Chapter 4

WEIGHTED-AVERAGE AND MEAN ENSEMBLE MODELS

4.1 Introduction

In Chapter 3, Network science was applied to understand the trends in diverse scenarios such as (i) social networking websites (Twitter, Google+), (ii) blogs (Blog.com), (iii) photo-sharing websites (Flickr), (iv) protein-protein interactions (Protein-Net), (v) citation networks (CORA and CiteSeer), (vi) transportation networks (High-Net), (vii) sexual contact network (Grey-Net), (viii) trade network (Trade-Net) and (ix) bill co-sponsorship network (Bill-Net).

Trends identified using network analysis were the (i) rate of information diffusion, (ii) entities' dominance, (iii) the number of social contacts, and (iv) community structure. These trends would not be easily identifiable using relational databases. However, calculating network statistics such as gini index, average path length, and diameter was computationally intensive. Earlier, network analysis required network statistics to build features for training the learning model. However, as network increased in size, this feature engineering step created bottlenecks. Hence, the latent space representation of such networks was used. With L.S.R., it was possible to develop downstream machine learning applications without feature engineering [120].

In Chapter 3, the application of statistical models such as Stochastic Block Model (S.B.M.) and latent variable model for latent space representation revealed certain drawbacks such as (i) limited applicability, (ii) computation cost, (iii) reliance upon expensive and often unstable methods for probabilistic inference [70, 73]. To overcome these drawbacks, a type of latent space representation techniques named network representation learning was proposed. Chapter 2 discussed the taxonomy of L.S.R. techniques present in literature. Three broad categories of L.S.R. techniques for networks exist: Probabilistic models, Statistical models, and Network representation learning models. Statistical models were found to be feasible for datasets with nodes in the range of 10^{1-3}. This was due to the computationally costly stochastic calculations in such models. N.R.L. methods are not computationally costly and are suitable for network analysis of the networks with nodes beyond 10^3.

Contributions in the chapter are summarized as: (i) Highlight Research gaps in state of the art N.R.L. techniques, (ii) Development of a mathematical model for the proposed ensemble model for latent space representation of networks, (iii) Experimental study on the proposed weighted average and mean based ensemble model for latent space representation of networks, (iv)The proposed weighted-average and mean ensemble methods were evaluated on network datasets to understand its performance vis-a-vis other N.R.L. techniques, and (v) Provided criteria to decide an N.R.L. technique most suitable for a network dataset.

Section 4.2 provides a background of N.R.L techniques and states their drawbacks. The motivation for the proposed ensemble approaches is provided in Section 4.2.1. Section 4.3 describes the mathematical model followed by experimental study in Section 4.4 and the chapter is concluded in Section 4.5.

4.2 State of the Art

N.R.L. frameworks consist of an encoder that learns to preserve a similarity measure between nodes in a network. This similarity measure can be (i) adjacency in the network, (ii) neighbourhood overlap between the nodes, (iii) reachability within multiple hops or

(iv) co-occurrence on random walks. Several authors have proposed different versions of similarity. This has consequently led to a proliferation of N.R.L. frameworks in the literature. A majority of these frameworks can be categorized into three groups. Groupings are based on the similarity measure that is captured in the latent representations. There groups are Adjacency preserving methods [25, 26, 27, 31, 32, 85, 86, 87, 88, 89], Multi-hop distance preserving methods and Random walk occurrence preserving methods [90, 91, 92, 93, 94, 95, 96, 97, 98, 99].

The frameworks in the literature preserve one of the three proximity measures defined on the graph. This could lead to vector embedding that is favourable for specific applications but unsuitable for others [49, 50, 51]. There is strong evidence that an ensemble of models [121] can outperform a single model. Two primary approaches for ensembling models are Combine by learning and Combine by consensus [121]. Combined by consensus approach is suitable for unsupervised learning tasks such as L.S.R. In combination by consensus technique, the base models' representations are unified by consensus to obtain the final representations. The three base models are MF (adjacency preserving methods), LINE (multi-hop distance preserving), and DeepWalk (random walk occurrence preserving) methods. The selection of base models is after a literature survey (see Chapter 2) based on their time complexity to ensure scalability to network datasets with nodes beyond 10^3.

4.2.1 Motivation

This chapter describes experiments with an ensemble model for generating effective L.S.R. The aim is to avoid a single model's shortcoming by relying on a combination of multiple frameworks. An extensive survey of N.R.L. techniques did not reveal the use of ensemble models for representation learning. As a single model is unlikely to capture the entire underlying structure of the data to achieve optimal representations, integrating multiple models may improve the accuracy of representation learning significantly. This gap serves as the principal motivation for this study. The current inquiry focuses on an ensemble model in which representation learning of base models is combined by consensus using a weighted average.

4.3 Ensemble Network Representation Learning Framework

Three different network representation learning frameworks are used to obtain the initial node level embeddings.

4.3.1 Calculate node embeddings through adjacency preserving similarity (M.F.)

The adjacency matrix A_{m*n} is factored into low rank representations B_1, B_2 such that $A \sim B_1 B_2$ [32]. A non-negative matrix factorization approach (Eq. 4.1) ensures topological information is captured in vector embedding.

$$B_1, B_2 = ||A - B_1 B_2||_F^2 + \alpha_1 ||B_1||_F^2 + \alpha_2 ||B_2||_F^2 \tag{4.1}$$

where,

1. $||.||_F^2$ - Frobenius norm
2. $\alpha_1, \alpha_2 \geq 0$ - Weight parameters

4.3.2 Calculate node embeddings through multi-hop distance preserving similarity (LINE)

The objective function is to obtain node embeddings that capture second-order proximity. The L2-norm of the loss function, which needs to be minimized is given in Eq. 4.2.

$$min||S - U^s.U^{t^T}||_F^2 \tag{4.2}$$

$$S = M_g^{-1}.M_t \tag{4.3}$$

$$M_g = I - \beta.A \tag{4.4}$$

$$M_t = \beta.A \tag{4.5}$$

where,

1. S = high order proximity matrix in Eq. 4.2, Eq. 4.3

2. U^s, U^t = source and target embedding vectors
3. M_g, M_t = polynomial of adjacency matrix A in Eq. 4.4, Eq. 4.5
4. I = identity matrix, β = decay parameter of Katz index

4.3.3 Calculate node embeddings through random walk occurrence preserving similarity (DeepWalk)

Random walks on the graph are performed and the results are provided to skip-gram neural network to obtain the node embeddings. The objective of the skip-gram model is as shown in Eq. 4.6:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c;j\neq 0} logp(w_{t+j}|w_t) \tag{4.6}$$

The basic skip-gram formulation defines $p(w_{t+j}|w_t)$ (probabilities of other vertices in the social network occurring within the context window of the current vertex) using the softmax function as shown in Eq. 4.7:

$$p(wout|win) = \frac{exp(v_{wout}^T * v_{win})}{\sum_{w=1}^{W} exp(v_{wout}^T * v_{win})} \tag{4.7}$$

where, v_{win} and v_{wout} are the "input" and "output" vector representations of w and W is the number of words in the vocabulary.

4.3.4 Ensemble network representation learning framework

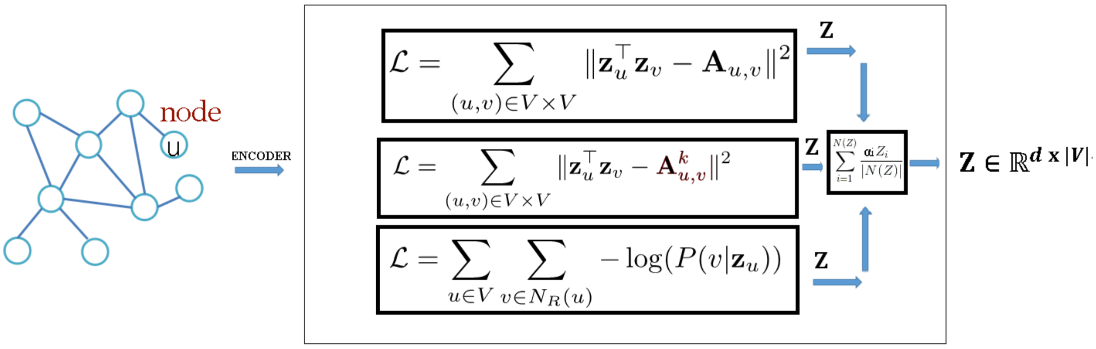
For obtaining a robust representation learning framework, an ensemble mechanism that takes the weighted mean of the individual models (MF, LINE and DeepWalk) is used. The final node level embeddings are obtained as given in Eq. 4.8. Z is the final L.S.R. of the network obtained by using combine by consensus technique, i.e., the representations of the base models (MF, LINE and DeepWalk) Z_i are unified by consensus to obtain the final representations. Algorithm 4 is a minimized algorithm of the proposed mechanism, and an abstract block diagram for the same is given in Fig. 4.1. Depending on the values of α, two consensus models are possible. If α is equal for all base models (MF, LINE and DeepWalk), then the ensemble model is a mean ensemble, and if α is different for each model, then it is a weighted-average ensemble.

$$Z = \sum_{i=1}^{N(Z)} \frac{\alpha_i Z_i}{|N(Z)|} \tag{4.8}$$

where,

1. $N(Z)$ = Number of base models.
2. α = weights of base models. Obtained for weighted-average ensemble through tuning.

```
Algorithm 4: Ensemble Network representation learning framework
  Input: G = V, E
  Output: Z = R^{$d*|V|$}
  Calculate Cost function for adjacency preserving similarity as:
  $L = \sum_{(u,v)\in(V*V)} ||z_u^T z_v - A_{u,v}||^2$
  Calculate Cost function for multi-hop distance preserving similarity as:
  $L = \sum_{(u,v)\in(V*V)} ||z_u^T z_v - S_{u,v}||^2$
  Calculate Cost function for random walk occurance preserving similarity as:
  $L = \sum_{(u\in V)} \sum_{(v\in N_r(u))} -log(P(v|z_u))$
  Optimize parameters of $Z$ using stochastic gradient descent
  Combine individual learned representations to final learned representation $Z$:
  $\sum_{i=1}^{N(Z)} \frac{\alpha_i Z_i}{|N(Z)|}$
```



node
u
ENCODER
$\mathcal{L} = \sum_{(u,v)\in V\times V} \|\mathbf{z}_u^\top \mathbf{z}_v - \mathbf{A}_{u,v}\|^2$
$\mathcal{L} = \sum_{(u,v)\in V\times V} \|\mathbf{z}_u^\top \mathbf{z}_v - \mathbf{A}_{u,v}^k\|^2$
$\mathcal{L} = \sum_{u\in V} \sum_{v\in N_R(u)} -\log(P(v|\mathbf{z}_u))$
Z
Z
Z
$\sum_{i=1}^{N(Z)} \frac{\alpha_i Z_i}{|N(Z)|}$
$\mathbf{Z} \in \mathbb{R}^{d \times |V|}$

Figure 4.1: Ensemble network representation learning framework

4.4 Experimental Study

The proposed weighted average and mean ensemble methods are evaluated on network datasets to understand its performance vis-a-vis other N.R.L. techniques. The weighted-average and mean ensemble methods convert network data of Blog-Net, Cora-Net, Cite-Net, Flickr-Net, Protein-Net and Wiki-Net from their original dimensions into vector space of dimension = 32. The quality of the vertex embeddings is measured using distance-based statistics obtained from clustering the representations. The performance metrics used are separation index, widest within-cluster gap, average silhouette width, the average distance between clusters and Dunn index. Twenty iterations are performed on vertex embeddings for calculating each performance metric. The mean values along with limits of two standard deviations $2SD$ of the mean (upper and lower bounds of the 95% confidence interval) are given for quantitative comparison. Configuration of the system on which experimental study was conducted is Intel(R)Core(T.M.) i5-6402P CPU@2.8GHz with four cores and 8GB DDR3 RAM. ProNet-Core C++ framework was used for implementation network embedding techniques (M.F., LINE and DeepWalk).

4.4.1 Performance metrics

4.4.1.1 Separation index

It is computed based on the distances for every point to the closest point not in the same cluster. A lower value indicates that clusters are well separated.

4.4.1.2 Widest within-cluster gap

Largest link in the within-cluster minimum spanning tree is computed. A more considerable value indicates useful clustering.

4.4.1.3 Average silhouette width

It is a measure of how similar an object is to its cluster compared to other clusters. The silhouette ranges from -1 to +1, where, a high value indicates that the object is well matched to its cluster and poorly matched to neighbouring clusters.

4.4.1.4 Average distance between clusters

For two clusters r, s average distance between clusters should be significant to indicate well-separated clusters.

$$L(r,s) = \frac{1}{n_r * n_s} \sum_{i=1}^{n_r} \sum_{i=1}^{n_s} D(x_{ri}, x_{sj}) \tag{4.9}$$

4.4.1.5 Dunn index

A lower Dunn index indicates better clustering. $\delta(C_i, C_j)$ is inter-cluster distance metric between clusters C_i and C_j, Δ_k is the maximum within cluster variance.

$$DI_m = \frac{\min_{1 \le i < j \le m} \delta(C_i, C_j)}{\max_{1 \le k \le m} \Delta_k} \tag{4.10}$$

4.4.2 Data-sets

Network datasets with $|V| > 10^3$ were chosen to evaluate the ensemble techniques.

Table 4.1: 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
c	0.08	0.1	0.09	0.43	0.14	0.13

4.4.3 Results

DeepWalk [90], LINE [25] and MF [28] are the baseline techniques used for comparison. Two ensemble methods: mean ensemble and weighted-average ensemble (WeightedAvg) are evaluated to understand the latent space representations obtained by them. Table 4.2 gives the weights of the individual models of the weighted-average ensemble. α_1, α_2, and α_3 are the weights of adjacency based similarity, multi-hop based similarity, and random walk-based similarity.

Table 4.2: Weights of the base models of WeightedAvg

Data-set	**Mean** α_1 $(\pm 2SD)$	**Mean** α_2 $(\pm 2SD)$	**Mean** α_3 $(\pm 2SD)$
Blog-Net	0.34154 (±0.0068)	0.2963 (±0.0059)	0.36216 (±0.0024)
Flickr-Net	0.57146 (±0.0021)	0.22486 (±0.0082)	0.20368 (±0.0043)
Protein-Net	0.53127 (±0.0038)	0.24513 (±0.0029)	0.22360 (±0.0046)
Wiki-Net	0.26481 (±0.0029)	0.52432 (±0.0046)	0.21087 (±0.0018)
Cora-Net	0.19219 (±0.0037)	0.32149 (±0.0028)	0.48632 (±0.0043)
Cite-Net	0.27416 (±0.0018)	0.18256 (±0.0027)	0.54328 (±0.0034)

Table 4.2 indicates that for Blog-Net, the WeightedAvg ensemble model has combined the learnings of the base models in equal proportions. In Flickr-Net and Protein-Net, the weights of adjacency based similarity, multi-hop based similarity, and random walk based similarity were $\sim 2 : 1 : 1$. In Wiki-Net the weights of adjacency based similarity, multi-hop based similarity, and random walk based similarity were $\sim 1 : 2 : 1$. Finally, for Cora-Net and Cite-Net, the weights of adjacency based similarity, multi-hop based similarity and random walk based similarity were $\sim 1 : 1 : 2$.

Using eigen-gap heuristic, graph of singular values of revealed that Blog-Net had five clusters (Figure 3.5b), Flickr-Net had three clusters (Figure 3.7b), Protein-Net had two clusters (Figure 3.9b), Wiki-Net had three clusters (Figure 3.11b), Cora-Net had three clusters (Figure 3.13b) and Cite-Net had five clusters (Figure 3.15b).

Table 4.3 gives the results obtained on the parameter separation index. Blog-Net, Flickr-Net, and Wiki-Net have diam(G) in the range of $2-4$. This indicates that nodes are located close to each other. Whereas, for Protein-net, Cora-Net, and Cite-Net, the diam(G) is $6-8$ indicating the comparatively longer distance between nodes. Thus, the separation index of Blog-Net, Flickr-Net, and Wiki-Net will have a higher range then Protein-net and Cora-Net. As Cite-Net has a larger number of clusters, its separation index shall have a higher range. The L.S.R. obtained using mean ensemble method were higher (+) or lower (-) on separation index compared to the closest technique as mentioned: Wiki-Net (-23.5%), Cora-Net (-9%), Cite-Net (-6%), Blog-Net (-15%), Flickr-Net (-17%) and Protein-Net (-52%). The L.S.R. obtained using weighted-average ensemble method were higher (+) or lower (-) on separation index compared to the closest technique as mentioned: Wiki-Net (-4%), Cora-Net (4%), Cite-Net (5%), Blog-Net (-5%), Flickr-Net (-28%) and Protein-Net (-38%).

Table 4.3: Comparison of WeightedAvg, Mean with DeepWalk, LINE, MF on separation index

Data-set	**DeepWalk**	**LINE**	**MF**	**WeightedAvg**	**Mean**
Blog-Net	**3.89** (±0.16)	4.88 (±0.25)	5.23 (±0.46)	4.09 (±0.17)	4.38 (±0.07)
Flickr-Net	2.57 (±0.24)	**2.25** (±0.34)	3.08 (±0.19)	2.91 (±0.21)	2.71 (±0.04)
Protein-Net	1.85 (±0.27)	**1.36** (±0.12)	1.82 (±0.18)	1.73 (±0.07)	2.02 (±0.06)
Wiki-Net	2.85 (±0.31)	3.27 (±0.29)	3.19 (±0.28)	3.02 (±0.12)	**2.65** (±0.14)
Cora-Net	2.21 (±0.16)	2.59 (±0.31)	2.82 (±0.19)	**2.06** (±0.13)	2.43 (±0.04)
Cite-Net	4.23 (±0.58)	4.58 (±0.41)	5.07 (±0.49)	**4.02** (±0.21)	4.64 (±0.26)

Table 4.4 gives the results obtained on the parameter widest within-cluster gap. The L.S.R. obtained using mean ensemble method were higher (+) or lower (-) on widest within-cluster gap compared to the closest technique as mentioned: Flickr-Net (-2%), Wiki-Net (0.2%), Cite-Net (22%), Blog-Net (-32%), Protein-Net (-12%) and Cora-Net (-31%). The L.S.R. obtained using weighted-average ensemble method were higher (+) or lower (-) on widest within-cluster gap compared to the closest technique as mentioned: Flickr-Net (5%), Wiki-Net (-5%), Cite-Net (11%), Blog-Net (-21%), Protein-Net (-5%) and Cora-Net (-22%).

Table 4.4: Comparison of WeightedAvg, Mean with DeepWalk, LINE, MF on widest within-cluster gap

Data-set	**DeepWalk**	**LINE**	**MF**	**WeightedAvg**	**Mean**
Blog-Net	2.68 (±0.14)	1.85 (±0.16)	**3.23** (±0.34)	2.73 (±0.21)	2.53 (±0.11)
Flickr-Net	5.62 (±0.31)	6.20 (±0.56)	5.87 (±0.18)	**6.54** (±0.59)	6.08 (±0.21)
Protein-Net	**8.92** (±0.76)	7.65 (±0.68)	8.82 (±0.46)	8.39 (±0.37)	7.89 (±0.29)
Wiki-Net	6.45 (±0.48)	5.68 (±0.51)	6.48 (±0.49)	6.17 (±0.37)	**6.53** (±0.57)
Cora-Net	5.93 (±0.46)	6.59 (±0.61)	**8.22** (±0.78)	6.85 (±0.32)	6.36 (±0.44)
Cite-Net	2.09 (±0.18)	2.59 (±0.31)	1.59 (±0.17)	2.88 (±0.27)	**3.02** (±0.3)

Table 4.5 gives the results obtained on the parameter average silhouette width. The

L.S.R. obtained using mean ensemble method were higher (+) or lower (-) on average silhouette width compared to the closest technique as mentioned: Blog-Net (9%), Cora-Net (-2%), Cite-Net (-5%), Flickr-Net (-6%), Protein-Net (-3%) and Wiki-Net (-11%). The L.S.R. obtained using weighted-average ensemble method were higher (+) or lower (-) on average silhouette width compared to the closest technique as mentioned: Blog-Net (5%), Cora-Net (2%), Cite-Net (3%), Flickr-Net (-12%), Protein-Net (-1%) and Wiki-Net (-16%).

Table 4.5: Comparison of WeightedAvg, Mean with DeepWalk, LINE, MF on average silhouette width

Data-set	**DeepWalk**	**LINE**	**MF**	**WeightedAvg**	**Mean**
Blog-Net	0.03 (±0.0011)	-0.13 (±0.0016)	0.19 (±0.0046)	0.24 (±0.0027)	**0.28** (±0.0019)
Flickr-Net	0.19 (±0.0022)	**0.31** (±0.0037)	0.26 (±0.0028)	0.19 (±0.0016)	0.25 (±0.002)
Protein-Net	0.46 (±0.0024)	**0.47** (±0.0031)	**0.47** (±0.0036)	0.46 (±0.0021)	0.44 (±0.0019)
Wiki-Net	0.37 (±0.0042)	0.33 (±0.0026)	**0.42** (±0.0034)	0.26 (±0.0016)	0.31 (±0.0021)
Cora-Net	0.46 (±0.0025)	0.43 (±0.0037)	-0.21 (±0.0019)	**0.48** (±0.0036)	0.44 (±0.001)
Cite-Net	0.13 (±0.0011)	0.21 (±0.0016)	-0.06 (±0.009)	**0.24** (±0.0017)	0.16 (±0.0015)

Table 4.6 gives the results obtained on the parameter average distance between clusters. The L.S.R. obtained using mean ensemble method were higher (+) or lower (-) on the average distance between clusters compared to the closest technique as mentioned: Flickr-Net (8%), Wiki-Net (6%), Cora-Net (14%), Cite-Net (-8%), Blog-Net (-43%) and Protein-Net (-15%). The L.S.R. obtained using weighted-average ensemble method were higher (+) or lower (-) on the average distance between clusters compared to the closest technique as mentioned: Flickr-Net (-5%), Wiki-Net (10%), Cora-Net (8%), Cite-Net (4%), Blog-Net (-52%) and Protein-Net (-11%).

Table 4.6: Comparison of WeightedAvg, Mean with DeepWalk, LINE, MF on average distance between clusters

Data-set	DeepWalk	LINE	MF	WeightedAvg	Mean
Blog-Net	2.08 (±0.24)	0.49 (±0.089)	**1.83** (±0.15)	0.93 (±0.13)	1.08 (±0.09)
Flickr-Net	2.39 (±0.27)	2.82 (±0.54)	3.08 (±0.21)	2.96 (±0.14)	**3.22** (±0.16)
Protein-Net	4.97 (±0.38)	**6.82** (±0.59)	5.82 (±0.64)	6.28 (±0.57)	5.91 (±0.37)
Wiki-Net	2.85 (±0.34)	3.74 (±0.29)	3.18 (±0.42)	**4.12** (±0.36)	3.94 (±0.41)
Cora-Net	2.94 (±0.27)	3.07 (±0.25)	1.85 (±0.11)	3.25 (±0.18)	**3.56** (±0.31)
Cite-Net	0.82 (±0.18)	1.28 (±0.25)	0.79 (±0.037)	**1.32** (±0.14)	1.19 (±0.09)

Table 4.7 gives the results obtained on the parameter Dunn index. The L.S.R. obtained using mean ensemble method were higher (+) or lower (-) on Dunn index compared to the closest technique as mentioned: Protein-Net (16%), Cora-Net (-12%), Blog-Net (18%), Flickr-Net (36%), Wiki-Net (2%) and Cite-Net (30%). The L.S.R. obtained using weighted-average ensemble method were higher (+) or lower (-) on Dunn index compared to the closest technique as mentioned: Protein-Net (-8%), Cora-Net (-3%), Blog-Net (12%), Flickr-Net (14%), Wiki-Net (7%) and Cite-Net (11%).

Table 4.7: Comparison of WeightedAvg, Mean with DeepWalk, LINE, MF on Dunn index

Data-set	DeepWalk	LINE	MF	WeightedAvg	Mean
Blog-Net	0.26 (±0.006)	**0.11** (±0.0042)	0.32 (±0.0027)	0.23 (±0.0034)	0.29 (±0.0013)
Flickr-Net	**0.21** (±0.0028)	0.41 (±0.0031)	0.31 (±0.0046)	0.35 (±0.0038)	0.57 (±0.0028)
Protein-Net	0.28 (±0.0031)	0.21 (±0.0037)	0.42 (±0.0051)	**0.13** (±0.0036)	0.37 (±0.0021)
Wiki-Net	0.39 (±0.0019)	**0.27** (±0.0024)	0.47 (±0.0031)	0.34 (±0.003)	0.29 (±0.0018)
Cora-Net	0.46 (±0.0024)	0.61 (±0.0028)	0.52 (±0.0031)	0.43 (±0.0019)	**0.34** (±0.001)
Cite-Net	**0.13** (±0.0054)	0.26 (±0.0037)	0.34 (±0.0061)	0.24 (±0.0059)	0.43 (±0.0037)

4.5 Discussion of Results and Summary

Network representation learning has emerged as a preferred method of transforming social network data to make it more amenable to analysis. Multiple N.R.L. techniques are available in the literature; however, individual techniques preserve one of the many proximity measures. Hence, in this inquiry, N.R.L. was approached as a multi-objective optimization process, i.e., to preserve the three main proximity measures in a network. The resultant ensemble model was applied on network datasets with nodes $> 10^3$.

1. Blog-Net had Gini index $G = 0.39$ and clustering coefficient $c = 0.08$. This indicated the presence of more hubs (popular nodes) in the network. This conclusion is consistent with the observation of five clusters in the network, of which three are well-separated (Figure 3.5b). Each hub has many followers $d = 66.3$, and the value of c indicates that followers did not form relationships with each other. When represented as vector embeddings, such networks would lead to popular nodes and their followers occupying disjoint regions in the latent space. In such a scenario, any proximity preserving N.R.L. technique would be efficient. The WeightAvg model assigns $\sim 1 : 1 : 1$ weights to its base models and its performance resembles that of the mean ensemble model for this scenario.

2. Flickr-Net and Protein-Net have a high Gini index $G = 0.63 - 0.67$ and low clustering co-efficient $0.09{-}0.1$. This indicates the presence of dominant nodes in the network. The plot of singular values for both networks (Figure 3.7b and Figure 3.9b) also indicates well-separated regions in the network. In such a scenario, the adjacency preserving technique may achieve effective results. WeightedAvg ensemble was observed to assign weights of adjacency based similarity, multi-hop based similarity, and random walk based similarity in proportion of $\sim 2 : 1 : 1$.

3. Wiki-Net data-set has high transitivity $c = 0.43$, it also has presence of dominant nodes $G = 0.62$. The graph of singular values (Figure 3.11b) indicates three well-separated clusters in the network. As $c = 0.43$ is higher, the second-order proximity between nodes (neighbourhood overlap) could have a higher impact on the L.S.R. Therefore, in Wiki-Net, the weights of adjacency based similarity, multi-hop based similarity and random walk based similarity were $\sim 1 : 2 : 1$.

4. Cora-Net and Cite-Net have low diameter and average path length $\mathrm{diam}(G) = 6{-}8$ and $\bar{P} = 1.74 - 1.81$. There are few popular nodes in the network $G =\sim 0.42$. A random walk on the nodes would traverse more nodes and effectively represent them in the latent space due to low diameter. Hence, for Cora-Net and Cite-Net, the weights of adjacency based similarity, multi-hop based similarity and random

walk based similarity were $\sim 1 : 1 : 2$.

5. In summary, from the experiments, it could be concluded that all three base models had near equal weight in networks where, the Gini index and transitivity were low. In networks that had a high Gini index, the adjacency preserving model had higher weightage in the ensemble. The networks with high transitivity gave more weightage to the Multi-hop distance preserving model, and networks with high diameter gave more weightage to Random walk occurrence preserving models.

Table 4.8 gives the results of transitivity on the data-sets. The L.S.R. obtained using mean ensemble method were higher (+) or lower (-) on transitivity compared to the closest technique as mentioned: Protein-Net (3%), Cora-Net (5%), Blog-Net (-5%), Flickr-Net (4%), Wiki-Net (2%) and Cite-Net (1%). The L.S.R. obtained using weighted-average ensemble method were higher (+) or lower (-) on transitivity compared to the closest technique as mentioned: Protein-Net (-1%), Cora-Net (-3%), Blog-Net (2%), Flickr-Net (2%), Wiki-Net (-1%) and Cite-Net (-4%).

Table 4.8: Comparison of WeightedAvg, Mean with DeepWalk, LINE, MF on transitivity

Data-set	**Actual** c	**DeepWalk**	**LINE**	**MF**	**WeightedAvg**	**Mean**
Blog-Net	0.08	0.11 (±0.0042)	0.12 (±0.0068)	0.09 (±0.0037)	0.11 (±0.0034)	0.14 (±0.0028)
Flickr-Net	0.1	0.14 (±0.0061)	0.13 (±0.0054)	0.06 (±0.0038)	0.09 (±0.0043)	0.1 (±0.0033)
Protein-Net	0.09	0.12 (±0.0063)	0.06 (±0.0057)	0.08 (±0.0061)	0.07 (±0.0031)	0.12 (±0.0028)
Wiki-Net	0.43	0.27 (±0.0056)	0.46 (±0.0051)	0.34 (±0.0049)	0.39 (±0.0029)	0.38 (±0.0031)
Cora-Net	0.09	0.12 (±0.0037)	0.19 (±0.0053)	0.16 (±0.0027)	0.15 (±0.003)	0.14 (±0.0032)
Cite-Net	0.13	0.09 (±0.0012)	0.16 (±0.0024)	0.14 (±0.0006)	0.08 (±0.0012)	0.13 (±0.0004)

A key drawback observed in the proposed ensemble techniques was the computation complexity. Even though the computation complexity was significantly better than statistical models such as Stochastic Block Model (S.B.M.) and latent variable model. The proposed ensemble techniques had a computation complexity of $O(|E|d) + O(|V|d) + O(|E|d)$. This was the summation of the computation complexities of the base models. Constant time was required for performing the combination by consensus operation, i.e., weighted-average or mean.

The proposed weighted-average or mean ensemble N.R.L. techniques required that all nodes in the network be present during the encoder's training. Hence, these approaches were inherently transductive and would not generalize to unseen nodes (non-inductive). The proposed methods discarded attribute data associated with the network nodes. Hence, "deep encoder" based techniques could be investigated for situations where, all network nodes are not available at the time of training, and also network nodes are associated with attribute data.

Discussions and results in this chapter are communicated in:

1. Nerurkar, P., Chandane, M. and Bhirud, S. (2019). Ensemble methods for Latent Space Representations. International Journal of Networks. (Inderscience)