Review Article

A Comparative Analysis of Community Detection Agglomerative Technique Algorithms and Metrics on Citation Network

Sandeep Kumar Rachamadugu[1,2,*] and Pushphavathi Thotadara Parameshwarappa[1]

[1]M.S. Ramaiah University of Applied Sciences, Bangalore, India
sandy.racha@gmail.com; pushphavathi.cs.et@msruas.ac.in
[2]G. Pulla Reddy Engineering College, Kurnool, Affiliated to JNTUA, Ananthapuramu, AP, India
sandeep.cse@gprec.ac.in
*Correspondence: sandy.racha@gmail.com

Received: 3rd April 2023; Accepted: 7th August 2023; Published: 1st October 2023

Abstract: Social Network Analysis is a discipline that represents social relationships as a network of nodes and edges. The construction of social network with clusters will contribute in sharing the common characteristics or behaviour of a group. Partitioning the graph into modules is said to be a community. Communities are meant to symbolize actual social groups that share common characteristics. Citation network is one of the social networks with directed graphs where one paper will cite another paper and so on. Citation networks will assist the researcher in choosing research directions and evaluating research impacts. By constructing the citation networks with communities will direct the user to identify the similarity of documents which are interrelated to one or more domains. This paper introduces the agglomerative technique algorithms and metrics to a directed graph which determines the most influential nodes and group of similar nodes. The two stages required to construct the communities are how to generate network with communities and how to quantify the network performance. The strength and a quality of a network is quantified in terms of metrics like modularity, normalized mutual information (NMI), betweenness centrality, and F-Measure. The suitable community detection techniques and metrics for a citation graph were introduced in this paper. In the field of community detection, it is common practice to categorize algorithms according to the mathematical techniques they employ, and then compare them on benchmark graphs featuring a particular type of assortative community structure. The algorithms are applied for a sample citation sub data is extracted from DBLP, ACM, MAG and some additional sources which is taken from and consists of 101 nodes (n_c) with 621 edges € and formed 64 communities. The key attributes in dataset are id, title, abstract, references SLM uses local optimisation and scalability to improve community detection in complicated networks. Unlike traditional methods, the proposed LS-SLM algorithm is identified that the modularity is increased by 12.65%, NMI increased by 2.31%, betweenness centrality by 3.18% and F-Score by 4.05%. The SLM algorithm outperforms existing methods in finding significant and well-defined communities, making it a promising community detection breakthrough.

Keywords: *Citation Network; Community Detection; Directed Graph; Modularity; SLM*

1. Introduction

Clusters of nodes which is a part of the same group will have many edges between them, while those belonging to other communities will have fewer. Researchers focus on finding communities of interest in citation networks. The approach allows us to refine the network and reveal its community formations.

Community detection analyses topology, finds hidden rules, and predicts behaviour [1]. Each document in a citation network acts as a vertex in a graph, and the edges represent the citations that each

Sandeep Kumar Rachamadugu and Pushphavathi Thotadara Parameshwarappa, "A Comparative Analysis of Community Detection Agglomerative Technique Algorithms and Metrics on Citation Network", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 1-13, Vol. 7, No. 4, 1st October 2023, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2023.04.001, Available: http://aetic.theiaer.org/archive/v7/v7n4/p1.html.

document has to others. Citation graph is the directed graph where all the references of respective articles act as an in-link to that article. Each paper in this graph serves as a node, and the citations between them are shown as edges. Citation networks can improve community-based recommender systems using social network analysis. If communities are identified for citation networks, the nodes in the directed graph state that the source node influences over target node. Maximizing the paper cited network's influence can help new works reach more scholars faster. Finding the most influential nodes in a network is a challenging problem, when there are lot of nodes involved. The major task is to discover which users in large networks have the most significant impact [2].

The evolution of the citation network has an impact on the process of locating articles that are comparable and provides recommendations for locating appropriate articles personalized to the user's individual profile and preferences. Most of the areas fall under the category of social network which is a complex network to find the similarities among nodes or to share some information from the nodes. The objective of social network analysis is to map and examine the connections between groups of people. The links in the social network or graph are directed or undirected. From figure (1) Citation network is a directed graph where network that can be visualized as a directed acyclic graph with nodes that represent papers from $\boldsymbol{A_1}, \boldsymbol{A_2}, \boldsymbol{A_3}, \dots \dots, \boldsymbol{A_i}$ and edges that show a co-citation relationship between two nodes when set of papers cites paper $\boldsymbol{B_1}$ **[3].**

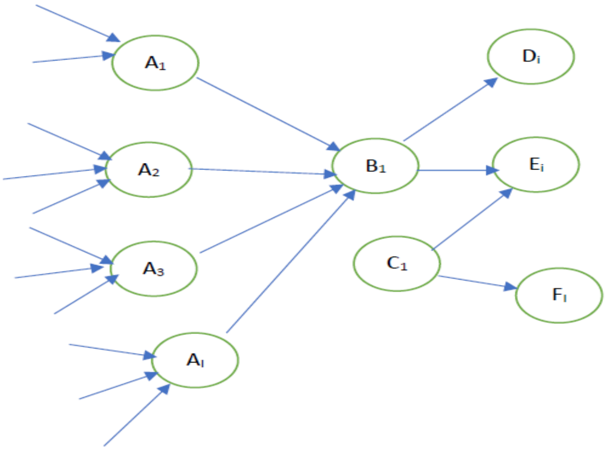

Figure 1. Citation Network

The network topology is used to understand network infrastructure, identify interesting insights, predict links, and broaden the area. Social network analysis analyses network data to reveal underlying community structures. The goal of community detection is to divide a graph into communities made up of tightly interconnected nodes and other, less closely-knit groups. The purpose of graph community detection is to locate substructures such as modules and hierarchies using only the topological data available in the graph. Communities have a proper understanding of how the network is organized. Communities have been identified to determine whether or not their publications are consistent with expectations for that community [4].

Communities in citation networks are used to identify scholars with similar research interests and to represent groups of papers that share a common topic [5]. With the increase in size and complexity of complex networks, it is essential to understand the related literature and key findings. The Louvain algorithm is broadly recognised for its efficiency in detecting communities in large-scale networks, leveraging a modularity optimization approach. The Fast-Greedy algorithm, on the other hand, employs a hierarchical agglomerative technique, iteratively merging communities based on modularity gain. In contrast, the SLM algorithm introduces a novel approach that combines local optimization and scaling techniques, offering enhanced accuracy and efficiency in community detection. While the Louvain algorithm and Fast Greedy algorithm have been extensively studied, the SLM algorithm presents a novel perspective by incorporating local optimization and scaling, promising improved detection of meaningful and well-defined communities. This study aims to provide a comparative analysis of these algorithms to assess their performance and efficacy in uncovering community structures in diverse networks. These techniques mostly optimize objective functions. Modularity optimization is one of them. Modularity optimization is one of them. Optimizing modularity is NP-hard [6]. In order to reveal the hidden structure, the main areas of study are from a complex network and how to quantify its characteristics. Community detection is used to figure out the underlying architecture and features of large and complex networks [7]. The motivation for generating community detection for a citation network is to discover knowledge, research trend analysis, concerned areas, co-authorship, and collaboration. Highly cited papers within a community can suggest influential research and recommendation systems, and publications within the same community tend to have comparable citation patterns.

The contribution of this paper is to identify the communities of a citation network which will be useful for the user to collect and acquire knowledge on a specific domain. In order to do so, it is crucial that the network be of appropriate strength. Section 2 discusses about the literature review about the background work involved in identifying the community detection and the difficulties identified at each stage. Section 3 introduces the agglomerative technique algorithms and metrics for a community, in Section4 the proposed algorithm is introduces to overcome the difficulty or to improve the performance of a community. In Section 5, results and discussion where compared the entire traditional algorithm with proposed algorithm and Section 6 and 7 exhibit the conclusions and future scope, respectively.

2. Literature Review

2.1. Background Work

The graph partitioning algorithm is used to extract clusters of important terms and phrases that form latent communities to uncover social topics. In directed graph, the directionality from source to destination represents that the source node transmits information to the target node and is asymmetric in nature. For directed graphs, to solve the clustering problem the spectrum algorithm uses a Laplacian matrix with a mixing rate of random walks. The various methodologies in constructing the directed graph are modularity, spectral based clustering cu-based measures, page rank and random walk-based methods and clique percolation method and local density clustering [5].

Node attributes include user network profiles, author articles, and articles published histories. This helps find similar nodes and the node's community module. Nodes reveal friendships, author collaborations, followers, and topic interactions.

2.1.1. Newman Girvan Method

In this, the communities are formed based on the Girvan method and link analysis. Newman Girvan Method is a divisive hierarchical clustering technique. This algorithm automatically generates the number of communities, i.e. no need to give implicitly. The algorithm works by removing edges and recomputing at each and every step based on the betweenness centrality. The time complexity is very high, i.e. O (mn) with 'm' nodes and 'n' edges [8]. The author proposed a divisive clustering using Girvan and Newman method and an edge betweenness as a similarity measure. In this, the loose similarity and cosine similarity are computed first between each pair of nodes, and this is the same process iteratively performed. Once a certain number of edges have been removed from a graph, the extraction of communities is complete, and the resulting structure is considered definite. In that case, the iteration process starts over from the beginning. In contrast to divisive hierarchical clustering, in which edges connecting nodes with low similarity are removed, and there is no prior guarantee that the removed inter-cluster edges connect nodes with least similarity. In this, the author introduces a divisive clustering using Girvan and Newman method and an edge betweenness as a similarity measure. In this, the loose similarity and cosine similarity is computed first between each pair of nodes and this is the same process iteratively performed.

Once a certain number of edges have been removed from a graph, the phase of extracting communities has ended. In that case, the iteration process starts over again from the beginning. However, unlike divisive hierarchical clustering, in which edges between pairs of vertices with low similarity are removed, in this case inter-cluster edges are removed, and there is no prior guarantee that inter-cluster edges connect vertices with minimal similarity. Instead of removing a single edge, it may be necessary to remove an entire vertex or sub graph [9]. The time complexity of Newmann Girvan Method is O (m^2n) where m is the number of edges and n is the total number of nodes. Only used for maximum of 10,000 nodes.

2.1.2. The Kernighan-Lin algorithm

Partitions of arbitrary size can be extracted using an extension of the Kernighan-Lin algorithm, but the algorithm's runtime and storage requirements grow as the number of clusters. Using the Laplacian matrix's spectral properties, a method called spectral bisection is developed. The benefit of hierarchical clustering is that it does not demand prior knowledge on the number and size of the clusters. Spectral clustering, where nodes are correctly expressed as the product of eigenvectors of the connectivity matrix. In this way, we can reframe the community detection technique as one of data mining's clustering challenges [10].

In agglomerative technique algorithms, initially one node is collected and later the nodes are combined based on the similarity of nodes using similarity-based metrics. The disadvantage with Agglomerative algorithm is i) It is not always possible to correctly classify a community's vertices, and sometimes important vertices are ignored even when they are present.

The difficulty of scaling up to very large datasets is a major drawback of agglomerative hierarchical clustering. To use distance as a measure of dissimilarity among points, they must be embedded in space, which increases the computational complexity to O (n^2) for a single linkage and O (n^2 log n) for the total and average linkage schemes. Spectral clustering, where nodes are represented as the product of eigenvectors of the connectivity matrix. In this way, we can reframe the community detection technique as one of data mining's clustering challenges**.**

3. Comparative Analysis

In order to adapt community detection techniques to a word-based graph structure, scientists developed algorithms for discovering groups that consistently discuss shared interests [11]. This section describes the methodologies like a) Louvain Algorithm, b) Fast Greedy Algorithm, c) Stochastic Block Model, and d) Smart Local Moving algorithm and metrics like i) Modularity ii) Normalized Mutual Information (NMI) iii) Betweenness Centrality and iv) Purity and F-Measure, which are suitable in identifying the communities for a directed graph.

3.1. Existing Contrast Techniques

3.1.1. Louvain Algorithm

This algorithm comprises of two phases, namely the modularity optimization phase and the community aggregation phase. The Louvain method is a greedy heuristic algorithm that finds small communities by locally maximizing modularity. Until Q is maximized, and a community hierarchy is formed, this procedure repeats itself [12]. The working environment of Louvain algorithm is explained as follows.

Step1. Initialize each and every node as one community.

Step2. Find all the communities that are linked to node 1, and the transition in modularity can then be computed after node 2 has been moved to each of the neighbouring communities. Transfer node 1 to the community, which will maximise modularity. i.e. allowing only local changes to node-community memberships optimizes modularity.

Step3. Repeat this process for each node and perform step2. As long as there are no nodes to relocate, community partition will not occur.

Step4. In step 3, each community should be turned into a new node. The connections between the node are identical to the connections between the earlier communities. The recognized communities are grouped into super nodes to create a network.

Continue with step 1 until all nodes form one community. The optimal modularity partition is the result of a multilevel community partition [13].

The Louvain community detection algorithm is a method for finding communities in networks that maximizes modularity. This method can quickly and efficiently detect communities in massive networks. It's Sci2's resolution parameter customizes community detection granularity. Existing approaches for community detection aren't limited to a directed weighted network. Most existing modularity methods or algorithms have a resolution limit [14].

This algorithm is an approximation algorithm, and it does not return perfect results. There are two pitfalls using this algorithm namely resolution and randomness. The time complexity of Louvain technique is O (nlogn). This method supports for size of a network up to 100 million nodes and billions of links.

3.1.2. Fast Greedy Technique

Decomposition from bottom up iteratively merges two communities to achieve maximum modularity at local optimal. Fast greedy detects modular communities efficiently. This strategy uses a subnetwork of highly connected nodes. The algorithm adds random links that strengthen the sub network's modularity. This is repeated until modularity improves. The sub network's connected components determine the communities. This procedure takes O(n^2) or O(n+m) time, where n refers to the sum of nodes and m is the sum of linkages. Greedy algorithm maximizes modularity at each step [15].

Step1. At first, every node is part of a different community;

Step2. The two nodes or communities that, when merged, enhance modularity and most of them become members of the same community.

Step3. Repeat Step 2 until one community remains.

Although modularity is mathematically suitable and accurate for community detection, there are two pitfalls, namely resolution limit and modularity maxima.

3.1.3. Stochastic Block Model (SBM)

Stochastic block models are a type of random graph model studied in the social sciences and computer science [16]**.** SBM can cluster and discover the latent network structure. SBMs are related to latent space models and community detection.

First, a block model section divides nodes into group membership vectors (Vector membership).

In the second step, a block matrix with each edge representing the probability of two nodes is created.

0.8 edge probability if both nodes belong to the same group, 0.05 otherwise.

To introduce G= (N, E), where N is the n-node set and E is the M-edge list. For the directed graph G, Y_{pq}=1(0) represents a p to q edge. Undirected graph$Y_{pq} = Y_{qp}$, but directed graphs are independent. SBM groups every node (K). Z= $(Z_1, Z_2, Z_3, \ldots \ldots Z_n)^{\mathrm{T}}$ where Z_{pi} is the i^{th} element of Z_p.N=($N_1, N_2, N_3, \ldots \ldots N_k)^{\mathrm{T}}$ denotes the size of each group; Z and Y derive the K*K edge matrix between groups. E_{ij} denotes the amount of connections that exist between the groupings i and j.

Assuming they all belong to the same group, Z, the block densities in matrix C are conditionally independent. The Y_{pq} pair value is computed from the probable outcomes from the Bernoulli distribution $Z_p^T C Z_q$.

The concept called assertiveness where the links with in the cluster should have high density than the links which are connected outside to this means C_{ii} (i = 1, 2. . . K) is high while C_{ij} is low for j≠i. Instead of modularity the SBM model uses modularity density which solves the hurdle of resolution limit [17]**.**

3.1.4. Smart Local Moving (SLM)

In the second stage of the SLM, the Louvain algorithm is applied to a cluster of neighbouring communities. Throughout the phase 2 process, every small community will function as a network vertex. Modularity is improved by splitting the network and relocating nodes when the SLM method is executed repeatedly. More SLM algorithm iterations can improve the community structure.

The Louvain algorithm has high execution efficiency, unsupervised and easy to implement. Even though the algorithm has a low time complexity, it can only produce a near-optimal community layout [18]. SLM is derived from the Louvain algorithm. The SLM algorithm is run over and over again, and the likelihood of increasing modularity is always being looked for by splitting the community into smaller groups and moving the nodes from one group to another. Because of this, the community structure can always be improved by doing more repetitions of the method in the SLM algorithm. Specifically, SLM maximizes modularity by partitioning communities and relocating nodes among them. The SLM Algorithm derives from the Louvain algorithm by changing the second stage i.e., at reduced network stage but the SLM algorithm changes the step of building a smaller network by doing the following:

i) Iterates over all first-step communities. Each community is copied into its own sub network.

ii) After assigning each sub network node to a singleton community, it uses local moving optimization.

SLM creates a reduced network after the local moving which creates a community structure for each sub network, with sub network communities as nodes. SLM divides networks into smaller communities. Each node joins the community of the subnetwork. Each sub network has a defined community and detects sub network communities become nodes in the reduced network. The iteration continues until no further reduction is possible [19].

All the above algorithms are the existing algorithms. The proposed algorithm, which is an extension of SLM algorithm that moves a node from one community to another, improves modularity for each node. After iterations, the lowest and highest modularity observations are recorded. Iterate until no node affects the modularity change [20].

3.2. Evaluation Measurements

Each and every metric it has its own specification. The following are the metrics for community detection introduced here:

3.2.1. Modularity

In a network, the degree of modularity indicates the strength of the connections between nodes. The modularity can be increased when a node is combined into a community. So, the gain in modularity for directed networks is computed in the equation (1) as

$$\varDelta Qd = \frac{d_i^c}{m} - \frac{d_i^{in}.\varSigma_{Out}^{in}+d_i^{in}.\varSigma_{tot}^{Out}}{m^2} \tag{1}$$

Where d_i^c means how much a node i is involved in the network C, $\sum_{in}$ the number of edges confined in community C and $\sum_{out}$the total number of edges incident to community C. degree of vertex i, m=sum of edges. d_i^{in}stands for in-degree of i and 'm' is the sum of edges, $\sum_{tot}^{in}(\sum_{tot}^{out})$ represents the number of in-going arcs incident to C [21].

Modularity measures how many edges were actually present against how many were predicted. Another definition for modularity for directed networks is given in equation (2) as

$$Q = \sum_{c_{i\in C}}\left[\left(\frac{|E_{ct}^{in}|}{|E|} - \frac{(|E_{ct}^{in}|+|E_{out,ct}|)(|E_{ct}^{in}|+|E_{ct,out}|)}{|E|^2}\right)\right] \tag{2}$$

Where |E| total of the weights carried by each and every edge in the network, $|E_{ct}^{in}|$is equal to the sum of the weights of all of the edges that connect the nodes that make up the community c_i , and $|E_{ct}^{out}|$ is the sum of the weights of the edges connecting the nodes that are part of community c_i to the nodes that are not part of community c_i , $|E_{out,ct}|$ is the total of the edges from outside c_i to c_i and $|E_{ct,out}|$ counts the edges from community c_i to outside nodes [22].

Modularity evaluates the quality of network communities. The proportion of edges that fall within specified groups compared to the fraction predicted if the edges were distributed at random. A network with high modularity has sparse links between nodes in different communities but dense links between nodes within a community. The range of the Modularity Q measure is 0 to 1. The modularity report using the Gephi tool, a network visualization tool, generates the following result: From the graph, it was observed that there are 64 communities formed, where each and every community holds the number of nodes [23] [24].

The sample result obtained in figure (2) from the Gephi tool has a modularity=0.975, and number of communities of 64. The Gephi tool's modularity report details the size distribution and total sum of nodes within the identified communities in the network. The modularity report about size distribution may provide community sizes. The minimum, maximum, average, and perhaps median and standard deviation are provided. This shows community sizes and node count outliers. The modularity report usually lists the overall number of nodes and how many are assigned to each community. This information shows the proportion of nodes given to each community and the community size balance. Large communities: Communities with many nodes may suggest cohesive groups or strongly interconnected locations.

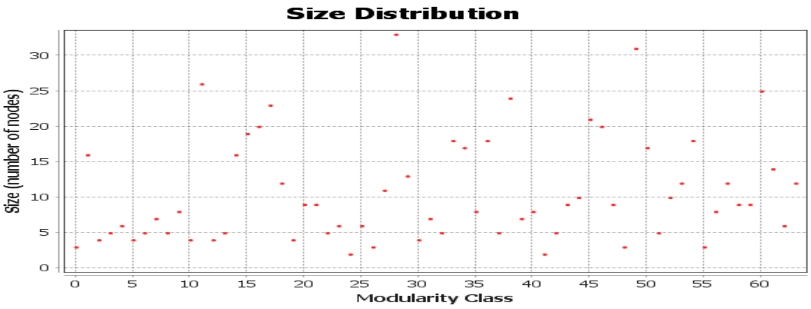

Figure 2. Modularity Report

A Citation network with 101 nodes with 621 edges is converted into 64 communities where each community consists of some nodes that have similarity in characteristics or behaviour. It is given in the following figure (3).

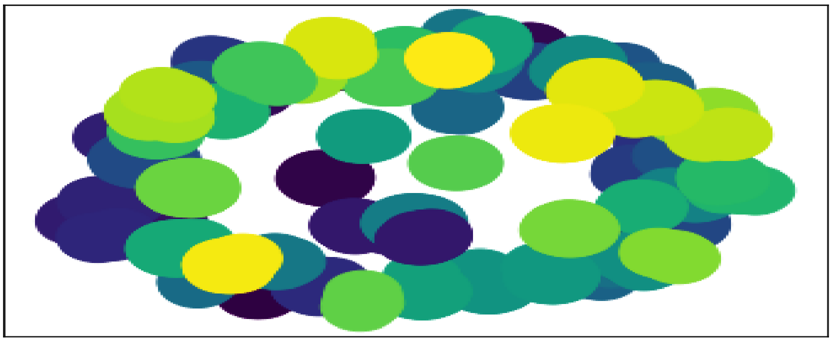

Figure 3. A Citation Graph with 64 Communities

The following figure (4) depicts about all articles were distributed, based on the type of the article. The dataset comprises different types of venues, where an article is published in a journal, conference, patent, repository, or some other related type of venue.

Other
Patent
nan
Repository
Journal
Conference

Figure 4. Type of an article from the set of citation network sample database

3.2.2. Normalized Mutual Information (NMI)

NMI is defined as the fraction of mutual information with conditional entropy. Equation for NMI is given as in equation (3)

$$NMI(U,V) = \frac{I(U,V)}{\sqrt{H(U).H(V)}} \tag{3}$$

U and V are clusters; I is mutual information; and H is conditional entropy.

$$NMI(X,Y) = \frac{-2\sum_{i=1}^{Cx}\sum_{j=1}^{Cy}\log\left(\frac{z_ijN}{z_iz_j}\right)}{\sum_{i=1}^{Cx} z_i\log\left(\frac{z_i}{N}\right)+sum_{j=1}^{Cy} z_j\ \log\left(\frac{z_j}{N}\right)} \tag{4}$$

To evaluate the similarity of two clustering's, this metric is applied. It also evaluates how much information from one cluster is used to generalize about the other is given in equation (4). The more important NMI is, the more information can be shared between communities.

Assume there are two partitions X and Y, and Z is the confusion matrix where rows represent the real community and columns represent the found community and Z_{ij} represent the intersection of number of nodes in community i and community j. The equation for NMI is given in equation (4). In this equation, N includes the number of nodes, C_x the actual number of communities, and C_y the total number of communities that were discovered. In the matrixz_{ij}, the symbol z_i denotes the sum over the ith row, and the symbol z_j denote the sum over the jth column. NMI increases indicate that the discovered communities are getting more and more in line with the true one [25].

3.2.3. Betweenness Centrality

It indicates which nodes will be protected first, and it seems to be the most reliable option in this instance. One measure of centrality is betweenness centrality, which identifies the hub of a network by counting how many shortest paths originate from and terminate at that node. Nodes with high betweenness centrality are the ones that connect two different groups. They form the shortest pathways of communication within the network and questions like who controls network information flow most effectively. Who or what would interrupt flow the most if removed?

Betweenness Centrality measured as the proportion of shortest paths between any two points. It compensates for degree and closeness of centrality, node's shortest paths. Betweenness centrality measures are given in equation (5) to show how often a node acts as a bridge between two others. It measures a person's influence on social network communication. High betweenness vertices are likely to be on a random shortest path between two vertices [26].

$$Betweeness\ Centrality(V) = \sum\nolimits_{S\neq v\neq t\epsilon V} \frac{\sigma_{st}(V)}{\sigma_{st}} \qquad (5)$$

where $\sigma_{\ st}$ shortest paths from node 's' to node 't' and $\sigma_{st\ (V)}$

3.2.4. Purity and F-Measure

The fraction of correctly specified vertices is said to be purity. The purity of part $x_i$which is relatively interacts with some partition 'Y' is said to be purity [27]. It is computed in equation (6) as

$$Pur(x_i, Y) = max_j \frac{n_{ij}}{n_i+} \qquad (6)$$

The greater the degree of intersection, the greater the degree of purity.
The harmonic mean of purity is said to be F-Measure-Measure can be computed as given in equation (7)

$$F(X, Y) = \frac{2.Pur(X,Y).Pur(Y,X)}{Pur(X,Y)+Pur(Y,X)} \qquad (7)$$

4. Proposed Algorithm

By using the LS-SLM (Large Scale-Smart Local Moving) algorithm on large-scale citation networks, researchers can learn about the structure and organisation of the scholarly landscape, find important articles, and look into communities that are focused on a certain topic. The goal is to find clusters of related articles inside the citation web, where the articles are highly connected to one another and frequently cite one another.

Algorithm 1. Linear Scale Smart Local Moving Algorithm (LS-SLM Algorithm)

Input: Directed graph G = (V, E), Maximum number of times max_iter will be used.

Output: Final community assignment C, Modularity score Q, NMI (Normalized Mutual Information), Betweenness centrality F-Score.

```
Begin
1.  Initialize the community assignment
        C = Initial Assignment (G).
    2.  Set iteration counter i = 0.
    3.  Set the maximum modularity score Q_max = -∞.
    4.  Set the best community assignment C_best = C.
    5.  Continue until convergence or maximum iterations.
        (a) Increment i by 1.
        (b) For each node v in V, calculate the change in modularity ΔQ (v) by moving v to its neighbouring
            communities.
        (c) For each node v in V:
            For each community c in C:
                i. Compute the modularity gain ΔQ (c, v) by moving v to community c.
            Find the community c_max that maximizes the modularity gain ΔQ (c_max, v).
            If ΔQ (c_max, v) > 0, move node v to community c_max.
        (d) Compute the modularity score Q using the updated community assignment C
```

$$Q = \left(\frac{k\_in}{2 * m}\right)\left(\frac{(k\_total * k\_out)}{(4 * m\hat{}2)}\right) \tag{8}$$

From equ (8) Q = modularity of a community in a network, k_in = the overall degree of every node in the network, m=the overall edges in the network, k_total=The total degree of the network's nodes., k_out= The total degree of all nodes outside the community.

(e) Compute the NMI between the ground truth and the obtained community.

$$MI = sum\left(sum\left(\frac{P[i,j]*log2(P[i,j])}{(P\_ground\_truth[i])*P\_obtained[j]}\right)\right) \tag{9}$$

From equ (9) P [i, j] = probability distribution of witnessing nodes i and j belonging to the same community,

P_ground_truth= The likelihood that node i will be seen to belong to a certain community, as determined by the ground truth community assignments.

P_obtained= The acquired (predicted) community assignments are used to calculate the likelihood that node belongs to a given community.

$$NMI = \frac{(2*MI)}{(H_{ground\_truth\_norm} + H_{obtained\_norm})} \tag{10}$$

(f) Compute the betweenness centrality for each node in the graph.

B(j)=sum($fraction\ of\ shortest\ paths\ passing\ through\ node\ j\ for\ all\ nodes\ i\ \neq j$)

(g) Compute the F-score based on the community assignment C.

$$F - Score = 2 * \frac{(Precision*Recall)}{Precision+Recall} \tag{11}$$

Where precision is the percentage of a community's nodes that are accurately classified to the total count of C-assigned nodes and Recall is defined as the proportion of correctly identified nodes to the total number of nodes that belong to community C.

(h) h. If Q > Q_max, update Q_max = Q and store the current community assignment as C_best.

6. Return the final community assignment C_best, the modularity score Q_max, NMI, betweenness centrality, and F- Score.

5. Results and Discussion

5.1. Dataset

The dataset used in this experiment is the 1citation network dataset, a directed network which consists of the attributes like id, title, authors name, venue, year, keywords, abstract, volume, references, author, volume, page number, publisher, titles, type, venue, year, etc. The paper fields were taken from of Microsoft Academic Graph (MAG) papers**.** The dataset can be used for multiple purpose like clustering for relevant information, identifying the most influential papers, topic modelling analysis [28][29].

From the following Table1 and figures 5 to 8 discussed about the comparison of all suitable algorithms and the modified SLM algorithm, proposed SLM algorithm provides better results in all aspects that predicts the outcome of a community detected, i.e., strength, i.e., gain in modularity. The results were obtained using NetworkX package and implemented using python in PyCharm.

5.2.1. Modularity

It is a metric that assesses how well-organized the communities are in a network. When referring to the process of community detection, modularity measures how well a network may be divided into strongly connected groups. Modularity compares community edges to the expected number if the network were randomly connected in community detection, the x-axis often reflects community detection methodologies or parameter settings, while the y-axis shows modularity levels. A priority with higher modulation offers a better representation of the network's community structure. For the 101 nodes with 601 edges, from figure (5) it is observed that the proposed SLM algorithm has a 95.35% value, which provides the information in terms of quality to assess the strength of the significance of the detected community.

5.2.2. Normalized Mutual Information (NMI)

It is a measure to evaluate network partitioning. It measures the network partitioning using community finding algorithms. Mutual information refers to the amount of information that may be learned about two distributions through the exchange of data. NMI is a metric for evaluating how closely two-community detection methods are alike. After factoring in the size and composition of each community, it calculates the percentages shared by the expected and actual communities. NMI is a useful tool for comparing and evaluating different techniques and choosing the best way to divide citation networks into coherent and useful communities. NMI is essential in determining the accuracy and quality of the communities detected in the context of citation networks. From figure (6), the proposed SLM algorithm shared 96% mutual information between the expected and actual communities, denoted by the µ in NMI. It measures how much information is shared between the projected community assignments and the ground-truth community assignments.

5.2.3. Betweenness Centrality

It refers to a method for determining how much of an impact a certain node has on the way information moves across a graph. It also figures out how much a node affects the way information flows through a graph. Nodes with high betweenness values connect communities. The centrality betweenness score implies influential nodes help communities share knowledge. In figure (7) The y-axis shows centrality betweenness levels, while the x-axis shows community detection methodologies or parameter settings. By plotting multiple algorithms or parameter settings on the same plot, we can compare their respective centrality betweenness values and visually analyse their performance in identifying influential nodes or community connectors within a network. By incorporating centrality and betweenness into account when analysing a citation network, researchers can learn more about the network's structure, how information flows through it, and which articles are most important. This knowledge can be useful for many things, like finding key papers, understanding research trends, and looking into ways to work with people from different fields. This metric is used to find publications in a citation network that connect different research groups or play a critical role in the flow of information. The proposed SLM algorithm yields 94.3%, which yield the highest centrality betweenness value, which shows which nodes, are important for connecting different communities or controlling the flow of information in a network. High centrality betweenness values show that the nodes in question act as bridges between different parts of the network, which could affect the structure of the community and the way the network works as a whole.

5.2.4. F-Measure

It measures the accuracy of a network. The F-Measure value for each number of communities would be the value for the y variable. This graph can help you see how the F-Measure changes depending on how small the groups are that are being found. The names of the methods can be used as x-variable. The F-Measure numbers for each algorithm would be the y-variable. From figure (8) the proposed SLM algorithm achieves 95.79%, which helps us to understand the importance of various nodes in connecting communities. It helps researchers understand community creation, stability, and dynamics by revealing the network's structure.

Table 1. The comparative table of methodologies with metrics

Metrics & Methodologies	Modularity (%)	NMI (%)	Betweenness Centrality	F-Measure
Louvain Algorithm	81.18	83.86	82.45	83.98
Stochastic Block Model	86.77	87.54	85.92	84.22
Fast Greedy Algorithm	89.38	90.11	87.23	88.6
Existing SLM Algorithm	82.71	93.95	91.12	91.74
Proposed SLM Algorithm	95.35	96.26	94.3	95.79

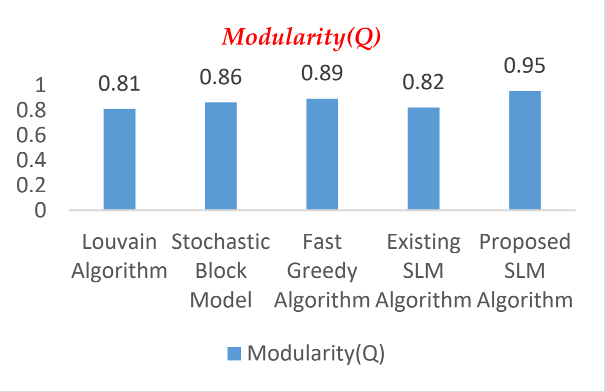

Figure 5. Comparison of Modularity of Communities generated by different community detection algorithms.

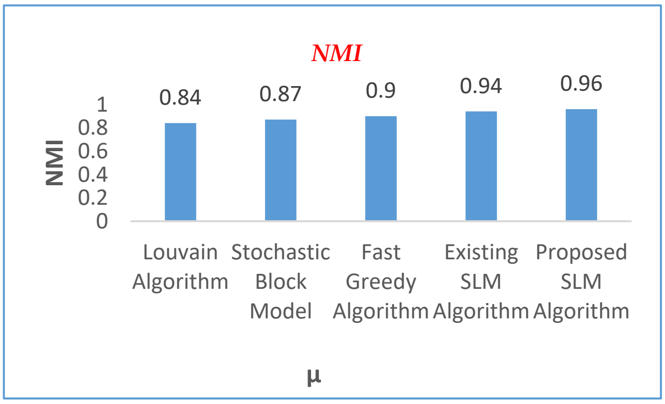

Figure 6. Comparison of NMI Values obtained for the results of the community detection algorithms for the citation network sample dataset.

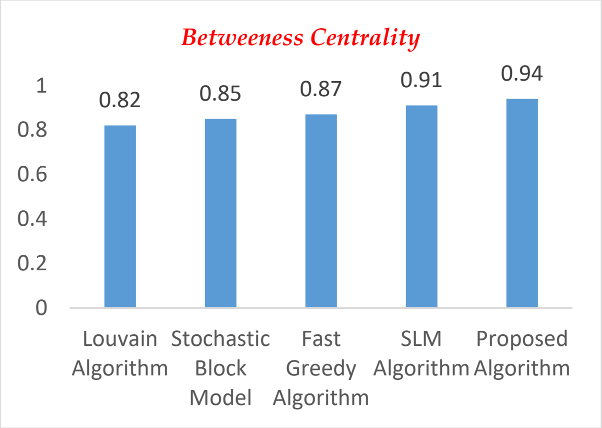

Figure 7. Comparison of Betweenness Centrality of Communities generated by different community detection algorithms.

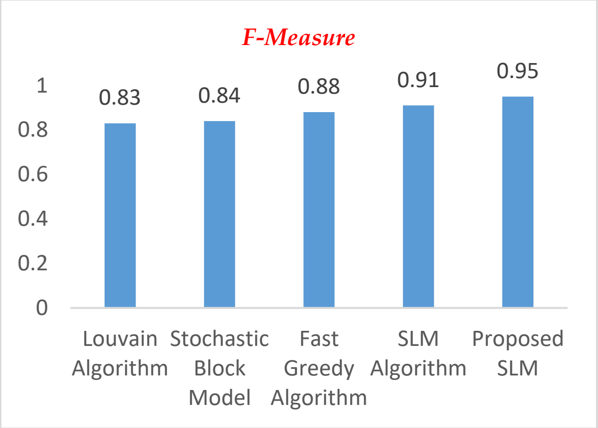

Figure 8. Comparison of F-Measure of Communities generated by different community detection algorithms.

Using scientometrics evaluations identify influential, central, and active nodes. The idea is to find new trends in addition to using network techniques to look at how the field is changing. Scientific metric is used to find who is highly cited author, pivot node with high centrality, strongest citation (citation count) using centrality score, and categories based on frequency and centrality.

6. Conclusion

The LS-SLM algorithm demonstrated a significant improvement in modularity compared to existing techniques, with a computed modularity score of 95.35%. Finding the modularity indicates the ability to identify highly cohesive and well-defined communities in the network. The LS-SLM algorithm showcased remarkable performance in terms of NMI, achieving a score of 96.26%. This result signifies the algorithm's capability to accurately capture the underlying community structure and align it with ground truth information. Regarding the betweenness centrality metric, the LS-SLM algorithm showcased a noteworthy enhancement reducing the average betweenness centrality by 94%. This improvement indicates the optimization and identifying the influential nodes with in the communities leading to more cohesive and hierarchical community structures. Finally, LS-SLM algorithm exhibited superior performance in terms of F-Measure achieving a score of 95.79%. This indicates the balance precision and recall in community detection. By fine-tuning the modularity metric, LS-SLM strives for accurate outcomes in community detection.LS-SLM finds coherent groups in the citation network by optimising the modularity score. SLM's resilience and scalability make it a useful tool for discovering hidden structures and patterns in real-world networks. Overall, the LS-SLM method is a major development in community discovery, providing academics and practitioners with a practical and dependable approach for comprehending network complexity.

7. Future Directions

Current algorithms suffer from high computations and lack of accuracy when detecting communities on a large scale. Once the communities are formed from the citation network, injecting topic modelling into each generated community will be useful in identifying domain-related articles and further useful for recommending the relevant scientific articles for new users.

Acknowledgement

We thank Jie Tang (Tsinghua University), Jing Zhang (Tsinghua University), Limin Yao (University of Illinois at Urbana-Champaign), Juanzi Li, Li Zhang, and Zhong Su. (University of Illinois at Urbana-Champaign) for providing the citation network dataset.

References

[1] Shen Gui-Lan and Yang Xiao-Ping, "A topic community detection method for information network based on improved label propagation", *International Journal of Hybrid Information Technology*, Print ISSN: 1738-9968,Vol. 9,

pp. 299-310, 2016, Published by Science & Engineering Support Society(SERSC) , DOI: 10.14257/IJHIT.2016.9.2.27, Available: https://gvpress.com/journals/IJHIT/vol9_no2/27.pdf.

[2] Jun Ge, Lei-Shi, Yan Wu and Jie Liu, "Human-Driven Dynamic Community Influence Maximization in Social Media Data Streams", *IEEE Access*, Print EISSN: 2169-3536, Vol. 8, pp. 162238-162251, 2020, Published by IEEE, DOI: 10.1109/ACCESS.2000.3022096, Available: https://ieeexplore.ieee.org/document/9187341.

[3] Iztok Fister and Matjaz Perc, "Toward the discovery of citation cartels in citation networks", *Frontiers in Physics,* Print ISSN: 2296424X, Vol. 4, pp. 49, 2016, Published by Frontiers, DOI: 10.3389/fphy.2016.00049, Available: https://www.frontiersin.org/articles/10.3389/fphy.2016.00049/full.

[4] Satiro Baskoro Yudhoatmojo and Muhammad Arvin Samuar, "Community detection on citation network of DBLP sample set using link rank algorithm", *Procedia Computer Science*, Print ISSN: 18770509, Vol. 124, pp. 29-37, 2017, Published by Elsevier, DOI: 10.1016/j.procs.2017.12.126, Available: https://www.sciencedirect.com/science/article/pii/S1877050917328946.

[5] Zhenqi Lu, Johan Wahlstrom and Arye Nehorai, "Community Detection in Complex Networks via Clique Conductance", *Scientific reports*, Print ISSN: 20452322, Vol. 8, pp. 5982, 2018, DOI: 10.1038/s41598-018-23932-z, Available: https://pubmed.ncbi.nlm.nih.gov/29654276/.

[6] Menta Sai Vineeth, Krishnappa RamKarthik, M. Shiva Phaneendra Reddy, Namala Surya and L.R.Deepthi, "Comparative analysis of graph clustering algorithms for detecting communities in social networks", *in Proceedings of the Ambient Communications and Computer Systems, Advances in Intelligent Systems and Computing*, Singapore, Print ISBN: 978-981-15-1517-0, Online ISBN: 978-981-15-1518-7, Vol. 1097, pp. 15-24, 2020, Published by Springer, DOI: 10.1007/978-981-15-1518-7_2, Available: https://link.springer.com/chapter/10.1007/978-981-15-1518-7_2.

[7] Fragkiskos D Malliaros and Michalis Vazirgiannis, "Clustering and community detection in directed networks: A survey", *Physics reports*, Print ISSN: 03701573, Vol. 533, pp. 95-142, 2013, DOI: 10.1016/j.physrep.2013.08.002, Available: https://www.sciencedirect.com/science/article/abs/pii/S0370157313002822.

[8] K. Sathiya Kumari and M. S. Vijaya, "Community Detection Based on Girvan Newman Algorithm and Link Analysis of Social Media", *Digital Connectivity Social Impact, CSI 2016, Communications in Computer and Information Science*, Singapore, Print ISBN: 978-981-10-3273-8, Vol. 679, pp. 223-234, 2016, Published by Springer, DOI: 10.1007/978-981-10-3274-5_18, Available: https://link.springer.com/chapter/10.1007/978-981-10-3274-5_18.

[9] Konstantinos Georgiou, Christos Makris and Georgiou Pispirigos, "A distributed hybrid community detection methodology for social networks", *Algorithms*, Print ISSN: 03701573, Vol. 12, pp. 175, 2019, Published by MDPI, DOI: 10.3390/a12080175, Available: https://www.mdpi.com/1999-4893/12/8/175.

[10] Santo Fortunato, "Community detection in graphs", *Physics Reports*, Print ISSN: 0370-1573, Vol. 486, pp. 75-174, 2010, Published by Elsevier, DOI: 10.1016/j.physrep.2009.11.002, Available: https://www.sciencedirect.com/science/article/abs/pii/S0370157309002841.

[11] Wenchuan Mu, Kwan Hui Lim, Junahu Liu, Shanika Karunasekara, Lucia Falzon *et al.*, "A clustering-based topic model using word networks and word embeddings", *Journal of Big Data*, Electronic ISSN: 2196-1115, Vol. 9, pp. 1-38, 11 April 2022, Article No. 38 (2022), DOI: 10.1186/s40537-022-00585-4, Available: https://www.journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00585-4.

[12] Tommy Dang and Vinh The Nguyen, "ComModeler: Topic Modelling Using Community Detection", in *Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA),* Print ISBN: 978-3-03868-064-2, pp. 1-5, Published by The Eurographics Association, 2018, DOI: 10.2312/eurova.20181104, Available: https://diglib.eg.org/handle/10.2312/eurova20181104.

[13] Jicun Zhang, Jiyou Fei, Xueping Song and Jiawei Feng, "An improved Louvain algorithm for community Detection ", *Mathematical Problems in Engineering,* Print ISSN: 1024123X, Vol. 2021, pp. 1-14, 2021, Published by Hindawi Limited, DOI: 10.1155/2021/1485592, Available: https://www.hindawi.com/journals/mpe/2021/1485592.

[14] Pravin Chopade and Justin Zhan, "A Framework for Community Detection in Large Networks Using Game-Theoretic Modelling", *IEEE Transactions on Big Data,* Print ISSN: 2332-7790, Vol. 3, pp. 276-288, 2016, Published by IEEE, DOI: 10.1109/tbdata.2016.26287, Available: https://ieeexplore.ieee.org/document/7745890.

[15] M.E.J Newman, "Fast algorithm for detecting community structure in networks", *Physical Review E*, Print ISSN: 24700045, Vol. 69, pp. 066133, 2004, Published by American Physical Society, DOI: 10.1103/PhysRevE.69.066133, Available: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.69.066133.

[16] Brian Karrer and M. E. J Newman, "Stochastic block models and community structure in networks", *Physical Review E,* Print ISSN: 2470-0045, Vol. 83, pp. 016107, 2011, Published by American Physical Society, DOI: 10.1103/PhysRevE.83.016107, Available: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.83.016107.

[17] Clement Lee and Darren J. Wilkinson , "A review of stochastic block models and extensions for graph clustering", *Applied Network Science,* Print ISSN: 2364-8228, Vol. 4, pp. 1-50, 2019, Published by Springer Open, DOI: 10.1007/s41109-019-0232-2, Available: https://appliednetsci.springeropen.com/articles/10.1007/s41109-019-0232-2.

[18] Jingyi Zhang, Zhixin Ma, Qijuan Sun and Jun Yan, "A Research Review on Algorithms of Community detection in Complex Networks", *Journal of Physics*, Print ISSN: 1742-6588, Vol. 1069, p. 012124, 2018, DOI: 10.1088/1742-6596/1069/1012124, Available: https://iopscience.iop.org/article/10.1088/1742-6596/1069/1/012124.

[19] Riza Aktunc, Ismail Hakki Toroslu, Mert Ozer and Hasan Davulcu, "A dynamic modularity-based community detection algorithm for large scale networks: DSLM", *in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* Paris, France, Print Electronic ISBN: 978-1-4503-3854-7, pp. 1177-1183, 2015, DOI: 10.1145/2808797.2808822, Available: https://ieeexplore.ieee.org/document/7403695.

[20] Ludo Waltman and Nees Jan van Eck, "A smart local moving algorithm for large-scale modularity-based community detection", *The European Physical Journal B,* Print ISSN: 1434-6036, Vol. 86, pp. 1-14, 2013, DOI: 10.1140/epjb/e2013-40829-0, Available: https://link.springer.com/article/10.1140/epjb/e2013-40829-0.

[21] Mingming Chen, Konstantin Kuzmin and Boleslaw Karol Szymanski, "Community Detection via Maximization of Modularity and its Variants", *in IEEE transactions on Computational Social Systems*, Print Electronic ISSN: 2329-924X, Vol. 1, pp. 46-65, 2014, Published by IEEE, DOI: 10.1109/TCSS.2014.2307458, Available: https://ieeexplore.ieee.org/document/6785984.

[22] Mingming Chen, Tommy Nguyen and Boleslaw K. Szymanski, "On Measuring the Quality of a Network Community Structure", *in Proceedings of 2013 International Conference on Social Computing,* Los Alamitos, CA, USA, Electronic ISBN: 978-0-7695-5737-1, pp. 122-127, 2013, Published by IEEE Computer Society, DOI: 10.1109/SocialCom.2013.25, Available: https://ieeexplore.ieee.org/document/6693322/.

[23] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre, "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment,* Print ISSN: 1742-5468, Vol. 2008, pp.100008, 2008, Published by IOP Publishing Ltd., DOI: 10.1088/1742-5468/10/P10008, Available: https://www.iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008.

[24] Renaud Lambiotte, Jean-Charles Delvenne and Mauricio Barahona, "Random walks, Markov processes and the multiscale modular organization of complex networks", *IEEE Transactions on Network Science and Engineering,* Electronic ISSN: 2327-4697, Vol. 1, pp. 76-90, 2014, Published By IEEE, Available: https://ieeexplore.ieee.org/document/7010026/.

[25] Simrat Kaur, Sarbjeet Singh, Sakshi Kaushal and Arun Kumar Sangaiah, "Comparative analysis of Quality metrics for community detection in social networks using genetic algorithm", *Neural Network World*, Print ISSN: 2336-4335, Vol.26, pp. 625-641, 2016, Published by Czech Technical University in Prague, DOI: 10.14311/NNW.2016.26.036, Available: http://www.nnw.cz/doi/2016/NNW.2016.26.036.pdf.

[26] Attila Mester, Andrei Pop, Bogdon-Eduard-Madalin Mursa, Horea Greblia, Laura Diosan *et al*., "Network Analysis Based on Important Node Selection and Community Detection", *Mathematics,* Electronic ISSN: 2227-7390, Vol. 9, p. 2294, 2021, Published by MDPI AG, DOI: 10.3390/math9182294, Available: https://www.mdpi.com/2227-7390/9/18/2294.

[27] Vincent Labatut, "Generalised measures for the evaluation of community detection methods", *International Journal of Social Network Mining,* Print ISSN: 1757-8485, Vol. 2, pp. 44-63, 2015, Published by Inderscience, DOI: 10.1504/ijsnm.2015.069776, Available: https://www.inderscienceonline.com/doi/pdf/10.1504/IJSNM.2015.069776.

[28] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, Zhing Su *et al.*, "ArnetMiner: Extraction and mining of academic social networks", *in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* New York, USA, Print ISBN: 9781605581934, pp. 990-998, 2008, Published by Association for Computing Machinery, DOI: 10.1145/1401890.1402008, Available: https://dl.acm.org/doi/10.1145/1401890.1402008.

[29] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide *et al.*, "An Overview of Microsoft Academic Service (MAS) and Applications", in *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, Print ISBN: 9781450334730, pp. 243-246, 2015, Published by Association of Computing Machinery, DOI: 10.1145/2740908.2742839, Available: https://dl.acm.org/doi/10.1145/2740908.2742839.