coefficients with the ground-truth we can conclude that none of the simulations gave accurate levels of gender homophily. Homophily was predicted higher in the ego-centered samples compared to the actual levels of homophily in the complete Hyves data.

According to the results, the bias in the network data that could result from way egos select the alters they report does not play a big role in the accuracy of the model. Even though there is some variation between the methods, the overlap of the density plots of the coefficients was large. We added a high level of bias as the alters were selected strictly based on their properties. In this article, we included a level of bias that assumes people to have complete information about the degree of their friends and the number of friends they have in common. In real life data-collection, participants presumably have limited information about the properties of their alters and likely would not have these extreme levels of bias in their alter selection. Thus, the levels of bias in survey data about networks is expected to be smaller than it was in the simulations. The differences in predicted coefficients is expected to be even lower when a form of noise would be added.

This implies that the way network data is collected and the sensitivity for biases in the way egos nominate alters during data collection is not of great influence on the results when using ego-centered ERGMs (to investigate gender homophily).

Additionally, the number of alters selected is found to have a larger influence on the outcomes of the ERGM compared to the alter selection bias. Still the differences that follow from the number of alters included in the data seem small. A suggestion can be made that not the bias an ego may have, but the number of alters an ego can fill in during data collection influences the results of an ego centered ERGM. This could result in biased results when a researcher doesn't take into account the bias that can result from the number of alters that can be filled in during data collection. Additionally, it needs to be noted that these types of problems may also occur when other analysis methods are used. More research is needed to discover how the influence of the alter selection size on the predicted coefficients occurs and whether this could also be problematic when using other methods.

The use of a large, complete online social network revealed the consequences of previously described selection methods (i.e., degree . . . , variation in numbers) for the accuracy of ego-centered ERGMs using gender homophily models. However, more research is needed to see if the results hold for other types of models or data. Though, large complete network data is rare which makes this a great challenge. Besides, the Hyves social network data that was used in this study was also not perfect. There was imputation needed for missing values on age and gender in order to use the ego-centered ERGMs which are not capable of dealing with missing values (yet). Nevertheless, the number of missing values that needed to be imputed was about 10 percent of the original data and thus not the majority of the data. It could be considerable to check the way the missing data is spread through the network.

In order to gain more knowledge about the influence of alter selection bias in ego network data collection and how this can influence results, more research is needed. Focus in this research can be the way models treat the data and deal with the number of alters included. How does the ego-centered ERGM treat the number of alters exactly, and how can the number of alters included be taken into account when using the model in research? We need to discover how we can close the gap between the results from the ERGM and the ground-truth and solve the validity problem that now exists, and thereby improve models for analyzing ego-centered network data for future research. As every minute of a participant is costly, researchers will often want to minimize the number of alters that can be filled in in a survey. However, the validity of findings is important and the sweet spot between cut costs and a high enough validity of the results needs to be found. The search for this sweet spot requires supportive research into the models that are used to analyze the data that is collected. Next to the more technical research and innovations that are needed, also more research needs to be done about understanding the internal processes an ego goes through during data collection. Although the type of selection biases tested here doesn't seem to influence the results significantly, more insight about possible bias in network data could help researchers interpret not only model results, but

also more descriptive results. As we have seen in section 6.1, large differences in descriptive statistics result from bias in the process where egos select their alters. Furthermore, similar studies could be done on more complex social processes, such as alter-alter relationships or more general properties of overall networks.

We gain more knowledge about how to properly analyze social networks by doing more research on network data collection, analysis methods, and how they all interact. Research will be more valid and reliable if we develop more accurate research methods and models, and this will provide us tools to get a better understanding of the world around us.

8 References

- Bojanowski, Michał, and Rense Corten. 2014. “Measuring Segregation in Social Networks.” *Social Networks* 39: 14–32.
- Butts, Carter T. 2018. “A Perfect Sampling Method for Exponential Family Random Graph Models.” *The Journal of Mathematical Sociology* 42 (1): 17–36.
- Campana, Paolo. 2016. “Explaining Criminal Networks: Strategies and Potential Pitfalls.” *Methodological Innovations* 9 (January): 2059799115622748. <https://doi.org/10.1177/2059799115622748>.
- Corten, Rense. 2012. “Composition and Structure of a Large Online Social Network in the Netherlands.” *PLOS ONE* 7 (4). <https://doi.org/10.1371/journal.pone.0034760>.
- Cranmer, Skyler J., Philip Leifeld, Scott D. McClurg, and Meredith Rolfe. 2017. “Navigating the Range of Statistical Tools for Inferential Network Analysis.” *American Journal of Political Science* 61 (1): 237–51. <https://doi.org/10.1111/ajps.12263>.
- Crossley, Nick, Elisa Bellotti, Gemma Edwards, and Martin G Everett. 2015. *Social Network Analysis for Ego-Nets*. <https://doi.org/10.4135/9781473911871>.
- Dodds, P. S., Roby Muhamad, and Duncan J. Watts. 2003. “An Experimental Study of Search in Global Social Networks.” *Science* 301 (5634): 827–29. <https://doi.org/10.1126/science.1081058>.
- Field, Andy. 2013. *Discovering Statistics Using IBM SPSS Statistics*. sage.
- Goodreau, Steven M., James A. Kitts, and Martina Morris. 2009. “Birds of a Feather, or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks.” *Demography* 46 (1): 103–25.
- Hermans, Frans, Murat Sartas, Boudy van Schagen, Piet van Asten, and Marc Schut. 2017. “Social Network Analysis of Multi-Stakeholder Platforms in Agricultural Research for Development: Opportunities and Constraints for Innovation and Scaling.” Edited by Frank van Rijnsoever. *PLOS ONE* 12 (2): e0169634. <https://doi.org/10.1371/journal.pone.0169634>.
- Khalilzadeh, Jalayer. 2018. “Demonstration of Exponential Random Graph Models in Tourism Studies: Is Tourism a Means of Global Peace or the Bottom Line?” *Annals of Tourism Research* 69 (March): 31–41. <https://doi.org/10.1016/j.annals.2017.12.007>.
- Knecht, Andrea. 2006. “Networks and Actor Attributes in Early Adolescence.” *ICS Codebook* 61.
- Krackhardt, David. 1988. “Predicting with Networks: Nonparametric Multiple Regression Analysis of Dyadic Data.” *Social Networks* 10 (4): 359–81.
- Krivitsky, Pavel N., M. S. Handcock, D. R. Hunter, C. T. Butts, C. Klumb, S. M. Goodreau, and M. Morris. 2003–2020. *Statnet: Software Tools for the Statistical Modeling of Network Data*. Statnet Development Team. <http://statnet.org>.

- Krivitsky, Pavel N., and Martina Morris. 2017. "Inference for Social Network Models from Egocentrically Sampled Data, with Application to Understand in Persistent Racial Disparities in HIV Prevalence in the US." *The Annals of Applied Statistics* 11 (1): 427–55. <https://doi.org/10.1214/16-AOAS1010>.
- Krivitsky, Pavel N., Martina Morris, and Michal Bojanowski. 2019. "Inference for Exponential-Family Random Graph Models from Egocentrically-Sampled Data with Alter–Alter Relations."
- Kuperman, Marcelo, and Guillermo Abramson. 2001. "Small World Effect in an Epidemiological Model." *Physical Review Letters* 86 (13): 2909.
- Lee, Seok Ho. 2019. "Diversity and Trust in the Newsroom: Examining the Role of Homophily on Establishing Trust Using ERGM." PhD thesis, Austin: University of Texas.
- Lewis, Kevin, and Andrew V. Papachristos. 2020. "Rules of the Game: Exponential Random Graph Models of a Gang Homicide Network." *Social Forces* 98 (4): 1829–58.
- Light, Ryan, and James Moody. 2020. *The Oxford Handbook of Social Networks*. Oxford University Press.
- Lusher, Dean, Johan Koskinen, and Garry Robins. 2013. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press.
- Marsden, Peter V. 1990. "Network Data and Measurement." *Annual Review of Sociology* 16: 435–63. <https://www.jstor.org/stable/2083277>.
- Marsden, Peter V. 2011. "Survey Methods for Network Data." *The SAGE Handbook of Social Network Analysis* 25: 370–88.
- McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27 (1): 415–44. <https://doi.org/10.1146/annurev.soc.27.1.415>.
- Newman, M. E. J. 2002. "Random Graphs as Models of Networks." *arXiv:cond-Mat/0202208*, February. <http://arxiv.org/abs/cond-mat/0202208>.
- Raub, Werner, Vincent Buskens, and Marcel A. L. M. Van Assen. 2011. "Micro-Macro Links and Micro-foundations in Sociology." *The Journal of Mathematical Sociology* 35 (1-3): 1–25. <https://doi.org/10.1080/0022250X.2010.532263>.
- Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher. 2007. "An Introduction to Exponential Random Graph (p^*) Models for Social Networks." *Social Networks*, Special section: Advances in Exponential Random Graph (p^*) Models, 29 (2): 173–91. <https://doi.org/10.1016/j.socnet.2006.08.002>.
- Shao, Zhi-Gang. 2010. "Network Analysis of Human Heartbeat Dynamics." *Applied Physics Letters* 96 (7): 073703.
- Smith, Edward Bishop, Tanya Menon, and Leigh Thompson. 2012. "Status Differences in the Cognitive Activation of Social Networks." *Organization Science* 23 (1): 67–82. <https://doi.org/10.1287/orsc.1100.0643>.
- Wellman, Barry. 1979. "The Community Question: The Intimate Networks of East Yorkers." *American Journal of Sociology* 84 (5): 1201–31. <https://doi.org/10.1086/226906>.

Appendix I

In this table, the descriptive statistics of the odds of homophily (figure 3) are shown.

N alter	Method	min	max	median	mean	sd
1	degree	0.92	1.29	1.10	1.10	0.08
1	random	1.05	1.52	1.24	1.24	0.08
1	triads	1.03	1.43	1.20	1.20	0.08
2	degree	0.98	1.29	1.11	1.12	0.07
2	random	1.06	1.40	1.24	1.24	0.07
2	triads	1.06	1.42	1.22	1.22	0.07
3	degree	1.00	1.29	1.14	1.14	0.06
3	random	1.10	1.46	1.22	1.23	0.07
3	triads	1.11	1.43	1.25	1.24	0.06
4	degree	1.02	1.31	1.15	1.15	0.06
4	random	1.08	1.40	1.23	1.23	0.06
4	triads	1.10	1.48	1.26	1.25	0.06
5	degree	1.03	1.29	1.15	1.16	0.05
5	random	1.12	1.40	1.22	1.23	0.05
5	triads	1.12	1.46	1.26	1.26	0.06
10	degree	1.08	1.30	1.19	1.19	0.05
10	random	1.13	1.39	1.23	1.23	0.05
10	triads	1.17	1.40	1.28	1.28	0.05
20	degree	1.12	1.32	1.22	1.22	0.04
20	random	1.12	1.37	1.24	1.23	0.05
20	triads	1.19	1.45	1.30	1.30	0.05
No sampling	N/A	1.16	1.35	1.26	1.26	0.04

Appendix II

In this table, the descriptive statistics of the ERGM gender homophily coefficient are shown separated by method and number of alters.

N alter	Method	min	max	median	mean	sd	bias
1	degree	-0.18	0.14	-0.03	-0.02	0.07	0.26
1	random	-0.17	0.14	-0.02	-0.02	0.06	0.25
1	triads	-0.19	0.13	-0.03	-0.03	0.07	0.26
2	degree	-0.15	0.12	0.01	0.01	0.05	0.23
2	random	-0.14	0.14	0.02	0.01	0.05	0.22
2	triads	-0.11	0.16	0.02	0.01	0.05	0.22
3	degree	-0.10	0.07	0.00	0.00	0.04	0.24
3	random	-0.12	0.08	0.00	0.00	0.04	0.24
3	triads	-0.11	0.11	0.00	0.00	0.05	0.23
4	degree	-0.10	0.05	-0.04	-0.03	0.04	0.26
4	random	-0.10	0.08	-0.02	-0.02	0.04	0.25
4	triads	-0.12	0.06	-0.03	-0.03	0.04	0.26
5	degree	-0.12	0.05	-0.05	-0.05	0.03	0.28
5	random	-0.13	0.04	-0.05	-0.05	0.04	0.28
5	triads	-0.17	0.05	-0.05	-0.05	0.04	0.29
10	degree	-0.12	0.01	-0.05	-0.05	0.03	0.29
10	random	-0.13	0.01	-0.05	-0.05	0.03	0.29
10	triads	-0.11	0.02	-0.05	-0.05	0.02	0.29
20	degree	-0.12	0.00	-0.05	-0.05	0.02	0.29
20	random	-0.10	-0.01	-0.06	-0.06	0.02	0.29
20	triads	-0.11	0.01	-0.05	-0.05	0.02	0.28
No sampling	N/A	-0.36	-0.02	-0.05	-0.06	0.05	0.29