

Exponential random graphs models for social networks

Lukas Hoffmann

Georg-August-University Göttingen and

Master in Applied Computer Science

Göttingen, Germany

Email: lukas.hoffmann2@stud.uni-goettingen.de

Abstract—This paper provides an overview about the class of exponential random graph models for social networks.

I. INTRODUCTION

In the recent years exponential random graph models are fast becoming acknowledged as one of the central approaches in analyzing social networks. The so called p^* -class models are probability models for networks on a given set of actors that allow the researcher to build them in a realistic way on the basis of social behavior. The network structure is being modeled by the presence and absence of network ties. Moreover, the network is being built of small local tie-based structures, called configurations. These include, for example, reciprocated ties or triangles. In the class of p^* -models it is suggested that configurations arise from local social processes, whereby actors in the network set up connections depending on other ties in their social environment. In addition, each configuration belongs to a parameter in the network.

A statistical approach means, that the researcher has to consider why the social ties in the network have arisen. Was the network created due to the process of homophily, reciprocity, transitivity or a combination of these? We can develop hypotheses about social processes that produce configurations by putting all the parameters together. In this case we test the effects of one parameter against the other and so infer the social processes that have constructed the network. Because of the statistical approach, a p^* -class model permits inferences about whether there are significantly more or less configurations like reciprocated ties or triangles, that we would expect in a network of interest.

In other words, a well-specified statistical model of an observed network enables us to understand the uncertainty connected with the observed outcomes. We can learn about the distribution of possible outcomes for a given specification of a model or we can estimate the parameters of the hypothesized model from which the data may have been generated [1, pp. 173-175].

This paper introduces the exponential random graph models beginning with an overview of the underlying logic of these kind of models including general framework for model construction. In Section 3 the general form of these models and is given and section 4 is about dependence assumptions. Section

5 outlines two different estimation techniques for parameter values and why they are necessary. Section 6 is presenting a concrete example of fitting a model to network data. In conclusion we summarize the principal aspects and present a short outlook to advanced topics beyond this academic paper.

II. THE LOGIC BEHIND EXPONENTIAL RANDOM GRAPH MODELS

Using Exponential random graph models is mainly about explaining patterns of ties in a social network. The researcher can learn about the social network by collecting data from the network he is interested in modeling. With this data it is possible to understand a given observed network and so obtain insight into the underlying process that create and sustain the network-based social system [2, p.1].

A. Parameters

Like many of the networks in the real world the observed network exists in only one realization from a set of possible systems of connections that we can study. Strictly speaking, the observed network is seen as one particular pattern of ties out of a large set of possible patterns [1, p.175]. We assume that the observed network is generated by an unknown stochastic process and referring to that, we propose a theoretically based hypothesis for this process. A common observed structural characteristic in social networks are reciprocated ties. One possible research question may be whether actors tend to reciprocate relationship choices. In this case a plausible hypothesis would be the following: The chance appearance of reciprocated ties in the observed network is higher probable than the chance appearance of a number of reciprocated ties if relationships occurred completely random. Structural characteristics like reciprocated ties are considered as a result of a social process. Here this social process is that actors choosing to reciprocate the choices of others [1, p.176]. In this small example two parameters for the stochastic model are determined. The first one mirrors the tendency for ties to occur at random and the second one mirrors the additional tendency for reciprocated ties to occur. Therefore we are able to suggest a model where the level of reciprocity is a parameter.

B. Probability distribution of all possible networks

By determining parameters like it was done in chapter 2.1, we are able to state whether a network is more probable than others. Robins et. al. define that a statistical model for a network on a given set of actors is assigning a probability to all possible networks on those actors. Note, that we represent networks as graphs, nodes and edges, where the set of actors is fixed. With this in mind, a probability distribution can be introduced which represents the range of all possible networks and their probability of occurrence under the model on the set of all possible graphs with this number of nodes [1, p.176]. Since the level of reciprocity is a parameter, graphs with a significant level of reciprocation have a higher probability than networks without reciprocity. Furthermore, the precise probabilities are depending on the values of the parameters. In this setup the observed network is a particular graph with a particular probability in the distribution.

At this point, we do not know the exact values of the parameters to assign probabilities to graphs in the distribution. The main goal is now to find the best values by estimating the model parameters using the observed network as a guide. The most common method to do that is the maximum likelihood criterion. The researcher chooses a parameter value and assumes that the most probable degree of that parameter is that which occurs in the observed network. This works, because a parameter is defined as zero when the corresponding structural characteristic occurs by chance. For instance, if a model has a reciprocation parameter and they are many reciprocated ties in the observed network, then a good model in terms of reciprocity will be one with a positive reciprocation parameter. If the researcher is confident that the parameter is positive, he infers that there are more reciprocated ties than expected by chance.

With a defined probability distribution on the set of all possible graphs the researcher is allowed to generate graphs at random conforming to their assigned probabilities. In the next step the graphs can be compared with the observed network by any other characteristic. If the model fits the data well, the sampled graph and the observed network will be quite similar in many different aspects. Furthermore the researcher could explain the rise of the network due to the modeled structural effects. Analyzing properties of the sampled graphs then may lead to a better understanding of networks that are likely to emerge from these effects.

C. Configurations and node-level attributes

Researchers assume that network ties can organize themselves into patterns because the presence of some ties is leading to more ties. Because of this the network is stated as a self-organizing system of relational ties. In theory there are local social processes that depend on the surrounding social environment like existing relations and generate dyadic relations. As we have indicated above a structural characteristic is a result of a social process. In addition, each parameter corresponds to a structural characteristic or in other words a local network pattern, termed a network configuration.

Remembering that we represent the model as a graph and then network configuration is a possible small sub-graph that may represent a local regularity in a social network structure [2, p.17]. Moreover possible attributes can be assigned to single nodes. Some examples of configurations in directed networks are given in Fig.1.

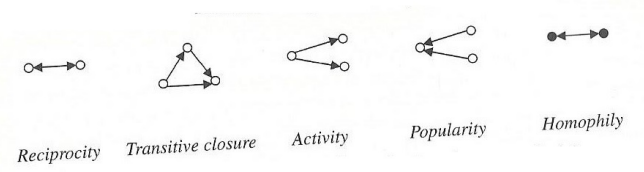


Fig. 1. network configurations and their underlying social process [2, p.18]

The first example on the left represents a reciprocated tie between two actors which is mentioned above. The next configuration represents transitive closure in form of a triadic structure. If two unconnected actors are connected to a third actor, it is likely to happen that they form a friendship tie between them. In fact there are other directed triadic forms where the direction of arrows is varying or ties in the triad might be reciprocated. The third configuration is called an out-2-star with two outgoing ties from one central node. Structures like that are often used to symbolize that an actor directs ties to many other network partners. For instance, a possible node-level attribute for this actor could be a superior, because he is assigning tasks to his staff members. Therefore this configuration is labeled as Activity. Complementary there are also in-star-configurations which are often described as network popularity. The last configuration in Fig.1 is a structure where two actors with the same attribute have a reciprocated tie, which is often used to show reciprocated homophily [2, p. 18].

D. General framework

[1] introduce a general framework for model construction where the researcher follows five steps in which he makes explicit choices that connect theoretical decisions to data analysis. In the following the essential points will be outlined and summarized.

1) *Each tie is regarded as a random variable*: The first step is that each tie is regarded as a random variable. Relationships do not happen spontaneously, but some relationships might be highly probable. With random variables the researcher knows that the model is not going to make the perfect predictions and the result is going to have some additional statistical noise, which we cannot explain. Now some basic notation which is also necessary for the next chapters will be presented. For each pair i and j of a set N of n actors, Y_{ij} is a network tie variable with $Y_{ij} = 1$ if there is a network tie from i to j and $Y_{ij} = 0$ otherwise. A realized value of Y_{ij} is y_{ij} with \mathbf{Y} the matrix of all variables and \mathbf{y} a matrix of realized ties in the network. \mathbf{Y} may be directed or non-directed. If \mathbf{Y} is directed there is a distinction between Y_{ij} and Y_{ji} and if \mathbf{Y} is non-directed Y_{ij} is equal to Y_{ji} .

2) *Dependence hypothesis*: In step two a dependence hypothesis is proposed which embodies the local social processes that are assumed to build up the network, especially the network ties. This assumption could also be that ties are independent from each other, meaning that people form social connections independently of their other social contacts. Mostly this is not a very realistic assumption, because friendships are usually created to connections to others like a triadic structure (see chapter 2.3). Ties may also depend on node level-attributes with for example homophily effects in a classroom or a sports club. More about dependency assumptions will be provided in chapter 3 where particular dependency assumptions are discussed.

3) *Connection between the dependence hypothesis and the particular form of the model*: Step three is about the implication of the dependence hypothesis to a particular form of the model. The model represents a distribution of graphs which are assumed to build up from network configurations. Examples for those configurations are described in chapter 2.3. The corresponding parameters that are related to presence or absence of the configurations in the observed graph have to be included in the model.

4) *Simplification of the parameters*: In step four a simplification of the parameters through different constraints is taking place. To define a model clearly it is crucial to reduce the number of parameters. Generally homogeneity effects are used to define suitable constraints. Applied to the model it means parameters are equated or related in other ways. For instance, there is only one parameter for reciprocity effects across the entire network. In this case it is assumed that the whole set of reciprocity parameters for each reciprocated tie are equal.

5) *Parameter estimation*: In the last step model parameters are estimated and interpreted. Parameter estimation is usually the focus of research. Nevertheless this step implies that the other four steps have already been executed. Empirically, for a given model and a given data set, it is possible to estimate the parameter values that are most likely to have generated the observed graph. This can be eminently complicated if the dependence structure is complex, as it should be for a good fitting and realistic model. "Having obtained parameter estimates, as well as estimates of the uncertainty of estimation, we may then take full advantage of having a statistical model for the network that is constructed from specifiable dependence assumptions and that is estimated from observed network data" [1, p.178]. In this paper parameter estimation is discussed in chapter 5.

III. OVERVIEW ABOUT DIFFERENT DEPENDENCE ASSUMPTIONS AND MODELS

A fundamental concept of exponential random graph models is the dependency between network ties. With dependency among ties it is possible to explain the tendencies for certain patterns of ties to arise. Furthermore a dependence assumption constraints the possible configurations in the model. This chapter is presenting four different assumptions in order to

understand the next chapter where the general form of exponential random graph models is presented. Each of them has has different definition of what counts as a local social substructure and for this reason implies a different family of an exponential random graph model.

A. Bernoulli Assumption

Bernoulli graphs are the simplest dependence assumption, because it is assumed that all possible distinct ties are independent of one another. Edges occur randomly according to a fixed probability. The next chapter will be showing that only configurations are relevant to the model are those in which all possible ties are conditionally dependent on each other. Therefore the only possible configurations are single edges. This assumption is considered as quite unrealistic for social networks, but it can be used as a baseline for comparison.

B. Dyadic-Independent Assumption

Dyadic models are a more complicated approach for directed networks. The assumption is that dyads and not only edges are independent from one another. That leads to the fact that there are two types of configurations in the model, single edges and reciprocated ties.

C. Markov Dependence Assumption

Frank and Strauss [3] proposed the "Markov dependence assumption" in which two tie-variables are conditionally dependent if they share a node. In contrast to the previous assumptions the Markov dependence assumption has different view. "If instead of considering the edges of the graph as connecting nodes, we think of the nodes of the graph as connecting the edges" (2, p.57). For instance, if a node connects the possible edges (i, j) and (i, k) , then the tie-variables related to (i, j) and (i, k) are conditionally dependent to the rest of the graph. Providing a real-world example here will help to understand the relations.

The relationship between Peter and Mary may be dependent on the presence or absence of a relationship between Mary and John. The two possible ties Y_{PM} and Y_{MJ} are conditionally dependent, because they share the same node "Mary".

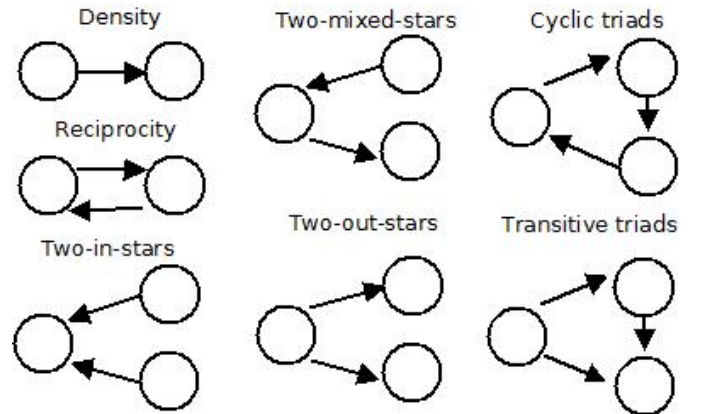


Fig. 2. Configurations and parameters for Markov random graph models

Fig.2 is presenting configurations and their associated parameters for Markov random graphs and for directed networks. The first two configurations (edge and reciprocity) were the possible one in the Bernoulli and dyadic independence model. In addition to that various two-star effects are conceivable, like the two-mixed stars or cyclic triads. The inclusion of all these parameters is a strength of Markov random graphs. The full parameter set also includes all possible higher order stars as well, but if all this stars are included there are too many parameters for the model to be estimable [1, pp. 182-184].

IV. GENERAL FORM OF EXPONENTIAL RANDOM GRAPH MODELS

[1] present the following form for exponential random graph models:

$$Pr(\mathbf{Y} = \mathbf{y}) = \left(\frac{1}{\kappa}\right) \exp\left\{\sum^A \eta_A g_A(\mathbf{y})\right\} \quad (1)$$

where:

- (i) the summation is over all configurations types A ; different sets of configuration types represent different models;
- (ii) η_A is the parameter corresponding to the configuration of Type A , which is non-zero only if all pairs of variables are expected to be *conditionally independent*;
- (iii) $g_A(\mathbf{y})$ is the *network statistic* related to configuration A . $g_A(\mathbf{y}) = 0$ if the configuration is observed in the network \mathbf{y} and is 0 if not;
- (iv) κ is a normalizing quantity to ensure that (1) is a proper probability distribution.

The model represents a probability distribution of graphs on n nodes, where the probability of observing any particular graph \mathbf{y} in this distribution is depending both on the network statistics $g_A(\mathbf{y})$ and the different non-zero parameters η_A for all the corresponding configurations A in the model. The configurations might include many different ones, like reciprocated ties, transitive closures or activities. Consequently the model makes it possible to check a variety of structural regularities.

At this point dependence assumptions (see chapter 3) are deciding. This is the case, because they have the property of picking out different types of configurations which are relevant to the model. Point (ii) from above says that parameters are equalled zero whenever variables in the configuration are independent on each other. That is the reason why only configurations are relevant to the model are those in which all possible ties in the configuration are mutually contingent on each other. [1, p.179]

In the previous chapter it was noted, that dependence assumptions always constraints the possible configurations in the model. If a set of possible edges is representing a configuration in the model, then (1) suggest that any subset of possible configurations is also a configuration. Due to this fact single edges are always configurations. Particular dependency assumptions will be presented in chapter 3.

A configuration A corresponds to a small network structure with a subset of tie variables. If a dyadic dependence assumption (see chapter 3.2) is applied, then reciprocity parameters are for certain part of the model. Consequently every single dyad has its own configuration with a set of variables, for instance $\{Y_{12}, Y_{21}\}$, $\{Y_{13}, Y_{31}\}$ and so on. Naturally if both of the ties are present in the observed graph for any of these configurations, a reciprocated tie can be seen. Because of that a configuration represents a structural characteristic that may be observed in a realized graph \mathbf{y} .

Furthermore configurations represent possibilities. It is clear that not all of the possible ties in a given configuration will exist in a graph \mathbf{y} . Only some of them will be observed. In the case of a dyadic assumption some ties will be reciprocated and some will not. Whether a configuration A is in fact observed in a network \mathbf{y} is given by the graph statistic $g_A(\mathbf{y})$. For a reciprocity configuration A , that statistic says whether there are reciprocated ties between a pair of nodes or not.

In addition, the strength and the direction of any parameter value will affect how frequently the corresponding configuration is observed [2, p.8]. If the parameter value is large and positive, the corresponding configuration in the distribution of (1) will be observed more frequently than if the value were zero. On the other hand, if the parameter value is large and negative it has to be assumed that the configuration is observed less frequently than if the parameter is zero.

Because (1) has an exponential term, such distributions have been named as *exponential random graph models* and the models which apply a Markov dependence assumption (see chapter 3.3), the so-called *Markov random graphs*, are one particular class of them.

V. ESTIMATION

Previously it was shown that information regarding configurations are enough to sign a probability to graphs in a distribution. If the strength of a parameter value is increased, graphs with that particular configuration become more likely in the resulting distribution. This fact is used in estimation and simulation techniques.

Empirically, for a given model and a given dataset, it is possible to estimate the parameter values that are most likely to have generated the observed graph. These estimates are named the "Maximum likelihood estimates. Without going into detail, it is discovered that, depending on the data, there may be serious problems with standard maximum likelihood estimates for Markov models. In fact, these kind of estimation is only tractable for very small networks, because of the difficulties in calculating the normalizing quantity κ in equation (1). To overcome this problem new techniques have been developed, the Monte Carlo maximum likelihood estimation techniques. Both types of approaches with an example will be outlined in the following.

A. Pseudo-likelihood estimation

Pseudo-likelihood estimation was first suggested by Strauss and Ikeda (1990) in order to estimate the parameters of

Markov models. In this kind of estimations the equation (1) is transformed into a equivalent *conditional* form which is noted as:

$$\log \left[\frac{Pr(Y_{ij} = 1 | \mathbf{y}_{ij}^C)}{Pr(Y_{ij} = 0 | \mathbf{y}_{ij}^C)} \right] = \sum_{A(Y_{ij})} \eta_A d_A(\mathbf{y}) \quad (2)$$

where:

- (v) the summation is over all configurations A that contain Y_{ij} ;
- (vi) η_A is the parameter corresponding to the configuration;
- (vii) $d_A(\mathbf{y})$ is the *change statistic*. This is the change of the value of the network statistic when y_{ij} changes from 1 to 0;
- (viii) \mathbf{y}_{ij}^C includes all the observations of ties in \mathbf{y} except the particular observation y_{ij} .

The change statistic can be calculated in many different ways and was discussed by a number of authors. This would go beyond the scope of this study, but with the calculation it is possible to produce pseudo-likelihood estimates. Each possible tie Y_{ij} becomes a case in a standard logistic regression procedure where y_{ij} is predicted from the set of change statistics, described above.

Regardless, the procedure as a whole is not a logistic regression, because it would assume independent observations. Markov and higher models do not make this assumption (see chapter 3.3). This is the reason why the parameters should be biased and standard errors should be approximated at best.

On the one hand, the advantage of the method is that it can be used to date a pragmatic convenience and is relatively easy to fit complicated models. On the other hand, the disadvantage is that the properties of the estimator are not well understood and the accuracy for many data sets is poorly. As a result Monte Carlo estimation procedures are the preferred option [1, pp.186-187].

B. Monte Carlo estimation

At this point a brief summary of the Monte Carlo estimation techniques for exponential random graph models is given. The core element for these techniques is simulation. For a given model by fixing parameter values it is possible to examine the features of graphs in the distribution through simulation to gain insight into the outcomes of the model [2, p.141]. This is done by a number of algorithms, for example algorithms which are well-known in statistics generally.

Although there are differences between various Monte Carlo estimation techniques, they are all based on the same central approach:

Execute a simulation of a distribution random graphs with a starting set of parameter values. After that the parameters are incrementally enhanced by comparing the distribution of graphs against the observed graph and this process will be repeated until the parameter estimates stabilizes.

VI. APPLIED CONCRETE EXAMPLE

The goal of this chapter is to make the explained framework for exponential random graph models (see chapter 2.4) clearer. To do so, a simple illustrative example of a network analysis is given and this example network is meant to point out a number work steps presented so far.

A. Network presentation

The observed network on which data has been collected is a communication network within an fictional organization in the entertainment industry. Furthermore, the network consists of thirty-eight executives and is binary and directed. In a survey all actors of the organization were asked who it was important to communicate with in order to get work completed effectively. A tie then represents the response of an actor. The graph of the network is shown in Fig.3 and some basic statistics about it in Tab.1.

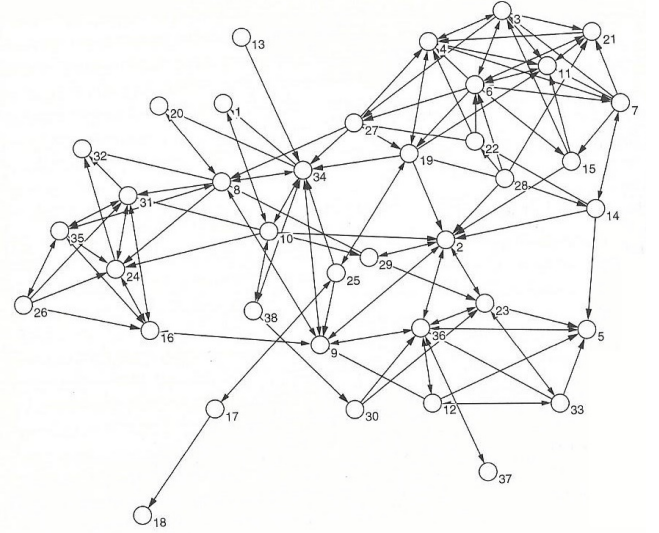


Fig. 3. Example communication network

TABLE I
NETWORK STATISTICS FOR THE EXAMPLE NETWORK

	Example network
Actors	38
Arcs	146
Reciprocated arcs	44
Transitive triads	212
In-2-stars	313
Out-two-stars	283






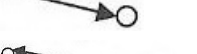


The organization is in competition with others and the researchers are interested in understanding the structure of informal communication ties. Noted in chapter 2 an exponential random graph model tries to model the effects of interest, like reciprocated ties or transitive triads, and want to find a distribution of graphs in relation to an observed graph where

the observed data is central in the distribution. A parameter in the statistical model is then corresponding to a configuration, but the exact values of the parameters to assign probabilities to graphs in the distribution are unknown. In the example network a Markov dependence (see chapter 3.3) was assumed and parameters were estimated with Monte Carlo estimation procedures (see chapter 5.2).

B. Interpretation of the modeled network

Accordingly, the model produces parameter estimates that indicate the strength and the direction of network patterns. The parameter estimates and the standard errors for one model of the communication network are shown in Tab.2. As mentioned earlier a positive (negative) estimate indicate more (less) of the number of configurations in the network than expected, given the other effects in the model. Next the results for the model will be presented by explaining and interpreting the estimates for the different parameters.

TABLE II
EXAMPLE MODEL PARAMETER ESTIMATES (AND STANDARD ERRORS)

Structural effects	Visualization	Estimate (SE)
Arc		-1.96 (0.73)
Reciprocity		2.88 (0.46)
Popularity (in-degree)		-0.27 (0.32)
Activity (out-degree)		-0.34 (0.34)
Mixed 2-star		-0.06 (0.08)
Multiple mixed 2-star		-0.06 (0.09)
Transitivity (multiple)		1.22 (0.19)
Cyclic closure (multiple)		-0.37 (0.17)

Arc: At first there is a negative *arc* effect. It can be interpreted as the baseline propensity for the occurrence of ties. Simultaneously the arc parameter is not a direct measure of network density.

Reciprocity: The reciprocity parameter is estimated positive and significant. Employees are likely to reciprocate communication.¹

¹If a parameter estimate is greater than two times the standard error in absolute value it is indicating that the effect is significant [2, p.43].

Popularity and activity: The popularity parameter is negative, but non significant suggesting that there are no exceptionally popular employees in the organization. The out-degree or activity estimate is negative and significant. That can indicate an absence of centralization in the network. People tend to not change the number of communication partners.

Mixed-two-star: This parameter is not significant demonstrating that neither more nor less mixed-two-stars are seen than might be expected given the other effects in the model. If this parameter were significantly positive, it could suggest that actors with the highest popularity are also the most active.

Multiple mixed-two-stars: The multiple mixed-two-star parameter is often used to show the *depth* of local connectivity between pairs of nodes. In contrast to the simple mixed-two-star parameter this one focuses on the connectivity between pairs and nodes at the end of paths and not on the node at the center of the two paths. Here, the parameter is not significant suggesting that the local connectivity is neither stronger nor weaker than expected given other effects in the model.

Transitivity: There is a significant and positive effect for transitivity indicating a hierarchical path closure in this model. Multiple transitive triads are used, because single triangle parameters are problematic in exponential random graph models and not result in understandable models. Further information is provided [2].

Cyclic closure: There is a significant and negative effect for cyclic closure indicating analogous to the transitivity parameter a lack of non hierarchical network closure in this model.

When examining the estimates in Tab.1 one could highlight the fact that there are significant effects for purely structural network effects. With more information like seniority and status within the organization it would be possible to assign actor-attributes to the nodes (see chapter 2.3). With that potential homophily effects might be included in the model [2, pp.41-45].

VII. CONCLUSION

This paper provides an introductory of the formulation and utilization of exponential random graph models for social networks. The focus was on presenting the underlying logic of these models. Given the limitation of space a summary of dependence assumptions, the general form and estimation was outlined.

REFERENCES

- [1] G. Robins, P. Pattison, Y. Kalish and D. Lusher *An introduction to exponential random graph (p*) models for social networks*, Department of Psychology, School of Behavioral Science, University of Melbourne Australia, 2006.
- [2] D. Lusher, J. Koskinen and G. Robins *Exponential random graph models for social networks - Theory, Methods and Applications*, Cambridge University Press England, Cambridge, 2012.
- [3] O. Frank and D. Strauss *Markov Graphs*, Journal of the American Statistical Association, Vol. 81, No. 395 1986.