VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

VIT CHENNAI

Vandalur –Kelambakam Road,Chennai-600127

Review III Report

Programme: Mtech in Artificial Intelligence and Machine Learning

Course: CSE NOSQL DATABASE

Faculty: Dr. A Bhuvaneshwari

Component: J

Title : COMMUNITY DETECTION IN NETFLIX DATASET TO MODEL RECOMMENDATION ENGINE

Team Members: SHILPA REJI (20MAI1010)

ALINA MARY SABU(20MAI1022)

ABSTRACT

Development of Internet and computer science, a lot of people join social networks. People communicate with each other and express their opinions on the social media, which forms a complex network relationship

Individuals in the social networks form a "relation structure" through various connections which produces a large amount of information dissemination. This "relation structure" is the community that we are going to research

Community detection is very important to reveal the structure of social networks, to know people's views, analyze the information dissemination and grasp as well as control the public sentiment

In recent years, with community detection becoming an important field of social networks analysis, a large number of academic literatures proposed numerous methods of community detection. Here, we first describe community detection in Netflix dataset. Then we classify the different communities obtained to a model recommendation system

1. INTRODUCTION

1.1 OBJECTIVE

Community detection in Netflix dataset aims to organize the nodes of the network in groups or communities such that nodes belonging to the same community are densely interconnected but sparsely connected with the remaining nodes in the network and to find hidden patterns in the communities to model a better recommendation system.

1.2 PROBLEM STATEMENT

- The lack of right data i.e., input data may not always be accurate because humans are not perfect providing ratings.
- User behavior is more important than ratings.
- So community based recommender systems for Netflix suggest resources such as movies and series based on group of users that manifest similar preferences and behaviors one another.

2. LITERATURE REVIEW

- **Review on Community Detection Algorithms in Social Networks**

  In the research of community detection, the academic community has made a lot of valuable achievements and proposed many effective methods, which characterized communities from different perspectives. However, with the development of the Internet, social networks are becoming more and more complex; there still exist many problems waiting further research.

- **A Review of Community Detection Algorithms in Signed Social Networks**

  It has presented an overview of the community detection algorithms in signed social networks. Some existing approaches are illustrated with the main focus on input parameters which are used to perform community detection while some of them are automatic approaches where input parameter is not required to perform community detection. In the future the applicability domain of the existing algorithms can be enhanced so that the algorithms can also find communities indynamic graphs also.

- **An Empirical Study of Community and Sub-Community Detection in Social Networks Applying Newman-Girvan Algorithm**

  In this paper, it has presented an empirical study of Newman-Girvan algorithm on various data sets. Our results differ from those presented earlier in the sense that we have defined a new concept of sub-communities. The main drawback of Newman-Girvan algorithm is the absence of a clear specification on the definition of what constitutes a community. A lot has been left out for individual interpretations. The problem increases many-folds in cases of unsupervised datasets. The user has to manually identify the major communities from the dendrograms structure. As a future work, we hope to apply the concept of multi-objective function to detect the stable communities in social networks.

- **An Adaptive Approximation Algorithm for Community Detection in Social Network**

  As Proposed Algorithm is more precise then Eigen vector based algorithm based on modularity and computation time. Aim to use this algorithm for complex and dense network as network contains many overlapping nodes and crossed edges. In future,we are looking to implement community detection algorithm for following properties:

  1. Overlapping Community
  2. Interest Based Community Detection

- **Influence Propagation Model for Clique-Based Community Detection in Social Networks**

  In this paper, it propose an approach to detect temporally active and dense communities, making use of the biased density metric and the influence of active users with the frequency of their interactions with the neighborhood. Here also propose an objective function to partition the graph by decomposing the data and distributing them evenly across the available processors. In the future, we aim to develop a time interval model that combines our influence propagation model and a time interval parameter. This will help in finding dense active communities. The model will incorporate the knowledge of graph structure and node attributes, in addition to time gaps.

3. PROPOSED ALGORITHM

In the field of complex networks, community detection is a hot topic, and numerous studies have been done on it. A widely accepted definition of a community in a network is a subset of nodes with high internal density but low exterior density. Modularity, as established by Newman and Girvan, is currently the most extensively used criterion for assessing the quality of communities in networks. Modularity can be thought of as the difference between the percentage of edges inside a community and the proportion of edges expected by a random version of the network while maintaining the degree distribution of the nodes in a broad sense.

Girvan and Newman develop an approach based on edge centrality that can handle small-scale networks in one of the earliest works with the goal of researching community patterns in networks. Later on, Newman develops a modularity-based heuristic method that can handle larger networks. Clauset et al. offer an approach for adapting the heuristic method to large-scale networks, allowing it to be conducted more efficiently.

Clauset, Newman, and Moore (CNM) approach is one of the most important ways for detecting communities in networks, and it is now one of the most studied approaches for this purpose. Some modifications, such as those proposed by Wakita and Tsurumi, Leon-Suematsu and Yuta, and Danon et al., have been proposed in the literature to speed up the execution and allow for the investigation of bigger networks. Blondel et al. propose another key heuristic method for community recognition in large size networks, which employs an agglomerative multistep process during execution. Nonparametric approaches, which try to modify networks to statistic models based on their structural qualities, are currently attracting a lot of attention.

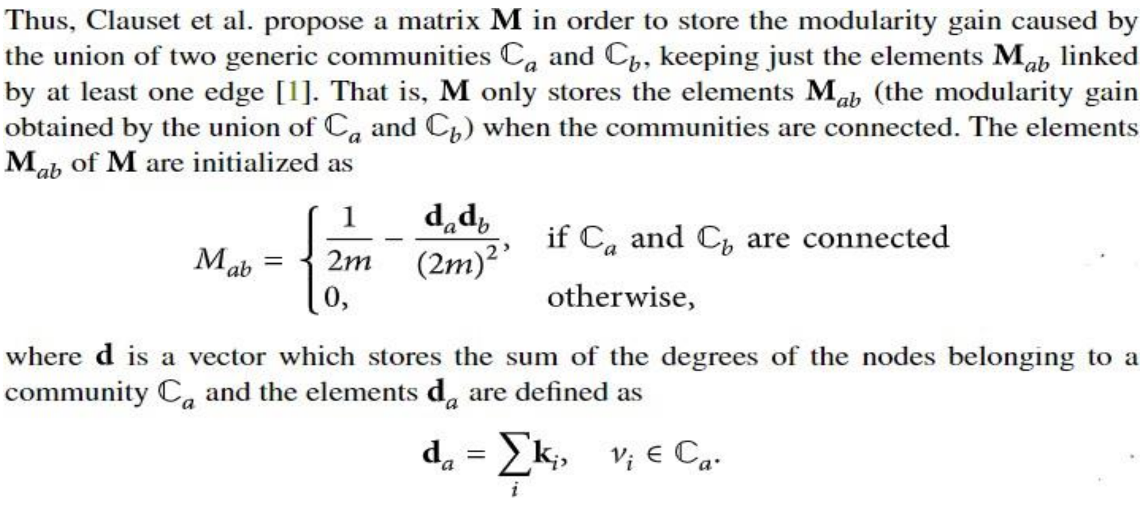
The goal of this research is to look at the computational challenges of community detection algorithms that can deal with massive networks. The spectrum technique of Newman combined with a version of the Kernighan-Lin approach, known as fine-tuning, and the Clauset, Newman, and Moore technique are two of the most widely used modularity Q based approaches for community detection. Some tweaks to the fine-tuning stage are also proposed in order to speed up the process without sacrificing the quality of the end result.

The Clauset, Newman, and Moore (CNM) technique is a heuristic method for quickly identifying communities in massive networks. CNM is one of the most often used strategies for dealing with big networks in the literature. Because CNM is a greedy heuristic method, it may produce partitions that differ from the ideal solution, and the modularity obtained is frequently lower than that obtained by other approaches.

According to its original definition, the method suggested by Newman initially associates each node of the network with a community. Then it repeatedly joins the communities whose union results in the greatest rise in the community structure's modularity Q.

The approach seeks to determine the community combination that results in the greatest increase in Q, and then conducts the operation. That is, the technique identifies the pair of communities Ca and Cb that, when joined, yields the highest modularity value. Such a value can be understood as an affinity measure between two generic communities Ca and Cb, with the goal of finding two communities that are similar enough to be combined.

The CNM approach employs an agglomerative technique, and when a network of nodes is combined, the result is only one community comprising all of the nodes, at which point the procedure terminates.

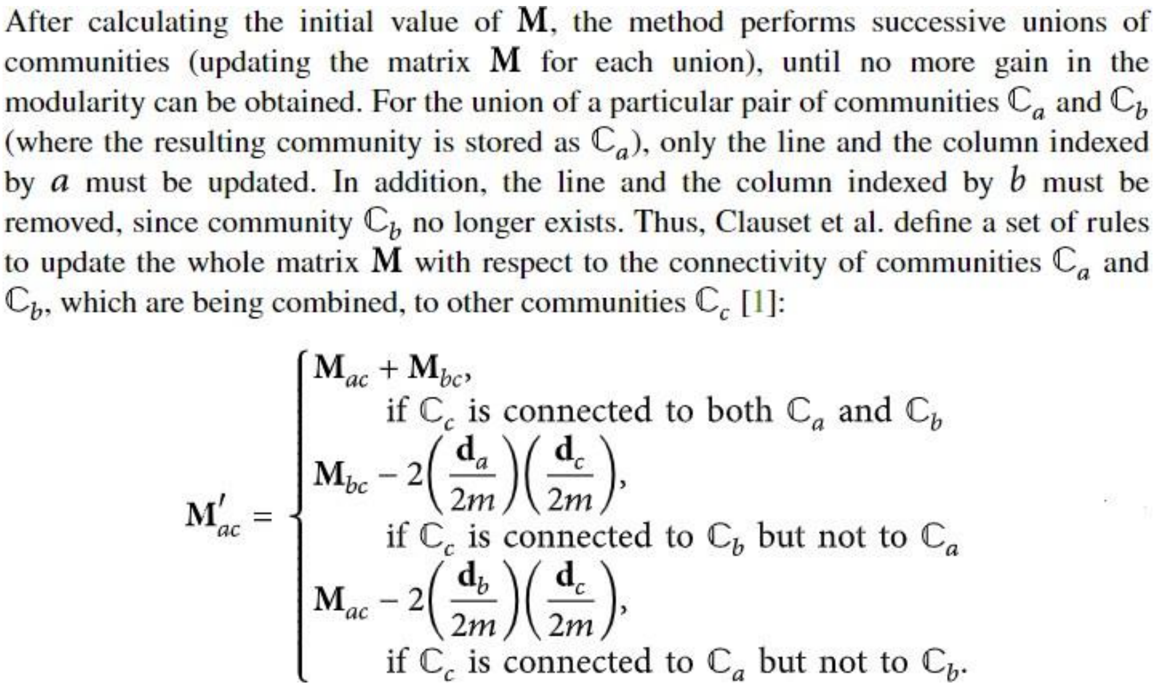


Thus, Clauset et al. propose a matrix $\mathbf{M}$ in order to store the modularity gain caused by the union of two generic communities $\mathbb{C}_a$ and $\mathbb{C}_b$, keeping just the elements $\mathbf{M}_{ab}$ linked by at least one edge [1]. That is, $\mathbf{M}$ only stores the elements $\mathbf{M}_{ab}$ (the modularity gain obtained by the union of $\mathbb{C}_a$ and $\mathbb{C}_b$) when the communities are connected. The elements $\mathbf{M}_{ab}$ of $\mathbf{M}$ are initialized as

$$M_{ab} = \begin{cases} \dfrac{1}{2m} - \dfrac{\mathbf{d}_a\mathbf{d}_b}{(2m)^2}, & \text{if } \mathbb{C}_a \text{ and } \mathbb{C}_b \text{ are connected} \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{d}$ is a vector which stores the sum of the degrees of the nodes belonging to a community $\mathbb{C}_a$ and the elements $\mathbf{d}_a$ are defined as

$$\mathbf{d}_a = \sum_i \mathbf{k}_i, \quad v_i \in \mathbb{C}_a.$$



After calculating the initial value of $\mathbf{M}$, the method performs successive unions of communities (updating the matrix $\mathbf{M}$ for each union), until no more gain in the modularity can be obtained. For the union of a particular pair of communities $\mathbb{C}_a$ and $\mathbb{C}_b$ (where the resulting community is stored as $\mathbb{C}_a$), only the line and the column indexed by a must be updated. In addition, the line and the column indexed by b must be removed, since community $\mathbb{C}_b$ no longer exists. Thus, Clauset et al. define a set of rules to update the whole matrix $\mathbf{M}$ with respect to the connectivity of communities $\mathbb{C}_a$ and $\mathbb{C}_b$, which are being combined, to other communities $\mathbb{C}_c$ [1]:

$$\mathbf{M}'_{ac} = \begin{cases} \mathbf{M}_{ac} + \mathbf{M}_{bc}, \\ \qquad \text{if } \mathbb{C}_c \text{ is connected to both } \mathbb{C}_a \text{ and } \mathbb{C}_b \\ \mathbf{M}_{bc} - 2\left(\dfrac{\mathbf{d}_a}{2m}\right)\left(\dfrac{\mathbf{d}_c}{2m}\right), \\ \qquad \text{if } \mathbb{C}_c \text{ is connected to } \mathbb{C}_b \text{ but not to } \mathbb{C}_a \\ \mathbf{M}_{ac} - 2\left(\dfrac{\mathbf{d}_b}{2m}\right)\left(\dfrac{\mathbf{d}_c}{2m}\right), \\ \qquad \text{if } \mathbb{C}_c \text{ is connected to } \mathbb{C}_a \text{ but not to } \mathbb{C}_b. \end{cases}$$

Algorithm for Clauset, Newman and Moore

RESEARCH CHALLENGES

Essentially, there are two kinds of communities: implicit and explicit communities [1]. Group members directly name explicit communities, such as whatsapp groups, facebook groups [2], and so forth. Implicit communities, on the other hand, lack a formal organisation but are founded on a shared set of values.

Members of a group have a common interest, property, or behaviour. The term "community detection" refers to the detection of such hidden communities.

There are numerous assessments on community detection accessible [3]. However, an assessment of the major concerns and challenges in this field is required. As indicated in Fig. 1, we have grouped the concerns into four categories: No Precise Definition, Dynamic Community, Overlapping Nature of Community, and Validation of Community Form.

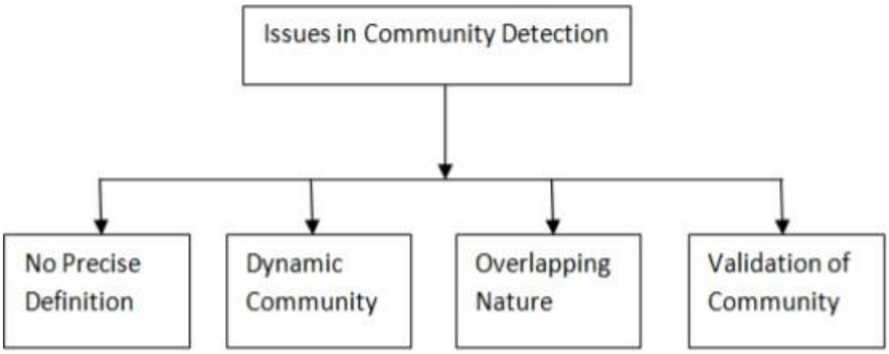
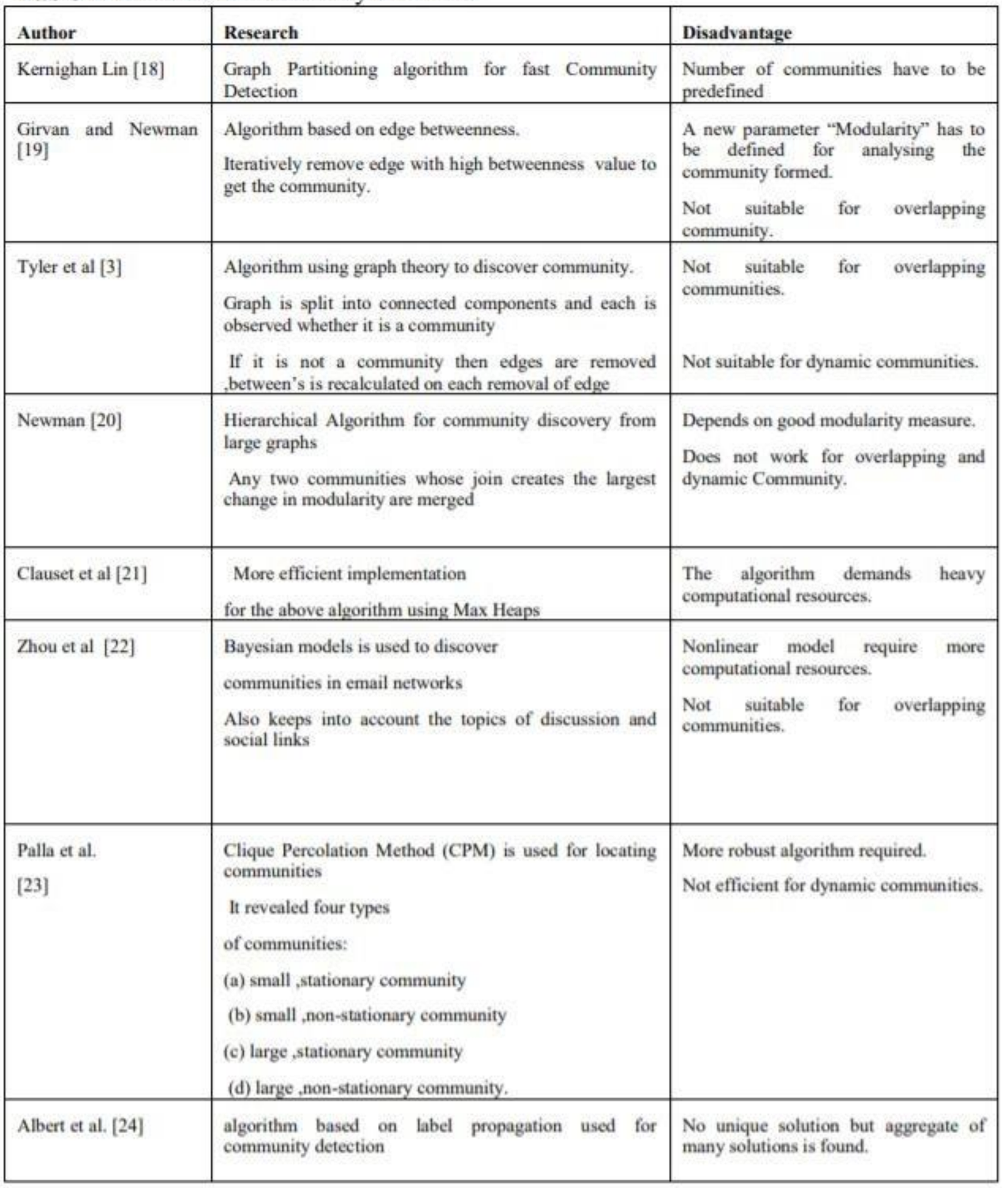


Fig.1. Issues in Community Detection Research

Due to the lack of a universally acknowledged definition of community, different researchers have varied perspectives on the subject, depending on their study needs [4]. Community was characterised by Kernighan Lin as sections of the network having minimal linkages to the rest of the system [5], while another researcher described Community as a set of vertices that are similar to each other [3]. Furthermore, the bulk of community algorithms believe communities to be discrete, when in reality, communities overlap [6]. Apart from their overlapping character, social media networks have constantly changing qualities that cannot be overlooked [7]. The challenge of validating different community detection algorithms is also a major roadblock in community detection research.

For community detection, a variety of techniques have been developed. Table 1 lists some of the key contributions in this field.

Table 1: Research in Community Detection



Author	**Research**	**Disadvantage**
Kernighan Lin [18]	Graph Partitioning algorithm for fast Community Detection	Number of communities have to be predefined
Girvan and Newman [19]	Algorithm based on edge betweenness. Iteratively remove edge with high betweenness value to get the community.	A new parameter "Modularity" has to be defined for analysing the community formed. Not suitable for overlapping community.
Tyler et al [3]	Algorithm using graph theory to discover community. Graph is split into connected components and each is observed whether it is a community If it is not a community then edges are removed ,between's is recalculated on each removal of edge	Not suitable for overlapping communities. Not suitable for dynamic communities.
Newman [20]	Hierarchical Algorithm for community discovery from large graphs Any two communities whose join creates the largest change in modularity are merged	Depends on good modularity measure. Does not work for overlapping and dynamic Community.
Clauset et al [21]	More efficient implementation for the above algorithm using Max Heaps	The algorithm demands heavy computational resources.
Zhou et al [22]	Bayesian models is used to discover communities in email networks Also keeps into account the topics of discussion and social links	Nonlinear model require more computational resources. Not suitable for overlapping communities.
Palla et al. [23]	Clique Percolation Method (CPM) is used for locating communities It revealed four types of communities: (a) small ,stationary community (b) small ,non-stationary community (c) large ,stationary community (d) large ,non-stationary community.	More robust algorithm required. Not efficient for dynamic communities.
Albert et al. [24]	algorithm based on label propagation used for community detection	No unique solution but aggregate of many solutions is found.

Challenges Involved

i. Determining which perspective of community is most advantageous under what circumstances [8].

ii. Choosing a certain community viewpoint to begin community detection.

The different views can be one of the following :-

- Cut based view
- Clustering view
- Stochastic block model view
- Dynamic view

Cut-based perspective:

This viewpoint describes community as a set of nodes with the fewest number of links between groups, disregarding the group's internal structure [9].

Kernighan-graph Lin's partitioning technique makes advantage of the community's methodology [3]. The user must select the size and number of groups to be produced in this method. It is necessary to optimise the difference between the edges within the group and the edges outside the group.

Challenges Involved

i. The group's internal relationship is not taken into account [3,9].

ii.Preferring tightly linked internal nodes group [3] is impossible.

iii.The number of communities must be determined beforehand [9].

Clustering view:

The goal of the clustering perspective is to optimise the group's internal density [10]. The primary idea is to organise the nodes so that nodes within a group have frequent connections with nodes within the group and sparse connections with nodes outside the group. The benefit of this method was that the number of groups did not have to be predetermined.

Newman Girvan's algorithm is the most important in this field[11]. The divisive technique is employed in this algorithm to eliminate edges with high betweenness.

Challenges Involved

i. A well-defined stopping criterion, such as modularity, must be defined.

ii. Using Modularity parameters, this problem becomes an optimization issue[12].

iii. Finding an optimal clustering algorithm is tough.

Stochastic Block Model View

Internal density isn't maximised, and external linkages aren't minimised. It makes use of the structural equivalency notion. This determines a set of nodes that connect to nodes in other communities in a similar manner. It has a number of benefits. Even from a bipartite graph, community may be determined, which was not achievable in the previous two ways.

Challenges Involved

i. It necessitates more complicated calculations than previous approaches [13].

Dynamic View

This method is distinct from other methods. The behaviour pattern between the nodes is more essential than the structure of the communities in this case. As a result, the focus is on how short-term dynamics affect the long-term behaviour of a system in a network. This is beneficial in situations when communities are well defined but internal dynamics are harder to comprehend.

Challenges Involved

i. It is still in its early stages and is primarily used in diffusion dynamics.

ii. More research is needed to adapt it to the dynamics of other complex systems.

Communities Overlap an Issue

The majority of early community detection research assumes that a community is a disjointed collection of strongly connected nodes. In the actual world, though, a person can be a member of multiple communities. This is also true in the case of social media. It means that the communities are not discontinuous in nature, but rather overlap [6], as illustrated in Fig.4. A, B, C, and D are communities that overlap. As a result, we must include the community's "Overlap" property.

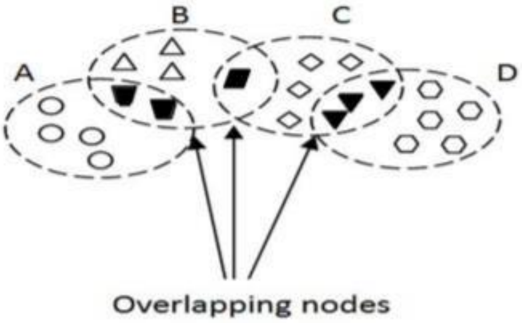


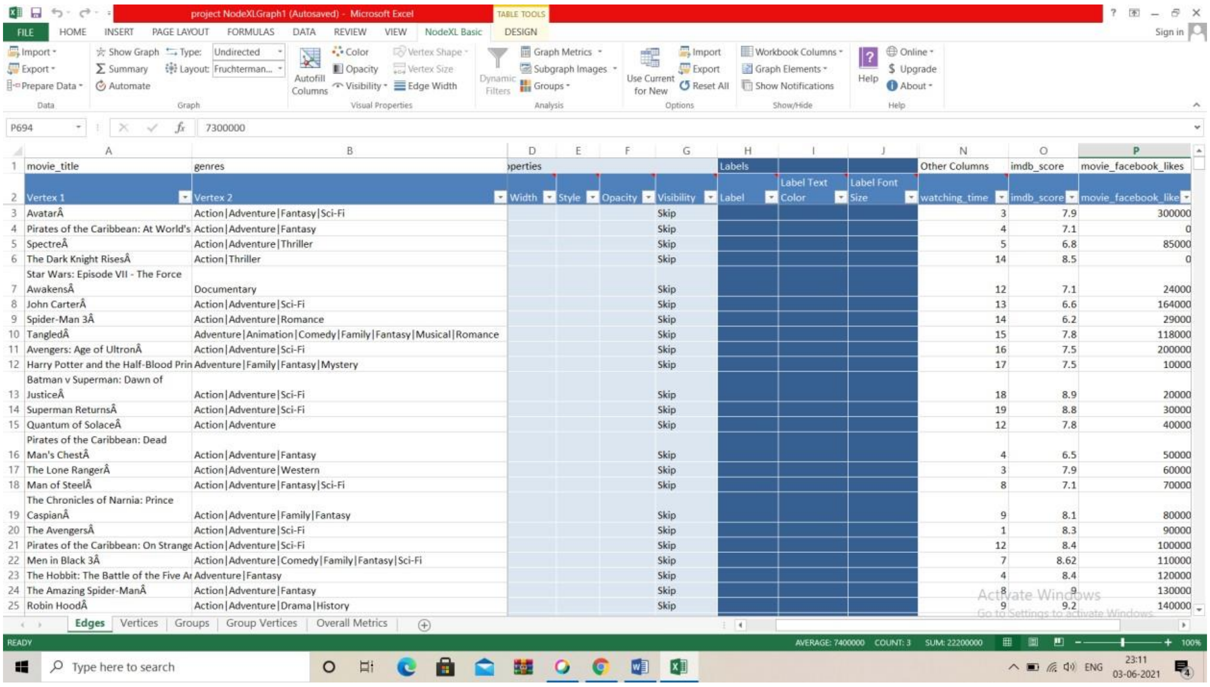
Fig.4. Overlapping Communities

The following are some of the difficulties encountered in the search for overlapping communities:

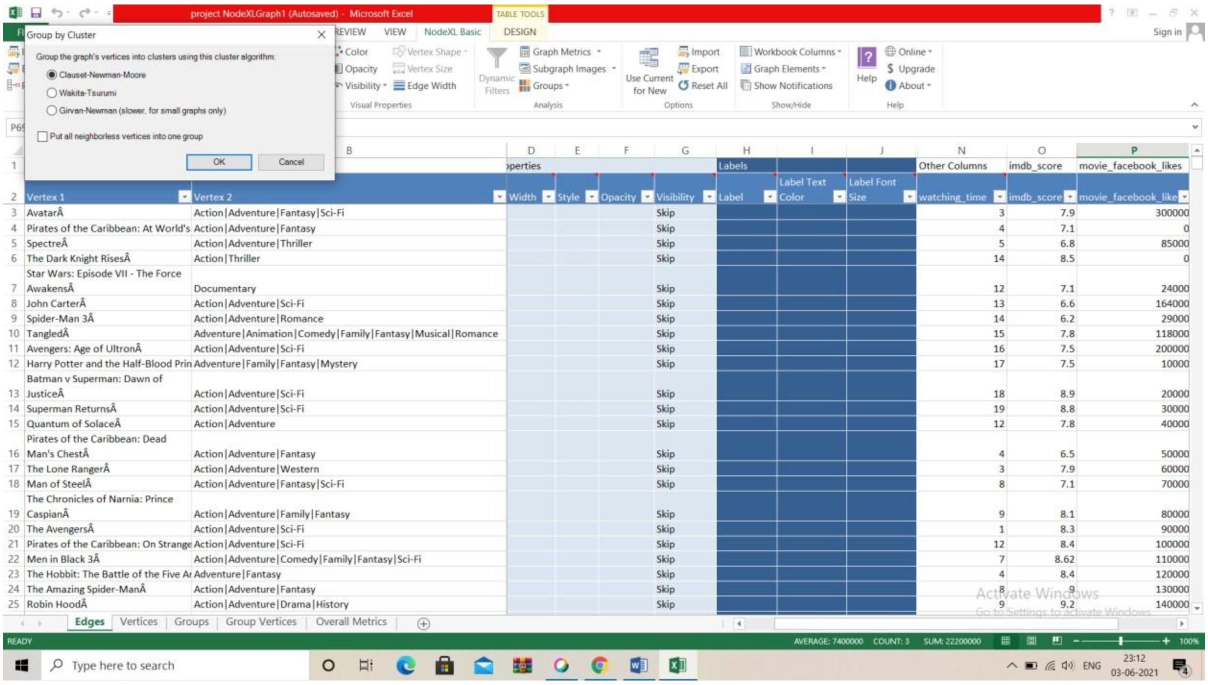
i. identifying the nodes that are shared by two or more groups.

ii. Determining the degree to which a node is linked to a specific community.

SCREENSHOTS



	A	B	D	E	F	G	H	I	J	N	O	P
1	movie_title	genres	›perties				Labels			Other Columns	imdb_score	movie_facebook_likes
2	Vertex 1	Vertex 2	Width	Style	Opacity	Visibility	Label	Label Text Color	Label Font Size	watching_time	imdb_score	movie_facebook_like
3	AvatarÂ	Action|Adventure|Fantasy|Sci-Fi				Skip				3	7.9	300000
4	Pirates of the Caribbean: At World's	Action|Adventure|Fantasy				Skip				4	7.1	0
5	SpectreÂ	Action|Adventure|Thriller				Skip				5	6.8	85000
6	The Dark Knight RisesÂ	Action|Thriller				Skip				14	8.5	0
7	Star Wars: Episode VII - The Force AwakensÂ	Documentary				Skip				12	7.1	24000
8	John CarterÂ	Action|Adventure|Sci-Fi				Skip				13	6.6	164000
9	Spider-Man 3Â	Action|Adventure|Romance				Skip				14	6.2	29000
10	TangledÂ	Adventure|Animation|Comedy|Family|Fantasy|Musical|Romance				Skip				15	7.8	118000
11	Avengers: Age of UltronÂ	Action|Adventure|Sci-Fi				Skip				16	7.5	200000
12	Harry Potter and the Half-Blood Prin	Adventure|Family|Fantasy|Mystery				Skip				17	7.5	10000
13	Batman v Superman: Dawn of JusticeÂ	Action|Adventure|Sci-Fi				Skip				18	8.9	20000
14	Superman ReturnsÂ	Action|Adventure|Sci-Fi				Skip				19	8.8	30000
15	Quantum of SolaceÂ	Action|Adventure				Skip				12	7.8	40000
16	Pirates of the Caribbean: Dead Man's ChestÂ	Action|Adventure|Fantasy				Skip				4	6.5	50000
17	The Lone RangerÂ	Action|Adventure|Western				Skip				3	7.9	60000
18	Man of SteelÂ	Action|Adventure|Fantasy|Sci-Fi				Skip				8	7.1	70000
19	The Chronicles of Narnia: Prince CaspianÂ	Action|Adventure|Family|Fantasy				Skip				9	8.1	80000
20	The AvengersÂ	Action|Adventure|Sci-Fi				Skip				1	8.3	90000
21	Pirates of the Caribbean: On Strange	Action|Adventure|Sci-Fi				Skip				12	8.4	100000
22	Men in Black 3Â	Action|Adventure|Comedy|Family|Fantasy|Sci-Fi				Skip				7	8.62	110000
23	The Hobbit: The Battle of the Five Ar	Adventure|Fantasy				Skip				4	8.4	120000
24	The Amazing Spider-ManÂ	Action|Adventure|Fantasy				Skip				8	9	130000
25	Robin HoodÂ	Action|Adventure|Drama|History				Skip				9	9.2	140000



Group by Cluster

Group the graph's vertices into clusters using this cluster algorithm:

● Clauset-Newman-Moore

○ Wakita-Tsurumi

○ Girvan-Newman (slower, for small graphs only)

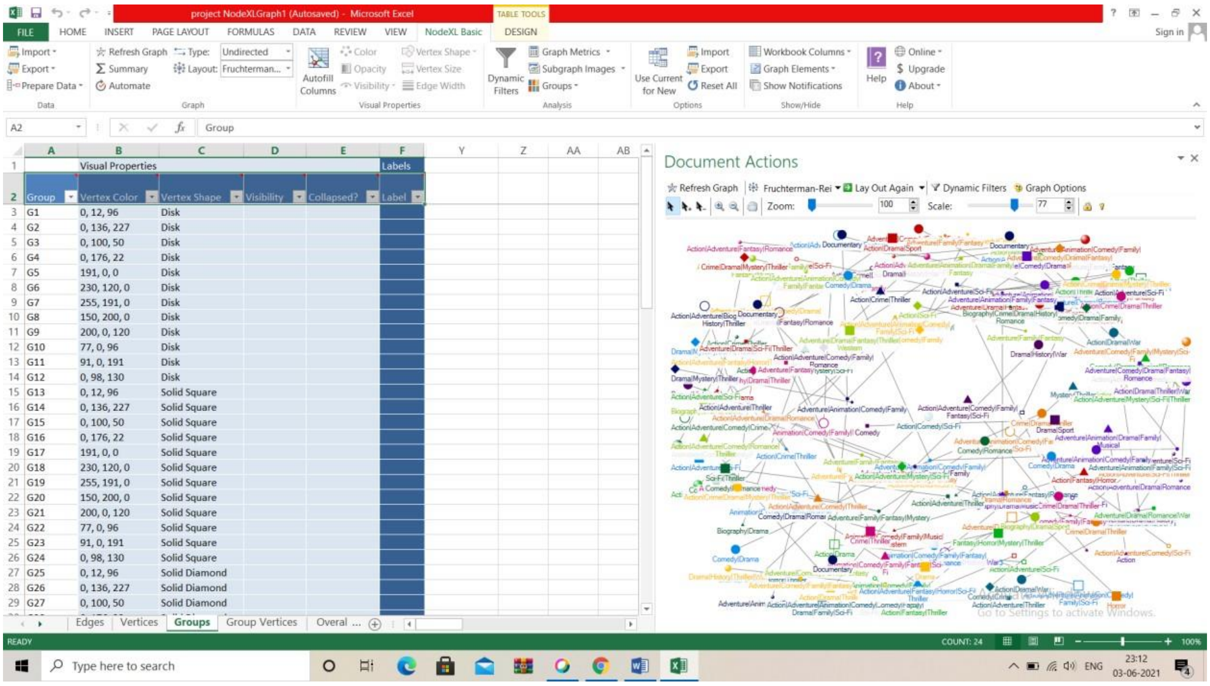
□ Put all neighborless vertices into one group

OK | Cancel

	A	B	D	E	F	G	H	I	J	N	O	P
1			perties				Labels			Other Columns	imdb_score	movie_facebook_likes
2	Vertex 1	Vertex 2	Width	Style	Opacity	Visibility	Label	Label Text Color	Label Font Size	watching_time	imdb_score	movie_facebook_like
3	AvatarÂ	Action|Adventure|Fantasy|Sci-Fi				Skip				3	7.9	300000
4	Pirates of the Caribbean: At World's	Action|Adventure|Fantasy				Skip				4	7.1	0
5	SpectreÂ	Action|Adventure|Thriller				Skip				5	6.8	85000
6	The Dark Knight RisesÂ	Action|Thriller				Skip				14	8.5	0
7	Star Wars: Episode VII - The Force AwakensÂ	Documentary				Skip				12	7.1	24000
8	John CarterÂ	Action|Adventure|Sci-Fi				Skip				13	6.6	164000
9	Spider-Man 3Â	Action|Adventure|Romance				Skip				14	6.2	29000
10	TangledÂ	Adventure|Animation|Comedy|Family|Fantasy|Musical|Romance				Skip				15	7.8	118000
11	Avengers: Age of UltronÂ	Action|Adventure|Sci-Fi				Skip				16	7.5	200000
12	Harry Potter and the Half-Blood Prin	Adventure|Family|Fantasy|Mystery				Skip				17	7.5	10000
13	Batman v Superman: Dawn of JusticeÂ	Action|Adventure|Sci-Fi				Skip				18	8.9	20000
14	Superman ReturnsÂ	Action|Adventure|Sci-Fi				Skip				19	8.8	30000
15	Quantum of SolaceÂ	Action|Adventure				Skip				12	7.8	40000
16	Pirates of the Caribbean: Dead Man's ChestÂ	Action|Adventure|Fantasy				Skip				4	6.5	50000
17	The Lone RangerÂ	Action|Adventure|Western				Skip				3	7.9	60000
18	Man of SteelÂ	Action|Adventure|Fantasy|Sci-Fi				Skip				8	7.1	70000
19	The Chronicles of Narnia: Prince CaspianÂ	Action|Adventure|Family|Fantasy				Skip				9	8.1	80000
20	The AvengersÂ	Action|Adventure|Sci-Fi				Skip				1	8.3	90000
21	Pirates of the Caribbean: On Strange	Action|Adventure|Sci-Fi				Skip				12	8.4	100000
22	Men in Black 3Â	Action|Adventure|Comedy|Family|Fantasy|Sci-Fi				Skip				7	8.62	110000
23	The Hobbit: The Battle of the Five Ar	Adventure|Fantasy				Skip				4	8.4	120000
24	The Amazing Spider-ManÂ	Action|Adventure|Fantasy				Skip				8	9	130000
25	Robin HoodÂ	Action|Adventure|Drama|History				Skip				9	9.2	140000

AVERAGE: 7400000 COUNT: 3 SUM: 22200000 100%

23:12
03-06-2021



Type: Undirected

Layout: Fruchterman...

A2 Group

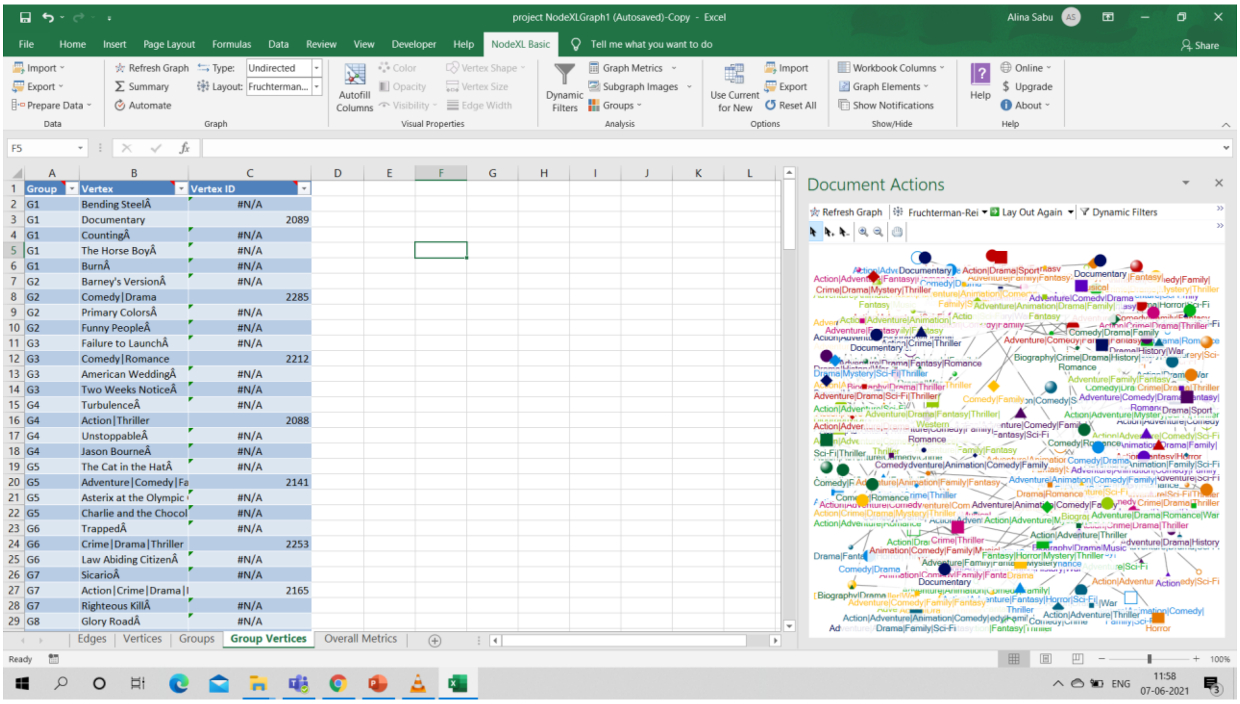
Fruchterman-Rei

Zoom: 100

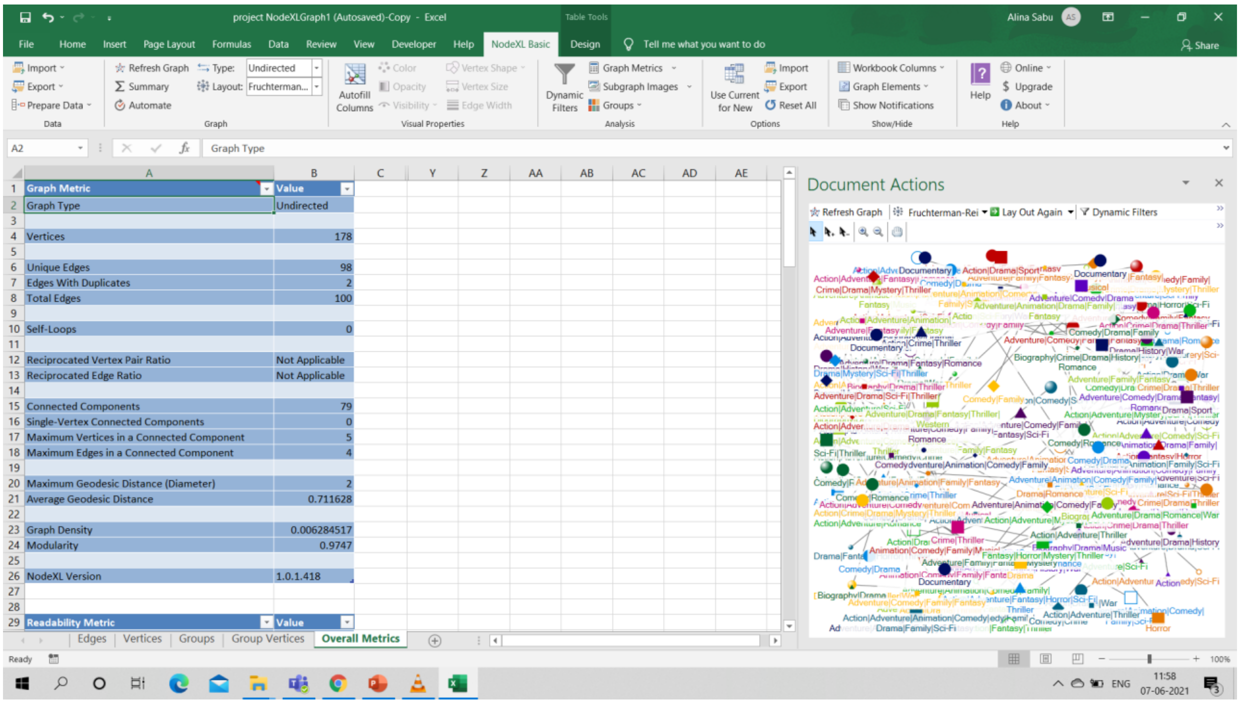
Scale: 77

	A	B	C	D	E	F
1		Visual Properties				Labels
2	Group	Vertex Color	Vertex Shape	Visibility	Collapsed?	Label
3	G1	0, 12, 96	Disk			
4	G2	0, 136, 227	Disk			
5	G3	0, 100, 50	Disk			
6	G4	0, 176, 22	Disk			
7	G5	191, 0, 0	Disk			
8	G6	230, 120, 0	Disk			
9	G7	255, 191, 0	Disk			
10	G8	150, 200, 0	Disk			
11	G9	200, 0, 120	Disk			
12	G10	77, 0, 96	Disk			
13	G11	91, 0, 191	Disk			
14	G12	0, 98, 130	Disk			
15	G13	0, 12, 96	Solid Square			
16	G14	0, 136, 227	Solid Square			
17	G15	0, 100, 50	Solid Square			
18	G16	0, 176, 22	Solid Square			
19	G17	191, 0, 0	Solid Square			
20	G18	230, 120, 0	Solid Square			
21	G19	255, 191, 0	Solid Square			
22	G20	150, 200, 0	Solid Square			
23	G21	200, 0, 120	Solid Square			
24	G22	77, 0, 96	Solid Square			
25	G23	91, 0, 191	Solid Square			
26	G24	0, 98, 130	Solid Square			
27	G25	0, 12, 96	Solid Diamond			
28	G26	0, 136, 227	Solid Diamond			
29	G27	0, 100, 50	Solid Diamond			

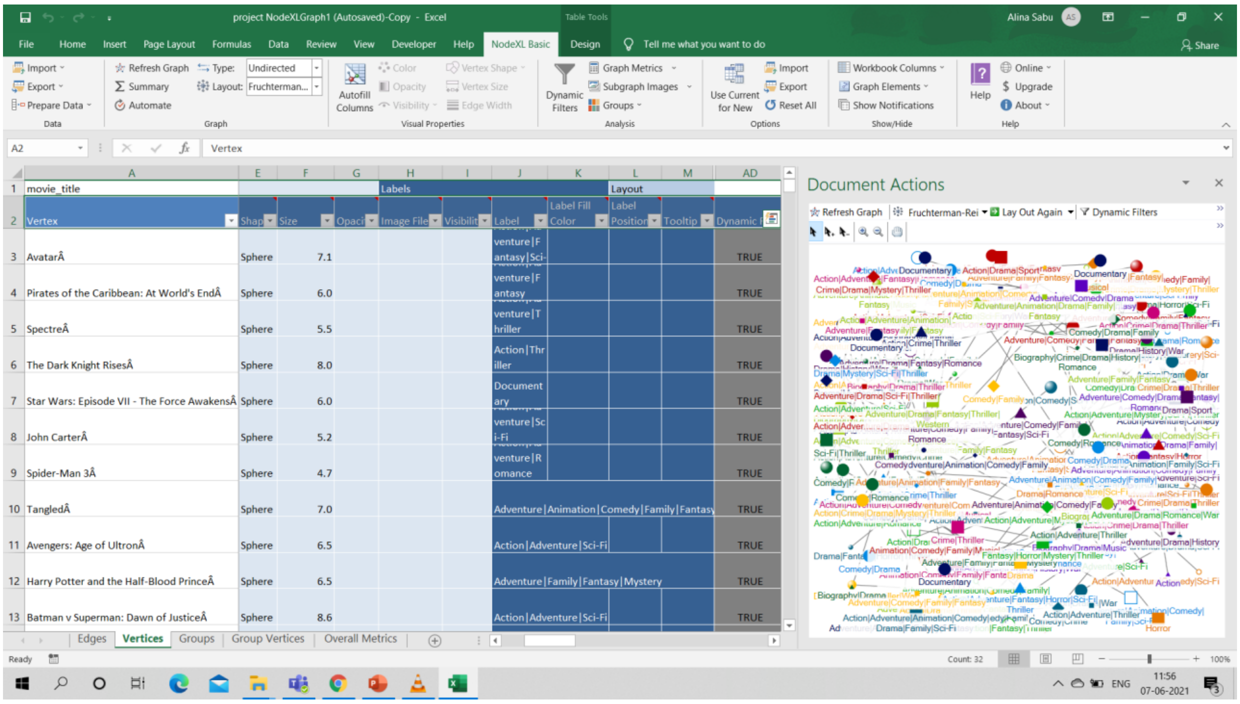
23:12
03-06-2021



	A	B	C
1	Group	Vertex	Vertex ID
2	G1	Bending SteelÂ	#N/A
3	G1	Documentary	2089
4	G1	CountingÂ	#N/A
5	G1	The Horse BoyÂ	#N/A
6	G1	BurnÂ	#N/A
7	G2	Barney's VersionÂ	#N/A
8	G2	Comedy|Drama	2285
9	G2	Primary ColorsÂ	#N/A
10	G2	Funny PeopleÂ	#N/A
11	G3	Failure to LaunchÂ	#N/A
12	G3	Comedy|Romance	2212
13	G3	American WeddingÂ	#N/A
14	G3	Two Weeks NoticeÂ	#N/A
15	G4	TurbulenceÂ	#N/A
16	G4	Action|Thriller	2088
17	G4	UnstoppableÂ	#N/A
18	G4	Jason BourneÂ	#N/A
19	G5	The Cat in the HatÂ	#N/A
20	G5	Adventure|Comedy|Fa	2141
21	G5	Asterix at the Olympic	#N/A
22	G5	Charlie and the Chocol	#N/A
23	G6	TrappedÂ	#N/A
24	G6	Crime|Drama|Thriller	2253
25	G6	Law Abiding CitizenÂ	#N/A
26	G7	SicarioÂ	#N/A
27	G7	Action|Crime|Drama|I	2165
28	G7	Righteous KillÂ	#N/A
29	G8	Glory RoadÂ	#N/A



	A	B
1	Graph Metric	Value
2	Graph Type	Undirected
3		
4	Vertices	178
5		
6	Unique Edges	98
7	Edges With Duplicates	2
8	Total Edges	100
9		
10	Self-Loops	0
11		
12	Reciprocated Vertex Pair Ratio	Not Applicable
13	Reciprocated Edge Ratio	Not Applicable
14		
15	Connected Components	79
16	Single-Vertex Connected Components	0
17	Maximum Vertices in a Connected Component	5
18	Maximum Edges in a Connected Component	4
19		
20	Maximum Geodesic Distance (Diameter)	2
21	Average Geodesic Distance	0.711628
22		
23	Graph Density	0.006284517
24	Modularity	0.9747
25		
26	NodeXL Version	1.0.1.418
27		
28		
29	Readability Metric	Value



A2 | Vertex

	A	E	F	G	H	I	J	K	L	M	AD
1	movie_title				Labels				Layout		
2	Vertex	Shap	Size	Opaci	Image File	Visibilit	Label	Label Fill Color	Label Position	Tooltip	Dynamic F
3	AvatarÂ	Sphere	7.1				venture|F antasy|Sci-				TRUE
4	Pirates of the Caribbean: At World's EndÂ	Sphere	6.0				venture|F antasy				TRUE
5	SpectreÂ	Sphere	5.5				venture|T hriller				TRUE
6	The Dark Knight RisesÂ	Sphere	8.0				Action|Thr iller				TRUE
7	Star Wars: Episode VII - The Force AwakensÂ	Sphere	6.0				Document ary				TRUE
8	John CarterÂ	Sphere	5.2				venture|Sc i-Fi				TRUE
9	Spider-Man 3Â	Sphere	4.7				venture|R omance				TRUE
10	TangledÂ	Sphere	7.0				Adventure|Animation|Comedy|Family|Fantasy				TRUE
11	Avengers: Age of UltronÂ	Sphere	6.5				Action|Adventure|Sci-Fi				TRUE
12	Harry Potter and the Half-Blood PrinceÂ	Sphere	6.5				Adventure|Family|Fantasy|Mystery				TRUE
13	Batman v Superman: Dawn of JusticeÂ	Sphere	8.6				Action|Adventure|Sci-Fi				TRUE

CONCLUSION

We observe from the above graph by applying the Clauset Newman-Moore algorithm. In this graph we can see different shapes for the nodes. Each shape corresponds to a specific genre. For example dark red square corresponds to a movie belonging to action|drama|sport , dark blue diamond stands for drama|mystery|thriller, yellow stand for action|crime|mystery|drama|thriller and so on and so forth. Each of these nodes are connected to another node which belongs to approximately the same category. This helps us to recommend to groups of users the movies that are similar to these genres. In the above resultant graph , the shape of the nodes and the vertex are based on the imdb score. The labels on the nodes correspond to the genre of the films. These genres can be seen under the column vertex2 under the edges sheet of the nodexl template. In the above graph we have displayed the genres of the movies having an imdb score of 9 and above. We went for an imdb score of 9 as we wanted to suggest some of the best movies to the users.This parameter can be changed based on the user’s interest.

TEAMS CONTRIBUTION:-

SHILPA REJI(20MAI1010) – Application of clause newman algorithm to cluster the groups based on genres. These movies can be recommended to the users. The nodes are displayed using different shapes and these correspond to the genres of the films.

ALINA MARY SABU(20MAI1022) – Displaying labels to the graph by exploring the autofill column option. Used options such vertex size and vertex shape. Displayed nodes having an imdb score above 9.

REFERENCES

[1] Zafarani, Reza, Mohammad Ali Abbasi, and Huan Liu. Social media mining: an introduction. Cambridge University Press, 2014

[2] Campan, Alina, Yasmeen Alufaisan, and Traian Marius Truta. "Community Detection in Anonymized Social Networks." EDBT/ICDT Workshops. 2014

[3] Fortunato, Santo. "Community detection in graphs." Physics reports 486.3-5 (2010): 75-174.

[4] Rosvall, Martin, et al. "Different approaches to community detection." arXiv preprint arXiv:1712.06468 (2017).

[5] Lin, Shen, and Brian W. Kernighan. "An effective heuristic algorithm for the traveling-salesman problem." Operations research 21.2 (1973): 498-516

[6] Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." nature 435.7043 (2005): 814.

[7] Cazabet, Remy, and Frédéric Amblard. "Dynamic community detection." Encyclopedia of Social Network Analysis and Mining. Springer New York, 2014. 404-414.

[8] Gulbahce, Natali, and Sune Lehmann. "The art of community detection." BioEssays 30.10 (2008): 934-938.

[9] Papadopoulos, Symeon, et al. "Community detection in social media." Data Mining and Knowledge Discovery 24.3 (2012): 515-554.

[10] Malliaros, Fragkiskos D., and Michalis Vazirgiannis. "Clustering and community detection in directed networks: A survey." Physics Reports 533.4 (2013): 95-142.

[11] Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the national academy of sciences 99.12 (2002): 7821-7826.

[12] Lancichinetti, Andrea, Santo Fortunato, and Filippo Radicchi. "Benchmark graphs for testing community detection algorithms." Physical review E 78.4 (2008): 046110

[13] Karrer, Brian, and Mark EJ Newman. "Stochastic blockmodels and community structure in networks." Physical review E 83.1 (2011): 016107.