Int. j. inf. tecnol. (March 2022) 14(2):1145–1159
https://doi.org/10.1007/s41870-019-00354-2

ORIGINAL RESEARCH

Understanding structure and behavior of systems: a network perspective

Pranav Nerurkar[1] · **Madhav Chandane**[1] · **Sunil Bhirud**[1]

Received: 17 April 2018 / Accepted: 7 August 2019 / Published online: 14 August 2019

Abstract Networks are interesting representation models for analysis of systems. The entities of the systems under review can be denoted as the nodes of the networks and the relationships between these entities as the edges connecting them. Such a representation has advantages in analysis as network theory has a rich collection of well defined concepts and methods. These concepts of can be applied on such networks to draw inferences about the systems. As digitization has penetrated almost all aspects of mankind, a wide variety of systems from diverse domains such as computer science, transportation, social science have become available in the form of networks. A network perspective provides valuable insights into their structure and behavior. In this inquiry networks representing real world systems from different domains are analyzed using concepts of network theory and statistical generative network models—SBM and LSM. This is done to various application scenarios to express the properties of these systems. The findings highlight the unique features and trends seen in each domain.

Keywords Statistical models · Graph representations · Latent variable models · Stochastic block models

✉ Pranav Nerurkar
panerurkar_p16@ce.vjti.ac.in

Madhav Chandane
mmchandane@it.vjti.ac.in

Sunil Bhirud
sgbhirud@ce.vjti.ac.in

[1] Department of CE & IT, VJTI, Mumbai, India

1 Introduction

"We will never understand complex systems unless we develop a deep understanding of the networks behind them"—Albert Laszlo Barabasi [1]. In scientific literature, different models have been developed to generate efficient representations and visualizations of data for its analysis [2]. However, the advantage of networks in representation of data is that they provide a general language for describing and modeling complex systems [3]. Hence, networks have become ubiquitous across various scientific disciplines and are being used to represent many real or artificial systems [4], for instance, internet, transportation systems, social networking websites, biological networks etc. [5–8].

Usefulness of network representations in the examination of complex systems is illustrated in the model given in Fig. 1. In this network, edges connect scientists that have coauthored at least one paper. Symbols indicate the research areas of the scientists. As seen in Fig. 1, density of edges between researchers in a particular domain are high compared to density of edges between researchers of different domains. The position of scientists (nodes) in the Santa Fe Institute network provides insights regarding their importance to the research community of the institution.

In addition to this, there are several other useful statistical concepts in graph theory which can be applied on network models to understand the structural characteristics of the systems they represent [10, 11]. Measurements of path length, diameter, connectivity, transitivity and density allow inferences on the structural characteristics of systems. Locations of nodes in a network model can be used to draw parallels about the importance of those entities in the overall scheme of the system [9]. For instance, nodes located at the boundaries in a network are important for

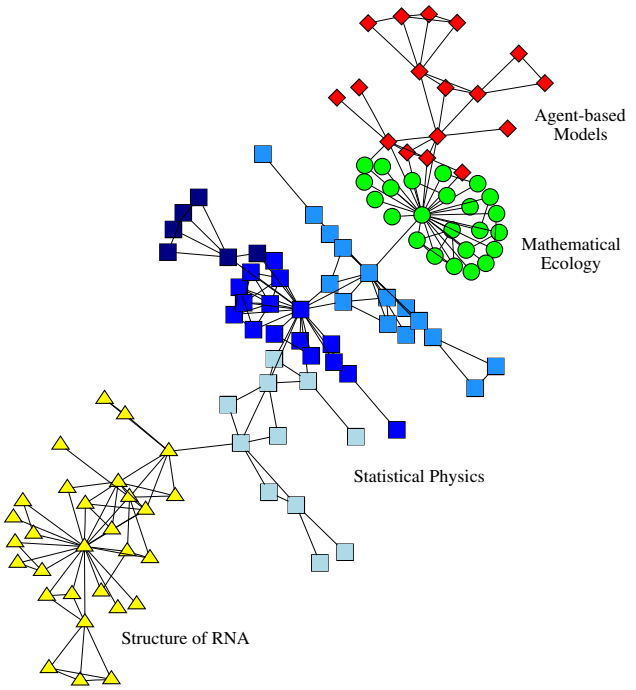


Fig. 1 A network model representing the collaborations between scientists working at the Santa Fe Institute (SFI). Reprinted figure from [9]

information exchange with nodes outside the network and nodes located at the center of a network are hubs which keep the network intact and play a key role in information interchange within the network.

Networks also capture the relationships or behaviors of the entities (nodes/actors) in the underlying systems [12–15]. Two distinct schools of thought are available for understanding this behavior: probabilistic models and statistical models. Probabilistic models assume that a network is formed as a result of the behavior of the actors (nodes) involved in them. Actors can have various tendencies to form relationships with other actors and as a result of these relationships a network is formed. Each type of probabilistic model assumes a different tendency to form relationships i.e. random attachment, preferential attachment etc. A drawback of probabilistic models is that the even though they are "generative", they do not generate networks that share many properties with the specific network they were fit to. Unlike probabilistic models, statistical generative models try to represent networks using a larger number of parameters to capture properties of a specific network. They provide a better fit to the network data and are preferred to understand the relationships or behaviors of entities of systems [16].

The focus of the current inquiry is on analysis of real-world systems from diverse domains to uncover behavioural and structural characteristics. With the use of statistical generative models, various domain specific phenomena that occur in these systems are highlighted. These phenomena are explained using intuition and empirical evidence is offered to support the findings. The key contributions of the inquiry are: providing a taxonomy of techniques and concepts of network science useful for network analysis and providing the trends and insights into the behavior of networks of various domains.

The rest of this inquiry is organized into three main sections. Section 2 describes statistical models for network analysis, Section 3 includes network theoretic concepts popular in network analysis [17, 18]. Section 4 presents the descriptions of the data-sets and their specifications. Section 5 discusses the results of the inquiry and summarizes the key findings. The concluding remarks of this inquiry are presented in Sect. 6.

2 Taxonomy of statistical models for network analysis

2.1 Stochastic blocks models

Stochastic blocks models (SBMs) are the natural enrichment of random networks (nodes form links with other nodes uniformly at random) [19]. In these models the probability of a link formation between actors is dependent on the characteristics of the actors i.e. latent or observed [20]. These models could be used to capture, the probability of linking with actors of the same type than with other types i.e. homophily. However, they fail to capture probability of link formation that is independent of characteristics of actors, for example, the probability of two actors linking together if they have a common friend i.e. Transitivity. Due to this drawback, SBMs are fit to systems where transitivity is not present. However, to fit systems where transitivity is likely Exponential Random Graph Models [p*/ERGM] were proposed.

2.2 Exponential Random Graph Models [p*/ERGM]

ERGM is an emerging statistical technique used to identify how the characteristics of the people or organizations in a network and larger social forces can explain or predict the observed patterns or ties in the observed network. They can also encode complex social theories such as transitivity, reciprocity etc. ERGM allows network data to be modeled in a way similar to least squares logistic regression but does not require observations to be independent [21]. Eqn. 1 provides a mathematical formulation of ERGM's [22, 23].

$$P(Y) = \frac{exp[\sum \beta s_k(Y)]}{\sum_{Y'} exp[\beta s_k(Y')]}] \quad (1)$$

where,

- s_k = graph statistics
- P(Y) = Probability of obtaining a particular graph
- β = Vector of coefficients for graph statistics
- Y' = Another realization of a graph such that $Y(|V|) = Y'(|V|)$

However, there are challenges in estimating ERGMs. The graph statistics s_k may not give a clear picture regarding their significance in the network. Hence, social theory regarding the domain is also needed to understand the results and draw inferences from them. A second challenge of ERGMs is estimating Y' which is a set of all possible graphs with equal number of nodes as Y. Thus the term in the denominator $\sum_{Y'}$ has to sum over $2^{\binom{n}{2}}$ possibilities. To avoid this computation, T. Snijders *et. al.* and M. Handcock *et. al.* proposed Markov chain Monte Carlo techniques for estimation of Y' [24, 25]. S. Bhamidi *et. al.* argued that for ERGMs, MCMC estimation of g' shall be efficient only if links in the graph are assumed to be formed independently [26]. But as this assumption is not applicable for ERGMs, MCMC leads to incorrect estimations [27, 28]. They also rely upon expensive and often unstable methods for probabilistic inference.

2.3 Statistical Exponential Random Graph Models [SERGMs]

Statistical Exponential Random Graph Models [SERGMs] were proposed to resolve the estimation issues of ERGMs. SERGMs assume that all network having same graph statistics s_k are equally likely. This reduces the exponential search space of Y' and the ERGM equation is modified to Eqn. 2.

$$P(Y) = \frac{exp[\beta s(Y)]}{\sum_{s_k} N(s_k) exp[\beta s_k]}] \tag{2}$$

where,

- N(s') = Number of networks that have a particular graph statistics
- s_k = graph statistics
- P(Y) = Probability of obtaining a particular graph

2.4 Latent variable models

Latent variable models represent data as an $n * n$ sociomatrix Y, with $y_{i,j}$ denoting an edge (relation) from node i to node j, and covariate information X. A conditional independence approach to modeling is given by Eq. 3. It assumes that the presence or absence of a edge between two nodes is independent of other edges in the system. Probability of an edge $P(y_{i,j}$ depends on the latent positions of the two nodes in social space z_i, z_j. These positions depend on $x_{i,j}$ and θ the vector of parameters to be estimated [2, 29]

$$P(Y|Z, X, \theta) = \prod_{i \neq j} P(y_{i,j}|z_i, z_j, x_{i,j}, \theta) \tag{3}$$

The Distance model, a variation of the latent variable model, is a logistic regression model in which the probability of a tie $P(y_{i,j}|z_i, z_j, x_{i,j}, \theta)$ depends on the Euclidean distance between z_i and z_j ($\psi_{i,j} = ||z_i - z_j||_2$). The probability of a tie $P(Y|Z, X, \theta)$ in this model is given by Eq. 4:

$$P(y_{i,j} = 1|\psi_{i,j}) = \sigma(\psi_{i,j}) \tag{4}$$

The above model is symmetric, $P(i \rightarrow j) = P(j \rightarrow i)$. However, in many networks such symmetry is not achieved. The shortcomings in the above model can be removed by supposing that a node i has an associated unit-length k-dimensional vector of characteristics v_i. These characteristics can be thought of as points on a k-dimensional sphere of unit radius. The angles between vector of characteristics of two actors can be: $v_i' v_j > 0$ (tie is likely), $v_i' v_j = 0$ (neutral), and $v_i' v_j < 0$ (unlikely). To this a parameter is added for each node to allow for different levels of activity viz. $a_i > 0$ be the activity level of actor i. Then model the probability of a tie from i to j as depending on the magnitude of $a_i v_i' v_j$ or, equivalently, $z_i' z_j = |zj|$, where $z_i = a_i v_i$. This is the signed magnitude of the projection of z_i in the direction of z_j and can be thought of the extent to which i and j share characteristics, multiplied by the activity level of i. The probability of a tie from i to j using the logistic regression model is by Eq. 5:

$$logodds(y_{i,j} = 1|z_i, z_j, x_{i,j}, \alpha, \beta) = \alpha + \beta' x_{i,j} - \frac{z_i' z_j}{|z_j|} \tag{5}$$

Latent variable models provide a visual and interpretable spatial representation of the network. They are also suitable to model transitivity in networks. Hence, they are preferably fit to systems where transitivity exists. Based on the review of statistical models, SBM and LSM were found to be more suitable for understanding the behavior of the entities in the network.

3 Definitions and data

3.1 Definitions :

3.1.1 Average clustering coefficient

The Average clustering coefficient or transitivity of a network is the probability that two incident edges are completed by a third edge to form a triangle

$$c = \frac{|\{u, v, w \in V \mid u \sim v \sim w \sim u\}|}{|\{u, v, w \in V \mid u \sim v \neq w \sim u\}|} \tag{6}$$

3.1.2 Diameter

Diameter of a graph G is defined as $\text{diam}(G) = \max \min d_G(x, y)$, where d is the distance function in G and the max min is taken over all vertices $x, y \in G$.

3.1.3 Assortativity coefficient

It is positive if vertices with high in-degrees tend to connect to each, and negative otherwise. Assortativity coefficient r for with edges $i = 1, 2, ..., M$ with j, k are degrees of the vertices at the ends of the i^{th} edge.

$$r = \frac{M^{-1} \sum_i j_i k_i - \left[M^{-1} \sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1} \sum_i \frac{1}{2}(j_i{}^2 + k_i{}^2) - \left[M^{-1} \sum_i \frac{1}{2}(j_i + k_i)\right]^2} \tag{7}$$

3.1.4 Edge density

For directed graphs $G(V, E)$ is $D = \frac{2|E|}{|V|(|V|-1)}$ and for undirected graphs is $D = \frac{|E|}{|V|(|V|-1)}$.

3.1.5 Gini index

It takes values between zero and one, with zero denoting total equality between degrees, and one denoting the dominance of a single node. Let $d_1 \leq d_2 \leq ... \leq d_n$ be the sorted list of degrees in the network. The Gini index G is twice the area between the Lorenz curve and its main diagonal. G is defined as:

$$G = \frac{2 \sum_{i=1}^{n} i d_i}{n \sum_{i-1}^{n} d_i} - \frac{n+1}{n} \tag{8}$$

3.1.6 Average degree d of G(V, E)

$$d = \frac{1}{|V|} \sum_{u \in V} d(u) \tag{9}$$

3.1.7 Variables in Network data:

Network data includes :

- a set of nodes (objects, actors, egos, individuals) and edges (links, ties, dyads)
- variables measured on nodes or pairs of nodes (edges)
- dyadic variables: measured on pairs of nodes (edges)
- nodal variables: measured on nodes

3.1.8 Types of node attributes or side information:

A binary (or dichotomous) relation takes only two values. A valued relation takes more than two values. A valued relation whose possible values have an order is called ordinal. A valued relation whose possible values lack an order is called categorical.

4 Network data:

4.1 Data-sets with no side information:

These are networks $G = (V, E)$ with vertex set V and edge set E and the vertex attribute set V^a and the edge attribute set E^a are null. Table 1 gives the description of data-set taken from online social networking websites Twitter.com (Twt-Net), Google+ (Gplus-Net) and a citation indexing website CiteSeer.com (Cite-Net). All data-sets are publicly available at Stanford Network Analysis Platform (SNAP) - a network data repository [30]. In **Twt-Net and Gplus-Net** the system that is modeled as a network is the online social networking website. The entities of the system are the users of these platforms, they are denoted as nodes of the network. The edges of the network are "follower" relationships between the users. If a user i follows user j, then this is denoted in the network by a directed edge from node i to node j. Similarly, in **Cite-Net** the nodes are the academic papers indexed in CiteSeer and if paper i cites paper j, then this is denoted in the network by a directed edge from node i to node j.

Table 1 Description of Network Data-sets with no side information

Description	Twt-Net	Gplus-Net	Cite-Net
Nodes	185	923	3327
Edges	5156	39400	4732
Avg Clustering Coeff	0.44	0.3	0.13
Diameter	8	7	8
Assortativity	−0.19	−0.23	0.12
Avg. path length	2.18	2.58	1.81
Edge density	0.15	0.05	0.0004
Gini index	0.41	0.52	0.43
Avg degree	55 ($\sigma = 41$)	85 ($\sigma = 106$)	2.8 ($\sigma = 3.41$)

4.1.1 Data-sets with binary attributes:

These are networks $G = (V, E)$ with vertex set $n = |V|$ and edge set E and the vertex attribute set $V^a = R^{n*f}$ and the edge attribute set $E^a = \phi$, f is number of features for each vertex. The feature matrix of G is denoted by F. If $F_{ij} = 1$ node i has feature j; otherwise we have $F_{ij} = 0$ (binary attributes). Table 2 gives the description of data-set taken from websites Flickr.com (Flickr-Net), Blog.com (Blog-Net), Wikipedia.com (Wiki-Net) and a protein-protein interaction network (Protein-Net). All data-sets are publicly available at Stanford Network Analysis Platform (SNAP) - a network data repository [30].

4.1.2 Data-sets with mixed attributes:

These are networks $G = (V, E)$ with vertex set $n = |V|$ and edge set E and the vertex attribute set $V^a = R^{n*f}$ and the edge attribute set $E^a = R^{|E|*k}$ where f and k are the number of features for the nodes and edges in the network respectively. Table 3 gives the description of data-sets such as High-Net—a Network of the highways in the state of Southern California as observed in 2016. It shows the cities connected by highways. Bill-Net is a bill co-sponsorship network in parliament of Slovakia in 2014. It shows information of legislators co-sponsoring bills together. Trade-Net is a Trade network of electrical automotive goods between Asia and European countries in the year 2016. It shows countries that have trade ties with each other. Grey-Net is a sexual contact network between characters of the television show Grey's Anatomy. It gives information of actors in relationships with other actors in the series. All data-sets are publicly available at Stanford Network Analysis Platform (SNAP)—a network data repository [30].

5 Expressing the structural and behavioral characteristics of networks

Algorithms and used in the analysis are based on frequentist inference. Parameters of stochastic block model (class memberships Z and block-dependent edge probabilities W and Latent space model parameters (latent node positions Z and scalar global bias θ) treated as having fixed but unknown values. These parameters are estimated by maximizing likelihood $\hat{\theta}_{MLE} = \text{argmax}_{\theta} P(X|\theta)$

Algorithm 1: Fit Stochastic blocks models to data

1. Load adjacency matrix Y;
2. Plot singular values of Y;
3. Choose number of latent classes (blocks) by use of Eigengap heuristic;
4. Fit selected model;
5. Analyze model fit: class memberships and block dependent edge probabilities;
6. Simulate new networks from model fit;
7. Check how well simulated networks preserve actual network properties (posterior predictive check);

Table 2 Description of Network Data-sets with binary attributes

Description	Blog-Net	Flickr-Net	Protein-Net	Wiki-net
Entities	users	users	proteins	articles
Relationships	users commenting on blogs by other users	users following other users	proteins interacting with other proteins	articles citing other articles
Type of relationship	directed	directed	un-directed	directed
Nodes	5196	7575	3890	4777
Node attributes	8189	12047	50	40
Edges	171743	239738	37845	54810
Edge attributes	–	–	–	–
Avg Clustering Coeff	0.08	0.1	0.09	0.43
Diameter	$\sim$2.03	$\sim$2.15	8	4
Assortativity	-0.02	-0.23	-0.09	-0.27
Avg. path length	$\sim$2.03	$\sim$2.15	3.09	2.15
Edge density	0.012	0.008	0.005	0.004
Gini index	0.39	0.67	0.63	0.62
Avg degree	66.30 ($\sigma = 54.8$)	63.3 ($\sigma = 131.52$)	19.45 ($\sigma = 34.29$)	22.95 ($\sigma = 105.92$)

Table 3 Description of network data-sets with mixed attributes

Description	High-Net	Grey-Net	Trade-Net	Bill-net
Entities	cities	actors	countries	legislators
Type of relationship	un-directed	un-directed	directed	un-directed
Nodes	205	44	99	139
Node attributes	9	17	12	4
Edges	203	46	725	471
Edge attributes	3	2	1	–
Avg Clustering Coeff	0.28	0	0.44	0.32
Diameter	16	8	0.17	
Assortativity	0.12	−0.22	−0.32	0.014
Avg. path length	6.8	3.49	2.31	3.71
Edge density	0.009	0.04	0.07	0.024
Gini index	0.54	0.37	0.61	0.46
Avg degree	1.97 ($\sigma = 2.12$)	2.09 ($\sigma = 1.72$)	14.64 ($\sigma = 18.74$)	6.77 ($\sigma = 6.23$)

Algorithm 2: Fit latent space model to data

1. Load adjacency matrix Y
2. Model selection: choose dimension of latent space
3. Fit selected model
4. Analyze model fit: examine estimated positions of nodes in latent space and estimated bias
5. Simulate new networks from model fit
6. Check how well simulated networks preserve actual network properties (posterior predictive check)

5.1 Twitter

Twt-Net is a data-set of 185 twitter users of a community and the "followers-following" relationships between them. The average clustering co-efficient i.e. transitivity between the members is 0.44. High transitivity indicates that twitter users prefer to link with "friends of friends". The high transitivity coupled with low average path length of 2.18 suggests that information diffusion between members could be rapid. As users prefer following popular users, social networks have negative assortativity and high inequality of degree (gini index = 0.41). Other characteristics commonly observed in social networks are large number of social contacts (average degree = 55). This leads to average path length and diameter being lower and edge density being higher.

For generating LSR of this network, LSM is a better choice than SBMs as the latter do not explicity model transitivity. Figure 3 shows the results of fitting SBM to Twt-Net using the procedure outlined in Algorithm 1. Figure 2a shows a dense adjacent matrix A but no presence of communities (latent classes) can be detected in the plot. Hence, to choose the latent classes the plot of the singular values of A needs to be obtained.

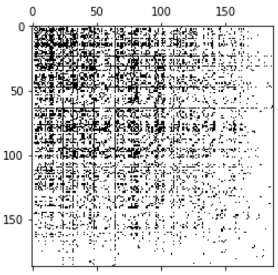
Using eigen-gap heuristic (gaps in the singular values correspond to latent classes in the network) four latent classes are observed in Fig. 2b. The nodes in these latent classes are assigned class-memberships and then the adjacency matrix is re-ordered. Figure 2c shows the re-ordered adjacency matrix with four latent classes. Once the class-memberships are assigned, the edge probabilities at the block level are calculated. Finally, new networks are simulated from edge probabilities to check model goodness of fit.

Figure 3a shows that SBM has generated LSRs that do not re-generate the original network (Twt-Net) effectively. The LSRs are not effectively capturing the transitivity of the original networks. LSMs were able to generate LSRs that could replicate the transitivity (as shown in Fig. 3b) and density (as shown in Fig. 3c) of the original network. Conclusions from posterior predictive check of SBM and LSM models is that LSRs generated using LSM are more effective than SBM in Twt-Net.

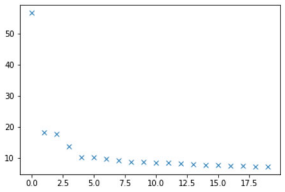
5.2 Google+

Gplus-Net is a data-set of 923 Google+ users of a community and the "followers-following" relationships between them. The average clustering co-efficient i.e. transitivity between the members is 0.3. As transitivity is lower as compared to Twitter, the users of Google+ have lower preference to "friends of friends". The high transitivity coupled with low average path length of 2.58 suggests possibility of rapid information diffusion between members. Users of Google+ prefer following popular users therefore the network has negative assortativity and high inequality of degree (gini index = 0.52). Other characteristics observed in Gplus-Net are large number of social contacts (average degree = 85). This leads to average path length and diameter

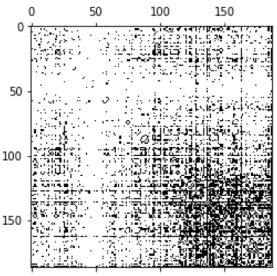
Fig. 2 Analysis of Twt-Net data-set using SBM



(a) Adjacency matrix of Twt-Net

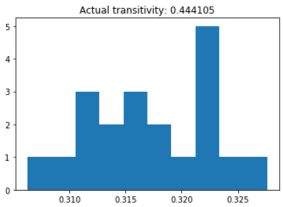


(b) Plot of singular values of Twt-Net

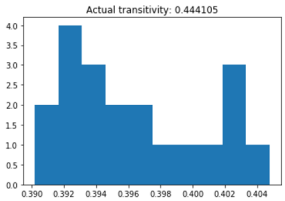


(c) Re-order nodes in adjacency matrix by class memberships

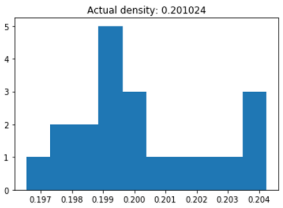
Fig. 3 Fitting SBM and LSM to Twt-Net data-set



(a) Simulate new networks from SBM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20.

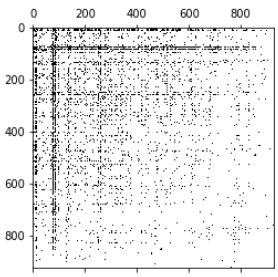


(b) Simulate new networks from LSM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20. Dimensions of latent space = 2.

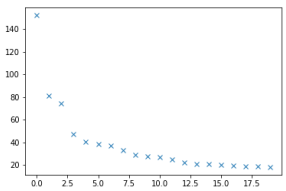


(c) Simulate new networks from LSM fit to check model goodness of fit on parameter density. Number of simulations = 20. Dimensions of latent space = 2.

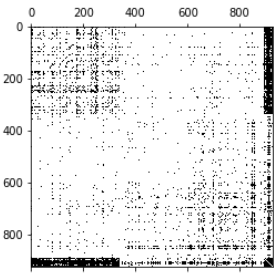
Fig. 4 Analysis of Gplus-Net data-set using SBM



(a) Adjacency matrix of Gplus-Net



(b) Plot of singular values of Gplus-Net



(c) Re-order nodes in adjacency matrix by class memberships

being lower and edge density being higher. Edge density is lower compared to Twt-Net which could imply that users are less active on Google+ than Twitter.

Applying LSM to obtain the LSR of Gplus-Net is infeasible due to the scale of the network. Therefore, only SBM was fit to the data following the procedure outlined in Algorithm 1.

Figure 5 shows that SBM has generated LSRs that do not re-generate the original network (Gplus-Net) effectively. The posterior predictive check reveals that LSRs are not effectively capturing the transitivity of the original network.

5.3 CiteSeer

Cite-Net is a data-set of 3327 academic papers and the citations between them. The average clustering co-efficient i.e. transitivity of the network is 0.13. Transitivity is lower for citation networks as i citing j and j citing k might not lead to k citing i. Unlike social networking websites, a

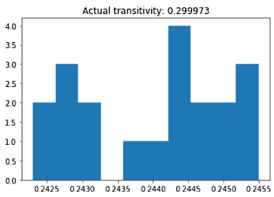


Fig. 5 Simulate new networks from SBM fit to check model goodness of fit on Gplus-Net for parameter transitivity. Number of simulations = 20.

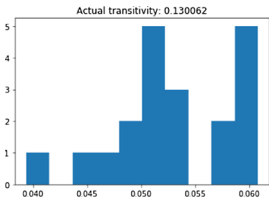


Fig. 7 Simulate new networks from SBM fit to check model goodness of fit on Cite-Net for parameter transitivity. Number of simulations = 20

citation network would have a positive assortativity due to the tendency of highly cited papers to cite each other. Edge density would be low as research papers are less inclined to form connections to each other. Gini index would be positive in citation networks as a few highly cited research papers would have majority of the incoming connections and large number of research papers would have very few connections. All these intuitions are consistent with the results obtained from the network analysis.

SBM are preferred for models with low transitivity. LSM will not be able to scale to Cite-Net as the nodes are in the order of $\sim 10^3$. The network is sparse i.e. $|V| \sim |E|$ and the adjacency matrix as shown in Fig. 6a does not reveal obvious possibilities of communities in the network.

Using eigen-gap heuristic five latent classes are observed in Fig. 6b. Figure 6c shows the re-ordered adjacency matrix with five latent classes. Using the edge probabilities at the block level, new networks are simulated to check model goodness of fit.

Figure 7 shows that SBM has generated LSRs that do not re-generate the original network (Cite-Net) effectively. The posterior predictive check reveals that LSRs are not effectively capturing the transitivity of the original network.

5.4 BlogCatalog

Blog-Net is a data-set of 5196 users and the relationships captured in the networks are of users commenting on blogs by other users. The average clustering co-efficient i.e. transitivity of the network is 0.08. As users will have low tendency to comment of blogs of each other, the network would have low tranitivity. The diameter and average path length are low even though the edge density of the network is less. This would be due to presence of popular blogs that see high user traffic. Such a network would also have high inequality i.e. gini-index would be high.

SBM are preferred for models with low transitivity. LSM will not be able to scale to Blog-Net as the nodes are in the order of $\sim 10^3$. The network has node attributes but SBM does not model attributes. The adjacency matrix as shown in Fig. 8a does not reveal obvious possibilities of communities in the network.

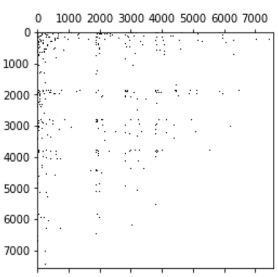
Using eigen-gap heuristic five latent classes are observed in Fig. 8b. Figure 8c shows the re-ordered adjacency matrix with five latent classes. Using the edge probabilities at the block level, new networks are simulated to check model goodness of fit.

Figure 9 shows that SBM has generated LSRs that do not re-generate the original network (Blog-Net) effectively. The posterior predictive check reveals that LSRs are not effectively capturing the transitivity of the original network. A second disadvantage of SBM is that node attributes in the original network were not considered by the model. Thus, SBM are not suitable for modeling networks with attribute information.

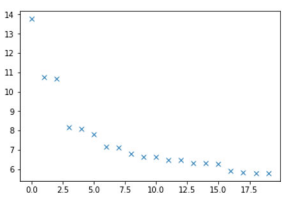
5.5 Flickr

Flickr-Net is a data-set of 7575 users and the relationships captured in the networks are of users "following" the profiles of other users. The average clustering co-efficient i.e. transitivity of the network is 0.1. Similar to online social networking websites like Twitter and Google+,

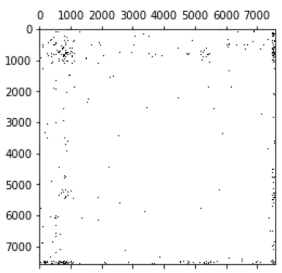
Fig. 6 Analysis of Cite-Net data-set using SBM



(a) Adjacency matrix of Flickr-Net

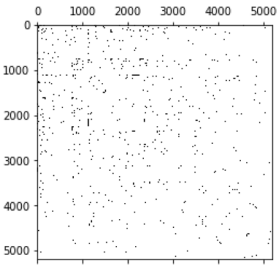


(b) Plot of singular values of Flickr-Net

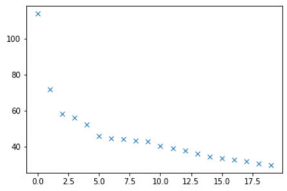


(c) Re-order nodes in adjacency matrix by class memberships

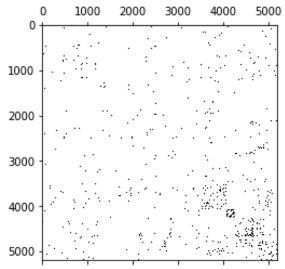
Fig. 8 Analysis of Blog-Net data-set using SBM



(a) Adjacency matrix of Blog-Net



(b) Plot of singular values of Blog-Net



(c) Re-order nodes in adjacency matrix by class memberships

Fig. 9 Simulate new networks from SBM fit to check model goodness of fit on Blog-Net for parameter transitivity. Number of simulations = 20

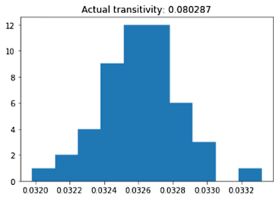
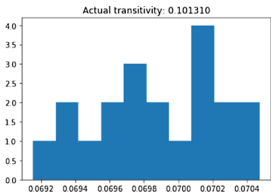


Fig. 11 Simulate new networks from SBM fit to check model goodness of fit on Flickr-Net for parameter transitivity. Number of simulations = 20.



Flickr-Net also has high gini-index, negative assortativity, high average degree, low diameter and low average path length.

LSM is not feasible for Flickr-Net due to the large number of nodes, so only SBM was fit to the data for analysis. The adjacency matrix as shown in Fig. 10a reveals several dense regions in the network.

Using eigen-gap heuristic five latent classes are observed in Fig. 10b. Figure 10c shows the re-ordered adjacency matrix with five latent classes. Using the edge probabilities at the block level, new networks are simulated to check model goodness of fit.

Figure 11 shows that SBM has generated LSRs that do not re-generate the original network (Flickr-Net) effectively. The posterior predictive check reveals that LSRs are not effectively capturing the transitivity of the original network.

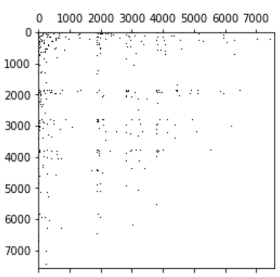
5.6 Protein protein interaction

Protein-Net is a data-set of 3890 proteins and their interactions with each other. The average clustering co-efficient i.e. transitivity of the network is 0.09. On an average, a protein interacts with upto 20 other proteins. But high gini-index indicates that majority of the interactions are concentrated amongst a section of proteins ($\sim 63\%$). Fitting LSM is not feasible for Protein-Net due to the large number of nodes and hence only SBM was fit to the data for analysis. The adjacency matrix as shown in Fig. 12a reveals a dense cluster of nodes in the network.

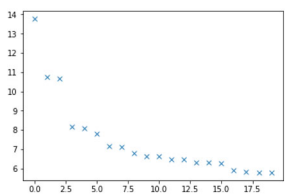
Using eigen-gap heuristic two latent classes are observed in Fig. 12b. Figure 12c shows the re-ordered adjacency matrix with two latent classes. Using the edge probabilities at the block level, new networks are simulated to check model goodness of fit.

Figure 13 shows that SBM has generated LSRs that do not re-generate the original network (Protein-Net)

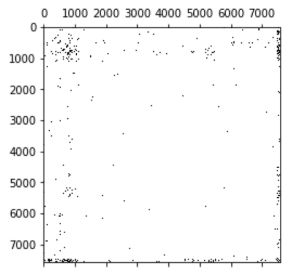
Fig. 10 Analysis of Flickr-Net data-set using SBM



(a) Adjacency matrix of Flickr-Net

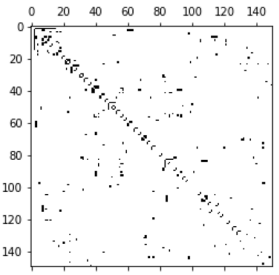


(b) Plot of singular values of Flickr-Net

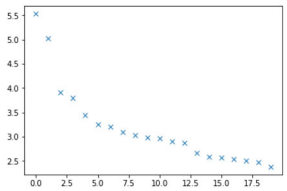


(c) Re-order nodes in adjacency matrix by class memberships

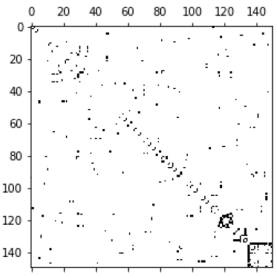
Fig. 12 Analysis of Protein-Net data-set using SBM



(a) Adjacency matrix of High-Net



(b) Plot of singular values of High-Net



(c) Re-order nodes in adjacency matrix by class memberships

Fig. 13 Simulate new networks from SBM fit to check model goodness of fit on Protein-Net for parameter transitivity. Number of simulations = 20

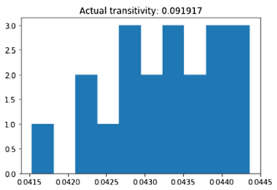
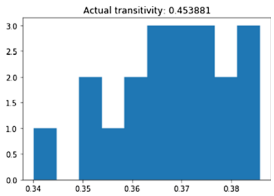


Fig. 15 Simulate new networks from SBM fit to check model goodness of fit on Wiki-Net for parameter transitivity. Number of simulations = 20



effectively. The posterior predictive check reveals that LSRs are not effectively capturing the transitivity of the original network.

5.7 Wikipedia

Wiki-Net is a data-set of 4777 and their hyperlinks to other web-pages in the network. The average clustering co-efficient i.e. transitivity of the network is 0.43. The adjacency matrix as shown in Fig. 14a reveals a a dense cluster of nodes in the network.

Using eigen-gap heuristic three latent classes are observed in Fig. 14b. Figure 14c shows the re-ordered adjacency matrix with three latent classes. Using the edge probabilities at the block level, new networks are simulated to check model goodness of fit.

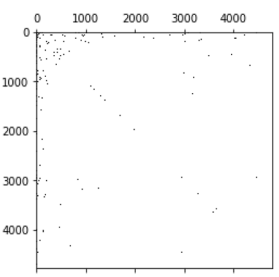
Figure 15 shows that SBM has generated LSRs that do not re-generate the original network (Wiki-Net) effectively. The posterior predictive check reveals that LSRs are not effectively capturing the transitivity of the original network.

5.8 Highway

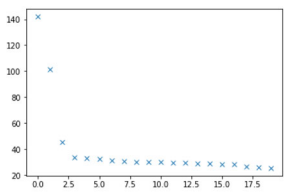
High-Net is a data-set of 205 cities and the highways that connect them to other cities in the network. The average clustering co-efficient i.e. transitivity between the members is 0.28. Figure 17 shows the results of fitting SBM to High-Net using the procedure outlined in Algorithm 1. Figure 16a shows presence of multiple dense regions (latent classes) in the plot of adjacency matrix A. Hence, to choose the latent classes we examine the singular values of A.

Using eigen-gap heuristic five latent classes are observed in Fig. 16b. The nodes in these latent classes are assigned class-memberships and then the adjacency matrix is re-ordered. Figure 16c shows the re-ordered adjacency matrix with five latent classes. Once the class-memberships are assigned, the edge probabilities at the block level are

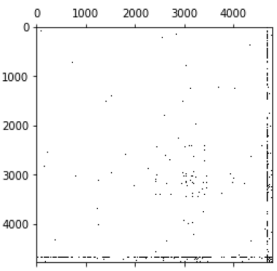
Fig. 14 Analysis of Wiki-Net data-set using SBM



(a) Adjacency matrix of Wiki-Net

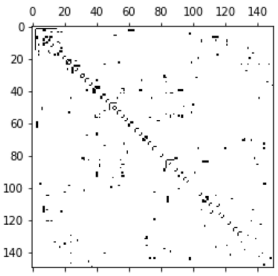


(b) Plot of singular values of Wiki-Net

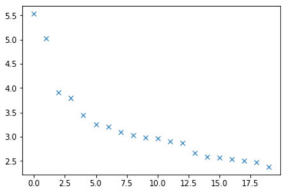


(c) Re-order nodes in adjacency matrix by class memberships

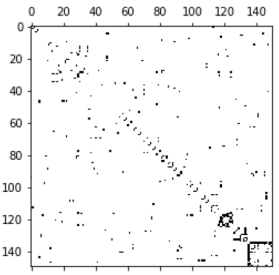
Fig. 16 Analysis of High-Net data-set using SBM



(a) Adjacency matrix of High-Net

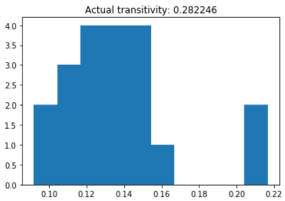


(b) Plot of singular values of High-Net

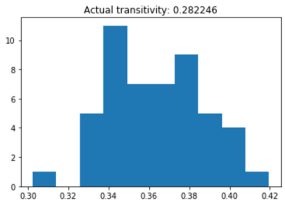


(c) Re-order nodes in adjacency matrix by class memberships

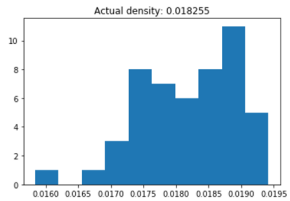
Fig. 17 Fitting SBM and LSM to High-Net data-set



(a) Simulate new networks from SBM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20.

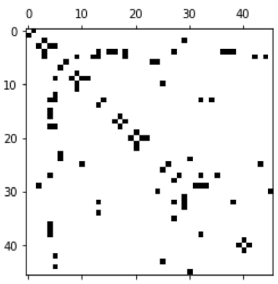


(b) Simulate new networks from LSM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20. Dimensions of latent space = 2.

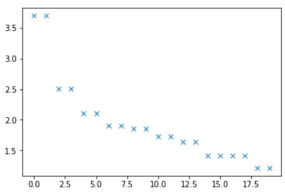


(c) Simulate new networks from LSM fit to check model goodness of fit on parameter density. Number of simulations = 20. Dimensions of latent space = 2.

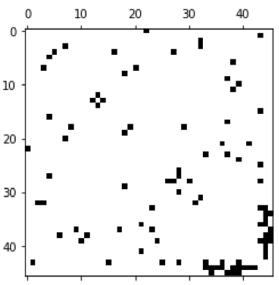
Fig. 18 Analysis of Grey-Net data-set using SBM



(a) Adjacency matrix of Grey-Net



(b) Plot of singular values of Grey-Net



(c) Re-order nodes in adjacency matrix by class memberships

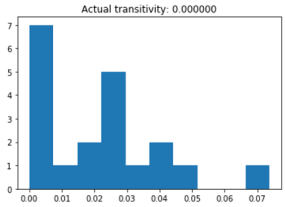
calculated. Finally new networks are simulated from edge probabilities to check model goodness of fit.

Figure 17a shows that SBM has generated LSRs that do not re-generate the original network (High-Net) effectively. The LSRs are not effectively capturing the transitivity of the original networks. LSMs were able to generate LSRs that could replicate the transitivity (as shown in Fig. 17b) and density (as shown in Fig. 17c) of the original network better than SBMs. Conclusions from posterior predictive check of SBM and LSM models is that LSRs generated using LSM are more effective than SBM in High-Net.

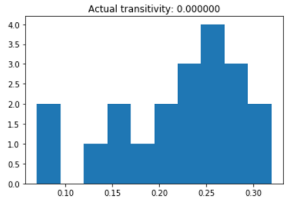
5.9 Grey's anatomy

Grey-Net is a data-set of 44 actors in the popular series Grey's Anatomy and the "sexual" relationships between them. Figure 19 shows the results of fitting SBM to Grey-Net using the procedure outlined in Algorithm 1. Figure 18a shows a dense adjacent matrix A but no presence of communities (latent classes) can detected in the plot. Hence, to choose the latent classes we examine the singular values of A.

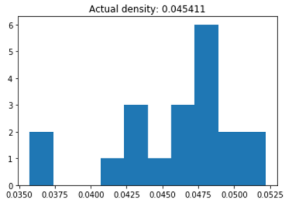
Fig. 19 Fitting SBM and LSM to Grey-Net data-set



(a) Simulate new networks from SBM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20.



(b) Simulate new networks from LSM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20. Dimensions of latent space = 2.



(c) Simulate new networks from LSM fit to check model goodness of fit on parameter density. Number of simulations = 20. Dimensions of latent space = 2.

Using eigen-gap heuristic seven latent classes are observed in Fig. 18b. The nodes in these latent classes are assigned class-memberships and then the adjacency matrix is re-ordered. Figure 18c shows the re-ordered adjacency matrix with seven latent classes. Once the class-memberships are assigned, the edge probabilities at the block level are calculated. Finally new networks are simulated from edge probabilities to check model goodness of fit.

Figure 19a shows that SBM has generated LSRs that re-generate the original network (Grey-Net) effectively. The LSRs are able to effectively capture the transitivity of the original network. LSMs were not able to generate LSRs that could replicate the transitivity (as shown in Fig. 19b) even though the fit to the density (as shown in Fig. 19c) of the original network was correct. Conclusions from posterior predictive check of SBM and LSM models is that LSRs generated using SBM are more effective than LSM in Grey-Net.

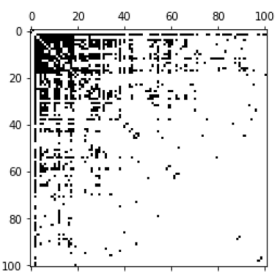
5.10 Trade

Trade-Net is a data-set of 99 countries and their trading ties. The average clustering co-efficient i.e. transitivity between the members is 0.44. Figure 21 shows the results of fitting SBM to Trade-Net using the procedure outlined in Algorithm 1. Figure 20a shows a dense adjacent matrix A. Hence, to choose the latent classes a plot of the singular values of A is created.

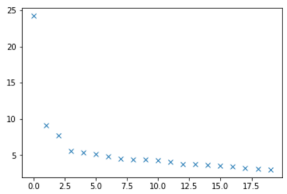
Using eigen-gap heuristic three latent classes are observed in Fig. 20b. The nodes in these latent classes are assigned class-memberships and then the adjacency matrix is re-ordered. Figure 20c shows the re-ordered adjacency matrix with three latent classes. Once the class-memberships are assigned, the edge probabilities at the block level are calculated. Finally new networks are simulated from edge probabilities to check model goodness of fit.

Figure 21a shows that SBM has generated LSRs that do not re-generate the original network (Trade-Net) effectively. The LSRs are not effectively capturing the transitivity of the original networks. LSMs were able to generate LSRs that could replicate the transitivity (as shown in Fig. 21b) and density (as shown in Figure 21c) of the original network better than SBM. Conclusions from posterior predictive check of SBM and LSM models is that LSRs generated using LSM are more effective than SBM in Trade-Net.

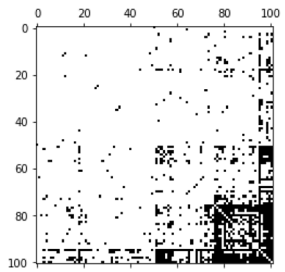
Fig. 20 Analysis of Trade-Net data-set using SBM



(a) Adjacency matrix of Trade-Net

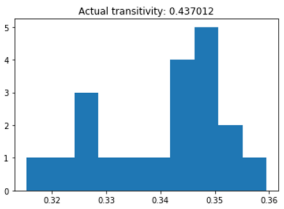


(b) Plot of singular values of Trade-Net

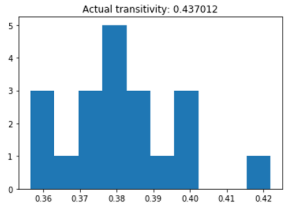


(c) Re-order nodes in adjacency matrix by class memberships

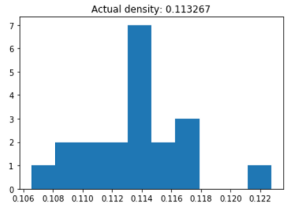
Fig. 21 Fitting SBM and LSM to Trade-Net data-set



(a) Simulate new networks from SBM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20.

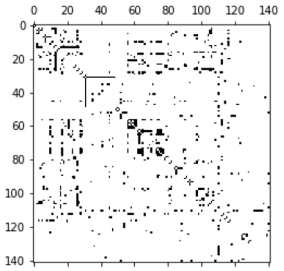


(b) Simulate new networks from LSM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20. Dimensions of latent space = 2.

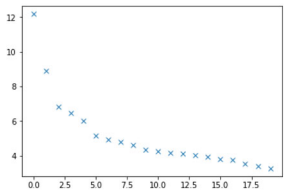


(c) Simulate new networks from LSM fit to check model goodness of fit on parameter density. Number of simulations = 20. Dimensions of latent space = 2.

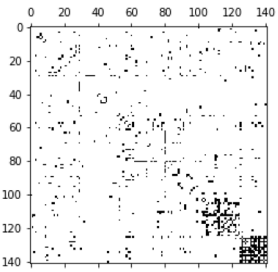
Fig. 22 Analysis of Bill-Net data-set using SBM



(a) Adjacency matrix of Bill-Net

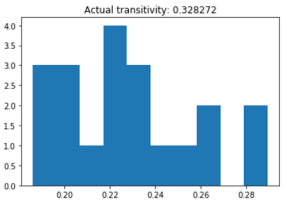


(b) Plot of singular values of Bill-Net

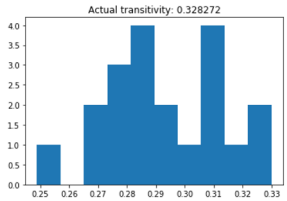


(c) Re-order nodes in adjacency matrix by class memberships

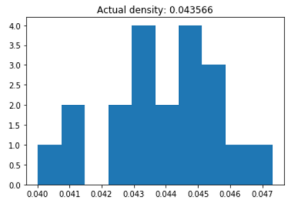
Fig. 23 Fitting SBM and LSM to Bill-Net data-set



(a) Simulate new networks from SBM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20.



(b) Simulate new networks from LSM fit to check model goodness of fit on parameter transitivity. Number of simulations = 20. Dimensions of latent space = 2.



(c) Simulate new networks from LSM fit to check model goodness of fit on parameter density. Number of simulations = 20. Dimensions of latent space = 2.

5.11 Bill co-sponsorship

Bill-Net is a data-set of 139 legislators that have co-sponsored legislations with each other. The average clustering co-efficient i.e. transitivity between the members is 0.32. High transitivity indicates that legislators sponsor bills of "friend of friends". Network is not connected and hence diameter is not calculated. Presence of isolates indicates inactive members in the assembly. Assortativity is neutral indicating a tendency of popular members colluding with each other and less popular figures with each other. A high gini-index indicates presence of a few active legislators that are key to several legislation's in the assembly. Activity of a legislative assembly is low compared to social networking sites and hence the edge density is low. Figure 3 shows the results of fitting SBM to Bill-Net using the procedure outlined in Algorithm 1.

Figure 22a shows a dense adjacent matrix A but to choose the latent classes the plot of singular values of A is needed.

Using eigen-gap heuristic four latent classes are observed in Fig. 22b. The nodes in these latent classes are assigned class-memberships and then the adjacency matrix is re-ordered. Figure 22c shows the re-ordered adjacency matrix with four latent classes. Once the class-memberships are assigned, the edge probabilities at the block level are calculated. Finally new networks are simulated from edge probabilities to check model goodness of fit.

Figure 23a shows that SBM has generated LSRs that do not re-generate the original network (Bill-Net) effectively. The LSRs are not effectively capturing the transitivity of the original networks. LSMs were able to generate LSRs that could replicate the transitivity (as shown in Fig. 23b) and density (as shown in Fig. 23c) of the original network better than SBM. Conclusions from posterior predictive check of SBM and LSM models is that LSRs generated using LSM are more effective than SBM in Bill-Net.

6 Conclusion

Network analysis is a crucial aspect of computation social science. The omnipresent nature of graphs (networks) in the world has further enhanced the importance of this field. Although networks are present in every domain, their analysis revealed that the structural characteristics shared by them are similar i.e. low average path, low diameter etc. It is further revealed that networks also capture the behavior of the entities present in them. Using concepts of graph theory it is possible to make statistically valid analysis of these systems and provide insights into their growth.

Networks across different domains saw low edge density and presence of inequality. Networks such as Twt-Net, Gplus-Net, Flickr-Net, Wiki-Net, Blog-Net, Grey-Net and Bill-Net are a particular type of networks called “social networks”. Social networks represent the sum of all professional, friendship or family ties of the actors involved in them. Social networks were observed to have higher edge density and average degree compared to other networks. They also had high transitivity, low diameter and negative assortativity.

Statistical models such as Stochastic Block Models (SBM) and Latent Space Model (LSM) were fit to various application scenarios. These are regarded as the “The most promising class of statistical models for expressing structural properties of networks observed at one moment in time”. However, these models ignore the attributes associated with the networks. SBM are applicable for networks with nodes in range of 10^3 whereas LSM are feasible for networks with a few hundred nodes. Hence, it is necessary to investigate models that can scale to large networks $(10^3 - 10^7)$.

References

1. Barabási A-L et al (2016) Network science. Cambridge University Press, Cambridge
2. Denny M (2014) Social network analysis. Institute for Social Science Research, University of Massachusetts, Amherst
3. Denny M (2015) Intermediate social network theory. Institute for Social Science Research, University of Massachusetts, Amherst
4. Jackson MO (2010) Social and economic networks. Princeton University Press, Princeton
5. Nerurkar P, Chandane M, Bhirud S (2019) A comparative analysis of community detection algorithms on social networks. Computational intelligence: theories, applications and future directions, vol I. Springer, New York, pp 287–298
6. Nerurkar P, Shirke A, Chandane M, Bhirud S (2018) A novel heuristic for evolutionary clustering. Procedia Comput Sci 125:780–789
7. Granell C, Darst RK, Arenas A, Fortunato S, Gómez S (2015) Benchmark model to assess community structure in evolving networks. Phys Rev E 92(1):12–19
8. Zou X, Yang J, Zhang J (2018) Microblog sentiment analysis using social and topic context. PloS one 13(2):119–163
9. Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44
10. Gomez Vicenc, Kaltenbrunner Andreas, Lopez Vicente (2008) Statistical analysis of the social network and discussion threads in slashdot. In: Proceedings of the 17th international conference on World Wide Web. ACM, pp 645–654
11. McGlohon Mary, Akoglu Leman, Faloutsos Christos (2011) Statistical properties of social networks. In: Social network data analytics. Springer, New York, pp 17–42
12. Golosovsky Michael (2018) Preferential attachment mechanism of complex network growth:” rich-gets-richer” or” fit-gets-richer”? arXiv preprint arXiv:1802.09786
13. Lopez FA, Barucca P, Fekom M, Coolen ACC (2018) Exactly solvable random graph ensemble with extensively many short cycles. J Phys A Math Theor 51:085101
14. Jackson MO, Rogers BW, Zenou Y (2017) The economic consequences of social-network structure. J Econ Lit 55(1):49–95
15. Leduc MV, Jackson MO, Johari R (2017) Pricing and referrals in diffusion on networks. Games Econ Behav 104:568–594
16. Cicala S, Fryer RG, Spenkuch JL (2017) Self-selection and comparative advantage in social interactions. J Eur Econ Assoc 16:983–1020
17. Ricci V (2005) Fitting distributions with r. *Contributed Documentation available on CRAN*, 96
18. Johnson RA, Johnson RA, Wichern DW (2003) Applied multivariate stastistical analysis. Prentice-Hall of India Private Limited, Delhi
19. Pattison P, Wasserman S (1996) Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p. Psychometrika 61(3):401–425
20. Park J, Newman MEJ (2005) Solution for the properties of a clustered network. Phys Rev E 72(2):026136
21. Prusaczyk B, Maki J, Luke DA, Lobb R (2018) Rural health networks: How network analysis can inform patient care and organizational collaboration in a rural breast cancer screening network. J Rural Health 35:222–228

22. Diaconis P, Chatterjee S (2013) Estimating and understanding exponential random graph models. Ann Stat 41(5):2428–2461
23. Kalish Y, Lusher D, Robins G, Pattison P (2007) An introduction to exponential random graph (p*) models for social networks. Soc Netw 29(2):173–191
24. Snijders TAB (2002) Markov chain monte carlo estimation of exponential random graph models. J Soc Struct 3(2):1–40
25. Snijders T, Moody J, Besag J, Handcock MS, Robins G (2003) Assessing degeneracy in statistical models of social networks. In J Am Stat Assoc, Citeseer
26. Sly A, Bhamidi S, Bresler G (2008) Mixing time of exponential random graphs. In: Foundations of Computer Science, 2008. FOCS'08. IEEE 49th annual IEEE symposium on. IEEE, Piscataway, pp 803–812
27. Butts CT, Goodreau SM, Pavel NK, Bender-deMoll S, Morris M, Handcock MS, Hunter DR (2016) statnet: Software Tools for the Statistical Analysis of Network Data. The Statnet Project
28. Butts CT, Goodreau SM, Morris M, Handcock MS, Hunter DR (2008) statnet: Software tools for the representation, visualization, analysis and simulation of network data. J Stat Softw 24:1–11
29. Hoff PD (2015) Dyadic data analysis with amen. arXiv preprint arXiv:1506.08237
30. Leskovec Jure, Krevl Andrej (June 2014) SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data. Accessed 21 Oct 2018