Improving Social Network Analysis to Enhance the Identification of Influential Nodes

By

Sadia Majeed
Fall 2015-MS-15 (CSE) 00000118433

Supervisor
Dr. Muhammad Usman Qamar

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD

August, 2019

Improving Social Network Analysis to Enhance the Identification of Influential Nodes

By

Sadia Majeed

Fall 2015-MS-15 (CSE) 00000 118433

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Software Engineering

Thesis Supervisor:

Dr. Muhammad Usman Qamar

Thesis Supervisor's Signature: ____________________________________

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

August, 2019

Declaration

I hereby certify that I have developed this thesis titled as "*Improving social network analysis to enhance the identification of influential nodes*" entirely on the basis of my personal efforts under the sincere guidance of my supervisor Dr. Muhammad Usman Qamar. All of the sources used in this thesis have been cited and contents of this thesis have not been plagiarized. No portion of the work presented in this thesis has been submitted in support of any application for any other degree of qualification to this or any other university or institute of learning.

Signature of Student

Sadia Majeed

Fall 2015-MS-15 (CSE) 00000118433

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student

Sadia Majeed

Fall 2015-MS-15 (CSE) 00000118433

Signature of Supervisor

Dr. Muhammad Usman Qamar

Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Signature of Student

Sadia Majeed

Fall 2015-MS-15 (CSE) 00000118433

Signature of Supervisor

Dr. Muhammad Usman Qamar

Copyright Statement

Acknowledgements

I thank Almighty Allah (SWT) my Creator Allah for his ultimate guidance throughout my research. Nothing would have been possible without his profound blessing. For all praise is due to God, the Sustainer of all the worlds. Also, my admirations be upon Prophet Muhammad (PBUH) and his Holy Household for being source of guidance for people.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life.

I would also like to express special thanks to my supervisor **Dr. Muhammad Usman Qamar** for his help throughout my thesis and also for Data Engineering course which he has taught me. I can safely say that I haven't learned any other engineering subject in such depth than the ones which he has taught. Without his kind advice, encouragement, guidance and support, it was impossible for me to carry out this task.

I would also like to express my gratitude to my very kind co- supervisor **Dr. Farhan Khan**, for all he has done for me to complete this work.

I would also like to thank **Dr. Wasi Haider Butt** and **Dr. Muhammad Abbas** for being on my thesis guidance and evaluation committee.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my parents whose tremendous support and cooperation led me to this wonderful accomplishment.

Abstract

In online social networks, social influence plays a vital role. Information in social networks propagates virally, as a consequence of this social networks are used to spread influence for multiple reasons including viral marketing, behavioral adoption, and opinion propagation. Numerous researchers are taking action to tackle this social influence study, including initial spreader detection, influence maximization, and influencer rankings, but there are numerous areas which are still quiet challenging. Detecting the influential nodes that occupy significant positions in social networks is a substantial problem as it relates to the effective distribution of information and has wide applications. Traditional ranking algorithms generally target only one out of global, local or community features. Global centrality is mostly measured in terms of betweenness centrality, which is deceptive as it assigns equal value to nodes of high degree scores which are central to local community and global bridges which connect different communities. Moreover, local centrality is usually measured by traditional degree centrality algorithm, which only considers the number of the nearest neighbors. We have used local centrality algorithm which take into consideration the number of the nearest and the next nearest neighbors of node. The thesis proposes a novel ranking framework in which we have taken into account both global and local features to measure influence. Global diversity is measured in terms of proposed bridge centrality method and local centrality is measured using local centrality method. The proposed approach is applied and tested on four different datasets. The thesis used a cross validation technique to measure the accuracy of proposed method. The hybrid classifier achieves 99% accuracy which is up to the mark. The overall aim of our research is to improve social network analysis to enhance the identification of influential nodes/ key players. The experimental results demonstrate that the proposed technique can rank influential nodes efficiently and accurately on social networks.

Key Words: *Social Network Analysis (SNA); Data Mining ; Social influence spreaders; Online Social Network (OSN) ; Centrality Measures ; Performance ; Social influence analysis*

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

In this chapter detailed introduction of research is presented. Section 1.1 shows the overview. Section 1.2 discusses the background and motivation for research work. Section 1.3 describes problem statement whereas Section 1.4 and Section 1.5 contain thesis contribution and thesis organizations.

1.1 Overview

Now days billions of users are using social networks as part of their routine. Social Influence is the behavioral change of a person due to the perceived relationship with other individuals. Ideas and information propagate through interactions between individuals on social networks. Influencers are renowned for playing an important role in many domains. In the marketing domain, the detection of influencers can be used to harness their viral power for spreading campaign relevant messages, maximizing the campaign's overall reach. Moreover, business organizations can use influencers to promote their product and services [11] [12]. There is wide research in this field.

1.2 Background and Motivation

Social influence plays a key role in online social networks. Information in social networks disseminate virally, due to this social networks are being used to spread influence for various purposes including behavior adoption, viral marketing, and opinion propagation. This has attracted many researchers to explore detection of influencer users on social networks as this can affect people decision making abilities. Ideas and information propagate through the interactions between individuals on social networks. An influencer is a person with the ability to persuade people, affecting their actions and behavior. Influence is defined as 'Change in an individual thoughts, feelings, attitudes, or behaviors that results from interaction with another individual or a group '[29]. Webster also interpret influence as 'the act or power of producing an effect without apparent exertion of force or direct exercise of command [30]. Influencer's are recognized for playing an important role in many domains. [53]

In the marketing domain, detection of influencer can be used to harness their viral power for spreading campaign relevant messages, maximizing the campaign's overall reach. In social-networking services, such as the most popular professional networking service, LinkedIn utilize influencer users for attracting the attention of regular users to promote contents and services. A motivating application is the viral marketing, opinion propagation and behavior adoption. Other real-world applications include recommendations systems and expert search engines. Moreover, business organizations can use influencer to promote their product and services. In this area many approaches are being adopted to measure influence in social networks. Researchers have looked upon this problem from two perspectives. The first group addresses the influence maximization problem which is defined as finding minimum number of users that can maximize diffusion of influence across social network. It is an observation based non deterministic technique usually represented as influence maximization [21] [27] [31]. The second group focuses on prediction-based technique to predict influencer users on social networks [25]. Some prediction-based measures take into account network topology to calculate centrality [20] – [27]. While other considered structure, as well as content to be spread, can be represented as topology and topic-based approach [28]. Similarly, some focus on content only usually represented as topic-based approach [27] [53].

In Pakistan, marketing on social networks is not taken seriously. If this medium is explored intelligently then the social network medium has a potential to provide many new ways to market the audience with the help of influencers. Our research will be introducing a novel approach for the identification of most influential individuals within a social network. Although massive research efforts are being made to measure influence epidemic in social networks however there are several areas which are still challenging. [53]

The main motivation of the thesis is to improve the business and marketing strategies of the small-scale businesses, monitoring terrorist groups and their suspicious activities in social networks, identifying central nodes for information flow in social networks and identifying local domestic centers and the global international bridge airports [50] [51][53]. The purpose of this research is to find out the influential nodes or the key-players in different networks.

Application domains such as social networks produce large amounts of graph data. Therefore, the management and processing of graph data is gaining importance. These days we widely use tables and graphs to describe data and turn data into knowledge. In fact, a graphical illustration may be more revealing than data in tabular form, and it allows viewers to discover insight in the data without intensive study. A doctor patient health care network [49] performs its visualization and analysis using a relational database. Although a relational database offers numerous advantages such as transactions, reliable storage and access control but graph analysis is still considered challenging. Our motivation is to bridge this gap by performing social network analysis in SQL Server.

1.3 Problem Statement

Information epidemic via social network is being used virally to spread influence, which is a complex task and needs more robust and efficient techniques. The aim of this research is to improve the identification of influential nodes in a social network by exploiting network topology.

1.4 Thesis Contribution

Our focus is to research on the problem of identifying key-players in social networks. This research will be introducing a novel approach for the identification of most influential individuals within a social network which is the central challenge for the social influence analysis.

We have proposed a two-step framework to identifying influential spreaders. The epidemic spreaders should satisfy two network topology conditions: high local centrality and high global diversity. Global diversity is calculated by proposed bridge centrality algorithm and local features are considered to measure local centrality. Our proposed approach fulfills both conditions to compute final influence. Hence this technique identifies influential nodes efficiently and accurately.

1.5 Thesis Outline

The report is structured as follows:

- Chapter 2 discusses about basics of Social Network Analysis, Layers of Social Networks, Social Network Metrics Measures, Detailed Literature Review, SNA approaches, their characteristics, and limitations.
- Chapter 3 includes the implementation of proposed approach for key-player detection using bridge centrality and local centrality algorithm.

- Chapter 4 contains results and discussion of the research, accuracy evaluation, Classification and Prediction.
- Chapter 5 includes the Future work and Conclusion.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The objective of this literature review is to emphasize social influence analysis (SIA) approaches, their characteristics, and limits in the past five years. Chapter 2 uses material from reference [53], in which I have published systematic literature review (SLR) on social influence analysis. Figure.3 provides a summary of research.

2.2 Background

2.2.1 Social Networks

Social Networks are one of the key ingredients of our daily life. Social media provides the platform where people around the globe interact with each other, several people join to form the groups and several groups together crafts the network.

These nodes in social network can co-exist as an individual's (persons) or organization. And the relationship among them are represented by the edges. The social network can be undirected or directed, either unweighted or weighted.

These terms with respect to the graph theory are explained later. All the graph theory concepts are valid and can be applied on the social network. Examples of social network include Facebook, LinkedIn, Instagram, GitHub, Email etc.

2.2.2 Single-layer Social Network

The social network that comprises only one layer is called single-layer social network. In social networks, accounts of the individuals are present rather than themselves. In single layer social network, the nodes or users are existed once in the whole network. A single node or single users have only one account in one network.

If a person with more than one account appears. All his accounts are merged or only the one is selected, and all the rest are ignored.

The Figure.1 depicts the single layer social network containing multiple communities/ modules. The vertexes denote the nodes or actors and the edges and arcs denote the relations between nodes.

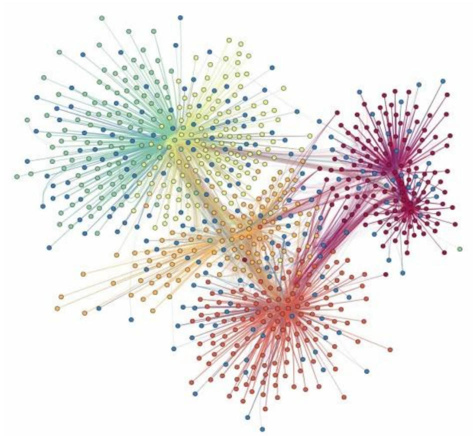


Figure 1: Single Layered Network

2.2.3 Multi-layer Social Network

The multi-layered social network is defined by the many names e.g. Multi-relational, Multi-dimensional term, multi- slice, multilevel and multi-type networks. Although the Multi-layered network is most commonly used. In social networks, accounts of the individuals are present rather than themselves. In multilayer social network the user (Node) can exist in more than one layer of the social network e.g. one user can exist on Facebook, the very same user can also have an account on Instagram, twitter and many more. The Connection of the user in the same layer is called interlayer edges or relations and the edges or the connection on other layers are called intra layer edges or relations. The most significant part of this analysis is to map the nodes of two different layer in multilayer network. The Figure 2.2 shows the graphical representation of the multilayer social network.

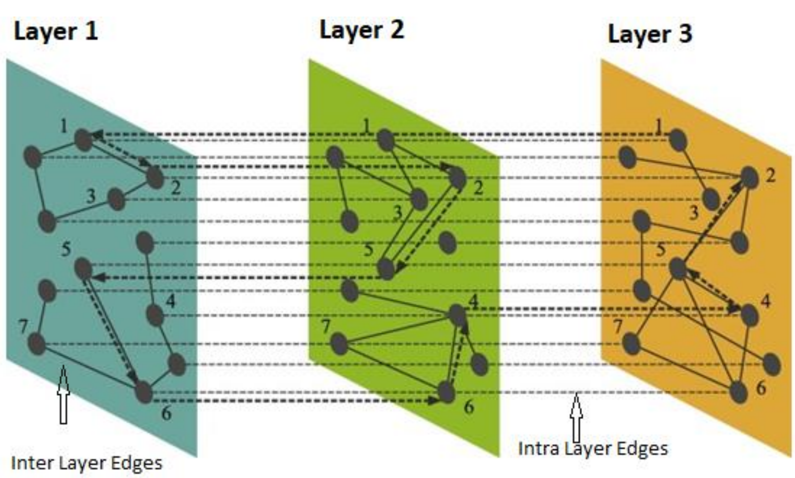


Figure 2: Multi-Layered Network

2.2.4 Social Network Analysis (SIA)

The common output of SIA is measuring the node's centrality and graphical visualization, which helps us to understand the extent and nature of connections between different nodes of the network.

The field of SIA is very vast and different scientists have explained the Social Network Analysis in different terms [34]. Likewise, according to Freeman, the techniques and procedures to uncover the hidden patterns of person interactions [35].

Usually, if we have to determine the most important and influential person from the company, we will be considering the persons and the managers that are sitting on the top of the company or the persons that are higher in the hierarchy.

Top managers are considered as the most influential as they are the decision makers. But that may not be true. The Social Network Analysis provides the right information regarding the importance. No matter where the person lies the importance of anyone is calculated how quickly the information will be spread via him/her or how accurately the information will be delivered.

2.2.4.1 Key-players

The most central area of the SIA is to identify the key-players which can help in epidemic. The exact definition of the key-player is dependent on the type of social network analysis. The simple and abstract definition of key-players is those nodes and individuals, which are more important/ influential in the network than the other. The importance of the key-player solely depends upon the type of analysis and the network.

2.2.5 Social Network Metrics Measures

2.2.5.1 Degree

The Degree identifies the connectedness of the nods in the graph by analyzing the direct contacts of the node with the rest of the network [36].

As mentioned earlier, Degree is the measure of the direct nodes or links of the selected node. The Degree Centrality is directly proportional to the connectivity. The equation for the calculation of the degree centrality is:

$$\boldsymbol{Dv} = \sigma iv\,{}_{i=1}^{n} \qquad (2.1)$$

Whereas, "σv" is the Degree Centrality (DC) measure of node v.

2.2.5.2 Closeness Centrality

The one drawback of the degree centrality is that, it only takes the direct attached nodes and links into the account.

While Closeness Centrality (CC) measures the shortest path of the selected node and the rest of the nodes in the network. [37] [38]

$$Cx = \frac{1}{y^{d(y,x)}} \qquad (2.2)$$

Where d(y,x) is a measure of distance between the x and y. Closeness Centrality (CC) can also be defined as a time taken to disperse the information from one node to whole network.

2.2.5.3 Betweenness Centrality

Two nodes in the social network that are not directly linked does not means they do not interfere.

Nodes that are not adjacent might interact to each other through other nodes in the network, especially through the nodes that lies on the paths between the two. The nodes Betweenness Centrality is said to be high if it is present between many other nodes. Betweenness Centrality can be represented as:

$$\mathrm{CB}v = \frac{\sigma\mathrm{st}\ (\mathrm{v})}{\sigma_{\mathrm{st}}} \qquad (2.3)$$

$$s \neq v \neq t \in V$$

Where "σ st" is the total number of "shortest paths" from a node s to a node t and "σ st(v)" represents the number of paths that passes through v.

Now let us consider the network in Figure 3. Node A has higher Degree centrality. Node B has higher closeness centrality and Node C has higher Betweeness centrality.

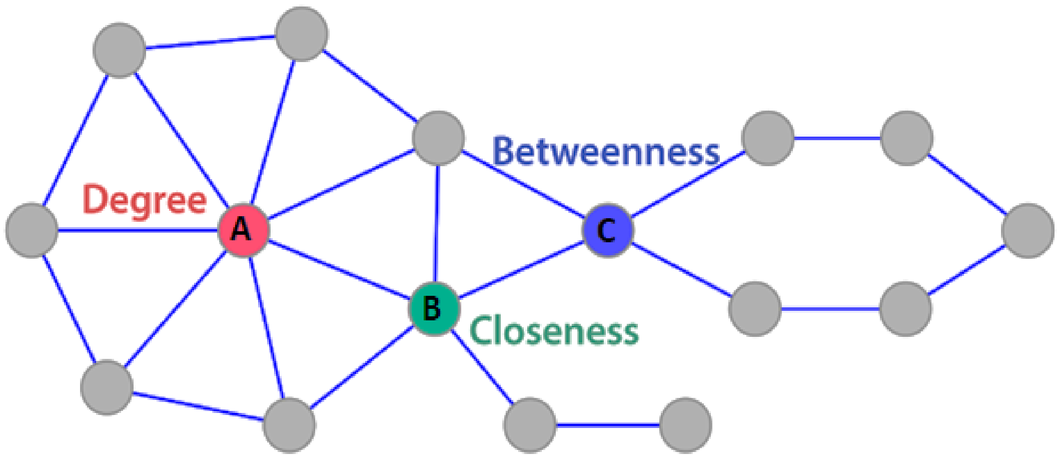


Figure 3: Comparison between Betweenness Centrality, Closeness Centrality and Degree Centrality measures

2.2.5.4 Eigenvector Centrality

Eigenvector Centrality is the measuring importance of a respective nodes in a social network significance depending on its links. [42] [43]

Vertex "v" whose Eigenvector Centrality is given as [39] [40]:

$$ECv = v_x = \frac{1}{\lambda_{\max}(A)} \sum_{j=1}^{n} a_{jx} v_j \qquad (2.4)$$

Whereas $v = (v1, 2)\, T$ refers an eigenvector for the max eigenvalue $\lambda_{\max\,(A)}$ of the adjacent matrix A.

2.2.5.5 Clustering Coefficient

Clustering Coefficient measures the likeliness of the nodes that are interrelated. The average local clustering coefficients for all vertices can be calculated [41].

$$C = \frac{1}{n} \sum_{i=1}^{n} C_i \qquad (2.5)$$

2.2.5.6 Page Rank

Lary and Brin were the first to introduce the Page rank algorithm which are displayed from zero to ten.

Page rank is calculated with the help of the web network, where a links are treated as edges and nodes are the webpages themselves [44].

2.2.5.7 Eccentricity

The longest route from a node to the rest of the nodes in the network is called the Eccentricity of that node [45]. According to Eccentricity measure the nodes with higher Eccentricity are less central in the network.

$$\text{Ec (v)} = \frac{1}{\max\{dist\ (v\ ,w) : w \in V\}} \qquad (2.6)$$

2.2.5.8 No. of triangles

Holland [46] demonstrated a new method that the organization of social networks can also be represented by the amount of triangles. This calculation is important in determining many friends of node are friends of each other also.

2.2.6 Visualization

By Visualization we mean the social networks can be drawn manually or with the help of the available SNA tools. There are various ways to denote the social network the most common ones among them are Adjacency Matrix, Edges List and many more. Different tools require different types of the input's formats. E.g. The Gephi and Node XL requires the Adjacency Matrix in excel formats or CSV files, on the other hand MuxViz requires the edges list format for the visualization and analysis.

A network can consist of more than millions of node and various layers. The network with more than one layer is called multilayer social networks. Visual of these networks are very difficult and almost impossible to evaluate the relationships of the nodes.

For these types of the networks the different software are available to evaluate the network. Gephi is one of the software to evaluate the large-scale network. The Figure 4 shows the visualization of the large-scale network visualization provided by the Gephi Visualization Software.

Figure 4: Visualization of the Large-Scale Network

2.3 Selection and Rejection Criteria

- **Subject Relevant and Latest Papers with Crucial Effects:** Choose the results which are related to SIA perspective and must be supporting answers to our research questions. The selected research study should be published from the year 2014 to 2018 and have essential effects that are positive concerning SIA.
- **Publishers:** The research work that has been selected must be available in one of the five famous scientific databases.
- **Result Oriented without Repetition:** The results that are chosen should be related to our research area. The proposal and results of the research paper must be supported by experimentation. The entire studies in a particular research perspective should not be included. [53]

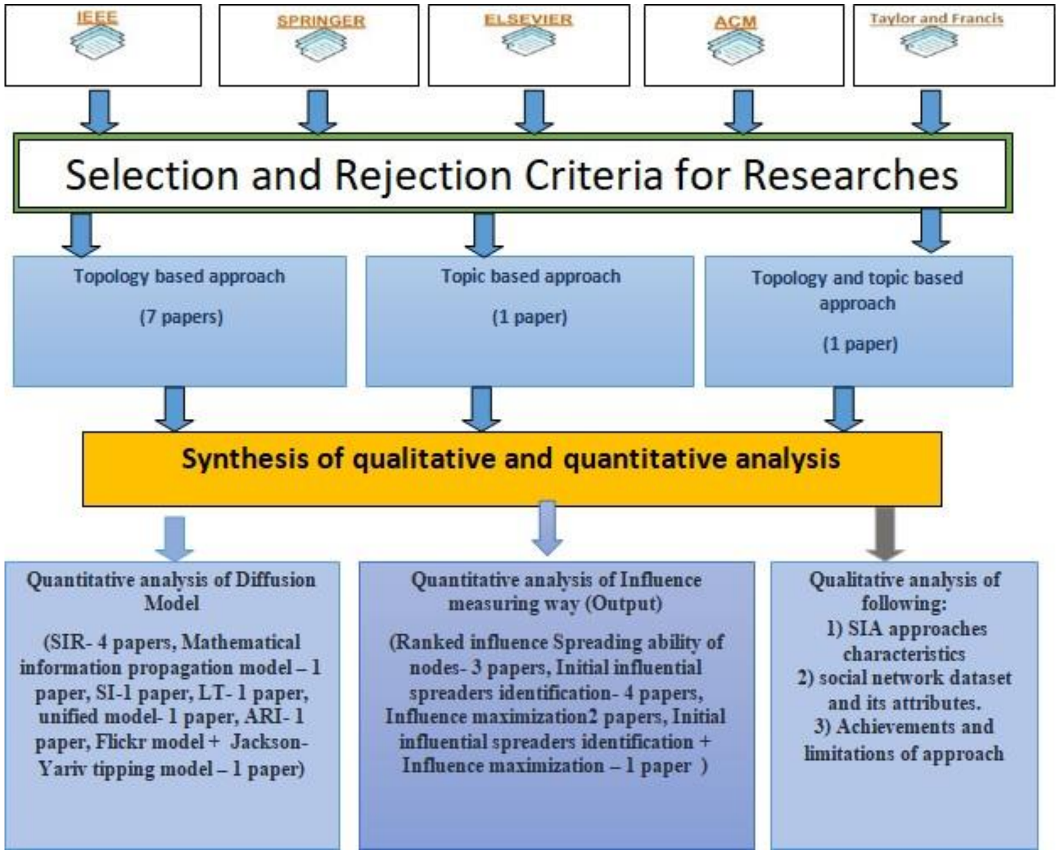


Figure 5: Overview of Research

2.4 Maintaining Quality Assessment

A quality criterion was developed that consists of quality assessment questions as described in the following TABLE 1 We have then provided ‘yes ’and ‘no ‘answers to the quality assessment questions. [53]

Table 1: Quality Assessment Checklist

Sr. No	Quality Assessment Checklist	
1	Is the data assessment showing facts that are concrete without any vague statements?	Yes/No
2	Have proper validation methods been used in this research and evaluation focus on its stated aims and purpose?	Yes/No

3	Is the study based on researches published in any of the five globally accepted scientific databases, and is conducted from the year 2014 to 2018?	Yes/No

2.5 Data Synthesis and Data Extraction

The data synthesis and extraction as shown in TABLE 2 is performed for specific researches for answering research questions. Data extraction is done on our specific researches byextracting a significant amount of data from them according to our inclusion/exclusion criteria. Whiles data synthesis is done by the detail study and analysis of our selected researches for proper identification of SIA approaches. TABLE 3 shows the research paper repositories names and their reference number. [53]

Table 2: Data Extraction & Data Synthesis Details

Sr. No	**Description**	**Details**
1	Bibliographic Information	Author, title, details of publisher, details of publishing year, research type
	Extraction of data	
2	Overview	The basic objective of our selected research and what it is about
3	Results	Results taken from the selected research
4	Data collection	Qualitative or Quantitative methods used
5	Assumptions	Assumptions that were made to authenticate the results used
6	Validation	Method of validation used to authenticate its proposal
	Synthesis of data	
7	Classification	Relevance to one of the predefined categories relevant to targeted social influence categories

8	Quantitative Analysis of SIA approaches	Statistical data about diffusion model used and influence measuring way; influence maximization, initial spreaders identification, ranking influencers
9	Qualitative Analysis of SIA approaches	Subjective information about specific characteristics including testing complexity, accuracy, limitations, achievements, dataset, comparative classical measures of SIA approach, as well as technique performance [53]

Table 3: No of Researches from their Corresponding Libraries

Sr. No	**Scientific Database**	**Type**	**Selected Research Method**	**No of Researches**
1	IEEE	Journal Conference	[1]	0 1
2	SPRINGER	Journal Conference	[2], [3]	2 0
3	ELSIVER	Journal Conference	[4]–[8]	5 0
4	ACM	Journal Conference	[9]	0 1
5	Taylor Francis	Journal Conference		0 0

2.6 Results

We performed a detailed analysis of each research in order to assign them to the corresponding category. Classification results for selected researches for targeted SIA approaches are given in TABLE 4. [53]

2.6.1 Quantitative analysis of SIA approach

The following section represents results based on quantitative analysis of SIA approaches. Statistical results of SIA approaches are applied to diffusion models. We have analyzed diffusion model used by SIA approaches in TABLE 5. Mostly Susceptible Infected Recovered (SIR) model is used by SIA approaches. [53]

Table 4: No of Researches from their Corresponding Categories

Sr. No	Social influence Categories	Relevant Research
1	Topology based approach	[1]–[7]
2	Topology and topic-based approach	[9]
3	Topic based approach	[8]

2.6.2 Comparative Quantitative analysis

This section describes the qualitative analysis that is subjective information about specific characteristics including limitations, the complexity of time and space, dataset and its attributes, positive and negative aspects of SIA approaches. [53]

Table 5: Quantitative Analysis of Diffusion Model Used

Sr. no	Diffusion Model	Relevant Research
1	Mathematical information propagation model	[4]
2	Susceptible Infected Recovered (SIR)	[5]–[8]
3	Susceptible-Infected (SI)	[1]
4	linear threshold (LT) influence propagation model	[3]
5	unified model	[2]
6	Flickr model + Jackson-Yariv tipping model	[2]

During the comparison of this SLR with existing surveys, we found out that literature review on this subject is scant. Kan Li et al. [16] recently published a review paper in which his focus was on microscopic and macroscopic models to study influence minimization, influence maximization and flow of influence along with influence evaluation metrics. Jimeng Sun et al. [14] in his review on SIA discussed the basic algorithms used for centrality measures such as betweenness, centrality and closeness, social influence models, influence maximization and its applications. Whereas our study has provided a comparative review of research discussing topology based, topic-based, topology and topic-based approach for influence measurement. These approaches measure influence in form of initial spreaders identification, influence maximization and ranking spreading ability of nodes using various techniques. Furthermore, it reveals which diffusion models are preferred by which type of approach. Qualitative analysis of computation complexity, accuracy, comparative technique, net- work (real/synthetic), dataset and its attributes, achievements, limitations and future directions are also explored. [53]

2.6.3 Qualitative analysis of SIA approach followed

Here is a brief description of techniques discussed in selected 9 research papers. TABLE 6 highlighted techniques, methods and models used by researchers in SIA. Moreover, strengths, limitations and future directions of research articles are depicted in TABLE 9.

Sheng Wen et al. [4] computed epidemic betweenness firstly based on presenting the propagation dynamics and then by computing the influence of each node reversely. This algorithm is suitable when information is sent from one node to multiple neighboring nodes through multi-paths to all reachable nodes. Furthermore, Jun Zhao [1] proposed community-based distributed algorithm to identify initial spreaders from both overlapping communities and non-overlapping communities. Gennaro Cordasco [3] presented a fast and simple Minimum target set algorithm which works by iteratively deprecating nodes from the input digraph until a specific condition occurs and the node is considered irrelevant, which results in the addition of that node to the output target set. It also takes into account the influence of depreciated node on the outgo- ing neighbors. Additionally, Ajitesh Srivastava [2] proposed greedy seed-set selection algorithm. It is a greedy solution to maximize influence based on a unified model. Jonathan Herzig [9] discussed author-Reader Influence (ARI) model in which author write attractive content which readers cite. Citation leads to further diffusion to other readers. Likewise, Chanhyun Kang [8] proposed Diffusion Centrality (DC) algorithm which finds top influencers nodes using semantic characteristics of the social network. HyperDC algorithm is used which computes top k vertices for the small social network while Coarsening Back and Forth (CBAF) algorithm is also presented to compute top k vertices having the highest diffusion centrality in the huge social network. Shuai Gaoet al. [6] considered both the number of nearest and then next nearest neighbors and the topological connections among neighbors.

Moreover Yu-Hsiang Fu [7] proposed two-step framework: combine global diversity and local features. Lei Guo et al. take statistical properties of the network in consideration. Firstly, all nodes whose distance between nodes is at least ‘r’ are colored. Secondly, nodes with the same color are classified into the same set considering that the distance between nodes in the same set is at least r. Thirdly the nodes at the topmost position of the ranking list with the maximum degree in the same set are chosen as multiple spreaders.

As shown in TABLE 9 different influence models are being used for the topic-based and topology-based approaches. Topology based approaches used mathematical information propagation model, SIR model, SI model, unified and LT model whereas topic-based approach used AIR model and Flickr model. The SIR model is appropriate when a node once recovered, it would receive lifetime immunity and is not appropriate if a node was infected but is not contagious. In LT model nodes are either active or inactive. Whereas in SI mod- els, nodes never leave the contagious state and have lifetime infections. The unified model provides an efficient method for influence maximization. Moreover, ARI model is used for the topic-based approach as the model is recognized by means of a topic-based citation graph, where nodes symbolize authors and edges represent reader-to-author citations. [53]

Table 6: Qualitative Analysis of SIA Approaches Characteristics

Ref	Author/Year	Technique	Diffusion models	Factors considered	Result
[4]	Sheng Wen et al. 2017	epidemic betweenness	Mathematical information propagation model	Topology based approach Ranked	influence spreading ability of nodes
[6]	Shuai Gao et al. 2014	Local structural centrality measure	Susceptible Infected Recovered (SIR)	Topology based approach	Ranked influence spreading ability of nodes
[7]	Yu-Hsiang Fu et al. 2015	k-shell and degree centrality	Susceptible Infected Recovered (SIR)	Topology based approach	Ranked influence spreading ability of nodes

[5]	Lei Guo et al. 2016	distance-based coloring method	Susceptible-Infected-Recovered (SIR)	Topology based approach	Initial influential spreaders identification
[1]	Jun Zhao 2016	community-based distributed algorithm	Susceptible-Infected (SI)	Topology based approach	Initial influential spreaders identification
[3]	Gennaro Cordasco et al. 2016	Minimum target set algorithm	linear threshold (LT) influence propagation model	Topology based approach	Initial influential spreaders identification
[2]	Ajitesh Srivastava et al. 2015	online seed set selection using unified model (OSSUM)	unified model	Topology based approach	Influence maximization
[9]	Jonathan Herzig et al. 2014	Extension of Topic-Sensitive PageRank algorithm	Author-Reader Influence (ARI) model	Topology and topic-based approach	Initial influential spreaders identification
[8]	Chanhyun Kang et al. 2016	Diffusion Centrality (DC)	Flickr model + Jackson-Yariv tipping model + SIR model of disease spread	Topic based approach	Initial influential spreaders identification + Influence maximization

2.6.4 Qualitative analysis About Complexity Level, Accuracy of SIA Approaches:

Level of complexity depends mainly on time and space required to compute influence. If the complexity level is high, then SIA approach becomes difficult and unfeasible to use for large-scale networks. Accuracy and performance are measured as compared to competitive techniques. Computational complexity is defined (low, high) as compared to comparative technique. Social networks which are selected as dataset can be real as well as synthetic

(artificially created.)

As depicted in TABLE 7. Epidemic betweenness [4] centrality has low computational complexity and higher accuracy. Local structural centrality measure [6] has a computational complexity greater than DC and LC but less than KS, BC, CC and gives more accurate results. K-shell and degree centrality when used together to measure local and global influence [7] [53]

Table 7: Qualitative Analysis about Complexity, Accuracy and Comparative Technique off SIA Approach Usage

Related researches	Comparative technique	Computation complexity	Social networks (Real/Synthetic)	Accuracy/ Performance
[4]	Random walk, flow shortest path	Low: O (n 2 T)	Synthetic + Real	High Accuracy
[6]	k-shell (KS), degree (DC), closeness (CC), betweenness (BC) and local centrality (LC)	Greater than DC and LC, Less than KS, BC, CC	real networks	High Accuracy
[7]	Betweenness, Degree, Coreness, Closeness k-core and PageRank	O ((k)2 +(k)). n)	Collaboration + Social (Real)	High Accuracy
[5]	Betweenness, Degree, Coreness, Closeness	Not defined	Synthetic + Real	Outperforms than traditional IS algorithm
[1]	Randomly selecting the initial spreaders from the social network, randomly selecting the initial diffusers	Not defined	Real	Better Performance than random selection

	from the community			
[3]	Greedy strategy, TIP_DECOMP	O(E| log |V|) Compareable with latest algorithms for MTS problem	Real	Outperforms than competitors
[2]	Degree, Degree Discount, Single Discount, CELF++, LDAG, SPS-CELF++	O(R|E|t)	Real	Better performance than competitors
[9]	TwitterRank, PageRank, Indegree centrality	Not defined	Real	14 % improvement than TwitterRank, improved on PageRank by 19%
[8]	Classical centrality measures	Liner but nonlinear for worst case	Real + synthetic	Better than classical measures

provided more accurate results than its competitors. Distance-based coloring method [5] outperforms than traditional IS algorithm. Community-based distributed algorithm [1] has better performance than random selection. The minimum target set algorithm [3] has comparable computational complexity with the algorithms for MTS problem and it outperforms than competitors. Seed selection algorithm [2] has better performance. Furthermore, Extension of Topic-Sensitive PageRank algorithm [9] also shown improvement. Diffusion Centrality (DC) has better performance than traditional measures and has Liner computational complexity but it becomes nonlinear for the worst case. [53]

2.6.5 Qualitative analysis about dataset and Meta data:

Furthermore TABLE 8 shows qualitative analysis about dataset and its attributes [53]

Table 8: Qualitative Analysis about Data Set and Its Meta Data Approach

Selected Reference	**Dataset**		**Attributes**
	Social networks (Synthetic)	***Social networks (real)***	
[4]	Erdos-Renyi (ER), scale-free network, small-world network	Enron Email network, the protein-protein interaction (PPI) network and the U.S. Power Grid network	Edges (Relation/links/ Communication) and vertices(nodes)
[6]	Not defined	Email, PGP, Twitter, Blog	Edges (Relation/links/ Communication) and vertices(nodes)
[7]	Not defined	Ca-AstroPh, ca-CondMat, ca-GrQc, C-HepTh, Jazz-Mucians, Email-contacts, Email-Enron, C eleganNeural, Dolphins, LesMis, NetScience, PloBlogs	Edges (Relation/links/ Communication) and vertices(nodes)
[5]	ER	Polblogs, Erdos and Routers networks	Edges (Relation/links/ Communication) and vertices(nodes) In Erdos EI network (Nodes:scientist), Polblogs network(Nodes: owner of blog,), and in Routers network(Nodes: Routers)
[1]	Not defined	Renren social network (Kind of Chinese Facebook)	Edges (Relation/links/ Communication) and vertices(nodes)

[3]	Not defined	Amazon, BlogCatalog, Ca-AstrpPh, Ca-HepTH, Facebook, PowerGrid	Edges (Relation/links/ Communication) and vertices(nodes)
[2]	Not defined	Ca-HepTH, : co-authorship network	Edges (Relation/links/ Communication) and vertices(nodes)
[9]	Not defined	Twitter	Edges (Relation/links/ Communication) and vertices(nodes) Nodes: Authors / Readers Edges: citations (Re tweets/Mentions/Replies)
[8]	GAME, STEAM data	BlogCatalog, Enron Email, wiki-Vote	Edges (Relation/links/ Communication) and vertices(nodes)

2.6.6 Qualitative analysis about achievements and limitations of SIA approach:

The section is about achievement, limitations and future recommendations of approaches. [53]

Table 9: Achievements and Limitations of SIA Approach

Selected Reference	Achievements	Limitations/Future Recommendations
[4]	Since epidemic can start from any node in the real network. This measure has an edge on other measures that each node can be starting node or an intermediary. Moreover, it is rapid for large-scale networks with low computational complexity [53]	After careful analysis, it was found that when most nodes were left with one neighbor, their influence was almost the same to each other, which results in variation of interaction percentage of nodes. This percentage decreases first and then increases. [53]
[6]	This method is also robust for large-scale networks with community	This work can be extended by incorporating information relevant to

	structure and is more accurate than traditional centrality measures. [53]	community structure and structural diversity. Furthermore, the structure of nodes with multi-hops from the target node can also be analyzed.
[7]	An efficient approach to finding spreaders with high global diversity in a large-scale network. This algorithm can calculate global as well as local influence [53]	Its limitation is dependent on the type of involved nodes. If maximum k shell value for a network is lower in case of smaller networks then global diversity cannot be calculated and influence measurement will be limited to local network layers. [53]
[5]	Performance of multiple spreaders is enhanced.	Properties specifically sparsity (scattered) of the network can affect algorithm performance. In the future relationship of the optimal distance between multiple spreaders and network structure can be analyzed. [53]
[1]	Considers overlapping communities. Identifying initial spreaders in distributed manner. Applicable on large networks	Only applicable on networks with community structure
[3]	Optimal solution. Support for directed graphs provided [53]	Performance correlates with network modularity. In case of High modularity: influence is hard to spread and vice versa
[2]	Better seed selection. Can evaluate different cases for large networks under different influence models. Consider time and budget consumed during diffusion	This algorithm provides better seed selection providing unified model is being used [53]

[9]	Considers both topology and topic relevance. Also, provide relatively good results even with small sized profiles [53]	ARI algorithm captures reader random behavior whose goal is to select relevant author thus it does not differentiate between already known users or newly discovered authors. Considering this limitation, an algorithm can be developed which considers various aspects of the model while calculating authors influence
[8]	This approach is highly efficient as it studies scalability of network and uses such approach for the large network which results in lower runtime and better spread	This research opens up future directions to develop such models which can incorporate semantic aspects like mutual trust property, political factors, the voting trend in case of predicting election outcome which can be associated with edges and vertices

2.7 Discussion

In total, nine research studies have been covered in our literature review out of which seven articles of research are relevant to topology-based approach, topology and topic-based approach covered by just one research. One research targeted the topic-based approach. We also mentioned various characteristics of approaches practiced by practitioners including computation complexity, diffusion model used, comparative techniques, accuracy and performance which can be seen through TABLE 4, TABLE 5, TABLE 6 and TABLE 7. TABLE 8 reveals datasets used for social influence analysis. The dataset for different synthetic and real social networks are being used by SIA approaches. This table reveals which dataset is supported by specific technique and what dataset attributes are being considered. In this literature review, we also identified achievements and limitations of approaches practiced by researchers as given in the TABLE 9. This information facilitates researchers in classification and selection of approach over one another based on characteristics for specific purposes. This will be helping practitioners in choosing the relevant approach for specific purposes. [53]

CHAPTER 3: RESEARCH METHODOLGY

This chapter presents suggested methodology. Our aim is to develop and test an approach with higher accuracy to identify key-players in online social networks while exploiting network topology. This research will be introducing a novel approach for the identification of most influential individuals within a social network which is the central challenge for the social influence analysis.

In social network analysis centrality algorithms are classified as global and local measures. Local measure like degree centrality [36] only considers local features to measure influence and ignores global features. While algorithms like closeness centrality [37] [38] and betweenness centrality [33] consider global features only.

In this work we have considered topological structure to identify influencers. What distinguish us from traditional approaches is that this research considers global as well as local features while computing influence.

Traditional Betweenness centrality algorithm measures the node capability to connect different regions of network, which is deceptive as it assigns equal value to nodes of high degree score which are central to local community and global bridge which connects different communities. This distinction should be considered to find bridging nodes in the graph, which connect densely connected modules in a network.

Thus, we have measured local influence using local centrality algorithm [25] and for global influence measurement we have proposed an efficient bridge centrality algorithm which identify critical global bridging nodes in on-line social networks. Here Bridging centrality is used to indicate most critical nodes interrupting the information flow in the network.

Finally, we analyze local and global centrality to compute final centrality and rank nodes accordingly and consider top k nodes as influential nodes having high global centrality and local centrality. The proposed framework to measure influence is shown in Figure 6.

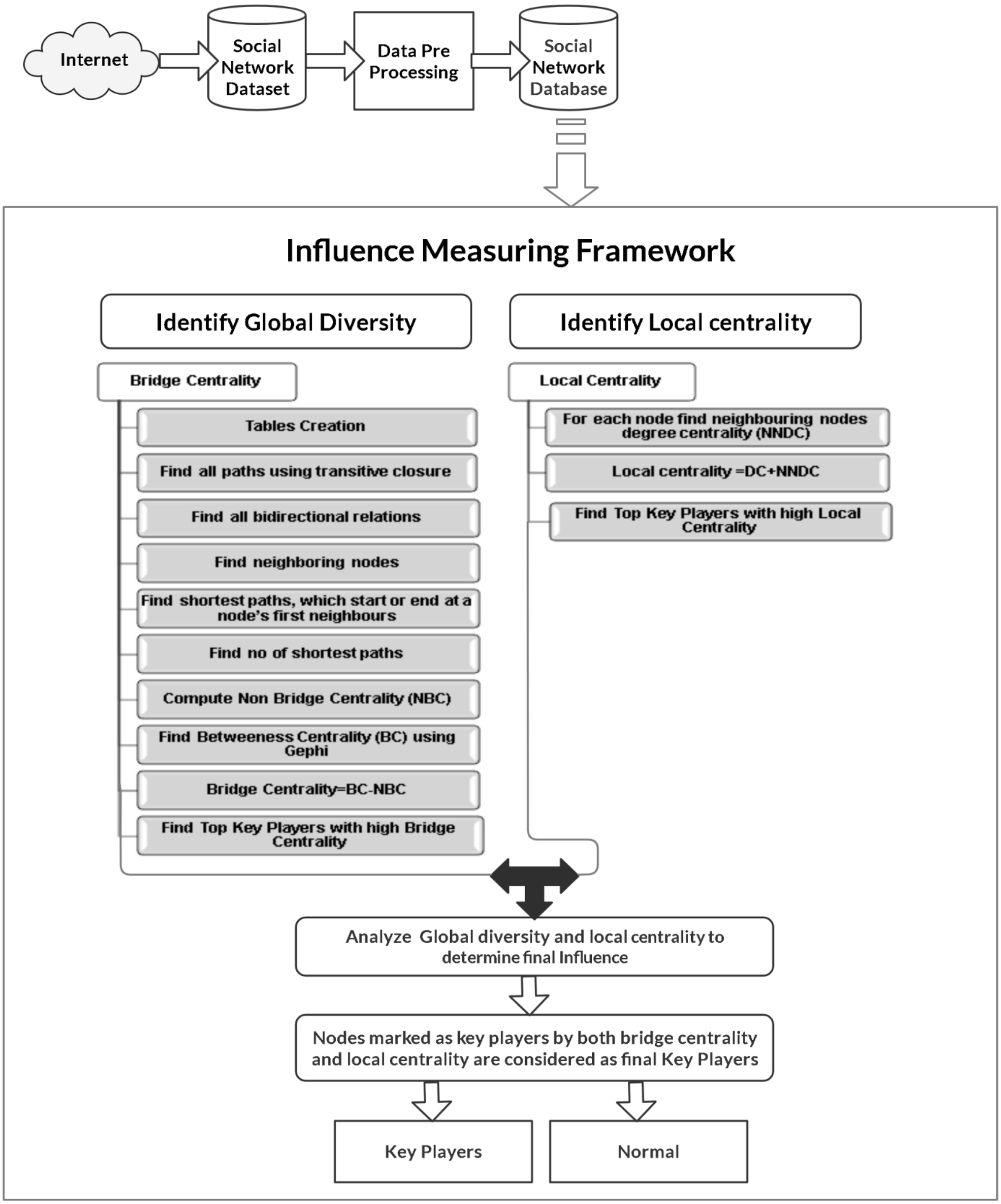
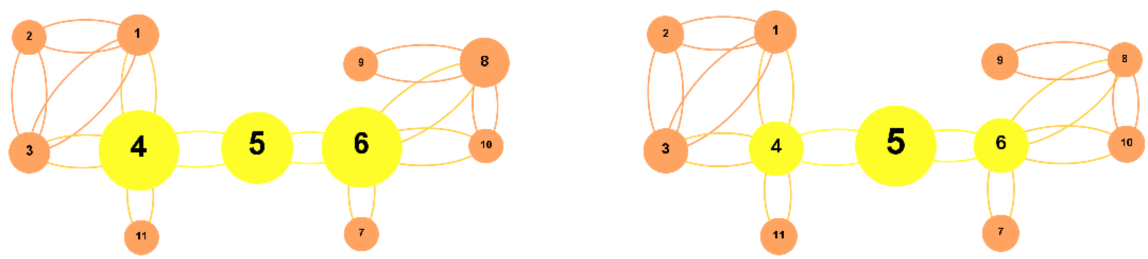


Figure 6: Our Proposed Framework to Measure Influence

3.1 Proposed Bridge Centrality Algorithm

The betweenness centrality (BC) [33] is the measure of centrality of node and is calculated as fraction of "shortest paths" (geodesic distance) which pass over the node of interest. BC has defect that it cannot distinguish between local central nodes (nodes which are central to its own community) and global central nodes (nodes that connect different communities) and thus attribute greater value to local centers than to global bridges as seen in figure 2. Whereas our proposed bridge centrality algorithm assigns higher scores to global node (a node connecting highly densely connected modules in a social network). For instance we want to highlight global nodes from local ones which are central to its own community, the simpler way is to discard betweenness centrality value for the node of interest, generated from the "shortest paths" which either start, or finish at a node's initial neighbors from the summation to calculate BC (Eq. 1.1)



(a) Betweeness Centrality **(b) Bridge Centrality**

Figure 7: The Figure shows (a) Betweeness Centrality and (b) Bridge Centrality for synthetic simple graph.

Betweeness centrality does not distinguish Global nodes (Node 5 acting as bridge between communities) from local ones (Nodes 4 and 5 with high degree scores). Rather Bridge Centrality give higher centrality score to node 5 that plays role of global bridge. Node size shows its centrality score.

Table 10: Centrality Analysis for Synthetic Simple Graph

Label	Degree Centrality	Betweeness Centrality	Bridge Centrality	Betweeness C Rank	Bridge C Rank	Bridge C Rank
5	2	25	16	2	1	1
4	4	27	3.5	1	2	2
6	4	27	3.5	1	2	2

The Table.10 depicts the centrality analysis for synthetic simple graph represented in Figure.7. It can be seen that node 4 and 6 have higher degree as well as betweeness centrality score, and are ranked as no 1 according to betweeness centrality, but according to bridge centrality they are ranked as second influential node. However, Bridge centrality has ranked node 5 as higher centrality node, although it has less degree and betweeness centrality. Thus we can conclude that betweeness centrality attributes higher score to nodes of high degree centrality, whereas bridgeness clearly discriminates between the global bridge and local centers by assigning higher score to the global bridge. More generally, Figure.7 and Table 10 clearly shows, as for the synthetic network, that "bridgeness" provides a better ranking than betweeness centrality method. More officially in a graph "G = (V;E)" where V assigns the set of nodes and E assigns the set of links to the definition of a node j's betweeness centrality stand as:

$$BC(j) = \sum_{i \neq j \neq k} \frac{\sigma_{ik}(j)}{\sigma_{ik}} = \sum_{\substack{i \notin N_G(j) \text{ and} \\ k \notin N_G(j)}} \frac{\sigma_{ik}(j)}{\sigma_{ik}} + \sum_{\substack{i \in N_G(j) \text{ or} \\ k \in N_G(j)}} \frac{\sigma_{ik}(j)}{\sigma_{ik}} \quad \textbf{(3.1)}$$

where the summation runs over any distinctive node pairs i and k; σ_{ik} symbolizes the number of shortest paths between i and k; while $\sigma_{ik}(j)$ is the number of such shortest paths running through j. Decomposing Betweeness Centrality into two parts the initial term indicates actually the global term, bridgeness centrality , where we consider shortest paths between nodes not in the neighborhood of j (N_G(j)) , whereas the next local term considers the shortest paths starting or ending in the neighbourhood of j . This definition also reveals that the bridgeness centrality value of a node j is always smaller or equal to the corresponding Betweeness Centrality value and they only vary by the local contribution of the first neighbours. Figure. 7 demonstrates the ability of bridgeness to highlight nodes that join different regions of a graph. Here the Betweeness Centrality (Fig. 7a) and bridgeness centrality values (Fig. 7b) calculated for nodes of the same network determine that bridgeness centrality allocates correctly the node , which is central globally, while Betweeness Centrality put forward a confusing image as it assigns greater centrality values to nodes with high degrees.

In our proposed bridge centrality measure we have calculated betweeness centrality (BC) score using Gephi tool and subtracted second term of equation 3.1 from BC to calculate bridge centrality value. The second term is calculated by following method using MS SQL.

- Step 1: Create table named "edges" with fields: Source node, target node, weight of edge connecting them and import dataset

- Step 2: Find all the paths between all nodes using transitive closure property and backtraceCTE recursive call
- Step 3: Find shortest paths which start or end in the neighborhood of node of interest. Calculate value which has to be subtracted from betweeness centrality to find bridge centrality

3.2 Local Centrality Algorithm

Local centrality (LC) is usually measured by traditional degree centrality algorithm, which only think through the number of the nearest neighbors". We have used local centrality algorithm which take into account the number of the nearby and the next nearest neighbors of node .
LC is calculated by following method using relational database.

- Calculate degree centrality using Gephi tool
- Calculate bidirectional edges of all links and add degree centrality of neighboring nodes for each node.
- Calculate final local centrality by adding neighboring nodes aggregated centrality and centrality of node of interest.

3.3 Influential nodes or key-players identification

- We have selected top k influential nodes from the network for both bridge centrality measure and local centrality measure separately after ranking them. To estimate how many key-players should be selected we use formula given in Eq 3.2 whereas k=10 in our case and N are total no of nodes. This gives estimated nodes to be selected as key-players for each dataset.

$$\mathbf{k\ (\ log(\ N\ /\ k\ )\ )} \qquad \textbf{(3.2)}$$

- Nodes which are marked as key-players by both bridge centrality as well as local centrality measure are considered as final Key-players or influential nodes, whereas others are marked as Normal ones.

In this way we can find the most influential users from social network which helps in influence dissemination. Thus, this methodology considers global diversity and local structure which is comparatively more efficient and accurate.

CHAPTER 4: IMPLEMENTATION OF PROPOSED FRAMEWORK

Our Proposed framework of global and local centrality measure is implemented for identification of the Key-players, and then classification methods are applied for evaluation of generated results. This process can be divided into 5 phases.

1. Social Network Selection (Dataset)
2. Pre-Processing of the Data
3. Influential nodes identification using Global and Local Centrality measures. (For Global diversity Bridge Centrality Algorithm is used and for calculation of Local Influence Local Centrality Method is used.)
4. Visualization
5. Feature Selection, and Accuracy evaluation using Cross Validation and voting method (discussed in results and discussion section)

3.1 Tool used

Gephi which is an open source application is used for calculation of the betweeness and degree metrics.

MS SQL Server, is also used to calculate bridge centrality and local centrality.

3.2 Datasets

Table 11: Network Specifications

Serial No.	Network	Nodes	Edges (Directed)
Case I	Noordin Terrorist Network	79	400
Case II	Online Huawei Social Network (Instagram)	1000	9865
Case III	Air Traffic Control	1226	2615
Case IV	Soc Firm Hi Tech	33	147

3.2.1 Data Pre-processing

Data pre-processing is essential part of the Social Media Network Analysis. Data sets are often noisy. Pre-processing of data has several steps some of them are removal of ambiguities and anomalies, either merge or remove the same nodes to remove data repetition. Pre-processing also includes the transformation of the data into desired format and saving the required file at desired location. [27].

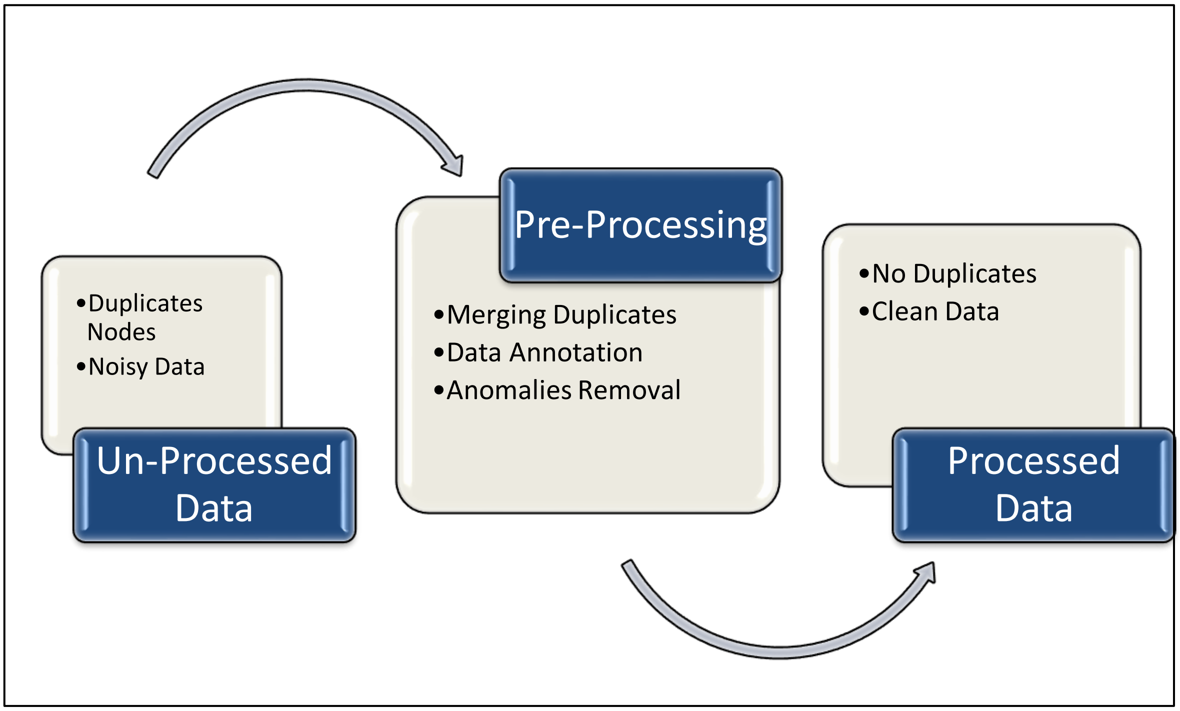


Figure 8: Data Pre-processing

3.2.2 Anomalies Removal

3.2.2.1 Missing Values

- Remove the valusses with the missing terms. Removal should not delete more than the 6% of the Data
- Fill the missing values. Either replace with frequent or average value
- Use the predicting algorithm to predict the missing values. Predicting algorithm include Decision Tree, Classification Model and Regression

3.2.2.2 Aggregation

The Purpose of the aggregation is to remove the irrelevant features and merge the same attributes. For example, date and time can be merged, similarly Month, day and year can be combined to create the single attribute.

Aggregation has several benefits including Reduction of the Data, Abstract view and data stability.

3.2.2.3 Irrelevant features

Irrelevant features, we mean sometime the data contain such information that is of no use in the analysis and visualization. Several attributes of an entity have information that is not useful in the analysis and prediction e.g. Students' ID is often irrelevant to the task of predicting student's grades

3.2.3 Merging Duplicates

If node with same id is repeated it is considered once to remove duplication by merging duplicate nodes.

3.3 Noordin Terrorist Network Dataset (CASE - I)

The Data of the Noordin Top Terrorist Network[24] includes data about associations of terrorist entities with terrorist / insurgent organisations, academic institutions, companies, and religious institutions.

It also labels which people are colleagues, relatives, friends, and co-religionists, and details which people offered logistical assistance or participated in training activities, terrorist activities, and conferences.

Dataset Collection

The "Noordin Top Terrorist Network" data was taken mainly from the International Crisis Group's "Terrorism in Indonesia: Noordin's Networks," which includes relational data on the 79 people mentioned in Appendix C of that publication .

3.3.1 Centrality Measures

Table 12: Metrics Values of Influential Nodes (Case I)

Id	Label	Bridge Centrality	Semi Local Centrality	Role
59	Noordin Mohammed Top	1408.10199	624	Key-player
4	Abdullah Sunata	460.831298	224	Key-player
13	Ahmad Rofiq Ridho	339.991069	392	Key-player
23	Azhari Husin	257.420997	424	Key-player
45	Iwan Dharmawan	252.664116	316	Key-player
33	Hambali	107.649999	202	Key-player
60	Purnama Putra	138.028210	254	Key-player
73	Ubeid	79.6827643	236	Key-player

3.3.2 Key Player Identification

Key-players:

Key-players are the nodes that are more important/ influential than the other ones in the network. The table below represents the most influential/ Key nodes of the Noor Din Terrorist Network with higher global and local metrics than that of other nodes in the Network.

Eight nodes are identified as key-players/ influential nodes and remaining nodes are considered as Normal ones in this dataset.

Table 13: Key Player Identified in Case I

Label	Role	Label	Role	Label	Role
Noordin Mohammed	Key-player	Azhari Husin	Key-player	Purnama Putra	Key-player
Hambali	Key-player	Iwan Dharmawan	Key-player	Ubeid	Key-player
Ahmad Rofiq Ridho	Key-player	Abdullah Sunata	Key-player		

3.3.3 Visualization

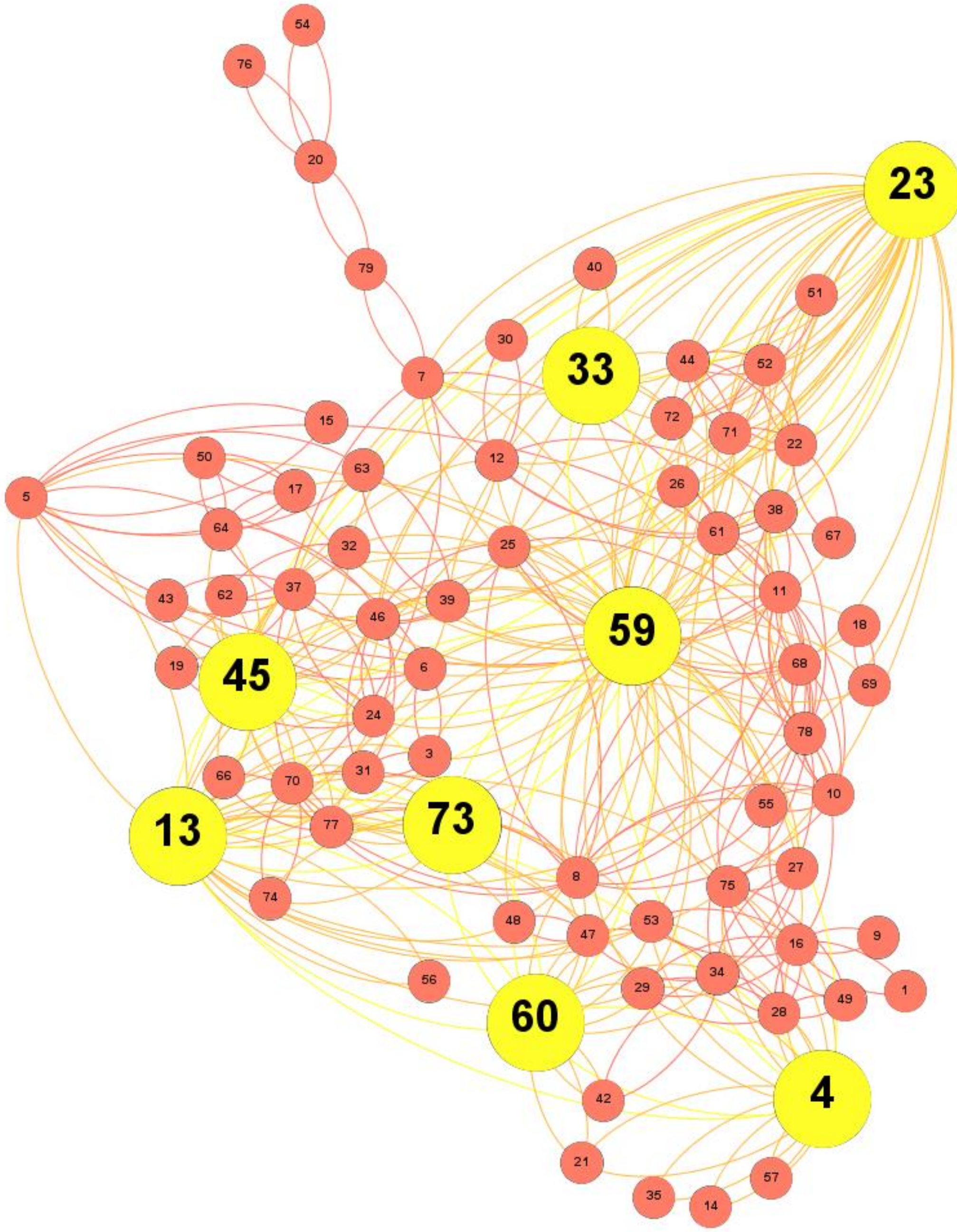


Figure 9: Noordin Top Terrorist Network (Case I) Visualization (Without Names)

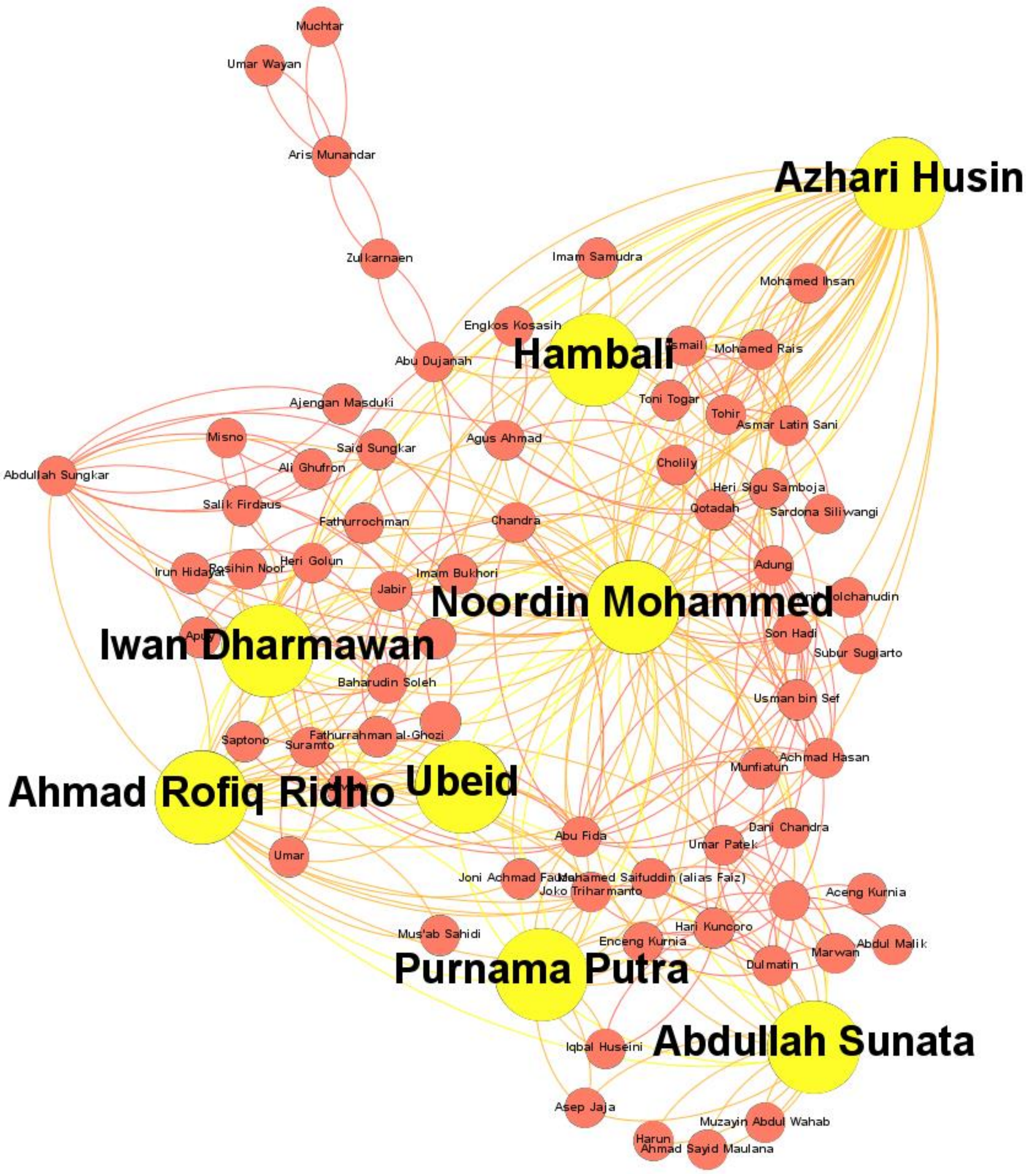


Figure 10: Noordin Top Terrorist Network (Case I) Visualization (With Names)

As seen in Figure above size of nodes depends on its metrics value, node with greater size have higher centrality metrics and vice versa.

The yellow larger dots represent Key-players identified in the network. They are more influential than the other individuals present in the network as they have higher global and local metrics values and other are considered as Normal nodes.

3.4 Huawei Online Business Dataset (CASE - II)

Huawei Technologies Co. Ltd. is a Chinese multinational networking company with headquarters in Shenzhen, Guangdong, with telecommunications facilities and services. It is the world's biggest manufacturer of telecommunications machinery, having surpassed Ericsson in 2012. In Fortune Magazine, Huawei became 83rd of Fortune Global 500 in 2017.

Application:

Research and Development: Huawei R&D is now using the instruments and methods of social network assessment to improve its company positions. Huawei Company's major achievement is that it promotes its goods through social media.

Dataset Collection

This network was collected by crawling the pages of the Instagram Huawei social media platform. Web crawlers API extract posts and remarks from Instagram.

3.4.1 Centrality Measures

Table 14: Metrics Values of Influential Nodes (Case II)

Id	Label	Bridge Centrality	Semi Local Centrality	Role
92	Farah Samad	7116.797	431	Key-player
698	Alveena	6510.52	510	Key-player
294	Alexis	6056.993	504	Key-player
305	Hallie	5709.086	424	Key-player
691	Ishku Ishku	5590.225	448	Key-player
112	Larissa	5517.392	392	Key-player
4	Porter Devries	5451.984	402	Key-player
188	Waseem	5366.571	406	Key-player
71	Muhammad Rehan	5154.732	422	Key-player
204	Kai	5084.549	404	Key-player
556	KH Hassan	4755.948	452	Key-player
115	Archie	4658.19	422	Key-player

500	Misno	4624.053	442	Key-player
559	Oliver	4483.264	394	Key-player
103	Robbie	4278.208	372	Key-player
408	Aleena	4214.568	414	Key-player
175	Naja	4188.324	370	Key-player
6	Ladawn Creason	4160.996	378	Key-player
776	Danish Saifullah	4098.818	450	Key-player
523	Erin	3935.482	404	Key-player
481	Ramos	3895.791	416	Key-player
914	Betsy	3866.713	390	Key-player
465	Evan	3851.907	422	Key-player

3.4.2 Key Player Identification

Key Players:

Key-players are the nodes in a social network that are more important/ influential than the other ones in the network. The table below represents the most influential/ Key customers of the Huawei Network with higher global and local metrics values than that of other nodes in the Network.

Twenty three nodes are identified as key-players/ influential nodes and remaining nodes are considered as Normal ones in this dataset.

Table 15: Key Player Identified in Case II

Label	Role	Label	Role	Label	Role
Farah Samad	Key-player	Waseem	Key-player	Robbie	Key-player
Alveena	Key-player	KH Hassan	Key-player	Aleena	Key-player
Alexis	Key-player	Kai	Key-player	Naja	Key-player
Ishku Ishku	Key-player	Archie	Key-player	Hallie	Key-player
Muhammad Rehan	Key-player	Misno	Key-player	Erin	Key-player

Porter Devries	Key-player	Oliver	Key-player	Ramos	Key-player
Ladawn Creason	Key-player	Evan	Key-player	Betsy	Key-player
Danish Saifullah	Key-player	Larissa	Key-player		

3.4.3 Visualization

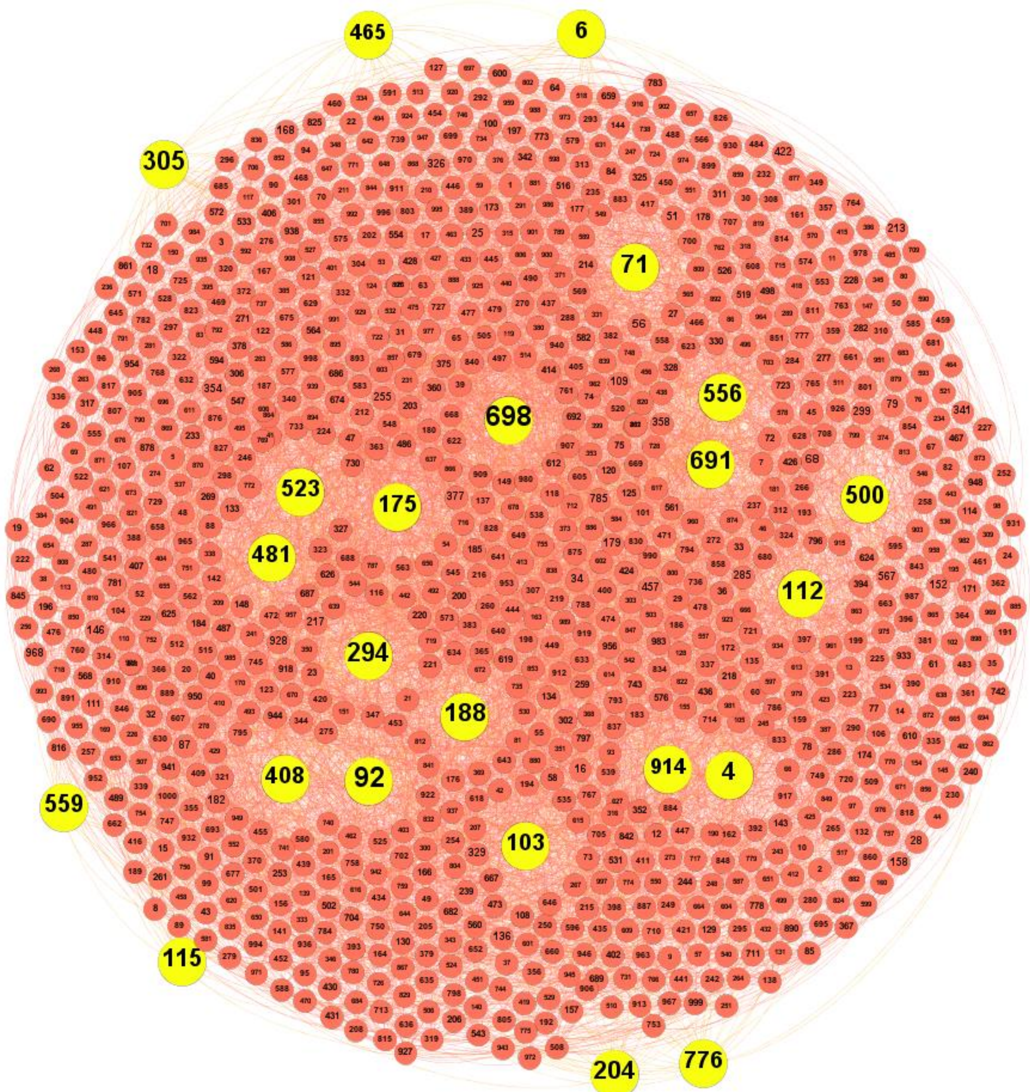


Figure 11: Huawei (Case II) Visualization (Without Names)

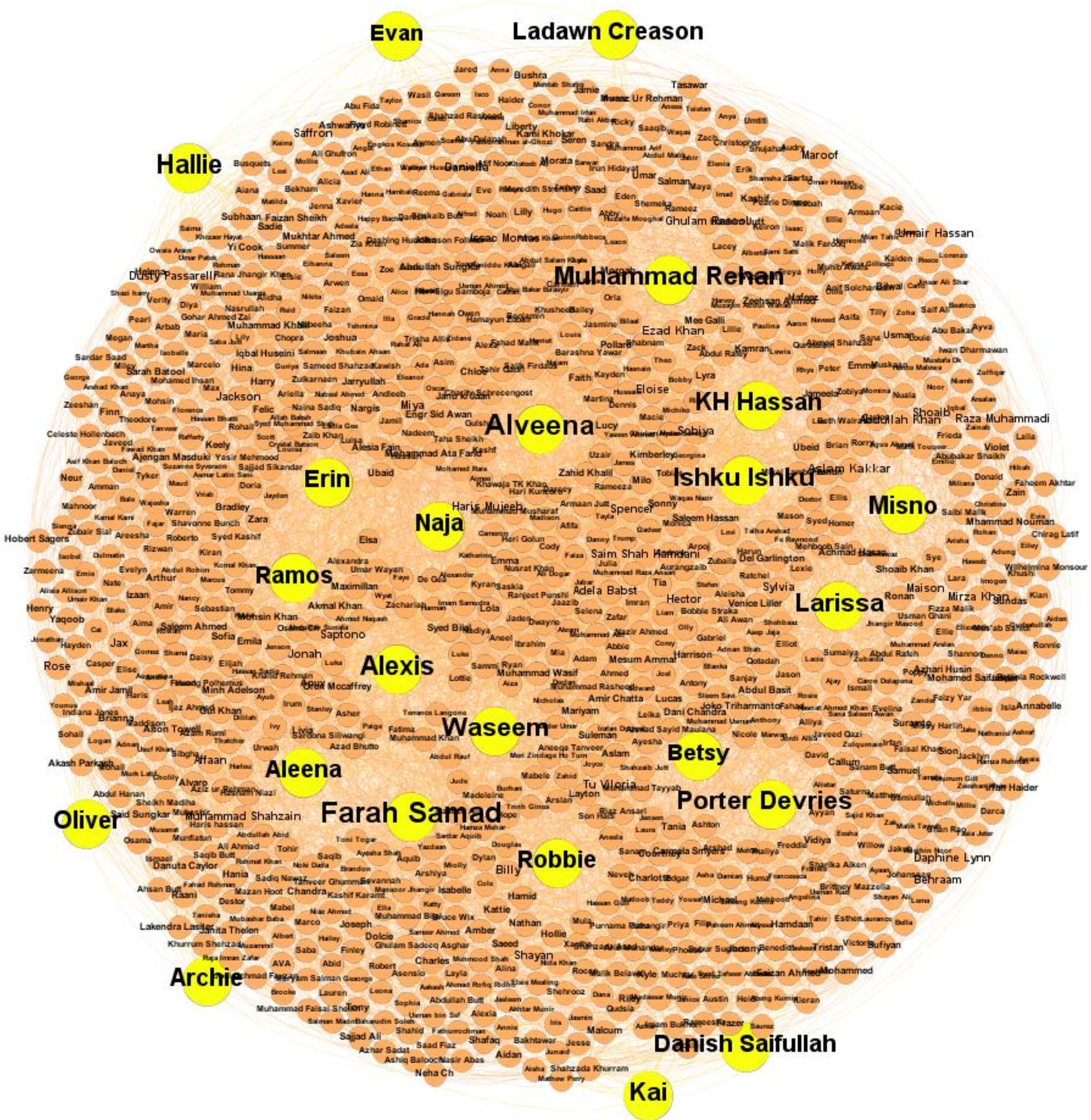


Figure 12: Huawei (Case II) Visualization (With Names)

As seen in Figure above size of nodes depends on its metrics value, node with greater size have higher centrality metrics and vice versa.

The Yellow larger Dots Represent the Key player identified in the network. They are more important than the other individuals present in the network as they have higher metrics values and remaining are considered as Normal nodes.

3.5 Air Traffic Control Dataset (CASE - III)

This network was constructed from the USA's FAA (Federal Aviation Administration) National Flight Data Center (NFDC), Preferred Routes Database. Nodes in this network represent airports or service centers and links are created from strings of preferred routes recommended by the NFDC [50] [51]. In the figure below yellow nodes represents the global international bridge airports in Air traffic control Network identified by bridge centrality algorithm. The node size increases with increase in bridge centrality value.

.

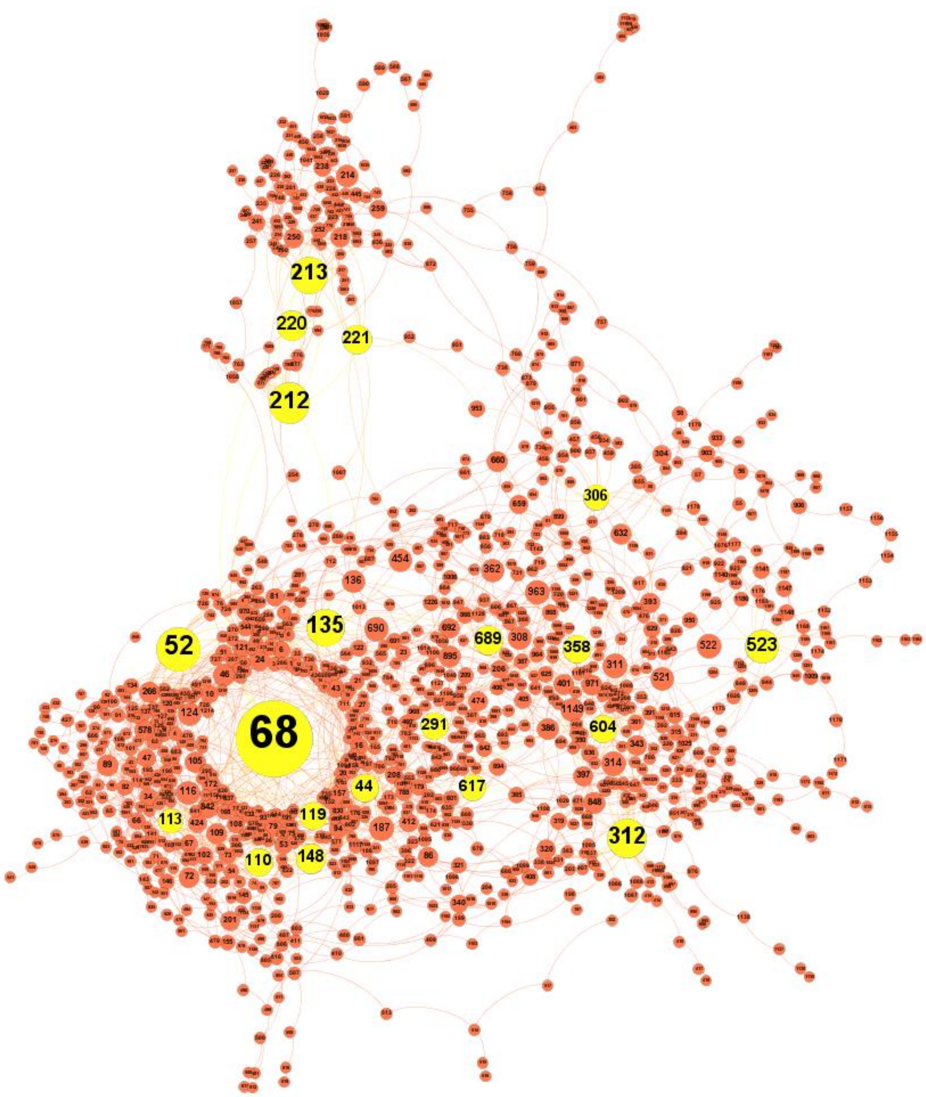


Figure 13: Global International Bridge Airports Visualization in Air Traffic Control

3.5.1 Centrality Measures

Table 16: Metrics Values of Influential Nodes (Case III)

Id	Label	Bridge Centrality	Semi Local Centrality	Role
68	68	204725.9	302	Key-player
52	52	103412.5	335	Key-player
312	312	85710.34	165	Key-player
135	135	83832.02	174	Key-player
148	148	58609.04	218	Key-player
110	110	51926.14	225	Key-player
119	119	51313.56	135	Key-player
44	44	50631.72	249	Key-player
113	113	42856.69	284	Key-player
187	187	42581.84	168	Key-player
116	116	39212.37	242	Key-player
308	308	38361.15	132	Key-player
124	124	36933.5	178	Key-player
46	46	34040.02	235	Key-player
89	89	31930.56	198	Key-player
266	266	30558.54	127	Key-player

3.5.2 Key Player Identification

Key-players are the nodes in a social network that are more central than the other ones in the network. The table below represents the most important/ Key nodes of the Air Traffic Network with higher global and local metrics than that of other nodes in the Network.

These central nodes represent domestic centers and the global international bridge airports in Air traffic control Network.

Table 17: Key Player Identified in Case III

Label	Role	Label	Role	Label	Role
68	Key-player	44	Key-player	308	Key-player

52	Key-player	113	Key-player	124	Key-player
312	Key-player	187	Key-player	46	Key-player
135	Key-player	116	Key-player	119	Key-player
148	Key-player	110	Key-player		Key-player

3.5.3 Visualization

Figure 14: Air Traffic Control (Case III) Visualization

As seen in Figure above size of nodes depends on its metrics value, node with greater size have higher centrality metrics and vice versa. The Yellow larger Dots Represent the Key-player identified in the network. They are more important than the other individuals present in the network as they have higher metrics values.

3.6 Soc_Firm_Hi_Tech Dataset (CASE - IV)

This dataset is taken from a graph and network repository containing hundreds of real-world networks and benchmark datasets.

"Soc_Firm_Hi_Tech" dataset is a social network dataset, which consists of 33 nodes and 147 edges.

3.6.1 Centrality Measures

Table 18: Metrics Values of Influential Nodes (Case IV)

Id	Label	Bridge Centrality	Semi Local Centrality	Role
29	29	178.5832	346	Key-player
33	33	72.33210	263	Key-player
24	24	76.16515	203	Key-player
4	4	62.72002	244	Key-player

3.6.2 Key Player Identification

Key-players are the nodes in a social network that are more central than the other ones in the network. The table below represents the most important/ Key nodes of the Soc Firm Hi Tech Network with higher global and local metrics than that of other nodes in the Network.

Node 29, 33, 4 and 24 are considered as Key-players or the most influential node in the network. The table 19 shows key-players identified for "Soc Firm Hi Tech" dataset.

Table 19: Key Player Identified in Case IV

Label	Role	Label	Role
29	Key-player	33	Key-player
4	Key-player	24	Key-player

3.6.3 Visualization

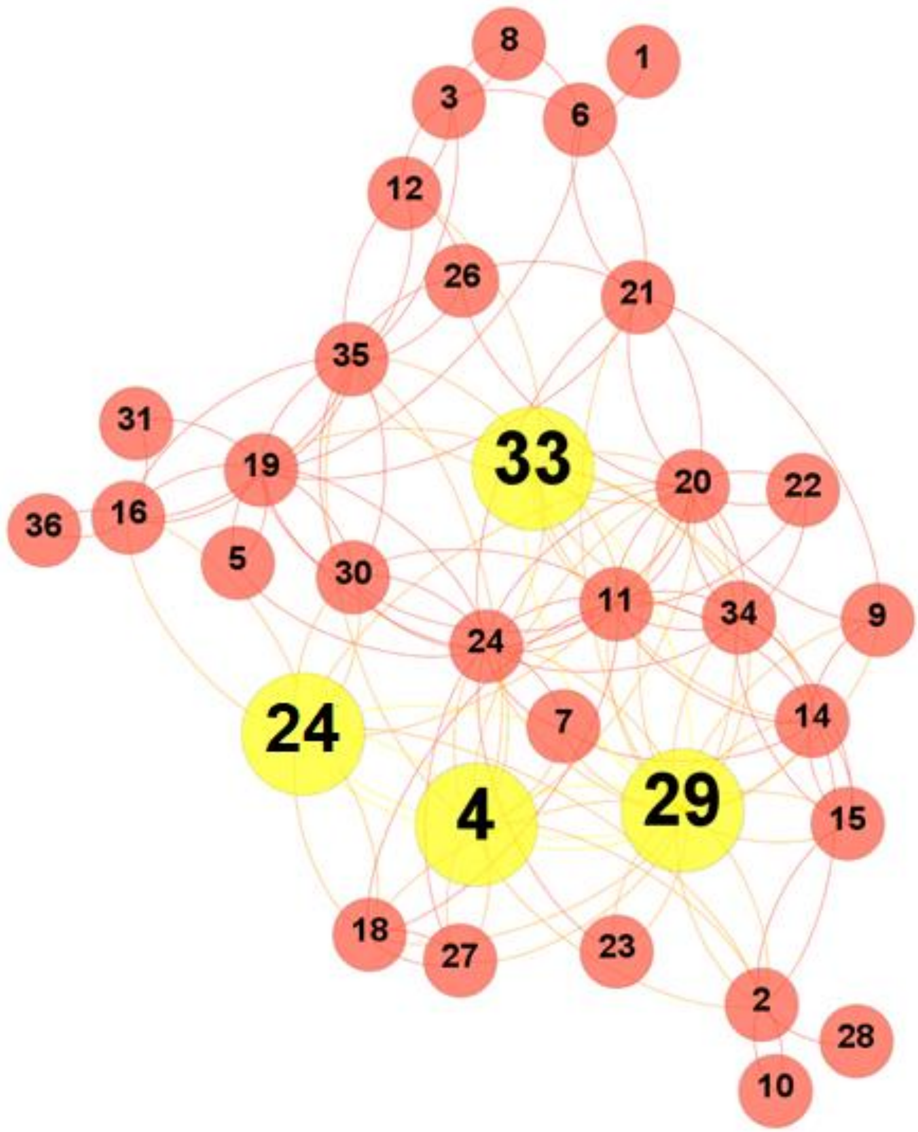


Figure 15: Soc_Firm_Hi_Tech (Case IV) Visualization

As seen in Figure above size of nodes depends on its metrics value, node with greater size have higher centrality metrics and vice versa.

The Yellow larger Dots Represent the “Key-player” identified in the network. They are more important than the other individuals present in the network as they have higher metrics values and remaining are considered as Normal nodes.

CHAPTER 5: RESULTS AND DISCUSSION

This Chapter of the Research covers the results and analysis of the proposed method for identification of the influential nodes (Key-players) in the network. The detailed analysis of proposed system is performed in this section. We carry out a series of experimentations to evaluate the accuracy of the suggested framework.

The performance of recommended system is measured using specificity, sensitivity, and accuracy. We simulate four real networks in the experiments, which include Noordin Terrorist Network, Online Huawei Social Network (Instagram), Air Traffic Control and Soc_firm_Hi_Tech.

5.1 Feature Selection

We have calculated betweeness centrality, degree centrality, bridge centrality and local centrality. Feature selection/ relevance analysis are done to select the feature of interest from the feature set. Betweeness centrality and degree centrality is of no use as it is irrelevant to our analysis. So, bridge centrality and local centrality is selected on the basis of the nature of the analysis being performed on the dataset as shown in table below.

Table 20: Selected Features Set

Features Selected	
Global Diversity	**Local Influence**
Proposed Bridge Centrality Algorithm	Local Centrality Algorithm

5.2 Classification and Prediction

Classification uses the given input to predict the outcome. Classification is used to solve out the wide range of the problems i.e. either simple or complex problems [28] [29].

5.2.1 Rapid Miner

Rapid miner is an open source software tool that provides the users the platform to train and test the classification models for various applications [33]. We have used Rapid miner tool to test and train the proposed hybrid classifier to predict the key-players in the networks.

5.2.2 Rapid Minor Process

The Figure 1 (Appendix) shows the main and level 0 Process having Retrieved Dataset and Cross Validation operated connected to Output for performance evaluation. While Figure 2 (Appendix) show the level 1 process. It reveals what is inside cross validation operator. Cross Validation operator contains Voting operator which combines the generated result of all three classifiers. As far as 3rd level process is concerned, it contains the operators to measure the performance and prediction classifiers as shown in Figure 3 (Appendix). Further figures show single classifiers and hybrid classifiers along with generated result example.

5.2.3 Classifiers

Classifiers use the phenomenon of training and testing. The division of the provided data into two disjoint sets is performed i.e. training and testing subsets. $f_{(x)} = f_{xtr} + f_{xte}$ Training f_{xtr} and testing f_{xte} subsets contains the objects of the both classes i.e. (Key-players and Normal). The objects in these subsets are added randomly, so that the subsets are biasness free. The selected classifier is trained using the "training subset" f_{xtr} and, performance of the model is evaluated using the "testing subset" f_{xte}.

5.2.3.1 Naïve Bayes

Naive Bayes is a low-variance, high-bias classifier , and even with a tiny information set, it can create a good model. It's easy to use and cost-effective computationally. Typical instances of use include categorization of text, including identification of spam, assessment of feelings, and recommendation systems.

5.2.3.2 Decision Tree

Decision Tree divides the entire dataset into smaller sub-sets and creates a decision tree. ID3 developed by "J. R. Quinlan" is the basis and primary algorithm used for decision trees.

5.2.3.3 SVM Algorithm

In machine learning, support-vector machines(SVMs) are supervised learning models with related learning algorithms that examine data used for classification and regression analysis.

5.2.4 Cross Validation

Cross validation is process which is used to estimate and assess the performance of a created and designed model. The "Cross-validation operator" is mostly used to test the performance of the trained operator to be used in the practice.

Cross validation is process that is dependent on two sub processes. These two processes are called the testing sub phase and the training sub phase. Throughout the training phase the model is trained on the labeled dataset. Then the trained model is applied in the testing phase. The results of the testing phase determine the performance of the trained model.

Example Set is divided into K subsets of equal sizes. The K-1 subsets are than used for the training and the remaining one is used for the testing of the trained model. The procedure is reiterated K many times such that the data once used for the testing is not tested again. The results from all the iterations are combined or either averaged to produce the single output estimation.

5.2.5 Voting

"Ensembles voting (Rapid Minor Operator)" is used to combine the predictive power of more than one classifier to achieve better outcomes than the single classifier outcomes. Three classifiers have been trained to predict the appropriate result, these are C_{SVM}, C_{dt}, C_{Naive}. For each test instance v_i a_i is calculated for "n" classifiers trained by means of feature vector f vi. Voting techniques based on prediction of majority lastly predict 1 (key-player) or 0 (ordinary member).

$$\boldsymbol{a = vote(a_{i(c)}, a_{i(c+1)}, .., a_{i(c+n)})} \qquad \text{Where}$$

$$a_{i(c)=}C_{SVM}(f_{vi}),$$

$$a_{i(c+1)=}C_{dt}(f_{vi}),$$

$$a_{i(c+n)=}C_{svm}(f_{vi}).$$

5.2.6 Performance Evaluation Model using Cross Validation

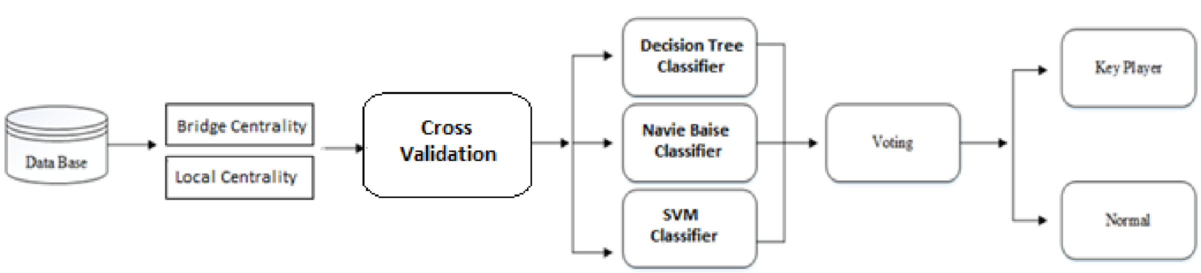


Figure 16: Hybrid Classifier Performance Evaluation Model using Cross Validation

5.3 Key Player Evaluation

The most significant area of the social network analysis is to identify the key-players. The exact definition of the key-player is dependent on the type of social network analysis. The simple and abstract definition of key-players are those nodes and individuals, which are more important in the network than the other. The importance of the key-player solely depends upon the type of analysis and the network. The Classifiers predict the role of the nodes on the basis of the trained dataset.

Cross Validation is a very useful technique for assessing the effectiveness of the model. K-fold Cross-Validation along with voting method is used for measuring accuracy of proposed method of influential nodes identification with the help of tool named Rapid Miner.

5.3.1 Classification

The detailed results of the proposed framework are depicted in this section. The performance rating and throughput of proposed method is calculated using various measures, which include:

- Sensitivity (SEN),
- Specificity (SPEC),
- Accuracy (ACC)

These measures are calculated using Equation (5.1), Equation (5.2) and Equation (5.3) respectively.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{5.1}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{5.2}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \tag{5.3}$$

- "TP are the number of the key-players that are recognized correctly by the model known as true positive
- TN is called the true negatives. It is the number of the normal nodes that are recognized correctly.

- FP are false positives , it defines the number of the normal nodes that are recognized as the key-players.
- FN also called False Negatives, it illustrates the number of the key-players that are recognized as the normal during the classification phase.

The trained models are tested with the help of cross validation technique. The Value of K is selected in such a way that the about 80% of data has been used as training and remaining 20% data as testing. The experimental procedures are repeated K=5 times. The Table.21 illustrates result set of proposed paradigm for key-player detection on all networks given in case studies.

Table 21: Statistical Performance Evaluation of Proposed Framework for key Player detection

Case study	Sensitivity	Specificity	Accuracy
I	62.50%	98.41%	94.29%
II	81.25%	100.00%	98.55%
III	83.33%	99.67%	99.43%
IV	100.00%	50.00%	93.81%

The hybrid classifier is compared with individual SVM, Naive Bayes and Decision Tree classifiers. The table.22 shows the comparison of all the terms of accuracy for key-player detection.

Table 22: Comparison of Hybrid Classifier with Existing classifiers

Methods	Case Study-I	Case Study-II	Case Study–III	Case Study-IV
SVM	94.29%	97.12%	99.10%	87.62%
Naive Bayes	92.86%	97.60%	97.47%	84.76%
Decision Tree	90.00%	99.52%	99.59%	90.95%
Proposed Hybrid Classifier	94.29%	98.55%	99.43%	93.81%

The proposed methodology has been tested by means of four case studies. The outcomes noticeably prove the validity and correctness of proposed frame work.

5.3.2 Comparison

The tables below illustrates the prediction outcome for top 40 nodes including key-players and normal nodes regarding the classification of all the 3 classifiers and the one proposed hybrid classifier for all four case studies.

Table 23: Comparison of Classifiers (CASE I)

Label	Role	Prediction (Role) Hybrid Classifier	Prediction (Role) Decision Tree Classifier	Prediction (Role) Naive Bayes Classifier	Prediction (Role) SVM Classifier
60	Key-player	Normal-nodes	Key-player	Normal-nodes	Normal-nodes
13	Key-player	Key-player	Key-player	“Key-player”	Key-player
73	Key-player	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
4	Key-player	“Key-player”	Normal-nodes	Key-player	Key-player
23	Key-player	“Key-player”	Key-player	Key-player	Key-player
59	Key-player	“Key-player”	Key-player	Key-player	Key-player
45	Key-player	“Key-player”	Key-player	Key-player	Key-player
33	Key-player	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
79	Normal-nodes	Normal-nodes	Normal-nodes	“Key-player”	Normal-nodes
20	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
75	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
24	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
29	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
27	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
53	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
69	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes

14	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
35	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
40	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
15	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
9	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
5	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
39	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
11	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
47	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
26	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
67	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
43	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
66	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
30	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
76	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
1	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
31	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
7	Normal-nodes	Key-player	Normal-nodes	Key-player	Key-player
60	Key-player	Normal-nodes	Key-player	Normal-nodes	Normal-nodes
13	Key-player	Key-player	Key-player	Key-player	Key-player
73	Key-player	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
4	Key-player	Key-player	Normal-nodes	Key-player	Key-player

Butt, Wasi Haider, M. Usman Akram, Shoab A. Khan, and Muhammad Younus Javed [54] in his research on covert network analysis used “Noordin Top Terrorist Network” for Key-players detection in social network and found out that only 8 nodes are influential enough to be considered as Key-players. Similarly as seen in Table.23 our proposed method has also ranked 8 nodes as Key-players and remaining ones are ranked as Normal-nodes nodes.

Table 24: Comparison of Classifiers (CASE II)

Label.	Role	Prediction (Role) Hybrid Classifier	Prediction (Role) Decision Tree Classifier	Prediction (Role) Naive Bayes Classifier	Prediction (Role) SVM Classifier
183	Key-player	Key-player	Key-player	Key-player	Key-player
72	Key-player	Key-player	Key-player	Key-player	Key-player
162	Key-player	Key-player	Key-player	Key-player	Normal-nodes
117	Key-player	Key-player	Key-player	Key-player	Key-player
80	Key-player	Key-player	Key-player	Key-player	Key-player
45	Key-player	Key-player	Key-player	Key-player	Normal-nodes
118	Key-player	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
27	Key-player	Key-player	Key-player	Key-player	Key-player
167	Key-player	Key-player	Key-player	Key-player	Key-player
74	Key-player	Key-player	Key-player	Key-player	Key-player
1	Key-player	Key-player	Key-player	Key-player	Key-player
116	Key-player	Key-player	Key-player	Key-player	Key-player
200	Key-player	Key-player	Key-player	Key-player	Normal-nodes
62	Key-player	Key-player	Key-player	Key-player	Key-player
78	Key-player	Normal-nodes	Key-player	Normal-nodes	Normal-nodes
132	Key-player	Normal-nodes	Key-player	Normal-nodes	Normal-nodes
130	Normal-nodes	Normal-nodes	Normal-nodes	Key-player	Normal-nodes
147	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
59	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
180	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
36	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
194	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
67	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
14	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes

23	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
183	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
72	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
162	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
117	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
80	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
45	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
118	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
27	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
167	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
74	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
1	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
116	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
200	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
62	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes

Table 25: Comparison of Classifiers (CASE III)

Label.	**Role**	**Prediction (Role) Hybrid Classifier**	**Prediction (Role) Decision Tree Classifier**	**Prediction (Role) Naive Bayes Classifier**	**Prediction (Role) SVM Classifier**
119	Key-player	Normal-nodes	Key-player	Key-player	Normal-nodes
46	Key-player	Key-player	Key-player	Key-player	Key-player
72	Key-player	Key-player	Key-player	Key-player	Normal-nodes
312	Key-player	Key-player	Key-player	Key-player	Key-player
113	Key-player	Key-player	Key-player	Key-player	Key-player
187	Key-player	Key-player	Key-player	Key-player	Normal-nodes
266	Key-player	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
135	Key-player	Key-player	Key-player	Key-player	Key-player

110	Key-player	Key-player	Key-player	Key-player	Key-player
116	Key-player	Key-player	Key-player	Key-player	Key-player
109	Key-player	Key-player	Key-player	Key-player	Key-player
52	Key-player	Key-player	Key-player	Key-player	Key-player
148	Key-player	Key-player	Key-player	Key-player	Key-player
44	Key-player	Key-player	Key-player	Key-player	Key-player
308	Key-player	Normal-nodes	Normal-nodes	Key-player	Normal-nodes
68	Key-player	Key-player	Key-player	Key-player	Key-player
124	Key-player	Key-player	Key-player	Key-player	Normal-nodes
89	Key-player	Key-player	Key-player	Key-player	Normal-nodes
291	Normal-nodes	Normal-nodes	Normal-nodes	Key-player	Normal-nodes
311	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
454	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
895	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
578	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
842	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
692	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
86	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
659	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
157	Normal-nodes	Normal-nodes	Normal-nodes	Key-player	Normal-nodes
6	Normal-nodes	Normal-nodes	Normal-nodes	Key-player	Normal-nodes
238	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
67	Normal-nodes	Normal-nodes	Normal-nodes	Key-player	Normal-nodes
16	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
206	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
179	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
319	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
1147	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
629	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes

176	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
788	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes

Table 26: Comparison of Classifiers (CASE IV)

Label.	Role	Prediction (Role) Hybrid Classifier	Prediction (Role) Decision Tree Classifier	Prediction (Role) Naive Bayes Classifier	Prediction (Role) SVM Classifier
4	Key-player	Normal-nodes	Key-player	Normal-nodes	Normal-nodes
13	Key-player	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
33	Key-player	Key-player	Key-player	Key-player	Normal-nodes
29	Key-player	Key-player	Key-player	Key-player	Normal-nodes
19	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
6	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
14	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
15	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
23	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
31	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
1	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
24	Normal-nodes	Normal-nodes	Normal-nodes	Key-player	Normal-nodes
21	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
9	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
18	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
27	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
5	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
2	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
3	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
10	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
8	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
36	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes

35	Normal-nodes	Normal-nodes	Normal-nodes	Key-player	Normal-nodes
12	Normal-nodes	Normal-nodes	Key-player	Normal-nodes	Normal-nodes
16	Normal-nodes	Normal-nodes	Key-player	Normal-nodes	Normal-nodes
26	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
7	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
34	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
20	Normal-nodes	Normal-nodes	Normal-nodes	Key-player	Normal-nodes
11	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
30	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
22	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes
28	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes	Normal-nodes

CHAPTER 6: CONCLUSION AND FUTURE WORK

It has become essential to study the influence spread mechanism owing to its immense impact on social networks but due to the complex nature of SIA it is actually challenging. A lot of practitioners have been working in this area. In this research, we have represented a novel approach to identify most influential nodes which can help in influence dissemination. Influence is mostly spread by nodes which have more global influence connecting communities of the social network. Moreover, locally central nodes spread influence in its own community. Thus, Global and local both features should be considered while recognizing influence spreaders. Our proposed approach computes influence by considering global diversity as well as local features of the network topology to identify influential nodes. For global diversity we have proposed bridge centrality algorithm, where as local is measured with the help of local centrality algorithm. The nodes with higher bridge centrality and local centrality metrics are considered top key-players/ Influential nodes.

Moreover, we have overcome limitation of Betweeness Centrality (BC) and Degree Centrality (DC) measure. BC has defect that it cannot distinguish between local central nodes and global central nodes and thus attribute greater score to local centers than to global bridges. Whereas our proposed bridge centrality algorithm assigns higher scores to global nodes . Traditional degree centrality algorithm, only considers the number of the nearest neighbors . We have used local centrality algorithm which take into consideration the number of the nearest and the next nearest neighbors of node.

The proposed approach is applied on Huawei Online Dataset of Instagram, Air Traffic control dataset, Noordin Top Network and Soc_Firm_Hi_Tech dataset. The thesis used a cross validation technique to measure the accuracy of proposed method in which the prediction of the influential nodes is done by the voting or various classifiers. The hybrid classifier achieves 99% accuracy which is up to the mark. The overall aim of our research is to improve social network analysis to enhance the identification of influential nodes/ key-players. The experimental results show that the proposed technique can rank influential nodes efficiently and accurately on social networks.

This manuscript has also discussed several opportunities for the future research. In the future, we recommend researchers to expand our approach to study “multi-layered social networks” and “dynamic social networks”. Most of the current research emphases on the single-type network, multi-type networks should be taken into consideration. In fact, users can have several social networks accounts like twitter, Facebook, LinkedIn, GitHub and many others. So, scholars can measure, combine influence for users on multiple networks in which nodes of different network layers are mapped in multi-type-network. Moreover, the research community can additionally address matters concerning scalability and complexity while including topic distribution and network structure under a single model and calculate approximately diffusion models being used on real large social networks for influence measurement.

REFERENCES

[1] Jun, Z.: 'Initial Spreaders in Online Social Networks', 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2016

[2] Srivastava, A., Chelmis, C., and Prasanna, V.K.: 'The unified model of social influence and its application in influence maximization ', Social Network Analysis and Mining, 2015, 5, (1)

[3] Cordasco, G., Gargano, L., and Rescigno, A.A.: 'On finding small sets that influence large networks ', Social Network Analysis and Mining, 2016, 6, (1)

[4] Wen, S., Jiang, J.J., Liu, B., Xiang, Y., and Zhou, W.L.: 'Using epidemic betweenness to measure the influence of users in complex networks ',

[5] Journal of Network and Computer Applications, 2017, 78, pp. 288-299

[6] Guo, L., Lin, J.H., Guo, Q., and Liu, J.G.: 'Identifying multiple influential spreaders in term of the distance-based coloring ', Physics Letters A, 2016, 380, (7-8), pp. 837-842

[7] Gao, S., Ma, J., Chen, Z.M., Wang, G.H., and Xing, C.M.: 'Ranking the spreading ability of nodes in complex networks based on local structure ', Physica A, 2014, 403, pp. 130-147

[8] Fu, Y.H., Huang, C.Y., and Sun, C.T.: 'Using global diversity and local topology features to identify influential network spreaders ', Physica A, 2015, 433, pp. 344-355

[9] Kang, C.H., Kraus, S., Molinaro, C., Spezzano, F., and Subrahmanian, V.S.: 'Diffusion centrality: A paradigm to maximize spread in social networks', Artificial Intelligence, 2016, 239, pp. 70-96

[10] Herzig, J., Mass, Y., and Roitman, H.: 'An author-reader influence model for detecting topic-based influencers in social media ', 2014, pp. 46-55

[11] Kitchenham, B.: 'Procedures for performing systematic reviews ', Keele, UK, Keele University, 2004, 33, (2004), pp. 1-26

[12] Clauset, Aaron, Mark EJ Newman, and Cristopher Moore. 'Finding community structure in very large networks.'Physical review E 70.6 (2004): 066111.

[13] LR Xiong Z L, Jiang W J, Wang G J. 'Evaluating user community influence in online social networks. Proceedings of 2012 IEEE 11thIn- ternational Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) , '2012: 640-64

[14] Aftab Farooq, Gulraiz Javaid, M. Uzair, M. Usman Akram, ‘Detection of influential nodes using social network analysis based on network metrics ’, Computing, Mathematics and Engineering Technologies (iCoMET), 2018 International Conference, Doi: 10.1109/ICOMET.2018.8346372

[15] G. Atwal and D. Bryson, ‘Luxury Brands in Emerging Markets (Bas- ingstoke: Palgrave Macmillan), ’2014

[16] Sun, J., Tang, J. ‘A survey of models and algorithms for social influence analysis. In Social network data analytics,’Springer, Boston, MA. 2011, pp. 177-214

[17] Adali, Sibel, Xiaohui Lu, and Malik Magdon-Ismail. ‘Attentive Be- tweenness Centrality (ABC): Considering Options and Bandwidth When Measuring Criticality.’SocialCom/PASSAT. 2012.

[18] COMTRADE, U. N. ‘Commodity Trade Statistics,’2016.

[19] Li K, Zhang L, Huang H. ‘Social Influence Analysis: Models, Methods, and Evaluation ’, Engineering, 2018,

[20] Jun, Z.: ‘Initial Spreaders in Online Social Networks’, 2016 54th An- nual Allerton Conference on Communication, Control, and Computing (Allerton), 2016

[21] Srivastava, A., Chelmis, C., and Prasanna, V.K.: ‘The unified model of social influence and its application in influence maximization ’, Social Network Analysis and Mining, 2015, 5, (1)

[22] Cordasco, G., Gargano, L., and Rescigno, A.A.: ‘On finding small sets that influence large networks ’, Social Network Analysis and Mining, 2016, 6, (1)

[23] Wen, S., Jiang, J.J., Liu, B., Xiang, Y., and Zhou, W.L.: ‘Using epidemic betweenness to measure the influence of users in complex networks ’,

[24] Journal of Network and Computer Applications, 2017, 78, pp. 288-299

[25] Guo, L., Lin, J.H., Guo, Q., and Liu, J.G.: ‘Identifying multiple influential spreaders in term of the distance-based coloring ’, Physics Letters A, 2016, 380, (7-8), pp. 837-842

[26] Gao, S., Ma, J., Chen, Z.M., Wang, G.H., and Xing, C.M.: ‘Ranking the spreading ability of nodes in complex networks based on local structure ’, Physica A, 2014, 403, pp. 130-147

[27] Fu, Y.H., Huang, C.Y., and Sun, C.T.: ‘Using global diversity and local topology features to identify influential network spreaders ’, Physica A, 2015, 433, pp. 344-355

[28] Kang, C.H., Kraus, S., Molinaro, C., Spezzano, F., and Subrahmanian, V.S.: ‘Diffusion centrality: A paradigm to maximize spread in social networks ’, Artificial Intelligence, 2016, 239, pp. 70-96

[29] Herzig, J., Mass, Y., and Roitman, H.: ‘An author-reader influence model for detecting topic-based influencers in social media ’, 2014, pp. 46-55

[30] Kitchenham, B.: ‘Procedures for performing systematic reviews ’, Keele,

[31] UK, Keele University, 2004, 33, (2004), pp. 1-26

[32] Newman, Mark EJ, and Michelle Girvan. ‘Finding and evaluating com- munity structure in networks.’Physical review E 69.2 (2004): 026113.

[33] Kempe, David, Jon Kleinberg, and Eva Tardos. ‘Maximizing the spread of influence through a social network.’Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[34] Wasserman, Stanley, and Katherine Faust. Social network analysis: Methods and applications. Vol. 8. Cambridge university press, 1994.

[35] Brandes, Ulrik. ‘A faster algorithm for betweenness centrality.’Journal of mathematical sociology 25.2 (2001): 163-177.

[36] "http://www.orgnet.com/sna.html," [Online]. Available: http://www.orgnet.com/sna.html [Accessed May 2018].

[37] L. C. Freeman, "The Study of Social Networks," [Online]. Available: http://www.insna.org/INSNA/na_inf.html.

[38] Laat, Maarten de; Lally, Vic; Lipponen, Lasse; Simons, Robert-Jan (2007-03-08). "Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis". International Journal of Computer-Supported Collaborative Learning. 2 (1): 87–103. doi:10.1007/s11412- 007-9006-4

[39] index of a graph". Psychometrika. 31: 581–603. Doi:10.1007/bf02289527.

[40] Dunne, C., Shneiderman, B.: Improving graph drawing readability by incorporating readability metrics: A software tool for network analysts , University of Maryland, Human-Computer Interaction Lab Tech Report HCIL-2009-13 (2009)

[41] P. Bonacich and P. Lloyd, ―Eigenvector-like measures of centralityfor asymmetric relations,‖ Social Networks, vol. 23, no. 3, pp.191–201, 2001

[42] Jungeun Kim and Jae-Gil Lee. 2015. Community Detection in Multi-Layer Graphs: A Survey. SIGMOD Rec. 44, 3 (December 2015), 37-48.

[43] Greenhow, C. and Robelia, B.: Old Communication, New Literacies: Social Network Sites as Social Learning Resources. Journal of Computer-Mediated Communication, 14:130–1161.doi: 10.1111/j.1083-6101.2009.01484.x (2009)

[44] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths." Social Networks, vol. 32, no. 3, pp. 245–251, 2010.

[45] S. P. Borgatti, "Centrality and network flow," Social Networks, vol. 27, no. 1, pp. 55–71, 2005.

[46] Tuhena Sen "Modified Page Rank Algorithm: Efficient Version of Simple Page Rank with Time, Navigation and Synonym Factor" 2017 International Conference on Computational Intelligence and Networks DOI 10.1109/CINE.2017.24

[47] Holland PW, Leinhardt S (1975) Local structure in social networks. In: Heise D(ed) Sociological methodology. Jossey-Bass, San Francisco

[48] "http://rapidminer.com/," [Online]. Available: http://rapidminer.com/. [Accessed 2018].

[49] https://csce.ucmss.com/cr/books/2017/LFS/CSREA2017/HIM3016.pdf

[50] Kunegis, Jérôme. "Konect: the koblenz network collection." Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013.

[51] Federal Aviation Administration. Air traffic control system command center. http://www.fly.faa.gov/

[52] Roberts N, Everton SF. Roberts and Everton Terrorist Data: Noordin Top Terrorist Network (Subset), Machine-readable data file

[53] Majeed, Sadia, Usman Qamar, and Aftab Farooq. "State of art techniques for social influence analysis: A systematic literature review." 2018 International Conference on Frontiers of Information Technology (FIT). IEEE, 2018.

[54] Butt, Wasi Haider, et al. "Covert network analysis for key player detection and event prediction using a hybrid classifier." The Scientific World Journal 2014 (2014).

[55] Gao, Shuai, et al. "Ranking the spreading ability of nodes in complex networks based on local structure." Physica A: Statistical Mechanics and its Applications 403 (2014): 130-147.

APPENDIX

Performance Evaluation:

Below is the image representation of the created classification model for performance evaluation/ accuracy measurement using the Rapidminer. These images are very helpful to set up the workflow.

Section 1: Level 0, Level 1and Level 2 classification process

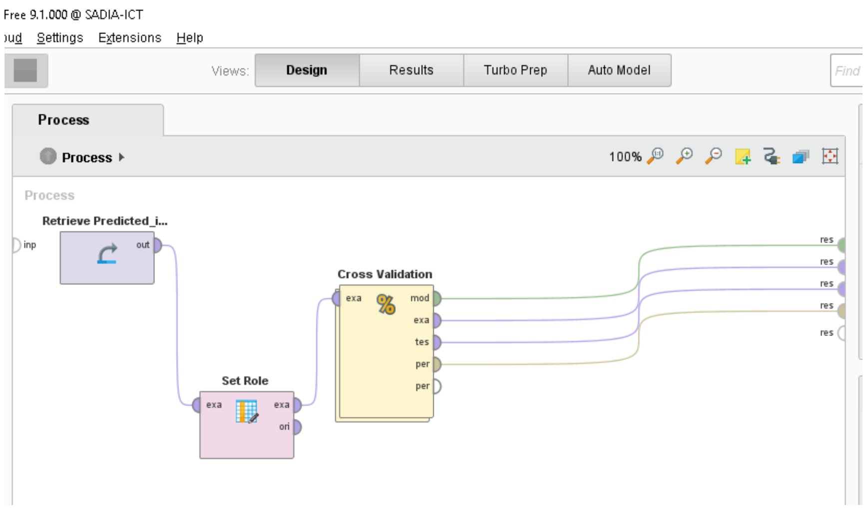


Figure 1. Level 0 Classification Process Diagram

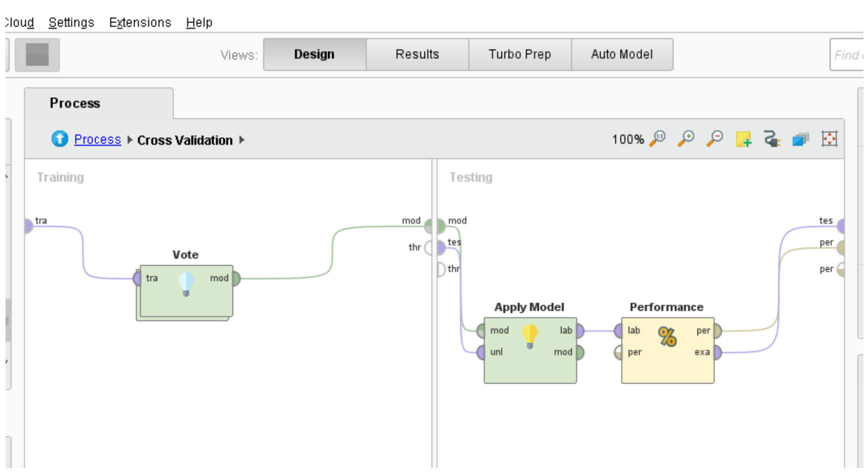


Figure 2. Level 1 Classification Process Diagram

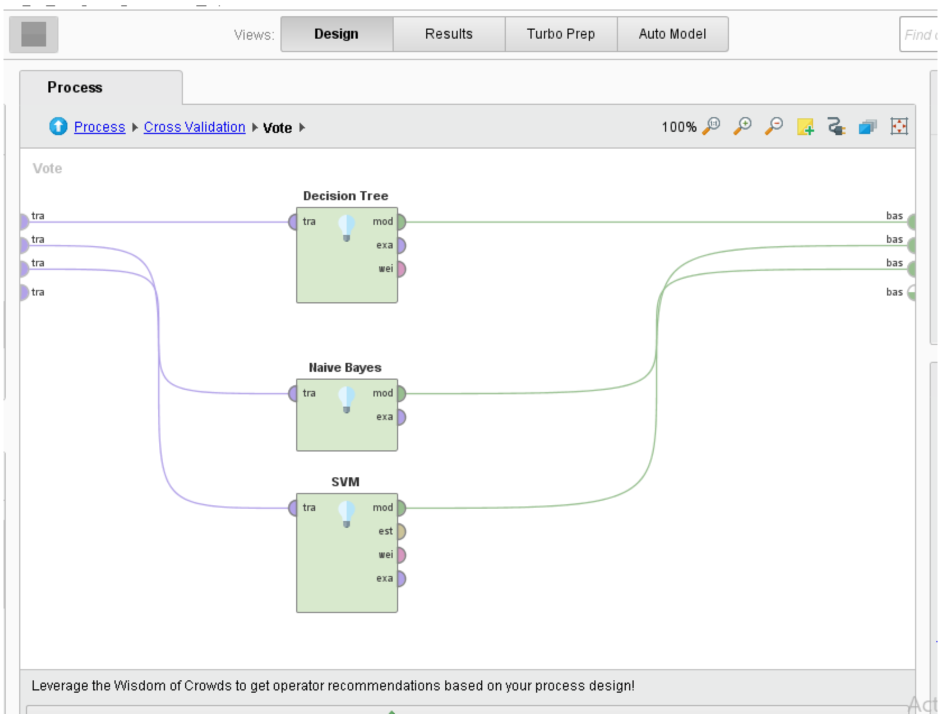
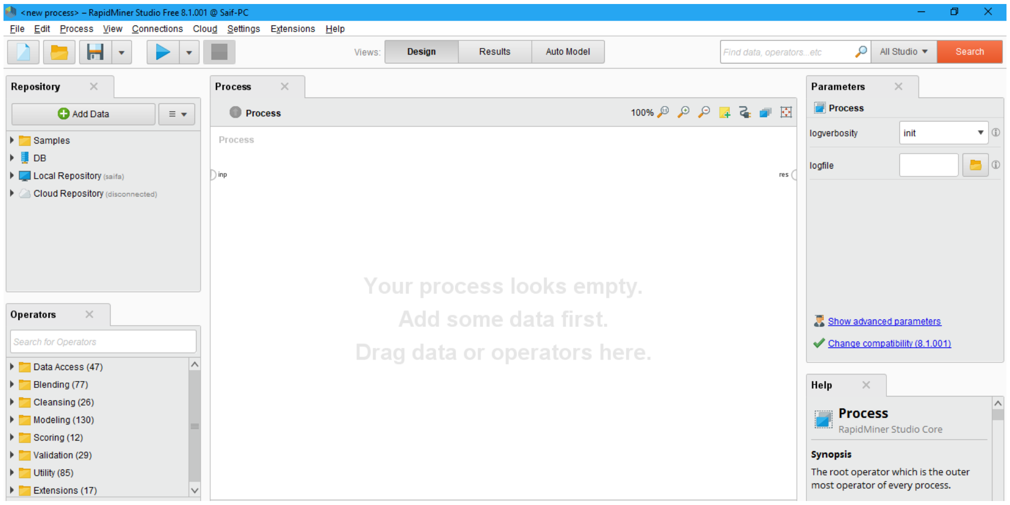


Figure 3. Level 2 Classification Process Diagram

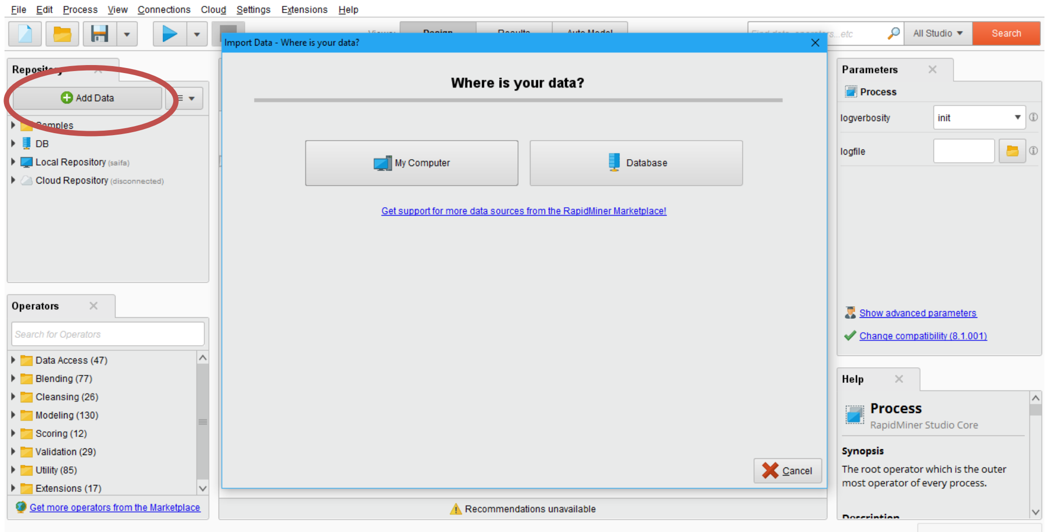
Section 2: Steps for detailed evaluation

Step 1: Open Blank Process.



Step 2: Load the Datasets

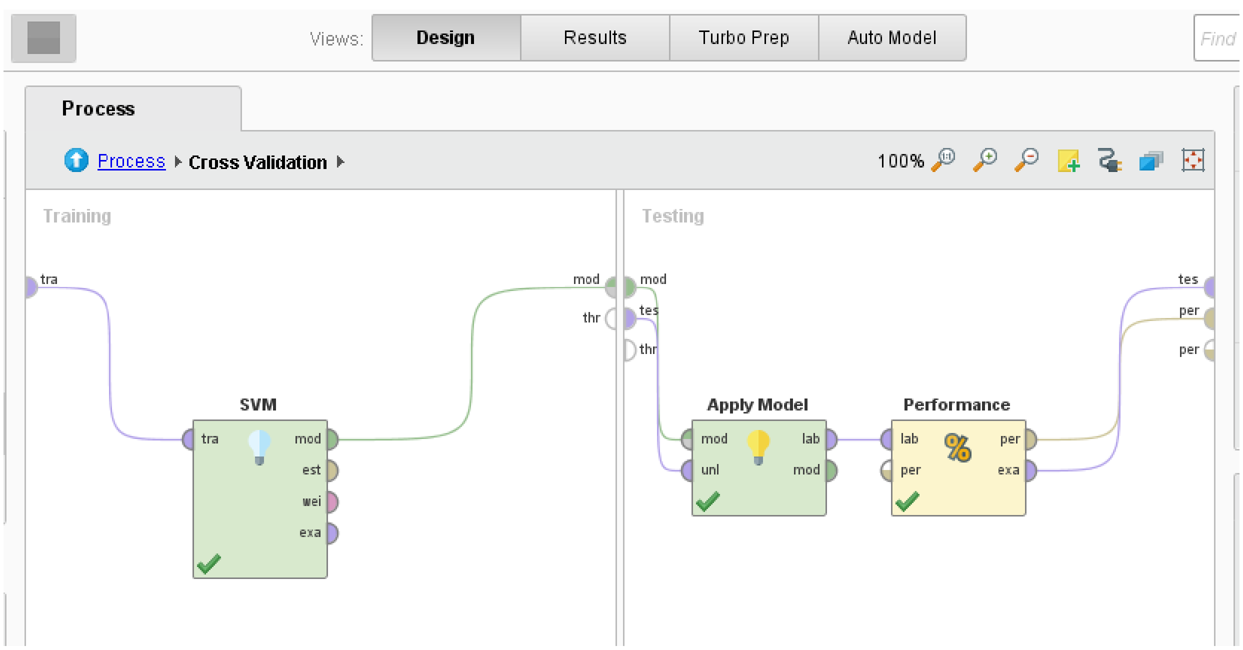
Press the Add Data Button to open a new dialogue box for the addition of the Dataset. Navigate to the path where dataset is located.



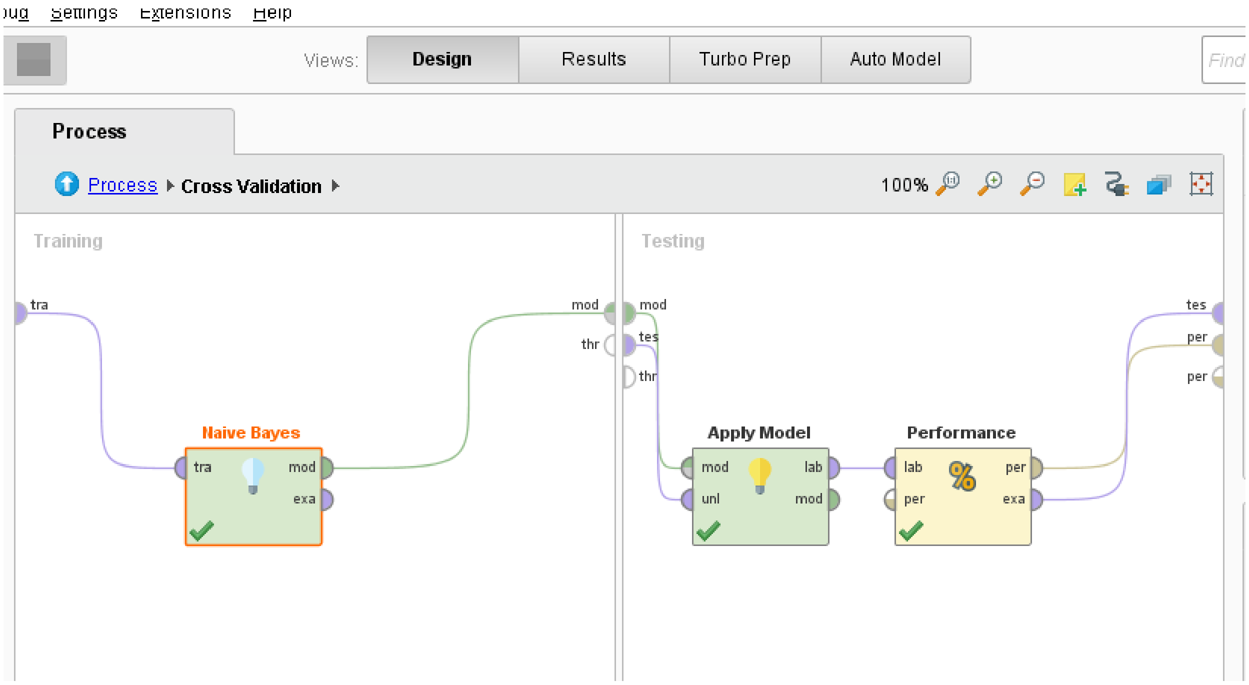
Step 3: Process Creation

Search for the required operators in the search bar. Join the wires to create the connection

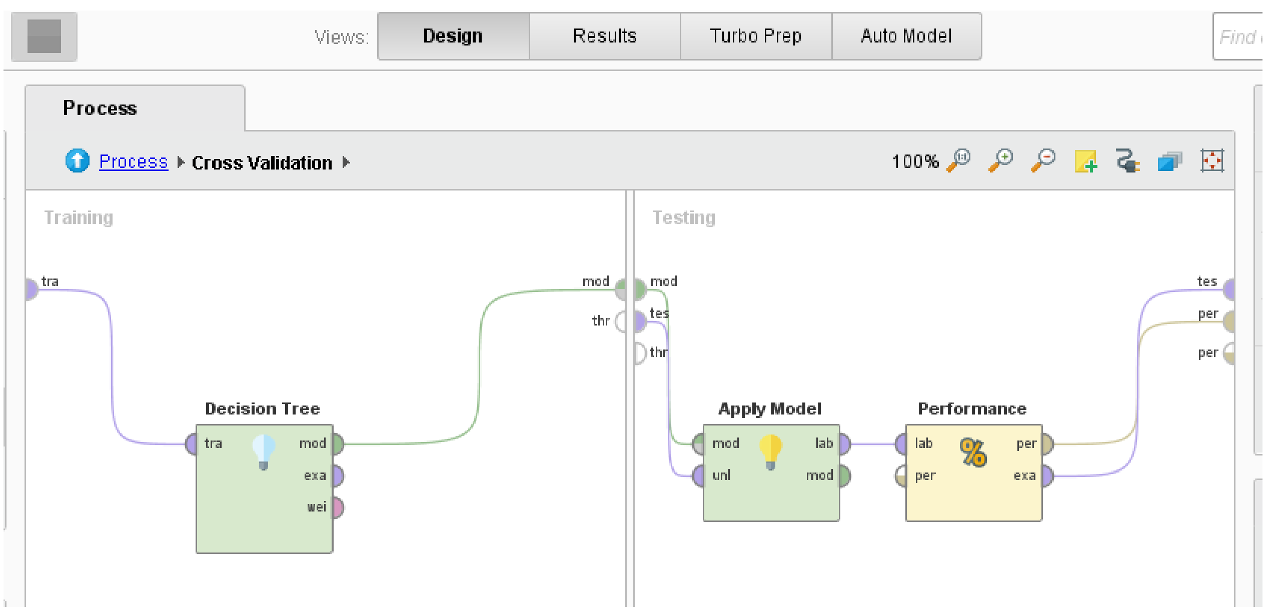
SVM Process



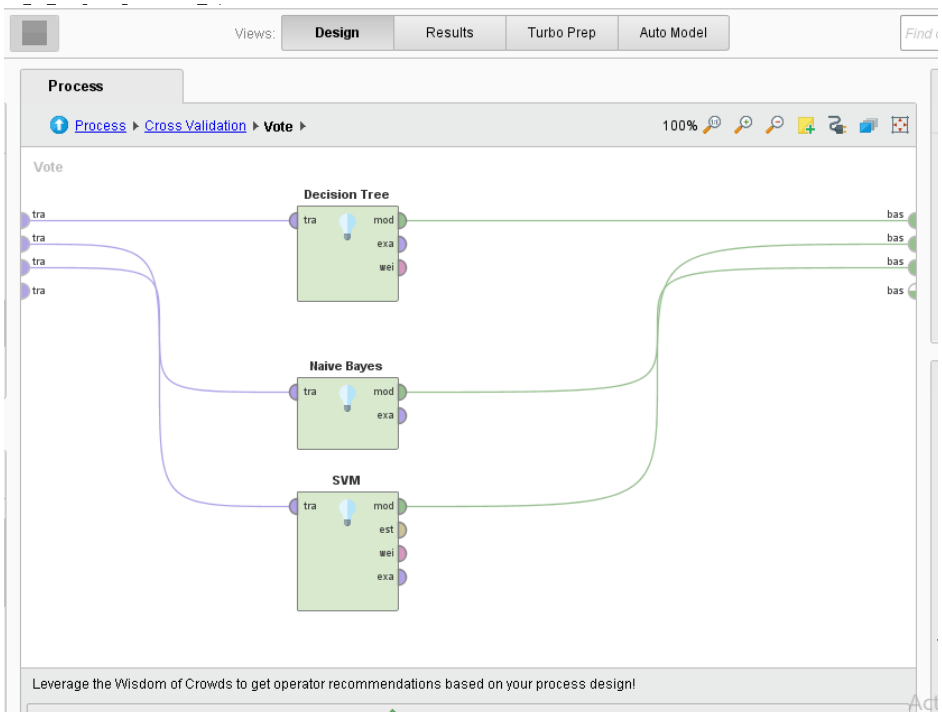
Naïve Bayes Process



Decision Tree Process

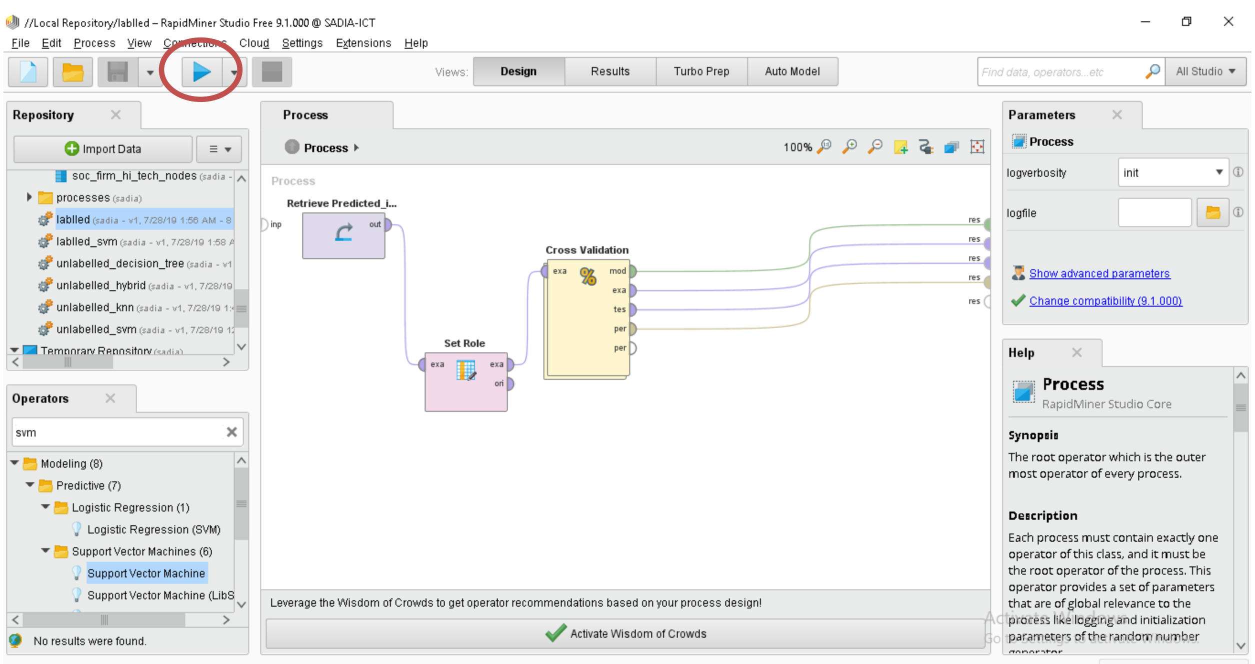


Hybrid Classifier Process

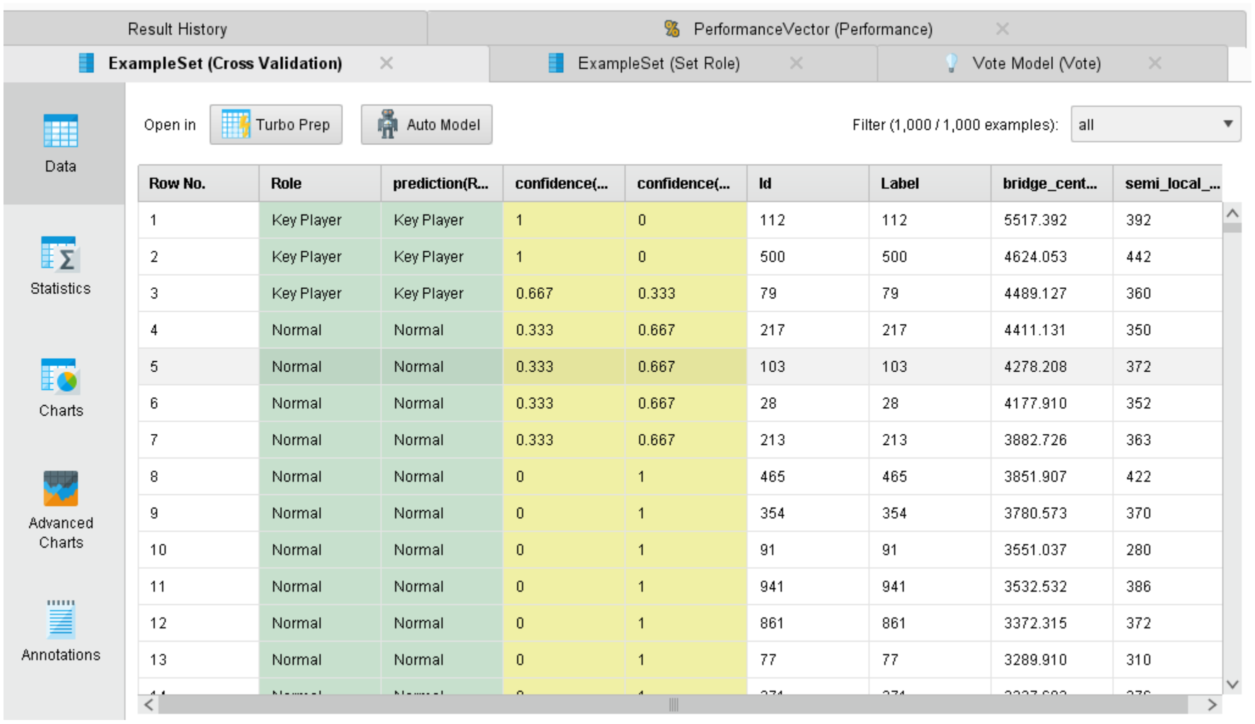


Step 4: Generating Results

After the Process is created press the Run Button to generate the results.



Generated Results:



Row No.	Role	prediction(R...	confidence(...	confidence(...	Id	Label	bridge_cent...	semi_local_...
1	Key Player	Key Player	1	0	112	112	5517.392	392
2	Key Player	Key Player	1	0	500	500	4624.053	442
3	Key Player	Key Player	0.667	0.333	79	79	4489.127	360
4	Normal	Normal	0.333	0.667	217	217	4411.131	350
5	Normal	Normal	0.333	0.667	103	103	4278.208	372
6	Normal	Normal	0.333	0.667	28	28	4177.910	352
7	Normal	Normal	0.333	0.667	213	213	3882.726	363
8	Normal	Normal	0	1	465	465	3851.907	422
9	Normal	Normal	0	1	354	354	3780.573	370
10	Normal	Normal	0	1	91	91	3551.037	280
11	Normal	Normal	0	1	941	941	3532.532	386
12	Normal	Normal	0	1	861	861	3372.315	372
13	Normal	Normal	0	1	77	77	3289.910	310