2.2 Deep Graph Representation

Let $G = (V, E)$ be a graph with vertex set $V = \{v_1, \ldots, v_n\}$ representing all the nodes. Two vertices have a similarity s_{ij} between the two nodes, and the edge is weighted by s_{ij}. To obtain new representation of normalized similarity matrix, we use the same protocol as in [2]. We like give a graph G with its similarity matrix S, then we send normalized S as n nodes to Autoencoder. Generally, the optimization target of Autoencoder is to minimize the reconstruction error for the output $h_{W,b}(x) \approx x^{(i)}$ can approximate the input training set $D^{-1}S$. We define ρ as a sparsity parameter and s_2 as the number of the hidden neurons. We use β controls the weight of the sparsity penalty term. The overall cost function of Sparse Auto-Encoder can be defining as:

$$J_{sparse}(W,b) = J(W,b) + \beta \sum_{j=1}^{s_2} KL(\rho||\hat{\rho}_j) \quad (3)$$

Where

$$\hat{\rho}_j = \frac{1}{m}\sum_{i=1}^{m}[a_j^{(2)}(x^{(i)})] \quad (4)$$

$$KL(\rho||\hat{\rho}_j) = \rho \log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \quad (5)$$

Eq. (4) is average activation of hidden unit j and we need to enforce the constraint $\hat{\rho}_j = \rho$ and Eq. (5) is the Kullback-Leibler divergence [7]. We consider stack Sparse Auto-Encoders layer by layer to form a whole deep neural network. And then we apply K-means to compute clusters of users.

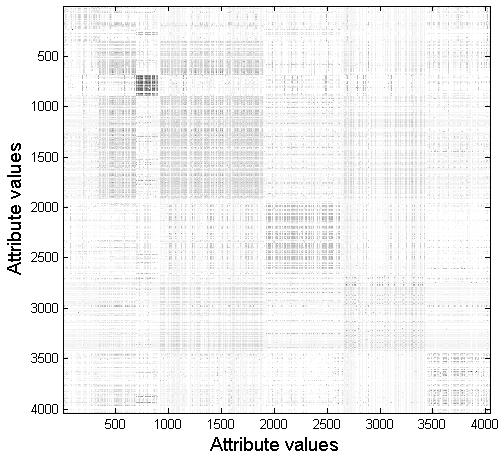

Figure 1. Calculated nodes similarity degree matrix with attribute relevance.

3. RESULTS AND DISCUSSION

The data which are used to clustering social network communities come from Stanford Large Network Dataset Collection [8]. It collected from Facebook and there are 4039 nodes, 88234 edges consist in this dataset. According to their description, the user profiles associated with users can amount to 175 which include birthday, education, languages work, etc. That is to say, each user can be encoded to 175 attributes as representation. Moreover, they also identified 193 known communities which mean our standard target clusters is 193. After running content relevance calculation, we have obtained the similarity matrix based on content relevance. In Fig.1 the calculated new nodes similarity is shown in a gray picture. NMI is an information-theoretic measure that can measure the degree of matching between the clusters discovered by proposed algorithms. We adopted NMI as the evaluation metric in our experiment. We also obtained performance of some clustering methods include FCAN [5], iTopicModel [3] and original Spectral Clustering [2] with the Facebook dataset. The experimental results on all methods are shown in Table 1. In addition, iTopicModel and FCAN also take consideration into both link information and content information. We can see that proposed approach performed better than iTopicModel and SGC and very close to FCAN in terms of NMI. In this regard, the proposed method just use content information and achieve a satisfactory performance.

Method	**DRAG**	FCAN	iTopicModel	SGC
NMI	**0.483**	0.49	0.34	0.39

Table 1. Comparison of NMI scores for Facebook

4. CONCLUSIONS

A graph based approach for social network community clustering is proposed. In summary, we introduce attributes relevance for content information to learn better nodes similarity and take Stacked Autoencoder to transform the calculated graph similarity matrix to the output graph embedding. Based on attributes measure, we can uncover some clues that mean vertices belong to the same cluster. Deep learning also subtly replace the step of find k largest eigenvalues of the normalized graph similarity matrix in spectral clustering. Experimental results on ego Facebook dataset demonstrate that DARG achieves a good performance under readily accessible content information.

5. REFERENCES

[1] M. Girvan and M. E. Newman. 2002. Community structure in social and biological networks. *Proc. Nat. Acad. Sci*., vol. 99, no. 12, pp. 7821–7826.

[2] U. Luxburg. 2007. A tutorial on spectral clustering. *Statist. Comput*, vol. 17, no. 4, pp. 395–426.

[3] Y. Sun, J. Han, J. Gao, and Y. Yu. 2009. iTopicModel: Information network integrated topic modeling. In *Proc. 9th IEEE Int. Conf. Data Mining*, pp. 493–502.

[4] Tian, F., Gao, B., Cui, Q., Chen, E. and Liu, T.Y. 2014. Learning Deep Representations for Graph Clustering. In *AAAI,* pp.1293-1299.

[5] Hu, L. and Chan, K.C., 2016. Fuzzy Clustering in a Complex Network Based on Content Relevance and Link Structures. *IEEE Transactions on Fuzzy Systems*, 24(2), pp.456-470.

[6] K. C. C. Chan, A. K. C. Wong, and D. K. Y. Chiu. 1994. Learning sequential patterns for probabilistic inductive prediction. *IEEE Trans. Syst., Man, Cybern*, vol. 24, pp.1532 -1547.

[7] Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science,* 313(5786):504–507.

[8] J. McAuley and J. Leskovec. 2012. Learning to Discover Social Circles in Ego Networks. *In Advances in neural information processing systems* (pp. 539-547).