Leveraging T-Distributed Stochastic Neighbor Embedding and K-Means For Effective Community Detection In Social Networks

Kurapati Sravanthi
University College of Engineering
Kakatiya University-Warangal, Telangana, India.

ABSTRACT

Community detection is a pivotal task in social network analysis, as it reveals the underlying structure and grouping of individuals within the network. Traditional methods often struggle with the high-dimensional nature of social network data, leading to suboptimal results. This paper introduces a novel approach that combines t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction with K-Means clustering for effective community detection. t-SNE excels at preserving local similarities while reducing the dimensionality of the data, making it suitable for handling the complex structures inherent in social networks. By applying K-Means clustering to the low-dimensional representation generated by t-SNE, we achieve accurate and meaningful community detection.

The t-SNE technique is fully formulated mathematically and shown to be able to preserve local structures in a lower-dimensional space. K-Means is then used to cluster this reduced representation, which successfully identifies communities throughout the network. Extensive studies on real-world social network datasets, such as the Zachary Karate Club Network, American College Football Network, Bottlenose Dolphins Network, US Political Books Network and Santa Fe Scientists Collaboration Network verify the effectiveness of the proposed approach. The outcomes demonstrate how much better our approach is in clustering, as measured by metrics like Normalized Mutual Information (NMI).

The results show that combining t-SNE with K-Means improves community detection accuracy while also offering insightful information about the underlying social dynamics. This method opens the door for more study and practical applications in the subject by providing a strong framework for evaluating intricate social networks.

Keywords: Community Detection, Social Networks, t-Distributed Stochastic Neighbor Embedding, t-SNE, K-Means Clustering, Dimensionality Reduction.

1. INTRODUCTION

Community detection in social networks is a fundamental problem that seeks to identify groups of nodes (or individuals) that are more densely connected internally than with the rest of the network. This task is crucial for various applications, such as understanding social dynamics, improving recommendation systems, and enhancing targeted marketing strategies. However, social networks typically exhibit high-dimensional, complex, and noisy data, posing significant challenges for traditional community detection methods.

Social networks can be represented as graphs where nodes correspond to individuals and edges represent relationships or interactions between them. Figure 1 depicts a social network graph G = (V, E) with two community structures [1] where V = {1,2,3,4,5,6,7,a,b,c,d,e,f} and E = {(1,2), (1,5), (1,7), (2,5), (2,7), (3,4), (3,5), (3,6), (4,6), (5,6), (5,7), (5,c), (6,c), (7,c), (7,d), (a,b), (a,d), (a,f), (b,c), (b,d), (b,e), (c,d), (d,f), (e,f)}such that |V| = 13 and |E| = 24.The discovery of community structure is crucial to comprehending and taking advantage of the structure of complex networks [18]. It has numerous applications in a variety of domains, including image segmentation, online social networking, molecular interaction networks, and circuit layout issues. Once identified, the communities show the members' relationships, associations, and behavioural patterns. As an illustration, the research community may identify domain-specific Special Interest Groups (SIGs) in social networks [2], which will then be utilized to facilitate productive member interactions for research purposes [17].Finding a community of websites that link to two or more websites inside the same community will yield a set of websites on comparable topics. Using this information, search engines and portals can then focus their search by looking for thematically related subsets of websites [3].

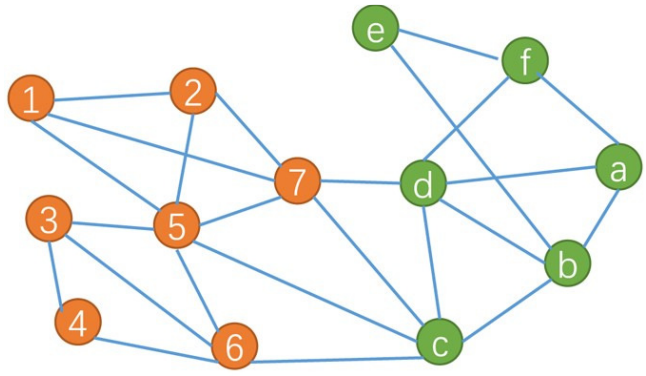


Fig. 1.A schematic diagram showing a social network with two communities

The dimensionality of social network data arises from various attributes such as user profiles, interaction histories, and content shared. Directly applying clustering algorithms to such high-dimensional data often results in poor performance due to the "curse of dimensionality," where the distance metrics become less informative as dimensions increase.

Dimensionality reduction techniques aim to transform high-dimensional data into a lower-dimensional space while preserving important structural properties. One such powerful technique is t-Distributed Stochastic Neighbor Embedding (t-SNE), which has gained popularity for its ability to maintain local similarities and visualize complex data structures effectively.

t-SNE is a non-linear dimensionality reduction method that converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between these joint probabilities in the high-dimensional and low-dimensional spaces. This technique is particularly effective for visualizing high-dimensional data, as it preserves the local structure of the data, making it ideal for subsequent clustering.

K-Means clustering is a widely used partitioning method that divides the dataset into K clusters by minimizing the within-cluster variance. When combined with t-SNE, K-Means can be applied to the lower-dimensional representation of the data, effectively identifying communities within the network.

The outline of the paper is as follows. In Section 2, we outline, related work concerning the literature on community detection in social networks. In Section 3, we present t-SNE along with K-means clustering algorithm. In Section 4, we describe the experimental setup and the resultsof our experiments. In Section 7, conclusions and suggestions for future work are presented.

2. RELATED WORK

In network science, community detection has been a thoroughly studied topic. Finding node clusters or groupings in a network that have stronger internal connections than external ones is the aim [15]. Statistical inference approaches, spectral clustering, and modularity optimization are examples of traditional methods for community detection.

Newman and Girvan [4] proposed modularity optimization, which is one of the most popular methods. The modularity function, which gauges the density of links inside communities relative to links between communities, is what it seeks to maximize. Resolution

restrictions plague modularity optimization despite its widespread use, which can cause big communities to split or tiny communities to merge [5].

Another well-known technique is spectral clustering, which divides the graph into communities based on the eigenvalues and eigenvectors of the network's using unnormalized graph Laplacian matrix [6]. Despite its effectiveness in identifying community structure, spectral clustering can be computationally costly for large networks and may not always accurately represent the underlying community structure because of its dependence on global data.

The network is modeled using statistical inference techniques like the Stochastic Block Model (SBM) as a combination of distributions, each of which represents a community [7]. These techniques can be difficult to use and computationally demanding, even if they offer a probabilistic foundation for community detection.

Community detection algorithms have substantial hurdles when dealing with high-dimensional data in social networks. The goal of dimensionality reduction approaches is to make the data simpler while maintaining its fundamental structure. A traditional linear technique called principal component analysis (PCA) maps data onto its principle components to capture the greatest amount of variance [8]. PCA can only capture a limited amount of non-linear features, though.To overcome these restrictions, non-linear dimensionality reduction methods have been created, including Isomap [9] and Locally Linear Embedding (LLE) [10]. Whereas LLE concentrates on preserving local neighbourhood relationships, isomap maintains global geometric characteristics. Although these techniques work well, they might be computationally demanding for large datasets.

Van der Maaten and Hinton [11] invented t-SNE, a powerful non-linear dimensionality reduction technique that has become popular for high-dimensional data visualization. In contrast to PCA, t-SNE minimizes the Kullback-Leibler divergence between the joint probability derived from pairwise distances in the high-dimensional and low-dimensional spaces, with the goal of maintaining local similarities. Because of this method, t-SNE is especially useful for grouping data points that are near to each other in the original high-dimensional space and for showing intricate structures.Numerous domains, including genomics [12], image processing, and natural language processing [13], have made extensive use of t-SNE. This study is motivated by the fact that, despite its potential, its application in community discovery inside social networks has not been thoroughly investigated.

Based on similarities, clustering algorithms divide data into groups. MacQueen [14] introduced K-Means clustering, which is one of the most popular partitioning techniques. By iteratively allocating data points to the closest centroid and updating centroids depending on the mean of allocated points, it seeks to minimize the within-cluster sum of squares (WCSS). K-Means is an effective and straightforward method, although it has limitations when dealing with non-globular clusters and necessitates knowing the number of clusters ahead of time.

Combining dimensionality reduction with clustering algorithms has been explored in this study to enhance clustering performance.However, the specific combination of t-SNE and K-Means for community detection in social networks has not been extensively studied.This hybrid approach leverages the strengths of both techniques: t-SNE's ability to preserve local structures and K-Means' efficiency in clustering. This study aims to fill this gap by thoroughly investigating the effectiveness of this combination in detecting communities within social networks.

We address the issues of high-dimensional data in social networks by examining this important research and laying the groundwork for our proposed solution, which combines t-SNE with K-Means for community detection.

3. PRELIMINARIES

3.1 t-Distributed Stochastic Neighbor Embedding (t-SNE):

t-SNE is a non-linear dimensionality reduction technique that maps high-dimensional data to a lower-dimensional space, maintaining the local structure of the data.

Given a high-dimensional dataset $X = \{x_1, x_2, \ldots, x_n\}$, t-SNE constructs a probability distribution P over pairs of high-dimensional objects such that similar objects have a higher probability of being picked.

The similarity between points x_i and x_j is given by:

$$p_{ij} = \frac{\exp(-||x_i - x_j||^2 / 2{\sigma_i}^2)}{\sum_{k \neq l} \exp(-||x_k - x_l||^2 / 2{\sigma_k}^2)} \quad (1)$$

The similarity between points y_i and y_j in the low-dimensional space is given by a Student-t distribution:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l} (1 + ||y_k - y_l||^2)^{-1}} \quad (2)$$

The cost function minimized by t-SNE is the Kullback-Leibler (KL) divergence between a joint probability distribution, *P*, in the high-dimensional space and a joint probability distribution, *Q*, in the low-dimensional space:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \ log \frac{p_{ij}}{q_{ij}} \tag{3}$$

where again, we set p_{ii} and q_{ii} to zero. We refer to this type of SNE as symmetric SNE, because it has the property that $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji}$ for $\forall i,j$.

The gradient of the Kullback-Leibler divergence between *P* and the Student-t based joint probability distribution *Q* (computed using Equation 2) is given by

$$\frac{\delta C}{\delta \gamma} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) \left(1 + \left||y_i - y_j|\right|^2\right)^{-1} \tag{4}$$

t-SNE employs a heavy-tailed distribution in the low-dimensional space to alleviate both the crowding problem and the optimization problems.

Algorithm 1: *Simple version of t-Distributed Stochastic Neighbor Embedding.*

Data: *data set* $X = x_1, x_2, \dots, x_n$,

Cost function parameters: perplexity Perp,

Optimization parameters: number of iterations T*, learning rate* η*, momentum* $\alpha(t)$.

Result: *low-dimensional data representation* $\gamma^{(T)} = y_1, y_2, \dots, y_n$.

begin

Compute pair wise affinities $p_{j|i}$ *with perplexity Perp (using Equation 1 above)*

Set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$

Sample initial solution $\gamma^{(0)} = y_1, y_2, \dots, y_n$ *from* $\mathcal{N}(0, 10^{-4}I)$

for *t=1* ***to*** *T* ***do***

Compute low-dimensional affinities q_{ij} *(using Equation 2 above)*

Compute gradient $\frac{\delta C}{\delta \gamma}$ *(using Equation 4)*

set $\gamma^{(t)} = \gamma^{(t-1)} + \eta \frac{\delta C}{\delta \gamma} + \alpha(t)(\gamma^{(t-1)} - \gamma^{(t-2)})$

end

end

3.2 The *K*-Means Clustering:

We are given a data set $(x^1, x^2, \ldots, x^n)$, where for $i \in \{1, 2, \ldots, n\}, x^i \in \mathbb{R}^d$. Here d ≥ 2 is the dimension of the data set. We are also specified an integer $k \geq 2$.The objective of k-means clustering is to partition the data set into k clusters, such that each cluster is as "tight" as possible. More precisely:

A clustering $C : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, k\}$ assigns one of k clusters to each point in the data set. Each cluster $k' \in \{1, 2, \ldots, k\}$ is also associated with a centre $\boldsymbol{\mu}_{k'} \in \mathbb{R}^d$. If we take a clustering C along with the sequence $\boldsymbol{\mu}$ representing the centres of its k clusters—$\boldsymbol{\mu}$ = ($\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$,..., $\boldsymbol{\mu}_k$)—we can define "tightness" in terms of the aggregate distance between the data points and the centres of the clusters to which they are assigned by C. If $C(i)$ is the cluster in $\{1, 2, \ldots, k\}$ to which C assigns input point *i*, the Euclidean distance between the point and its cluster center is $||x^i - \boldsymbol{\mu}_{C(i)}||$. The most common measure of the tightness of a clustering C (along with cluster centres $\boldsymbol{\mu}$) is the sum squared error (SSE), defined as

$$\sum_{i=1}^{n} ||x^i - \boldsymbol{\mu}_{C(i)}||^2$$

The*k*-means clustering problem is the problem of finding a clustering among the set of all clusterings, along with a sequence of cluster centres, such that the corresponding SSE is minimal. Unfortunately, even for k = 2, this problem is NP-hard for general *d* and *n* [16].

k-means Clustering Algorithm

Let C^0 *be an arbitrary clustering, and let* $\boldsymbol{\mu^0} = (\boldsymbol{\mu^1}, \boldsymbol{\mu^2}, \ldots, \boldsymbol{\mu^k})$ *be a sequence of centres such that for* $k' \in \{1, 2, \ldots, k\}$, $\boldsymbol{\mu}_{k'}^{\mathbf{0}}$ *is the centroid of the points in the* k' *-th cluster.*

$t \leftarrow 0$.

Converged ← *false.*

while ¬*converged*

converged ← *true.*

for $i \in \{1, 2, \ldots, n\}$

$$C^{t+1}(i) \leftarrow C^t(i) .$$

for $k' \in \{1, 2, \ldots, k\}$

if $k' \neq C^{t+1}(i)$ *and* $||x^i - \boldsymbol{\mu}_{k'}|| < ||x^i - \boldsymbol{\mu}_{C^{t+1}(i)}||$

$$C^{t+1}(i) \leftarrow k'.$$

Converged ← *false.*

for $k' \in \{1, 2, \ldots, k\}$

Set $\boldsymbol{\mu}_{k'}^{\boldsymbol{t+1}}$ *to be the centroid of all points i such that* $C^{t+1}(i) = k'$.

$$t \leftarrow t + 1$$

Return C^t, $\boldsymbol{\mu}^{\boldsymbol{t}}$.

Theorem: *The* $k-$*means clustering algorithm converges.*

Proof.Suppose that the algorithm proceeds from iteration t to iteration $t+1$. It suffices to show that SSE(C^{t+1},$\boldsymbol{\mu}^{\boldsymbol{t+1}}$) < SSE($C^t$, $\boldsymbol{\mu}^{\boldsymbol{t}}$). To see why, consider that if that was true, no clustering can be visited twice; since the number of possible clusterings is finite (k^n), the algorithm must necessarily terminate. By the construction of the algorithm, we know that it terminates when no point has a cluster centre closer than the centres of its current cluster; in other words, the current clustering is *locally* optimal.

We show that SSE(C^{t+1},$\boldsymbol{\mu}^{\boldsymbol{t+1}}$) < SSE($C^t$, $\boldsymbol{\mu}^{\boldsymbol{t}}$) in two steps. First we show that

$$\text{SSE}(C^{t+1},\boldsymbol{\mu}^{\boldsymbol{t+1}}) < \text{SSE}(C^t, \boldsymbol{\mu}^{\boldsymbol{t}}), \quad (1)$$

and next, we show that

$$\text{SSE}(C^{t+1},\boldsymbol{\mu}^{\boldsymbol{t+1}}) \leq \text{SSE}(C^{t+1}, \boldsymbol{\mu}^{\boldsymbol{t}}). \quad (2)$$

The first step follows directly from the logic of the algorithm: C^t and C^{t+1} are different only if

there is a point that finds a closer cluster centre in $\boldsymbol{\mu}^{\boldsymbol{t}}$ than the one assigned to it by C^t:

$$\text{SSE}(C^{t+1}, \boldsymbol{\mu}^{\boldsymbol{t}}) = \sum_{i=1}^{n} ||x^i - \boldsymbol{\mu}_{C^{t+1}(i)}^{t}||^2 < \sum_{i=1}^{n} ||x^i - \boldsymbol{\mu}_{C^{t}(i)}^{t}||^2 = \text{SSE}(C^t, \boldsymbol{\mu}^{\boldsymbol{t}}).$$

The second step:

$$\text{SSE}(C^{t+1},\boldsymbol{\mu}^{\boldsymbol{t+1}}) = \sum_{i=1}^{n} ||x^i - \boldsymbol{\mu}_{C^{t+1}(i)}^{t+1}||^2$$

$$= \sum_{k'=1}^{k} \sum_{i \in \{1,2,\ldots,n\}, C^{t+1}(i)=k'} ||x^i - \boldsymbol{\mu}_{C^{t+1}(i)}^{t+1}||^2$$

$$\leq \sum_{k'=1}^{k} \sum_{i \in \{1,2,\ldots,n\}, C^{t+1}(i)=k'} ||x^i - \boldsymbol{\mu}^t_{C^{t+1}(i)}||^2$$

$$= \sum_{i=1}^{n} ||x^i - \boldsymbol{\mu}^t_{C^{t+1}(i)}||^2$$

$$= \text{SSE}(C^{t+1}, \boldsymbol{\mu}^t).$$

4. EXPERIMENTAL RESULTS

We evaluate the effectiveness of the proposed technique t-SNE with K-means algorithm on real-world networks by comparing the results with the actual community structures and findings from other community detection techniques.

4.1. REAL WORLD NETWORKS WITH GROUND TRUTH

To evaluate the accuracy and efficiency of the proposed t-SNE with K-means algorithm, five real world networks with un-weighted and un direct links are used for experiments. Table 1 presents the ground truth of these five real world networks.

To assess the effectiveness of the proposed t-SNE with K-means algorithm, we evaluate its partitioning results against five well-known algorithms: Louvain, Girvan-Newman, Infomap, Label propagation (LPA), Fast Greedy Algorithm. The NMI index is employed to measure efficiency and accuracy and the results which are presented in Table 2.

Table 1. Statistics of real-world networks with ground truth communities

Network Dataset	**|V|**	**|E|**	**<k>**	**C**	**D**	***N***
Zachary Karate Club Network	34	78	4.59	0.56	5	2
American College Football	115	616	10.6	0.40	4	12
Bottlenose Dolphins Network	62	159	5.13	0.26	8	2
US Polbooks Network	105	441	8.40	0.48	--	3
Scientist's Collaboration Network	118	197	56	0.66	--	6

<k> - average degree of the dataset **C** - Clustering Coefficient of the dataset

D - Diameter of the dataset ***N*-** Number of Communities

Normalized mutual information (NMI) typically quantifies the resemblance between actual community formations and those identified in networks and is given by

$$NMI(X,Y) = \frac{2I(X,Y)}{H(X) + H(Y)}$$

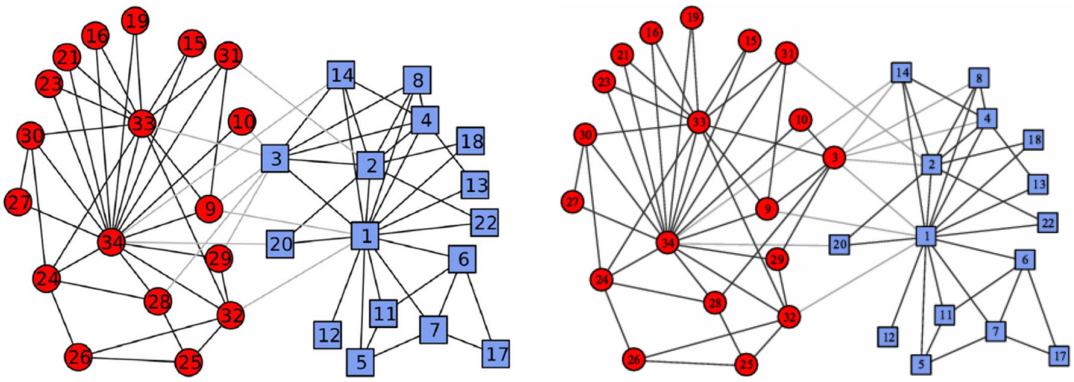
where I(X, Y), the mutual information that measures the informationshared by variables X and Y, H(X) is the entropy of communityof X. For example, if X and Y are independent, then knowingX does not given any information about Y and vice versa, soNMI(X, Y) = 0. At the other extreme, if X and Y are deterministicfor each other, then all information covered by X is shared withY and vice versa, so NMI(X, Y) = 1.

Results and Discussion

The networks we use for the evaluation must meet specific requirements in order for us to evaluate the results of our experiments in both qualitative and quantitative ways. Firstly, the ground truth community structures of the networks must be known beforehand, and their scales must be small enough to allow for easy interpretation and visualization of the data. Secondly, the networks must be publicly accessible in order to enable easy verification of the methods or algorithms. This resulted in the selection of five real-world network datasets i.e., Zachary Karate Club Network [4][1],American College Football Network [2][19], Bottlenose Dolphins Network [3][20], US Political Books Network [21],and a collaboration network of scientists working at the SantaFe Institute, which is an interdisciplinary research center in SantaFe, New Mexico [19].

A. Zachary Karate Club Network

In the early 1970s, at an American university, WayneZachary studied the members of a karate club for two yearsand recorded their social interactions. Based on their socialinteractions, he built a network dataset with 34 vertices and 78edges. In this dataset, the students were represented as verticesand two students are linked by an edge if they are good friends. By chance, a dispute arose during the course of his study between the club's administrator and the karate instructor. As a result, the club splits into two smaller communities with the administrator and the instructor being as the central persons accordingly. The original division of the club into 2 communities is shown in Figure 2(a) and the community findings by the proposed algorithm is shown in 2(b), respectively. The NMI values calculated by various comparison algorithms and the proposed algorithm are depicted in Table 2.

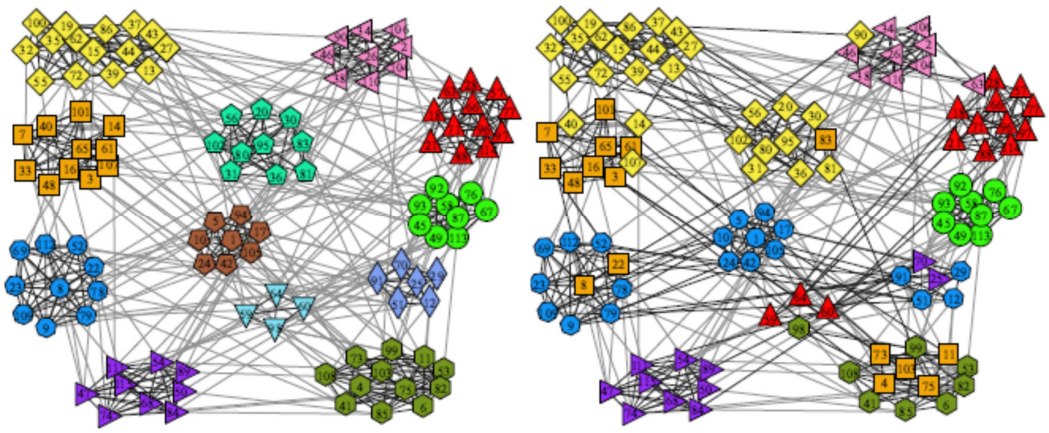


(a) (b)

Fig. 2. Zachary Karate Club Network (a) Original Community Structure (b) The community structure extracted bythe proposed method.

B. American College Football Network

The American College Football network dataset was developed from the United States college football games. The schedule of games between Division IA colleges during the season Fall 2000 is represented by this network. Teams are represented by vertices in the network and the regular season games between two teams are represented by edges. The total number of vertices in this dataset is 115 and the number of edges is 616. The teams are divided into conferences.Figure3 (a) shows the actual community structure of this dataset and figure 3(b) shows the communities identified by the proposed method.



(a) (b)

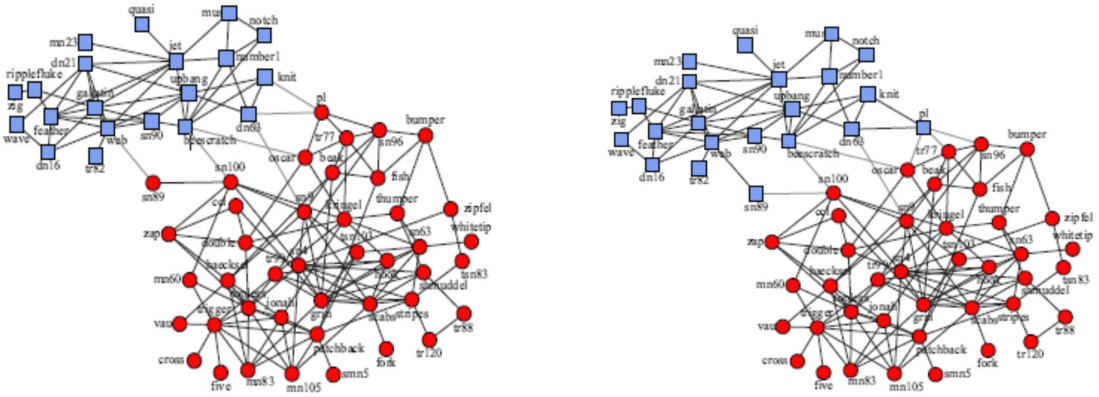
Fig. 3. American College Football Network. (a) Original Community Structure (b) The community structure identified by the proposed method.

Table 2. The NMI values for real-world datasets with ground truth.

Method	**Zachary Karate Club Network**	**American College Football Network**	**Bottlenose Dolphins Network**	**US Polbooks Network**	**Scientist's Collaboration Network**
Louvain	0.691	0.890	0.451	0.555	0.540
Girvan-Newman	0.732	0.359	0.888	0.668	0.547
Label Propagation	0.364	0.869	0.527	0.735	0.534
Infomap	0.568	0.924	0.481	0.493	0.546
fast greedy	0.564	0.697	0.572	0.531	
t-SNE with K-Means	**0.837**	**0.924**	**0.451**	**0.598**	**0.783**

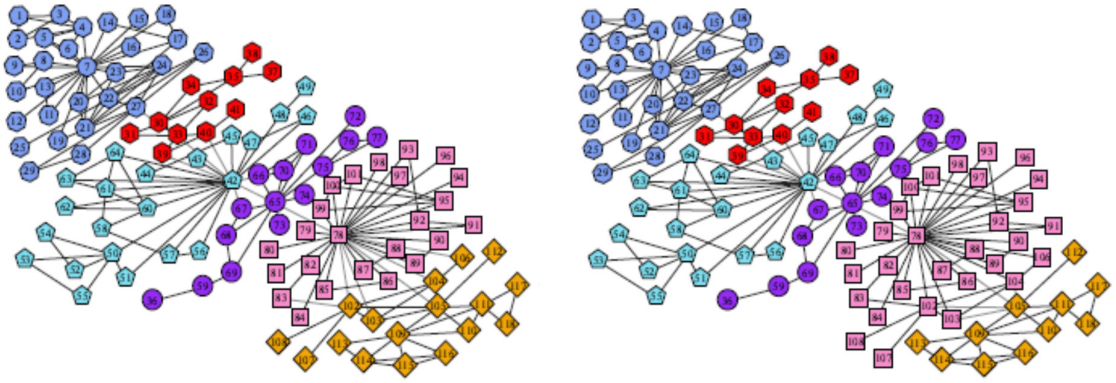
C. Bottlenose Dolphins Network

David Lusseau, a biologist, analyzed, for seven years, the behavior of bottlenose dolphins living in Doubtful Sound (New Zealand) and developed this network dataset. Based on the frequent association, a link is established between two dolphins if their association was significant. The total number of dolphins that were included in the study are 62 and 159edges were set between these dolphins that were seen together more often than expected by chance. Figure4(a) shows the original community structure of dolphin network and Figure4(b)shows the results obtained using the proposed approach. The NMI values calculated by various comparison algorithms and the proposed algorithm are depicted in Table 2.



(a) (b)

Fig. 4. Bottlenose Dolphins Network. (a) Ground truth Community Structure (b) The community structure identified by the proposed method.

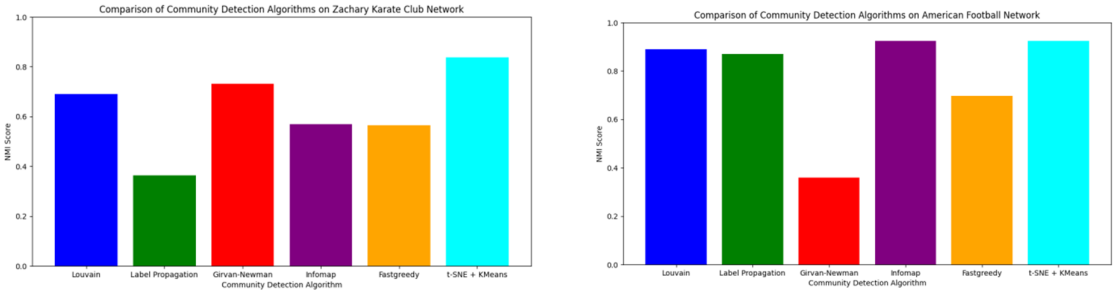


(a) (b)

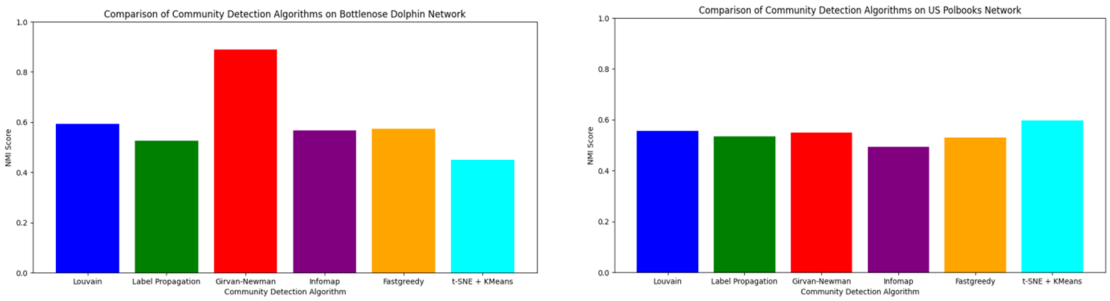
Fig. 5. Santa Fe Scientist's Collaboration Network. (a) Ground truth Community Structure (b) The community structure identified by the proposed method.

D. Santa Fe Scientist's Collaboration Network

This network is the biggest part of a network of scientists at Santa Fe Institute who collaborate together. This network contains 118 vertices and 197 edges. As per the scientists' specialties, it can be partitioned into6 communities. This is a coauthor network dataset between 118 scientists working at the Santa Fe Institute, in which each vertex represents a scientist, and each edge connects two scientists who have coauthored at least one article. Figure5(a) shows the original community structure of Scientist's Collaboration network and Figure4(b)shows the results obtained using the proposed approach. The NMI values calculated by various comparison algorithms and the proposed algorithm are depicted in Table 2. Figure 6(a) – 6(d) shows the comparison between state of art community detection algorithms and the proposed technique. From these graphs it is observed that overall the proposed approach performs better than the comparison algorithms.



(a) (b)



(c) (d)

Fig. 6. Comparison of Community Detection Algorithms and the proposed approach on (a) Zachary Karate Club Network (b) American College Football Network (c) Bottlenose Dolphin Network (d) US Pol books Network

5. CONCLUSION AND FUTURE WORK

This paper presents a novel approach for community detection in social networks using t-SNE for dimensionality reduction and K-Means clustering. The method shows promising results on real-world datasets, highlighting its potential for analyzing complex social networks. Future work will explore the integration of additional node attributes and the application to dynamic networks.

6. REFERENCES

1. E. Raju, Y. Rama Devi, and K. Sravanthi, ‘‘CCLPA: A clustering coefficient based label propagation algorithm for unfolding communities in the complex networks,’’ in Proc. 2nd Int. Conf. Commun. Electron. Syst. (ICCES), Coimbatore, India, 240–245, 2017.

2. R. Enugala, L. Rajamani, K. Ali and S. Kurapati, "Identifying Natural Communities in Social Networks Using Modularity Coupled with Self Organizing Maps", Computational Intelligence in Data Mining -Volume 1, Advances in Intelligent Systems and Computing 410, DOI10.1007/978-81-322-2734-2_37, Springer, India, 367 – 376, 2016.

3. Raju, E., Ramadevi, Y. & Sravanthi, K. CILPA: a cohesion index based label propagation algorithm for unveiling communities in complex social networks. *International journal of information technology,* 10, 435–445 (2018).

4. Newman, M. E. J., & Girvan, M. Finding and evaluating community structure in networks. Physical Review E, 69(2), 026113, 2004.

5. Fortunato, S., & Barthélemy, M. (2007). Resolution limit in community detection. Proceedings of the National Academy of Sciences, 104(1), 36-41.

6. Raju, E., Hameed, M. A., and Sravanthi, K. Detecting communities in social networks using un normalized spectral clustering incorporated with Bisecting K-means. *Proceedings of IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1–5, 2015.

7. Holland, P. W., Laskey, K. B., & Leinhardt, S. Stochastic block models: First steps. Social Networks, 5(2), 109-137, 1983.

8. Jolliffe, I. T. Principal Component Analysis. Springer Series in Statistics, 2002.

9. Tenenbaum, J. B., de Silva, V., & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500), 2319-2323, 2000.

10. Roweis, S. T., & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500), 2323-2326, 2000.

11. Van der Maaten, L., & Hinton, G. Visualizing Data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605, 2008.

12. Amir, E.-a. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., ... & Pe'er, D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature Biotechnology, 31(6), 545-552, 2013.

13. Lau, J. H., & Baldwin, T. An empirical evaluation of doc2vec with practical insights into document embedding generation. Proceedings of the 1st Workshop on Representation Learning for NLP, 78-86, 2016.

14. MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1(14), 281-297, 1967.

15. E. Raju and K. Sravanthi. Analysis of social networks using the techniques of web mining. International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 10, pp. 443–450, October 2012.

16. Sanjoy Dasgupta. The hardness of k-means clustering. Technical Report CS2008-0916,Department of Computer Science and Engineering, University of California, San Diego,2008.

17. Enugala R, Rajamani L, Kurapati S, et al. Detecting Communities in Dynamic Social Networks Using Modularity Ensembles SOM. International Journal of Rough Sets and Data Analysis (IJRSDA), 5(1): 34-43, 2018.

18. Raju Enugala, Lakshmi Rajamani, Kadampur Ali, Sravanthi Kurapati. Community Detection in Dynamic Social Networks: A Survey. International Journal of Research and Applications, 2(6): 278-285, 2015.

19. M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821-7826, 2002.

20. Lusseau D, Schneider K, Boisseau O, Haase P, Slooten E, et al. The bottle nose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology 54: 396–405, 2003.