Article

A comparative analysis on computational methods for fitting an ERGM to biological network data

Sudipta Saha[1], **Munni Begum**[2]

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, ONM5S 2J7, Canada
[2]Department of Mathematical Sciences, Ball State University, Muncie, IN47306, USA
E-mail: sudip.saha@mail.utoronto.ca, mbegum@bsu.edu

Received 16 October 2014; *Accepted 25 November 2014*; *Published online 1 March 2015*

Abstract

Exponential random graph models (ERGM) based on graph theory are useful in studying global biological network structure using its local properties. However, computational methods for fitting such models are sensitive to the type, structure and the number of the local features of a network under study. In this paper, we compared computational methods for fitting an ERGM with local features of different types and structures. Two commonly used methods, such as the Markov Chain Monte Carlo Maximum Likelihood Estimation and the Maximum Pseudo Likelihood Estimation are considered for estimating the coefficients of network attributes. We compared the estimates of observed network to our random simulated network using both methods under ERGM. The motivation was to ascertain the extent to which an observed network would deviate from a randomly simulated network if the physical numbers of attributes were approximately same. Cut-off points of some common attributes of interest for different order of nodes were determined through simulations. We implemented our method to a known regulatory network database of *Escherichia coli (E. coli)*.

Keywords biological networks; regulatory networks; exponential random graph models; Monte Carlo maximum likelihood estimation; maximum pseudo likelihood estimation; *E. coli*.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Over the last decade, there has been a growing interest in the study of biological interaction networks at the macro and micro molecular levels (Zhang 2012). Identifying basic structural relationships among micro components is the main goal in the field of systems biology (Li and Zhang, 2013). A formal basis for handling such complex networks includes computational tools to support the modelling and simulation through methods developed in mathematical biology and bioinformatics. Since 1960s, with some notable precursors in the

preceding decades, a variety of mathematical formalisms have been proposed to describe this kind of complex networking. During the last few years, modelling efforts targeted several distinct types of networks at the molecular level, such as gene regulatory networks (Pavlopoulos et al., 2011; Mason and Verwoerd, 2007), metabolic networks (Ideker et al., 2001), signal transduction networks (Stock, 1990) or protein-protein interaction networks (Pavlopoulos et al., 2011), transcription regulatory networks (Begum et al., 2014). Networks of interactions that are not restricted to a cell (intercellular communications) or take place at an altogether different level of detail (immunological networks, ecological networks) are also of immense interest.

In this paper, we considered a transcription regulatory network for the model organism *Escherichia coli* K-12 (*E. coli*) from *RegulonDB* (Salgado et al., 2006) version 7.4 (http://regulondb.ccg.unam.mx/). The *RegulonDB* contains information on transcription initiation and the regulatory network of *E. coli*. Downloadable experimental datasets are available on the regulatory network interactions *RegulonDB*. The transcription factor (TF) - transcription factor (TF) interaction network data are considered in this work. A transcriptional unit is defined as a set of one or more genes within an operon transcribed as a set through the utilization of a single promoter. In the original dataset (represented as a table) of *E.coli* in the *RegulonDB* website, there are four columns. The first column is the name of the Transcription Factor (TF), the second column is TF regulated by TF, third column is Regulatory effect of the TF on the regulated gene (+ activator, - repressor, +- dual, ? unknown) and the fourth column is the evidence of support of the existence of the regulatory interaction. The first two columns are considered and it created the TF-TF interaction network. The observed TF-TF network, which is a directed network with loops, is given in Fig. 1.

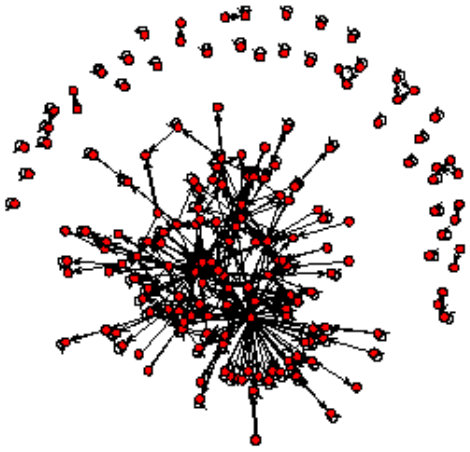

Fig. 1 Observed TF-TF network.

Each vertex is a TF and an edge between two TFs represents a regulation. An edge from a TF to another TF represents that the first TF regulates the second. We explored this observed network and counted the number of several network attributes i.e. edge, triangle and stars. In this observed network, there are *387* edges, *114* triangles, twenty *3-ostars*, thirty-four *3-istars*, ten *5-ostars*, and nine *5-istars* and the network has two big clusters and several small clusters. The basic definitions of some network attributes are given below,

Edges or arcs: This term adds one network statistic that is equal to the number of *edges* in the network. For undirected networks, an *edge* is same as*k-star (1)* [see below] whereas for directed networks, an *edge* represents both *ostar (1)* and *istar (1)* (Morris et al., 2008).

Triangles: This term adds one statistic to the model that is equal to the number of *triangles* in the network. For an undirected network, a *triangle* is defined to be any set $\{(i,j),(j,l),(l,i)\}$of three edges. For a directed network,

a *triangle* is defined as any set of three edges $\{(i \to j) and (j \to l)\}$ and either $(l \to i)$ or $(l \leftarrow i)$(Morris et al., 2008).

k-star: This term adds one statistic when there exists ties between one node and *k* number of other nodes. For a directed network the star statistics are replaced by outgoing stars (*k-ostar*) and incoming stars (*k-istar*) (Morris et al., 2008).

2 Exponential Random Graph Model (ERGM)

An Exponential Random Graph Model (ERGM) models the probability distribution (mass function / density function) for a given class of graphs. Given an observed graph and a set of local features of that graph, the probability distribution of the graph is estimated. The distribution provides a concise summary of the class of graphs to which the observed graph belongs, i.e. the probability distribution can be used to calculate the probability that any given graph is drawn from the same distribution as the observed graph (Fronczak, 2012; Robins et al., 2007; Saul and Filkov, 2007; Wasserman and Pattison, 1996).

ERGMs represent the generative process of tie formation in networks with two basic types of processes namely dyadic dependence and dyadic independence. A *dyad* refers to a pair of nodes and the relations between them. Dyadic dependent processes are those in which the state of one dyad depends stochastically on the state of other dyads. Dyadic independent processes exhibit no direct dependence among dyads. This distinction between these two types of processes affects the specification, estimation, and behaviour of ERGMs. Models with only dyadic independent terms have a likelihood function that simplifies to a form that can be maximized using standard logistic regression models. In contrast, models for processes with dyadic dependence require computationally intensive estimation and imply complex forms of feedback and global dependence that confound both intuition and estimation (Handcock et al., 2003; Hunter and Handcock, 2006).

Although an ERGM presents a flexible means to model complex networks, the likelihood function for parameter estimation involves a mathematically intractable normalizing constant. ERGMs generalize the Markov random graph models (Frank and Strauss, 1986), and edge and dyadic independence models. Several statistical computational methods had been proposed to address this difficulty in parameter estimation in an ERGM. These are the Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) method and the Maximum pseudo likelihood estimation (MPLE) method (Handcock et al., 2003; Robins et al., 2007; Snijders, 2002). We briefly discuss the general ERGM, which is also known asp^* model, to layout the theoretical background of such models.

The general log-linear form of p^* model is expressed as,

$$P(X = x) = \frac{\exp[\boldsymbol{\theta}'\mathbf{z}(\mathbf{x})]}{\kappa(\boldsymbol{\theta})} \tag{1}$$

here $\boldsymbol{\theta}$ is a vector of model parameters, $\mathbf{z}(\mathbf{x})$is a vector of network statistics, and $\kappa(.)$is a normalizing constant which is hard to compute for large networks. In order to simplify the estimation process of the model parameters, the log-linear model form of the p^* model can be re-expressed as a logit model. In particular, as per (Wasserman and Pattison, 1996), $\boldsymbol{X_{ij}^{+}}$ denotes an adjacency matrix where a tie from $i \to j$ is forced to be present. That is $\boldsymbol{X_{ij}^{+}} = \{X_{kl}, with\ X_{ij} = 1\}$. $\boldsymbol{X_{ij}^{-}}$ denotes an adjacency matrix where a tie from $i \to j$ is forced to be absent. That is $\boldsymbol{X_{ij}^{-}} = \{X_{kl}, with\ X_{ij} = 0\}$. Finally, $\boldsymbol{X_{ij}^{c}}$ denotes an adjacency matrix with complement relation for the tie from $i \to j$. That is, $\boldsymbol{X_{ij}^{c}} = \{X_{kl}, with\ (kl) \neq (i,j)\}$. The p^*model in Equation (1) can be turned to a logistic regression model by considering a set of binary random variables$\{X_{ij}\}$, where$X_{ij} = 1$ implying a tie from $i\ to\ j$ as follows.

$$\log\left\{\frac{P(X_{ij} = 1|\mathrm{X}_{ij}^{c})}{P(X_{ij} = 0|\mathrm{X}_{ij}^{c})}\right\} = \omega_{ij} = \boldsymbol{\theta'}\left[\mathbf{z}(\mathbf{x}_{\mathbf{ij}}^{+}) - \mathbf{z}(\mathbf{x}_{\mathbf{ij}}^{-})\right] \tag{2}$$

$$\omega_{ij} = \boldsymbol{\theta}'\boldsymbol{\delta}(\boldsymbol{x_{ij}}) \quad (3)$$

Here $\boldsymbol{z}(\boldsymbol{x_{ij}^{+}})$ is defined as vector of network statistic x_{ij}^{+}, $\boldsymbol{z}(\boldsymbol{x_{ij}^{-}})$ is the vector of network statistic x_{ij}^{-} and $\boldsymbol{\delta}(\boldsymbol{x_{ij}})$ is the vector of difference statistics obtained from the network statistics $\mathbf{z}(.)$ when the variable X_{ij} changes from *1* to *0*. The model in Equation (3) is referred to as the $logit\ p^{*}$ model for single binary relation. One can work with either the log-linear form of p^{*} model given in Equation (1) or the logit form given in equation (3).

3 Computational Methods

There are two methods commonly used to estimate the maximum likelihood fit to exponential random graph models. These are the maximum pseudo-likelihood estimation (MPLE) and the Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) (Handcock et al., 2003; Robins et al., 2007; Snijders, 2002). The pseudo likelihood function is simply the product of the probabilities of X_{ij} with each probability conditional on the rest of the data. The method avoids the technical difficulty inherent in the maximum likelihood approach. The maximum pseudo likelihood estimator (MPLE) for an ERGM, which maximize the pseudo likelihood, may easily be found (at least in principle) by using logistic regression as a computational device. However, when the ERGM in question is not a dyadic independence model, the statistical properties of pseudo likelihood estimators for a network are not well understood (Hunter and Handcock, 2006).

Monte Carlo maximum likelihood estimation (MCMCMLE) is preferred for dyadic dependent p^{*} models. The MCMCMLE of the parameter vector θ is obtained by maximizing the approximate likelihood. The MCMCML estimation algorithm is implemented to the software package *statnet* (Handcock et al., 2003) under the statistical computational environment R. We use these two packages *statnet* and *ergm* (Handcock et al., 2008) to fit the exponential random model given in equation (1).

4 Simulation Study

We conduct a simulation study for generating random network under varying conditions. We choose conditions by assigning different number of nodes and network statistics. The primary reason behind conducting the simulation is to determine the cut-off points for different number of nodes for specific attributes and also to compare our simulated models with an observed model. For the comparison part, we create two networks by imposing the same number of network attributes to the models and then compare the results of estimates with the TF-TF interaction network of *E. coli* by fitting ERGM.

We consider various network statistics such as *arc*, *stars*, and *triangles*. A *k-star* is defined where there exist ties between one node and *k* number of other nodes. For a directed network the star statistics are replaced by outgoing stars (*k-ostar*) and incoming stars (*k-istar*). In particular *arc*, *5-ostar*, *5-istar*, *6-ostar*, *6-istar*, and *triangles* are considered as our network attributes. We physically impose these attributes into the simulated network by keeping approximately the same number of attributes as the observed network. We also observe that if we simulate *triangles*, *ostars*, *istars*, and *arcs* are automatically created. We randomly assign these attributes to the simulated networks for different number of nodes (*n=20*, *50*, *100*) and determine the conditions for these statistics to become insignificant. The cut-off points for single attributes and for a combination of attributes are assessed. However, due to the convergence issues, we were unable to obtain the cut-off points for some cases. A cut-off point is defined as the value where network attributes become significant to insignificant and vice versa. The rationale is that if the biological network behaves almost the same as the random network, then if we have an observed network with different number of nodes, we can determine up to which point (approximately) certain statistics become insignificant.

We explore the TF-TF interaction network of *E. coli* from the *RegulonDB* database and found that there are ten*5-ostars*, nine*5-istars*, ten*6-ostars*, eight*6-istars* and *114 triangles*. The network contains *175* nodes with density *0.012*. Once we determine the number of attributes in the observed network, then we mimic this network and randomly simulate two networks. Then we consider different combinations of attributes (*ostars, istars* and *triangles*) and fit the models by ERGM. We fit the same models for the observed data by using ERGM and then compare the estimates of ERGM for both MCMCMLE and MPLE method.

We begin with networks with small number of nodes and move toward networks with higher number of nodes. With only *20* nodes, we consider reasonably smaller magnitude of network attributes such as *arcs, 3-ostars, 3-istars* and *triangles* as our attributes of interest and then fit the models with ERGM to get the estimates and also to determine the cut-off points. We start with smaller number of attributes, two*3-ostars*, two 3-*istars*, and two *triangles*. We increment each attribute one at a time to determine the cut-off points. Simulated network with 77 triangles, twelve*3-istars* and fourteen*3-ostars* is presented in Fig. 2.

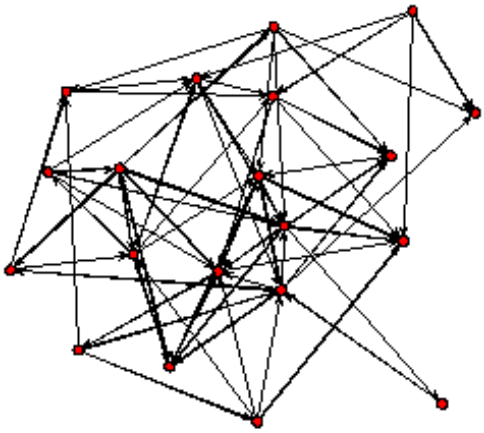

Fig. 2 Simulated network for *n=20*.

Next we increased the number of nodes to 50 and 100. A summary of the simulated networks with nodes 20, 50, and 100including the cut-off points for each network statistics is presented in Table 1.

Table 1 Summary of simulation studies for different numbers of nodes.

	For *n*=20			**For *n*=50**			**For *n*=100**		
	Trian-gles	3-Ostar	3-Istar	Trian-gles	3-Ostar	3-Istar	Trian-gles	3-Ostar	3-Istar
Lower cut-offs	-	7	6	-	-	3	-	4	5
% of*n* (apps)[a]	-	35%	30%	-	-	6%	-	4%	5%
Higher cut-offs	76-80	17	17	-	35	-	-	64	64
% of*n* (apps)	390%	85%	85%	-	70%	-	-	64%	64%

[a]% of n (apps) means that the lower cut-offs are the percentage of n (i.e. node). For example, for n=20, lower cut-off of 3-Ostar is 7 which is 35% of n=20.

It is to be noted that the cut-off points for *3-ostar* and *3-istar* are quite similar, although we could not find any conclusive answer when the number of nodes is *50*. For *3-ostar* and *3-istar*, we can say that, the cut-off points spread out with the increase in the number of nodes. That is if we move toward higher number of nodes, the lower cut-off points become smaller and the higher cut-off points become smaller. For *n=20*, the total spread of insignificant region is close to *(85-35) = 50%* and which is approximately *60%* for *n=100*. For triangles, cut-off points should be bigger than the number of nodes *n*. In summary, we can say that, for network data if we increase the order of the nodes, the spread of the insignificant region gradually becomes larger for any specific attributes. To determine the exact percentage of cut-off points, we have to do similar study for different other nodes, and then we can generalize the idea.

4.1 Comparisons of results under simulation schemes

In our observed TF-TF model, we have *175* nodes, *114* triangles, ten *5-ostars*, nine 5-*istars*, ten 6-*ostars* and eight *5-istars*. An *R*-script is written to count the number of attributes in the model. Then we randomly simulate two different network models to compare the estimates of these network attributes with the observed network. In both cases, we have very close estimates of network attributes from the simulated models compared to the actual model. In Fig. 3 and 4, we represent the observed TF-TF network with and without loops.

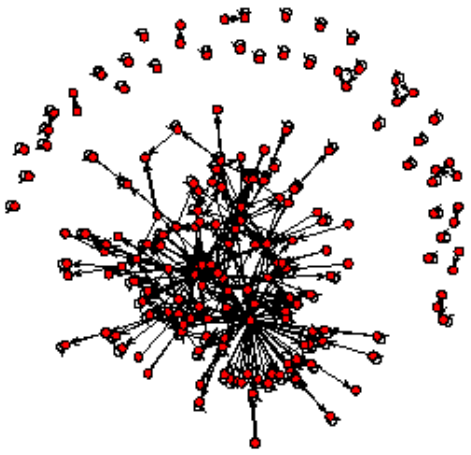

Fig. 3 Observed TF-TF network with looping.

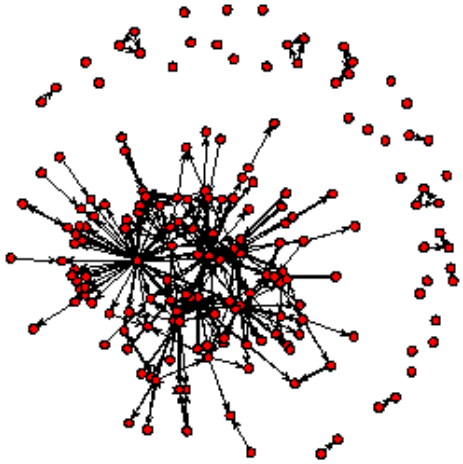

Fig. 4 Observed TF-TF network without looping.

The estimates of attributes of the observed network are presented in Table 2.

Table 2 Estimates of the observed network.

Network Attributes	MCMCMLE Estimates	MPLE Estimates
Edges	-5.3500647	-5.35
Triangle	0.9355000	9.355e-01
5-Ostar	0.0003851	1.564e-05
5-Istar	0.0022043	2.204e-01
6-Ostar	7.797e-05	1.676e-06
6-Istar	1.034e-01	1.034e-03

To compare the estimates of network attributes between the observed and simulated network, we randomly simulate two networks by imposing the same number of attributes as TF-TF, one with *5-ostars*, *5-istars*, and *triangles* (network-1) and the other with same number of *6-ostars*, *6-istars*, and *triangles* (network-2). The summary table of the number of network attributes is presented in Table 8.It is to be noted that we found very similar estimates for the common network attributes edge and triangle from network-1 and network-2. In Tables 3 and 4, we presented the estimates of the observed and simulated networks for both MCMCMLE and MPLE methods.

Table 3 Estimates from observed versus simulated networks with MCMC MLE.

Network Attributes	Estimates from observed networks	Estimates from simulated networks
Edges	-5.3500647	-5.73286
Triangle	0.9355000	2.90743
5-Ostar	0.0003851	-0.01720
5-Istar	0.0022043	-0.08434
6-Ostar	7.797e-05	-1.342e-01
6-Istar	1.034e-01	-8.702e-04

Table 4 Estimates from observed versus simulated networks with MPLE.

Network Attributes	Estimates from observed networks	Estimates from simulated networks
Edges	-5.35	-5.675632
Triangle	9.355e-01	2.905757
5-Ostar	1.564e-05	-0.016141
5-Istar	2.204e-01	-0.083797
6-Ostar	1.676e-06	-0.1343137
6-Istar	1.034e-03	-0.0006957

From Tables 3 and 4, we conclude that except triangles the rest of the estimates of network attributes are very close for both MCMCMLE and MPLE method. Therefore, from the biological point of view, if the observed network is available and the numbers of certain network attributes are known, then it behaves almost same as the random model for most of the cases. However, to generalize the case we need more experiment and more exploration among higher order of species. The simulated networks (1 & 2)are presented in Figs 5 and 6.

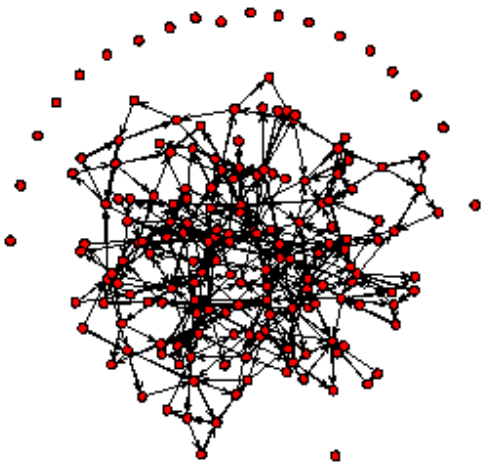

Fig. 5 Simulated network-1.

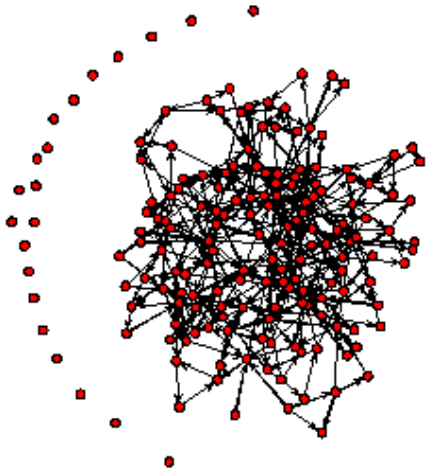

Fig. 6 Simulated network-2.

From this experiment, we observe that if we want to simulate a biological data, then one way would be to explore the observed data and count the number of statistics that we are interested and then physically impose the number of statistic and then compare. There are several other ways to simulate network models using several packages on R. The simplest one is to take the density of the observed model and simulate it using binomial distribution. Also, once a model is fitted by using ERGM package, it can be simulated from the fitted model. ERGM takes the estimates of the network attributes and simulates a similar type of model. However, in such a case the physical number of attributes differs substantially. Again, we can also simulate networks by using Erdos-Renyi model. The comparison of networks obtained using different simulation approaches is presented in the following section.

4.2 Comparison over simulation methods

In this section, we simulate several networks by the existing simulation schemes. We simulated a network by using Erdos-Renyi modelling scheme where we consider *175* nodes to create similarity with our observed TF-TF network and then consider the density of the TF-TF model. The summary of the estimates that we obtain under different approaches, are provided in Tables 5, 6, and 7 (for both MCMCMLE and MPLE).

Table 5 Estimates from Erdos-Renyi model

Network Attributes	MCMCMLE Estimates	MPLE Estimates
Edges	-4.438846	-4.42969
Triangle	-0.058951	-0.06431
5-Ostar	-0.007336	-0.00546
5-Istar	-0.120974	-0.15166
6-Ostar	-0.01302	-0.01366
6-Istar	-0.75202	-0.84257

Table 6 Estimates from Binomial simulated model

Network Attributes	MCMCMLE Estimates	MPLE Estimates
Edges	-4.33561	-4.30465
Triangle	-0.06395	-0.08794
5-Ostar	-0.01032	-0.04404
5-Istar	-0.07059	-0.07491
6-Ostar	-0.10063	-0.25368
6-Istar	-0.31184	-0.29370

Table 7 Estimates from fitted ERGM models

Network Attributes	MCMCMLE Estimates	MPLE Estimates
Edges	-5.3318479	-5.332e+00
Triangle	0.7194116	7.194e-01
5-Ostar	0.0001207	6.297e-06
5-Istar	0.0016440	1.644e-03
6-Ostar	1.484e-05	5.333e-07
6-Istar	-5.887e-02	-5.887e-02

We notice that as long as we consider the same network, estimates of certain attributes are always similar. Although some of the estimates we obtain in this simulation study are very close, the physical numbers of statistics differ substantially. As the simulation scheme takes the fitted estimates into account, the physical number of different attributes should be close to the observed model. It is important since the exact numbers of network statistics might have a significant influence on the overall process. The simulated networks using Erdos-Renyi modelling scheme, binomial density, and *ergm* package in *R* are presented in Figs 7, 8, and 9.

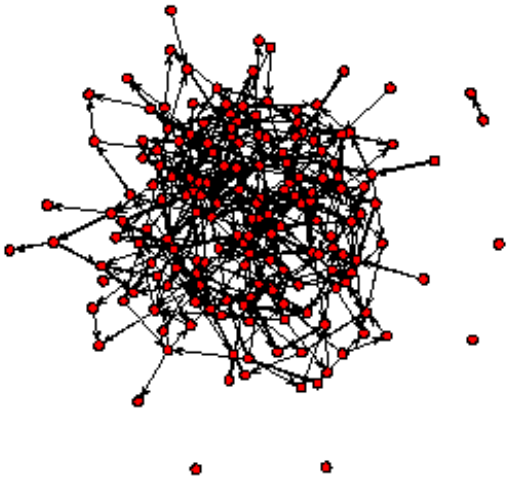

Fig. 7 Simulated from Erdos-Renyi model.

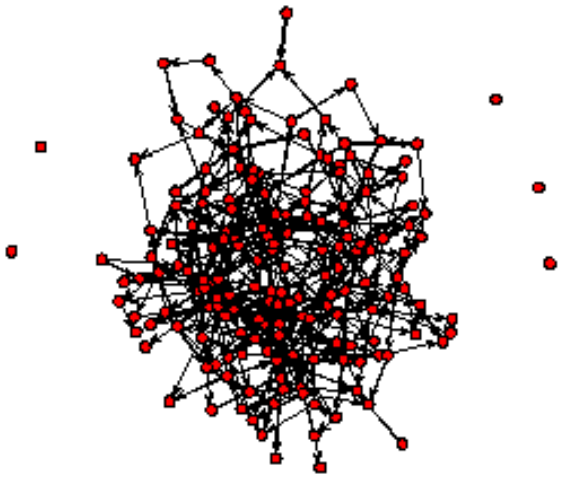

Fig. 8 Simulated network using binomial probability.

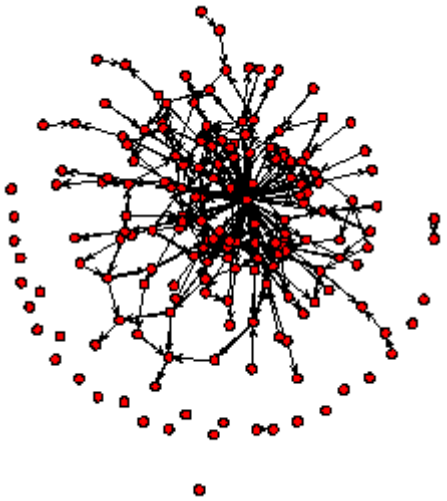

Fig. 9 Simulated network from fitted ERGM model.

The numbers of network attributes for different simulation models are presented in Table 8.

Table 8 Summary table of estimates observed versus simulated networks.

Network Attributes	Observed TF-TF network	Our Simulated network	Simulation using density	Erdos-Renyi simulation	ERGM fitted simulation
Edges	263	327	377	375	247
Triangle	114	115	12	9	82
5-Ostar	10	10	17	6	7
5-Istar	9	9	12	8	2
6-Ostar	10	10	6	2	1
6-Istar	8	8	3	1	3

From Table 8, we can see that the network attributes are different under all the simulation schemes. In our process as we are physically imposing the attributes, it is very close to the observed model. The only difference in the attributes is for the triangles which differ by just *1*. From Table 8, we can say that, in terms of number, the ERGM simulated network generates close result. However, the numbers of triangles substantially differ from the original observed model. For the simple binomial simulation, the edges do not even come close and the other attributes also significantly differ. We find similar characteristic for Erdos-Renyimodelling scheme. The reason behind this could be that both the binomial and Erdos-Renyi consider the density only while simulation. Thus, the number of attributes along with the edges is very close. However, other attributes such as the number of *5-istars* or *5-ostars* are not very close. In our random simulation, we emphasize on the number of attributes because a biological process is a very complicated process. A single edge might have significant influence over the entire process. Therefore, for biological simulation, we should always keep in mind the physical number of attributes that we are interested in.

5 Conclusions

The number of commonly used network attributes such as *k-istar, k-ostar* and *triangles* in the TF-TF regulatory network of *E. coli* is determined. These networks attributes statistically serve as the significant local structures for the *E. coli* regulatory network. An observed regulatory network of the model organism *E. coli* was exploredin terms of finding statisticallysignificant local structure in this study. Simulation of two network models, network-1 and network-2, and comparison of the estimates of the observed and simulated models are presented. In both cases, the estimates we obtain are very similar with the observed TF-TF network except for *triangles*. Networks simulated using existing methods are compared in terms of these estimates as well. At the end, our models provide close results and same number of network attributes, which is very important for biological network data. Therefore, it can concluded that for the*E. coil* regulatory network, the network can be reproduced by taking the counts for different attributes, and the simulated network will behave as the observed network.

Simulation of different networks with different number of nodes and network attributes were performed. The cut-off points were determined for a number of attributes at which point specific attributes become significant to insignificant, or vice versa. We observed that for smaller numbers of network attributes, the estimates usually become significant. If the number of attributes increases in a given model, the attributes become insignificant.

We also observe that the models in ERGM do not always converge. Addressing the convergence issue would be a desirable upgrade for the computational method. For the several models considered, convergence failure occurred while estimating parameters for any of the methods. For example, for our observed network, the model with *edges, 4-istars, 4-ostars* and *triangles* did not converge. Also, due to the convergence issue, cut-off points could not be determined for several network attributes. In addition, computation for networks with self loops demonstrates convergence problems. Therefore, while the ERGM provides flexible methodology, these issues remain in need of further analysis.

References

Begum M, Bagga J, Blakey A, Saha S. 2014. Network motif identification and structure detection with graphical models. Network Biology, 14(4): 155-169

Frank O, Strauss D. 1986. Markov graphs. Journal of the American Statistical Association, 81(395): 832-842

Fronczak A. 2012. Exponential Random Graph Models. ArXiv e-prints. Available at http://arxiv.org/pdf/1210.7828.pdf

Handcock MS, Hunter DR, Butts CT, et al. 2003. Statnet: Software Tools for the Statistical Modeling of Network data. Statnet Project. Available at http://statnetproject.org/

Handcock MS, Hunter DR, Butts CT, et al. 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. Journal of Statistical Software, 24. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2743438/

Hunter DR, Handcock MS. 2006. Inference in curved exponential family models for networks. Journal of Computational and Graphical Statistics, 15(3): 565-583

Ideker T, Thorsson V, Ranish JA, et al. 2001. Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. Science, 292(5518): 929-934

Li JR, Zhang WJ. 2013. Identification of crucial metabolites/reactions in tumor signaling networks. Network Biology, 3(4): 121-132

Mason O, Verwoerd M. 2007. Graph theory and networks in biology. IET Systems Biology, 1(2): 89-119

Morris M, Handcock MS, Hunter DR. 2008. Specification of exponential-family random graph models: Terms and computational aspects. Journal of Statistical Software, 24(4): 1-24

Pavlopoulos GA, Secrier M, Moschopoulos CN, et al. 2011. Using graph theory to analyze biological networks. BioData Mining, 4(10): 1-27

RobinsGL,Pattison PE,Kalish Y, LusherD. 2007. An introduction to exponential random graph (p*) models for social networks. Social Networks, 29(2): 173-191

Salgado H, Gama-Castro S, Peralta-Gil M, et al. 2006. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Research, 34(1): 394-397

Saul ZM, Filkov V. 2007. Exploring biological network structure using exponential random graph models. Bioinformatics, 23(19): 2604-2611

Snijders TAB. 2002. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models.Journal of Social Structure, 3: 1-40.

Stock JB, Stock AM, Mottonen JM. 1990. Signal Transduction in Bacteria. Europe PubMed Central, 344(6265): 395-400

Wasserman S, PattisonPE. 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. Psychometrika, 61(3): 401-425

Zhang WJ. 2012. Computational Ecology: Graphs, Networks and Agent-based Modeling. World Scientific, Singapore