Chapter 3

EMPIRICAL ANALYSIS USING STATISTICAL MODELS

3.1 Introduction

Real-world systems take diverse forms, and hence, networks that model such data are complex. Such high-dimensional data contains redundancies that can be exploited to learn compact representations. Frameworks of data compression seek only to reduce the size of data. However, latent space representation methods seek to infer an underlying geometry in the data [33, 40].

Network data has issues such as: (i) lack of independent data-points, (ii) computationally costly calculations for statistics, (iii) in-applicability of parallel or distributed algorithms and (iv) the curse of dimensionality. The latent space representation methods minimize drawbacks associated with network-structured data. In the previous chapter, statistical models such as Stochastic Block Model (S.B.M.) and latent variable model were discussed. These are "Promising class of statistical models for expressing networks into low-dimensional geometries" [41, 42, 43, 44].

This chapter aims to verify the usefulness of these two state of the art frameworks by performing extensive experimentations. Generative network models are fit to various application scenarios for obtaining L.S.R. This is done in order to understand the

suitability of these models. Concepts of network science from the literature for experimentation on real-world datasets. Analysis is performed on twelve standard datasets from diverse real-world systems. Concepts from network science are applied to network representations of real-world systems. Aim is to understand the networks' characteristics and infer the reasons for observing such characteristics. Extensive experiments are performed on network datasets from the Stanford Network Analysis Project (S.N.A.P.). Objective is for understanding the structure and behaviour of systems. Efficacy of the S.B.M. and latent variable model in obtaining L.S.R. of network representation models is highlighted.

Rest of the chapter is organized into three sections: Section 3.2 gives the definitions of network science concepts used for the analysis and the description of the datasets. Section 3.3 describes the methodology of the analysis and the experiments performed. Summary of the results and observations are presented in Section 3.4.

3.2 Concepts of Network Science

3.2.1 Definitions

3.2.1.1 Diameter (diam(G))

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 $x, y \in G$.

3.2.1.2 Average clustering coefficient (c)

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.

$$c = \frac{|\{u, v, w \in V \mid u \sim v \sim w \sim u\}|}{|\{u, v, w \in V \mid u \sim v \neq w \sim u\}|} \tag{3.1}$$

3.2.1.3 Assortativity coefficient (r)

It is positive if vertices with high in-degrees tend to connect to each, and negative otherwise. Assortativity coefficient r for edges $i = 1, 2, ..., M$ with j, k as degrees of

the vertices at the ends of the i^{th} edge.

$$r = \frac{M^{-1}\sum_i j_i k_i - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2}{M^{-1}\sum_i \frac{1}{2}({j_i}^2 + {k_i}^2) - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2} \tag{3.2}$$

3.2.1.4 Edge density (D)

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

3.2.1.5 Gini index (G)

It takes values between zero and one, with zero denoting total equality between degrees, and one denoting a single node's dominance. 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:

$$G = \frac{2\sum_{i=1}^{n} i d_i}{n\sum_{i-1}^{n} d_i} - \frac{n+1}{n} \tag{3.3}$$

3.2.1.6 Average degree (d)

$$d = \frac{1}{|V|}\sum_{u \in V} d(u) \tag{3.4}$$

3.2.1.7 Average path length ($\bar{P}$)

The average number of steps along the shortest paths for all possible pairs of network nodes

$$\bar{P} = \frac{1}{n(n-1)}\sum_{i \neq j} d(v_i, v_j) \tag{3.5}$$

3.2.1.8 Variables in network data

Network data includes :

- A set of nodes (entities, objects, actors, egos, individuals) and edges (links, ties, dyads)
- Variables measured on nodes are nodal variables, and those measured on pairs of nodes (edges) are dyadic variables

3.2.1.9 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 are ordered is called ordinal. A valued relation whose possible values lack an order is called categorical.

3.2.2 Real-world systems: network data

In this study twelve datasets were used in the experiments. Each dataset is a real-world system represented as a network. Three types of datasets used are: Data-sets with no side information (see Section 3.2.2.1), Data-sets with binary attributes (see Section 3.2.2.2) and Data-sets with mixed attributes (see Section 3.2.2.3).

3.2.2.1 Data-sets with no side information

These are networks $G = (V, E)$ with vertex set V and edge set E. The vertex attribute set V^a, and the edge attribute set E^a is null. Table 3.1 and 3.2 gives the description of the dataset crawled from online social networking websites Twitter.com (Twt-Net) and Google+ (Gplus-Net). All datasets are publicly available at Stanford Network Analysis Platform (S.N.A.P.) - a network data repository [119]. In Twt-Net and Gplus-Net, the system modelled as a network is the online social networking website. The system's entities are the users of these platforms, and 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.

Table 3.1: Description of network data-sets with no side information

Description	**Twt-Net**	**Gplus-Net**
$\|V\|$	185	923
$\|E\|$	5156	39400

Table 3.2: Data Analysis of network data-sets with no side information

Description	**Twt-Net**	**Gplus-Net**
c	0.44	0.3
diam(G)	8	7
r	-0.19	-0.23
$\bar{P}$	2.18	2.58
D	0.201	0.05
G	0.41	0.52
d	55 (σ = 41)	85 (σ = 106)

3.2.2.2 Data-sets with binary attributes

These are networks $G = (V, E)$ with vertex set $n = |V|$ and edge set E. The vertex attribute set is $V^a = R^{n*f}$ and the edge attribute set is $E^a = \phi$, such that 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 3.3 and 3.4 describes datasets obtained from Blog.com (Blog-Net), Flickr.com (Flickr-Net), a protein-protein interaction network (Protein-Net), Wikipedia.com (Wiki-Net), C.O.R.A. Research Paper Classification Data-set by Andrew McCallum of University of Massachusetts Amherst (Cora-Net) and citation indexing website Citeseer.com (Cite-Net). All datasets are publicly available at Stanford Network Analysis Platform (S.N.A.P.) [119]. Blog-Net is a directed graph of users commenting on blogs by other users on the website Blog.com. Flickr-Net is a directed graph of users following other users on Flickr.com. Protein-Net is an undirected graph of proteins interacting with other proteins. Wiki-net is a directed graph of articles citing other articles on Wikipedia.com. In Cora-Net and Cite-Net, the nodes are the academic papers. If paper i cites paper j, then this is denoted in the network by a directed edge from node i to node j.

Table 3.3: Description of network data-sets with binary attributes

Description	**Blog-Net**	**Flickr-Net**	**Protein-Net**	**Wiki-net**	**Cora-net**	**Cite-net**
$\|V\|$	5196	7575	3890	4777	2708	3312
V^a	8189	12047	50	40	1434	3704
$\|E\|$	171743	239738	37845	54810	5429	4732

Table 3.4: Data Analysis of network data-sets with binary attributes

Description	**Blog-Net**	**Flickr-Net**	**Protein-Net**	**Wiki-net**	**Cora-net**	**Cite-net**
c	0.08	0.1	0.09	0.43	0.14	0.13
diam(G)	~ 2	~ 2	8	4	6	8
r	-0.02	-0.23	-0.09	-0.27	0.11	0.12
$\bar{P}$	~ 2.03	~ 2.15	3.09	2.15	1.74	1.81
D	0.012	0.008	0.005	0.004	0.0002	0.0004
G	0.39	0.67	0.63	0.62	0.42	0.43
d	66.30 (σ = 54.8)	63.3 (σ = 131.52)	19.45 (σ = 34.29)	22.95 (σ = 105.92)	2.4 (σ = 2.63)	2.8 (σ = 3.41)

3.2.2.3 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 3.5 and 3.6 describes datasets such as High-Net - an undirected network of the highways in Southern California, as observed in 2016. It shows the cities connected by highways. Grey-Net is an undirected sexual contact network between characters of the television show Grey's Anatomy. It gives information about actors in relationships with other actors in the series. Trade-Net is an undirected trade network of automotive electrical goods between Asia and European countries in 2016. It shows countries that have trade ties with each other. Bill-Net was an undirected bill co-sponsorship network in the parliament of Slovakia in 2014. It shows information about legislators co-sponsoring bills together. All datasets are publicly available at Stanford Network Analysis Platform

(S.N.A.P.) [119].

Table 3.5: Description of network data-sets with mixed attributes

Description	**High-Net**	**Grey-Net**	**Trade-Net**	**Bill-net**
$\|V\|$	205	44	99	139
Node attributes	9	17	12	4
$\|E\|$	203	46	725	471
Edge attributes	3	2	1	1

Table 3.6: Data Analysis of network data-sets with mixed attributes

Description	**High-Net**	**Grey-Net**	**Trade-Net**	**Bill-net**
c	0.28	0	0.437	0.32
diam(G)	16	8	2	4
r	0.12	-0.22	-0.32	0.014
$\bar{P}$	6.8	3.49	2.31	3.71
D	0.018	0.04	0.11	0.043
G	0.54	0.37	0.61	0.46
d	1.97 (σ = 2.12)	2.09 (σ = 1.72)	14.64 (σ = 18.74)	6.77 (σ = 6.23)

3.3 Analysis of Networks using Statistical Models

This section provides the analysis of real-world systems discussed in Section 3.2.2 from a network perspective. The concepts of graph theory defined in Section 3.2.1 are used to generate insights about the structure of these systems. The systems under investigation belong to diverse scientific domains. The objective is to highlight the efficacy of the S.B.M. and latent variable model in obtaining L.S.R. of graph-structured data. The procedure followed for the analysis is given in Algorithm 3.

Networks such as Twt-Net, Gplus-Net, Flickr-Net, Wiki-Net, Blog-Net, Grey-Net and Bill-Net are a particular type of network called "social networks." Social networks represent the sum of all professional, friendship, or family ties of the actors involved in them. Study of such systems provides an understanding of the behaviour of the actors

in them.

Performing downstream network tasks such as clustering on these systems' network representations have issues due to their scale and high dimensionality. Therefore, we generate the Latent space representation (L.S.R.) of the original networks using generative statistical models such as the S.B.M. and latent variable model. This reduces their dimensionality and makes the downstream network task amenable.

Algorithm 3: Procedure to understand structure and behaviour of networks

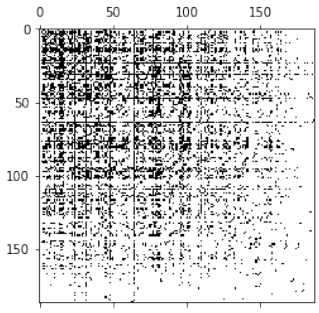
1. Load adjacency matrix Y of the network G;
2. Calculate descriptive measures $|V|$, $|E|$;
3. Calculate network statistics c, $\text{diam}(G)$, r, $\bar{P}$, D, G, d;
4. Plot the adjacency matrix and singular values;
5. Using eigen-gap heuristic observe latent classes;
6. Assign node membership according to the latent classes;
7. Re-order nodes in adjacency matrix by class memberships;
8. Fit latent variable model and S.B.M. Set dimensions of latent space;
9. Analyze model fit;
10. Simulate new networks from model fit. Set number of simulations;
11. Check how well simulated networks preserve actual network transitivity (posterior predictive check);

3.3.1 Twitter

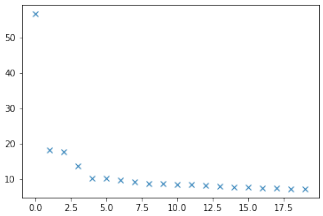
Twt-Net is a dataset of 185 twitter users of a community and the "follower-following" relationships between them. The average clustering coefficient, i.e., transitivity between the members is 0.44. The high transitivity, coupled with a low average path length of 2.18 suggests that information diffusion 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 a large number of social contacts (average degree = 55). This leads to a low average path length and diameter and high edge density.

Latent variable model was observed to be a better choice for L.S.R. than S.B.M. as

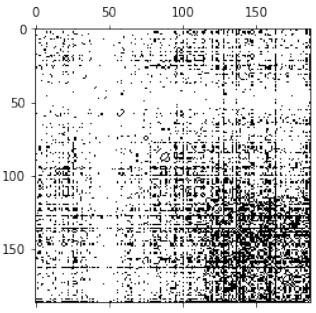
the latter does not explicitly model transitivity. Figure 3.2 shows the results of fitting S.B.M. to Twt-Net using the procedure outlined in Algorithm 1. Figure 3.1a shows a dense adjacent matrix A but no presence of communities (latent classes) can be detected in the plot. Hence, to choose the latent classes, the plot of the singular value of A needs to be obtained.



(a) Adjacency matrix of Twitter.com data-set Twt-Net



(b) Plot of singular values of Twitter.com data-set Twt-Net



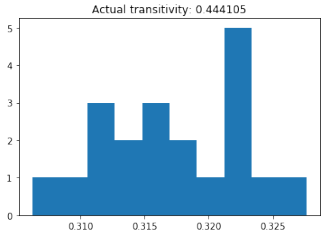
(c) Re-order nodes in adjacency matrix by class memberships

Figure 3.1: Analysis of Twitter.com data-set Twt-Net using S.B.M.

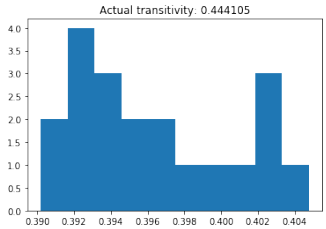
Figure 3.1b shows four latent classes were observed using eigen-gap heuristic (gaps in the singular values correspond to latent classes in the network). The nodes in these latent classes were assigned class-memberships, and then the adjacency matrix was re-ordered. Figure 3.1c shows the re-ordered adjacency matrix with four latent classes. Once the class-memberships were assigned, the edge probabilities at the block level were calculated. Finally, new networks were simulated from edge probabilities to check the goodness of fit of the model.

Figure 3.2a shows the transitivity of regenerated networks using S.B.M. The transitivity ranges from 0.3 - 0.33 (upper and lower bounds of the 95% confidence interval), which is 0.11 lower than the actual transitivity of Twt-Net. Figure 3.2b shows the transitivity of the regenerated networks using a latent variable model. Figure 3.2b shows that the transitivity of the regenerated networks is in the range of 0.4-0.404 (upper and lower bounds of the 95% confidence interval), which is 0.04 less than the actual transitivity of Twt-Net. Similarly, from Figure 3.2c, the edge density of the regenerated networks using the latent variable model is in the range of 0.19-0.204 (upper and lower

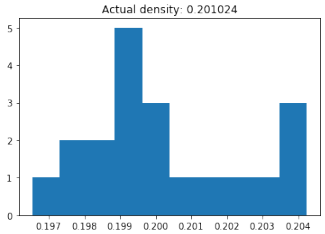
bounds of the 95% confidence interval). Whereas, the actual edge density of Twt-Net is 0.201. Thus, the latent variable model was able to generate L.S.R. that could replicate the transitivity (Figure 3.2b) and edge density (Figure 3.2c) of the original network. However, S.B.M. has generated L.S.R. that does not regenerate the original network (Twt-Net) effectively as a latent variable model, i.e. within ±0.05 from actual. Conclusions from the posterior predictive check of S.B.M. and latent variable model is that L.S.R. generated using the latent variable model are more effective than S.B.M. in Twt-Net.



(a) Simulate new networks from Stochastic Block Models (S.B.M.) fit to check model goodness of fit on parameter transitivity



(b) Simulate new networks from latent variable model fit to check model goodness of fit on parameter transitivity



(c) Simulate new networks from latent variable model fit to check model goodness of fit on parameter edge density

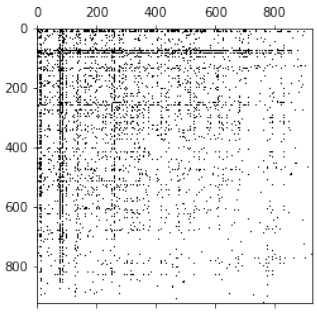
Figure 3.2: Fitting S.B.M. and latent variable model to Twt-Net data-set

3.3.2 Google+

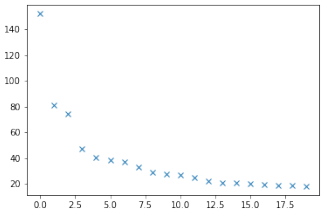
Gplus-Net is a dataset of 923 Google+ users of a community and the "follower-following" relationships between them. The average clustering coefficient, i.e., transitivity between the members is 0.3. The high transitivity, coupled with a low average path length of 2.58, suggests the possibility of rapid 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 a large number of social contacts (average degree = 85). This leads to low average path length and diameter and high edge density. Edge density is lower compared to Twt-Net, which could imply that users are less active on Google+ than Twitter.

Applying the latent variable model to obtain the L.S.R. of Gplus-Net is infeasible due

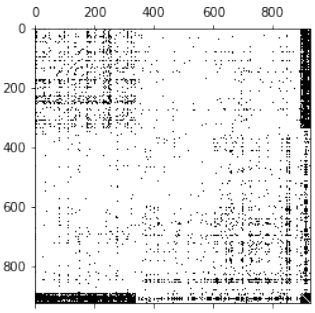
to the network's scale. Therefore, only S.B.M. was fit to the data to estimate the block probabilities following the procedure outlined in Algorithm 1.



(a) Adjacency matrix of Google+ website data-set Gplus-Net



(b) Plot of singular values of Google+ website data-set Gplus-Net



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

Figure 3.3: Analysis of of Google+ website data-set Gplus-Net using S.B.M.

Figure 3.3b shows three latent classes can be observed using eigen-gap heuristic. Figure 3.3c shows the re-ordered adjacency matrix with three latent classes. New networks are simulated to check model goodness of fit using the edge probabilities at the block level.

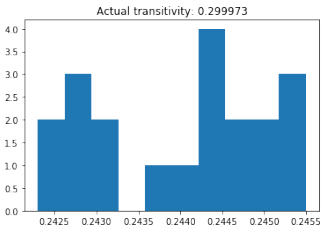
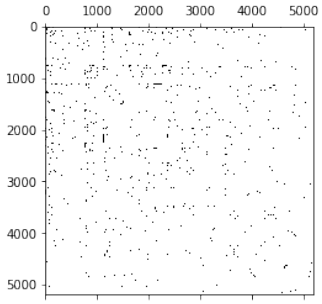


Figure 3.4: Simulate new networks from S.B.M. fit to check model goodness of fit on Gplus-Net for parameter transitivity

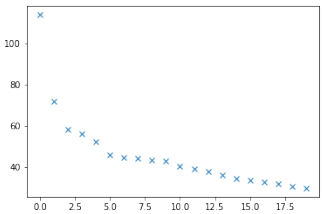
Block probabilities are estimated, followed by verification of the goodness of fit. Figure 3.4 shows that the range of transitivity of the simulated networks is 0.242-0.245 (upper and lower bounds of the 95% confidence interval). The network's actual transitivity is 0.3, which is 0.05 more than the range of transitivity of the simulated networks. S.B.M. has generated L.S.R. that regenerate the original network (Gplus-Net) effectively. The posterior predictive check reveals that L.S.R. adequately captures the transitivity of the original network.

3.3.3 Blog

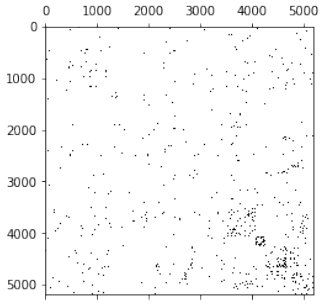
Blog-Net is a dataset of 5196 users and the relationships captured in the networks are of users commenting on blogs by other users. The average clustering coefficient, i.e., transitivity of the network is 0.08. S.B.M. is preferred for models with low transitivity. The latent variable model could not scale to Blog-Net as the nodes are in the order of $\sim 10^3$. The network has node attributes, however, S.B.M. could not model attributes.Figure 3.5a shows the adjacency matrix does not reveal the apparent possibilities of communities in the network.



(a) Adjacency matrix of Blog.com website data-set Blog-Net



(b) Plot of singular values of Blog.com website data-set Blog-Net



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

Figure 3.5: Analysis of Blog.com website data-set Blog-Net using S.B.M.

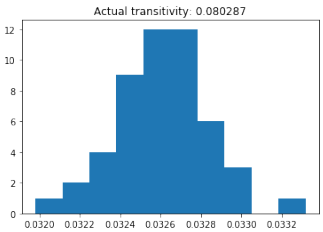


Figure 3.6: Simulate new networks from S.B.M. fit to check model goodness of fit on Blog-Net for parameter transitivity

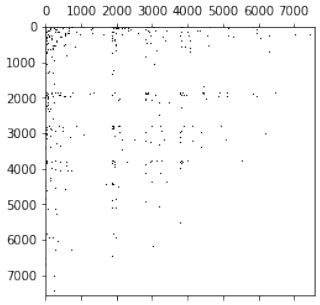
Figure 3.5b shows five latent classes can be observed using eigen-gap heuristic. Figure 3.5c shows the re-ordered adjacency matrix with five latent classes. New networks are simulated to check model goodness of fit using the edge probabilities at the block level.

Figure 3.6 shows that S.B.M. has generated networks that have transitivity in the range of 0.032-0.033 (upper and lower bounds of the 95% confidence interval), whereas the actual transitivity of the original network (0.08) is higher by 0.05. Thus, S.B.M. regenerates the original network (Blog-Net) effectively. The posterior predictive check reveals that L.S.R. is adequately capturing the transitivity of the original network. However, S.B.M. did not consider that node attributes in the original network. Thus, S.B.M. is not suitable for modelling networks with attribute information.

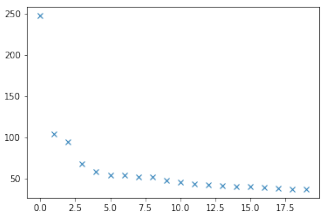
3.3.4 Flickr

Flickr-Net is a dataset of 7575 users and the relationships captured in the networks are of users "following" the profiles of other users. The average clustering coefficient, i.e., transitivity of the network is 0.1. Similar to online social networking websites like Twitter and Google+, Flickr-Net also has high Gini-index (0.67), negative assortativity (-0.23), high average degree (63.3), low diameter (2) and low average path length (2.15).

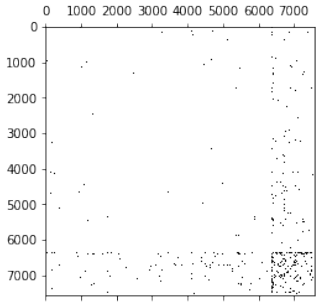
The latent variable model is not feasible for Flickr-Net due to a large number of nodes, so S.B.M. was fit to the data for analysis. Figure 3.7a shows the adjacency matrix which reveals several dense regions in the network.



(a) Adjacency matrix of Flickr.com website data-set Flickr-Net



(b) Plot of singular values of Flickr.com website data-set Flickr-Net



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

Figure 3.7: Analysis of Flickr.com website data-set Flickr-Net using S.B.M.

Using eigen-gap heuristic, three latent classes are observed in Figure 3.7b. Figure 3.7c shows the re-ordered adjacency matrix with five latent classes. New networks are

simulated to check model goodness of fit using the edge probabilities at the block level.

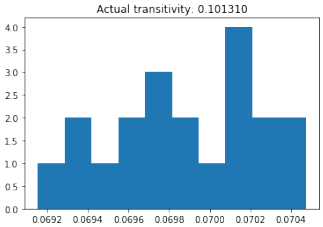
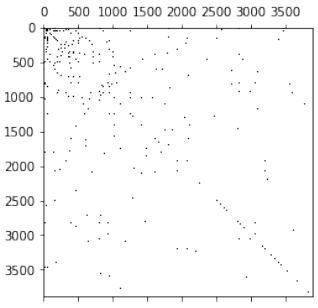


Figure 3.8: Simulate new networks from S.B.M. fit to check model goodness of fit on Flickr-Net for parameter transitivity

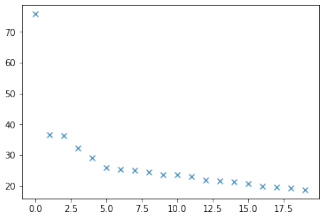
Figure 3.8 shows that the range of transitivity of the regenerated networks obtained from the simulation is 0.068-0.071 (upper and lower bounds of the 95% confidence interval), whereas the actual transitivity of the original network is 0.1 (higher by 0.03). Thus, S.B.M. has generated L.S.R. that regenerate the original network (Flickr-Net) effectively. The posterior predictive check reveals that L.S.R. is adequately capturing the transitivity of the original network.

3.3.5 Protein protein interaction

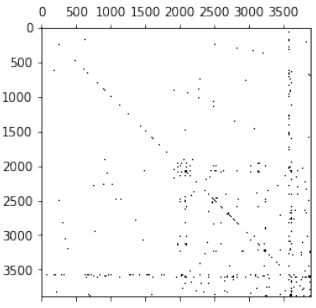
Protein-Net is a dataset of 3890 proteins and their interactions with each other. The average clustering coefficient, i.e., transitivity of the network is 0.09.



(a) Adjacency matrix of Protein-protein interaction Protein-Net



(b) Plot of singular values of Protein-protein interaction Protein-Net



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

Figure 3.9: Analysis of Protein-protein interaction Protein-Net data-set using S.B.M.

On average, a protein interacts with up to 20 other proteins. However, high Gini-index indicates that the majority of the interactions are concentrated amongst a section of proteins ($\sim$ 63%). Fitting latent variable model is not feasible for Protein-Net due to large number of nodes, and analysis was performed with S.B.M. Figure 3.9a illustrates the adjacency matrix which reveals a dense cluster of nodes in the network.

Using eigen-gap heuristic, two latent classes are observed in Figure 3.9b. Figure 3.9c shows the re-ordered adjacency matrix with two latent classes. New networks are simulated to check model goodness of fit using the edge probabilities at the block level.

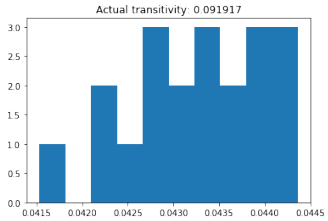


Figure 3.10: Simulate new networks from S.B.M. fit to check model goodness of fit on Protein-Net for parameter transitivity

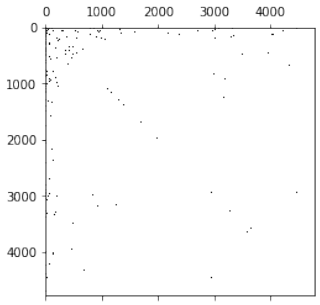
Figure 3.10 gives the simulation results that reveal the range of transitivity of the simulated networks as 0.04-0.044 (upper and lower bounds of the 95% confidence interval), which is 0.05 lower than the actual transitivity of the original network. Thus, S.B.M. has generated L.S.R. that regenerate the original network (Protein-Net) effectively. The posterior predictive check reveals that L.S.R. is adequately capturing the transitivity of the original network.

3.3.6 Wikipedia

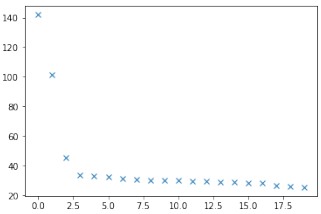
Wiki-Net is a dataset of 4777 and their hyperlinks to other web-pages in the network. The average clustering coefficient, i.e., transitivity of the network is 0.43. The adjacency matrix shown in Figure 3.11a reveals a dense cluster of nodes in the network.

Using eigen-gap heuristic, three latent classes are observed in Figure 3.11b. Figure 3.11c shows the re-ordered adjacency matrix with three latent classes. New networks

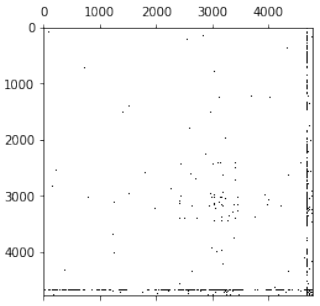
are simulated to check model goodness of fit using the edge probabilities at the block level.



(a) Adjacency matrix of Wikipedia.com data-set Wiki-Net



(b) Plot of singular values of Wikipedia.com data-set Wiki-Net



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

Figure 3.11: Analysis of of Wikipedia.com data-set Wiki-Net using S.B.M.

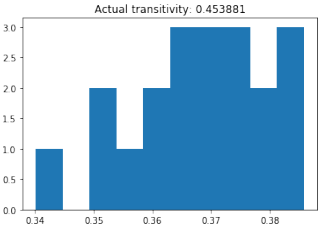


Figure 3.12: Simulate new networks from S.B.M. fit to check model goodness of fit on Wiki-Net for parameter transitivity

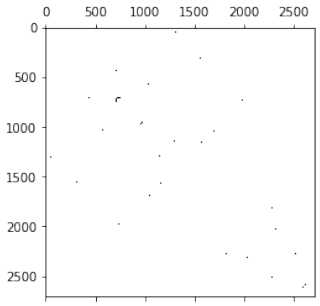
Figure 3.12 shows that regenerated networks from the simulations have transitivity in the range of 0.34-0.39 (upper and lower bounds of the 95% confidence interval), whereas the actual transitivity of the original network is 0.453 (0.08-0.12 higher). Thus, S.B.M. has generated L.S.R. that does not regenerate the original network (Wiki-Net) effectively. The posterior predictive check reveals that L.S.R. is not adequately capturing the transitivity of the original network.

3.3.7 Cora-Net

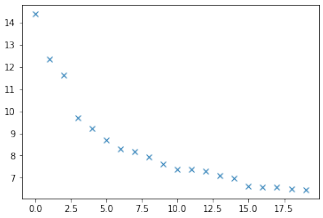
Cora-Net is a dataset of 2708 academic papers and the citations between them. The average clustering coefficient, i.e., transitivity of the network is 0.09. Transitivity is

lower for citation networks as compared to social networks as i citing j and j citing k might not lead to k citing i.

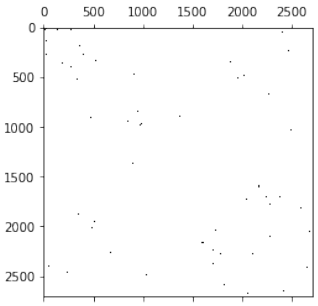
S.B.M. is preferred for models with low transitivity. The latent variable model will not be able to scale to Cora-Net as the nodes are in the order of $\sim 10^3$. The network is sparse, i.e., $|V| \sim |E|$ and the adjacency matrix, as shown in Figure 3.13a does not reveal apparent possibilities of communities in the network.



(a) Adjacency matrix of CORA Research Paper Classification Data-set Cora-Net



(b) Plot of singular values of CORA Research Paper Classification Data-set Cora-Net



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

Figure 3.13: Analysis of CORA Research Paper Classification Data-set Cora-Net using S.B.M.

Using eigen-gap heuristic, three latent classes are observed in Figure 3.13b. Figure 3.13c shows the re-ordered adjacency matrix with three latent classes. New networks are simulated to check model goodness of fit using the edge probabilities at the block level.

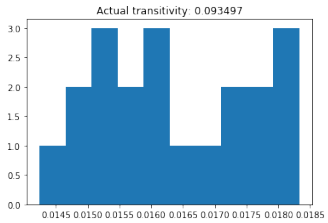


Figure 3.14: Simulate new networks from S.B.M. fit to check model goodness of fit on Cora-Net for parameter transitivity

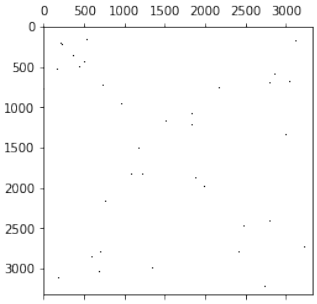
Figure 3.14 shows that the range of transitivity of the simulated networks as 0.0145-

0.0185 (upper and lower bounds of the 95% confidence interval), whereas the actual transitivity of the original network is 0.093 (higher by 0.08). Thus, S.B.M. has generated L.S.R. that does not regenerate the original network (Cora-Net) effectively. The posterior predictive check reveals that L.S.R. is not adequately capturing the transitivity of the original network.

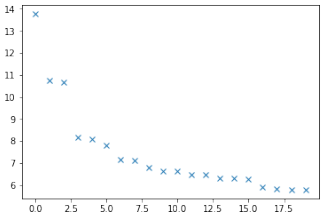
3.3.8 CiteSeer

Cite-Net is a dataset of 3312 academic papers and the citations between them. The average clustering coefficient, 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.

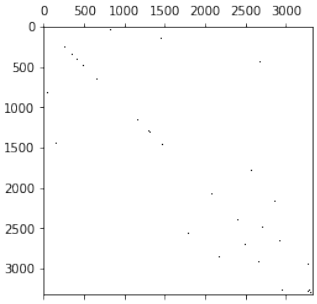
S.B.M. is preferred for models with low transitivity. The latent variable model will not be able to scale to Cite-Net as the nodes are in the order of $\sim 10^3$. The network is sparse, i.e., $|V| \sim |E|$ and the adjacency matrix, as shown in Figure 3.15a does not reveal apparent possibilities of communities in the network.



(a) Adjacency matrix of Citeseer.com data-set Cite-Net



(b) Plot of singular values of Citeseer.com data-set Cite-Net



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

Figure 3.15: Analysis of Citeseer.com data-set Cite-Net using S.B.M.

Figure 3.15b shows that five latent classes can be observed using eigen-gap heuristic. Figure 3.15c shows the re-ordered adjacency matrix with five latent classes. New networks are simulated to check model goodness of fit using the edge probabilities at the block level.

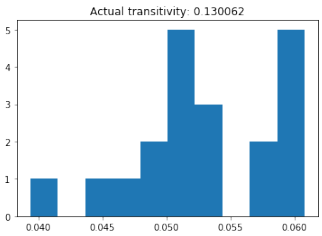
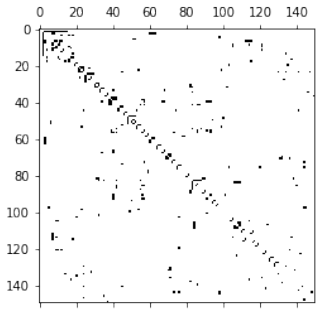


Figure 3.16: Simulate new networks from S.B.M. fit to check model goodness of fit on Cite-Net for parameter transitivity

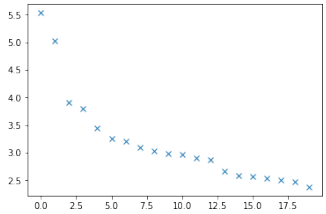
Figure 3.16 shows the range of transitivities of the simulated models as 0.04-0.06 (upper and lower bounds of the 95% confidence interval), which is 0.07 less than the actual transitivity of the original network. Thus, S.B.M. has generated L.S.R. that does not regenerate the original network (Cite-Net) effectively. The posterior predictive check reveals that L.S.R. is not adequately capturing the transitivity of the original network.

3.3.9 Highway

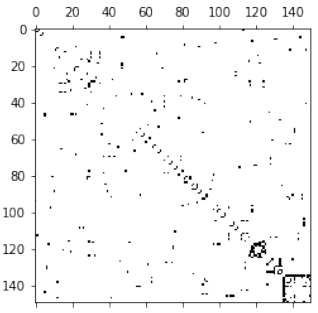
High-Net is a dataset of 205 cities and the highways that connect them to other cities in the network. The average clustering coefficient, i.e., transitivity between the members is 0.28. Figure 3.18 shows the results of fitting S.B.M. to High-Net using the procedure outlined in Algorithm 1. Figure 3.17a 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.



(a) Adjacency matrix of Transportation network data-set High-Net



(b) Plot of singular values of Transportation network data-set High-Net

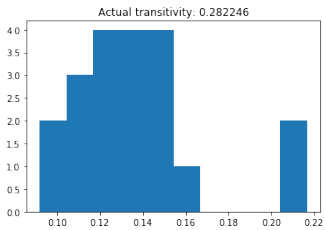


(c) Re-order nodes in adjacency matrix by class memberships

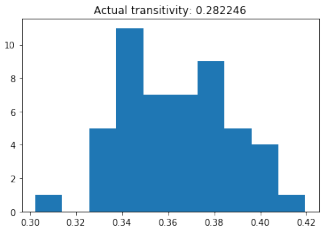
Figure 3.17: Analysis of Transportation network data-set High-Net using S.B.M.

Using eigen-gap heuristic, five latent classes are observed in Figure 3.17b. The nodes in these latent classes are assigned class-memberships, and then the adjacency matrix is re-ordered. Figure 3.17c shows the re-ordered adjacency matrix with five 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.

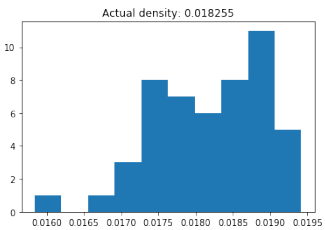
Figure 3.18a shows the range of transitivity of the simulated networks from S.B.M. as 0.09-0.16 (upper and lower bounds of the 95% confidence interval), which is 0.12 less than the actual transitivity of the original network. Thus, S.B.M. has generated L.S.R. that does not regenerate the original network (High-Net) effectively. The L.S.R. is not adequately capturing the transitivity of the original networks. The range of transitivities of the simulated networks obtained from the latent variable model is 0.3-0.42 (upper and lower bounds of the 95% confidence interval), which is 0.02 higher than the original network's actual transitivity, as shown in Figure 3.18b. Similarly, Figure 3.18c shows the range of densities of the simulated networks is 0.016-0.019 (upper and lower bounds of the 95% confidence interval). This is close to the actual edge density of the original network, which is 0.018. The latent variable model was able to generate L.S.R. that could replicate the transitivity (Figure 3.18b) and edge density (Figure 3.18c) of the original network better than S.B.M.s.



(a) Simulate new networks from data-set High-Net - S.B.M. fit to check model goodness of fit on parameter transitivity



(b) Simulate new networks from Latent variable model fit to check model goodness of fit on parameter transitivity



(c) Simulate new networks from Latent variable model fit to check model goodness of fit on parameter edge density

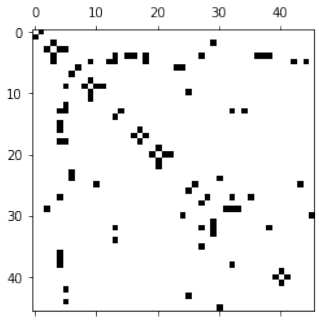
Figure 3.18: Fitting S.B.M. and Latent variable model to High-Net data-set

Conclusions from the posterior predictive check of S.B.M. and latent variable model is

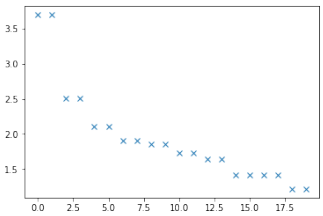
that L.S.R. generated using the latent variable model are more effective than S.B.M. in High-Net.

3.3.10 Grey's anatomy

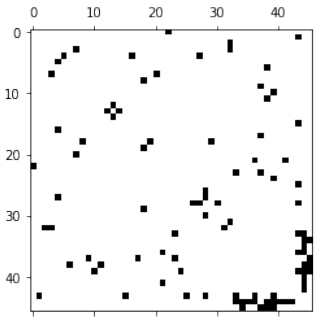
Grey-Net is a dataset of 44 actors in Grey's Anatomy and the "sexual" relationships between them. Figure 3.20 shows the results of fitting S.B.M. to Grey-Net using the procedure outlined in Algorithm 1. Figure 3.19a shows a dense adjacent matrix A but no presence of communities (latent classes) can be detected in the plot. Hence, to choose the latent classes, we examine the singular values of A.



(a) Adjacency matrix of Grey's anatomy data-set Grey-Net



(b) Plot of singular values of Grey's anatomy data-set Grey-Net



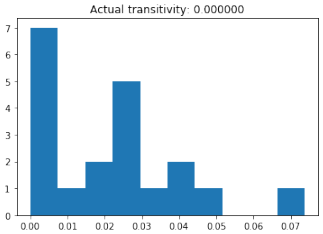
(c) Re-order nodes in adjacency matrix by class memberships

Figure 3.19: Analysis of Grey's anatomy data-set Grey-Net using S.B.M.

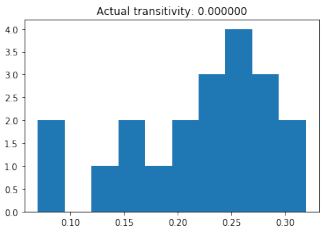
Using eigen-gap heuristic, seven latent classes are observed in Figure 3.19b. The nodes in these latent classes are assigned class-memberships, and then the adjacency matrix is re-ordered. Figure 3.19c shows the re-ordered adjacency matrix with seven latent classes. Once the class-memberships are assigned, the edge probabilities at the block level are calculated. Finally, new networks are simulated from edge probabilities to check model goodness of fit.

Figure 3.20a shows the range of transitivities of the simulated networks from S.B.M. to 0.0-0.05 (upper and lower bounds of the 95% confidence interval), which is 0.05 higher than the actual transitivity of the original network. Thus, S.B.M. has generated L.S.R. that regenerate the original network (Grey-Net) effectively. The L.S.R. can

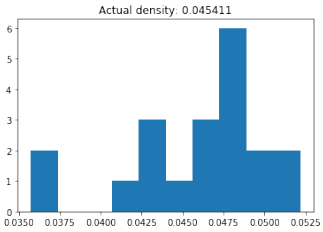
effectively capture the transitivity of the original network. Figure 3.20b shows the range of transitivities of the simulated networks from L.V.M. as 0.125-0.35 (upper and lower bounds of the 95% confidence interval), which is, on average, 0.35 higher than the actual transitivity of the original network. Figure 3.20c shows the range of densities of the simulated networks from the latent variable model to 0.04-0.05 (upper and lower bounds of the 95% confidence interval), which is close to the actual edge density (0.045) of the original network. The latent variable model was not able to generate L.S.R. that could replicate the transitivity (Figure 3.20b) even though the fit to the edge density (Figure 3.20c) of the original network was correct. Conclusions from the posterior predictive check of S.B.M. and L.V.M. models is that L.S.R. generated using S.B.M. are more effective than L.V.M. in Grey-Net.



(a) Simulate new networks of Grey-Net data-set from S.B.M. fit to check model goodness of fit on parameter transitivity



(b) Simulate new networks from Latent variable model fit to check model goodness of fit on parameter transitivity

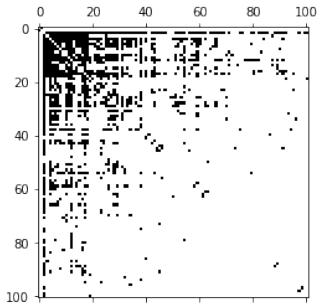


(c) Simulate new networks from Latent variable model fit to check model goodness of fit on parameter edge density

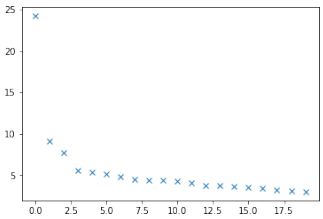
Figure 3.20: Fitting S.B.M. and Latent variable model to Grey-Net data-set

3.3.11 Trade

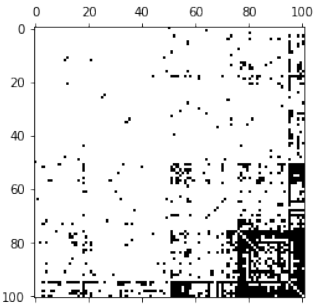
Trade-Net is a dataset of 99 countries and their trading ties. The average clustering coefficient, i.e., transitivity between the members is 0.44. Figure 3.22 shows the results of fitting S.B.M. to Trade-Net using the procedure outlined in Algorithm 1. Figure 3.21a shows a dense adjacent matrix A. Hence, to choose the latent classes, a plot of the singular value of A is created. Using eigen-gap heuristic, three latent classes are observed in Figure 3.21b. The nodes in these latent classes are assigned class-memberships, and then the adjacency matrix is re-ordered.



(a) Adjacency matrix of Trade network data-set Trade-Net



(b) Plot of singular values of Trade network data-set Trade-Net



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

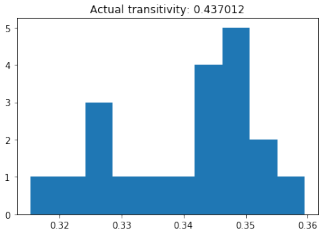
Figure 3.21: Analysis of Trade network data-set Trade-Net using S.B.M.

Figure 3.21c 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.

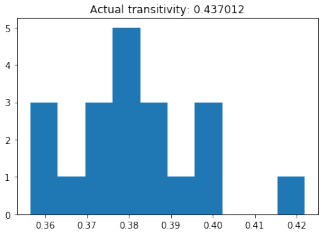
Figure 3.22a shows the range of transitivities of the simulated networks from S.B.M. as 0.31-0.36 (upper and lower bounds of the 95% confidence interval), which is 0.07 lower than the actual transitivity of the original network (0.43). Thus, S.B.M. has generated L.S.R. that does not regenerate the original network (Trade-Net) effectively. The L.S.R. is not adequately capturing the transitivity of the original networks. Figure 3.22b shows the range of transitivities of the simulated networks from the latent variable model to 0.36-0.42 (upper and lower bounds of the 95% confidence interval), which is 0.04 lower than the actual transitivity of the original network (0.43). Figure 3.22c shows the range of densities of the simulated networks from the latent variable model as 0.108-0.118 (upper and lower bounds of the 95% confidence interval), which contains the actual edge density of the original network (0.113).

The latent variable model was able to generate L.S.R. that could replicate the transitivity (Figure 3.22b), and edge density (Figure 3.22c) of the original network better than S.B.M. Conclusions from the posterior predictive check of S.B.M. and Latent variable model models is that L.S.R. generated using Latent variable model are more effective

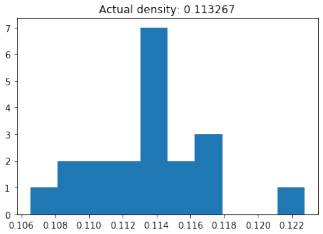
than S.B.M. in Trade-Net.



(a) Simulate new networks of Trade-Net data-set from S.B.M. fit to check model goodness of fit on parameter transitivity



(b) Simulate new networks from Latent variable model fit to check model goodness of fit on parameter transitivity

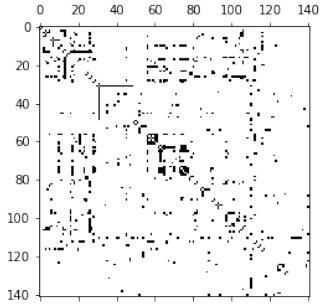


(c) Simulate new networks from Latent variable model fit to check model goodness of fit on parameter edge density

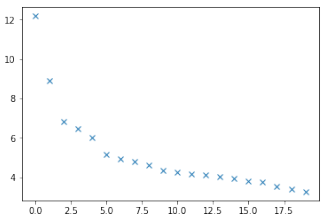
Figure 3.22: Fitting S.B.M. and Latent variable model to Trade-Net data-set

3.3.12 Bill co-sponsorship

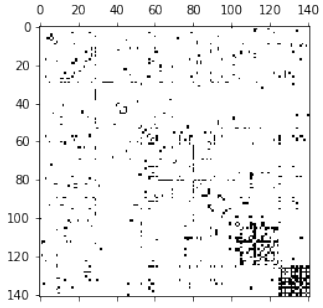
Bill-Net is a dataset of 139 legislators that have co-sponsored legislation with each other. The average clustering coefficient, i.e., transitivity between the members is 0.32. Figure 3.2 shows the results of fitting S.B.M. to Bill-Net using the procedure outlined in Algorithm 1. Figure 3.23a shows a dense adjacent matrix A but to choose the latent classes, the plot of singular values of A is needed.



(a) Adjacency matrix of Bill co-sponsorship data-set Bill-Net



(b) Plot of singular values of Bill co-sponsorship data-set Bill-Net



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

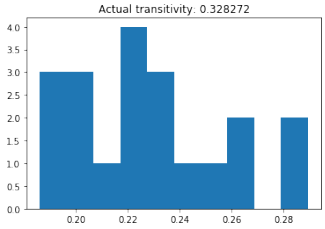
Figure 3.23: Analysis of Bill co-sponsorship data-set Bill-Net using S.B.M.

Using eigen-gap heuristic, four latent classes are observed in Figure 3.23b. The nodes

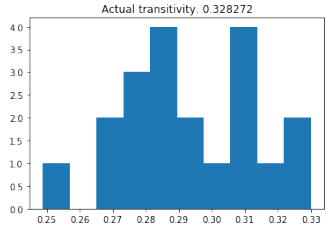
in these latent classes are assigned class-memberships, and then the adjacency matrix is re-ordered. Figure 3.23c shows the re-ordered adjacency matrix with four latent classes. Once the class-memberships are assigned, the edge probabilities at the block level are calculated. Finally, new networks are simulated from edge probabilities to check model goodness of fit.

Figure 3.24a shows the range of transitivities of the simulated networks from S.B.M. as 0.19-0.28 (upper and lower bounds of the 95% confidence interval), which is 0.10 lower than the actual transitivity of the original network (0.32). Figure 3.24b shows the range of transitivities of the simulated networks from the Latent variable model as 0.26-0.33 (upper and lower bounds of the 95% confidence interval), which is captures the actual transitivity of the original network (0.32). Figure 3.24c shows the range of densities of the simulated networks from Latent variable model as 0.04-0.047 (upper and lower bounds of the 95% confidence interval), which captures the actual edge density of the original network (0.043).

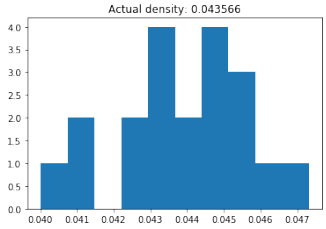
Thus, S.B.M. has generated L.S.R. that does not regenerate the original network (Bill-Net) effectively as a latent variable model. The L.S.R. is not adequately capturing the transitivity of the original networks as a latent variable model.



(a) Simulate new networks of Bill-Net data-set from S.B.M. fit to check model goodness of fit on parameter transitivity



(b) Simulate new networks from Latent variable model fit to check model goodness of fit on parameter transitivity



(c) Simulate new networks from Latent variable model fit to check model goodness of fit on parameter edge density

Figure 3.24: Fitting S.B.M. and Latent variable model to Bill-Net data-set

The latent variable model was able to generate L.S.R. that could replicate the transitivity (Figure 3.24b), and edge density (Figure 3.24c) of the original network better than

S.B.M. Conclusions from the posterior predictive check of S.B.M. and Latent variable model is that L.S.R. generated using Latent variable model are more effective than S.B.M. in Bill-Net.

3.4 Discussion of Results and Summary

Network analysis is a crucial aspect of computational social science. The omnipresent nature of 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. It is further revealed that networks also capture the behaviour of the entities present in them. Using concepts of network science, it is possible to make a statistically valid analysis of these systems and provide insights into their growth.

Networks across different domains saw low edge density and the presence of inequality. Social networks viz., Twt-Net, Gplus-Net, Flickr-Net, Wiki-Net, Blog-Net, Grey-Net and Bill-Net, 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 Model (S.B.M.) and Latent variable model were fit to various application scenarios. Table 3.7 provides a summary of the results of the Stochastic Block Model (S.B.M.) and Latent variable model on datasets. It provides the actual transitivity of the networks (Actual c) and the mean transitivities of the simulated networks from S.B.M. (Mean c_1) and latent variable model (Mean c_2). The values of Mean c_1 and Mean c_2 along with limits of two standard deviations $2SD$ of the mean (upper and lower bounds of the 95% confidence interval) are used to understand which model was comparatively more effective, i.e., within ± 0.05, in fitting the data. In Table 3.7, '*' is used to indicate that the Latent variable model was infeasible for the dataset. '-' is used to indicate that neither S.B.M. nor latent variable model could produce an effective fit on data.

Table 3.7: Summary of results of Stochastic Block Model (S.B.M.) and latent variable model on data-sets

Data-set	**Actual** c	**Mean** c_1 ($\pm 2SD$)	**Mean** c_2 ($\pm 2SD$)	**Effective fit** (± 0.05)
Twt-Net	0.44	0.31 (0.3 - 0.33)	0.4 (0.4-0.404)	Latent variable model
Gplus-Net	0.3	0.243 (0.242-0.245)	*	S.B.M.
Blog-Net	0.08	0.03 (0.032-0.033)	*	S.B.M.
Flickr-Net	0.1	0.07 (0.068-0.071)	*	S.B.M.
Protein-Net	0.09	0.04 (0.04-0.04)	*	S.B.M.
Wiki-Net	0.43	0.37 (0.34-0.39)	*	-
Cora-Net	0.09	0.01 (0.0145-0.0185)	*	-
Cite-Net	0.13	0.05 (0.04-0.06)	*	-
High-Net	0.28	0.13 (0.09-0.16)	0.32 (0.3-0.42)	Latent variable model
Grey-Net	0	0.02 (0.0-0.05)	0.2 (0.125-0.35)	S.B.M.
Trade-Net	0.43	0.33 (0.31-0.36)	0.38 (0.36-0.42)	Latent variable model
Bill-Net	0.32	0.24 (0.19-0.28)	0.31 (0.26-0.33)	Latent variable model

1. For datasets such as Twt-Net, High-Net, Trade-Net and Bill-Net where transitivity is >0.3, the latent variable model was observed as a better fit than S.B.M. as it is more suitable to model assortative mixing, i.e., transitivity.

2. Whereas, for Gplus-Net, Blog-Net, Flickr-Net, Protein-Net and Grey-Net, S.B.M. was more effective as these networks had low transitivity <0.3.

3. However, both models ignore the attributes associated with the networks. Hence, the results on networks with attributes were mixed.

4. None of the models were observed to regenerate networks with transitivity in the acceptable limits of ±0.05 for Wiki-Net, Cora-Net and Cite-Net.

5. The computational complexity of S.B.M. is $\Theta(|V| * |E| * d)$ and latent variable model is $\Theta(|V| * |V| * d)$.

Thus, S.B.M. was observed to be applicable for networks with nodes in a range of 10^3, whereas the Latent variable model was observed to be feasible for networks with a few hundred nodes. Hence, it is necessary to investigate models that can scale to larger networks (10^3+).

Discussions and results in this chapter are published in:

1. Nerurkar, P., Chandane, M. and Bhirud, S. (2019). Empirical analysis of synthetic and real networks. International Journal of Information Technology, pp. 1-13. (Springer)

2. Nerurkar, P., Chandane, M. and Bhirud, S. (2019). Understanding structure and behavior of systems: a network perspective. International Journal of Information Technology, pp. 1-15. (Springer)