Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/130418

Customized CWT-CCA: Discovery of Prominent Persons in the Crime Network

Ganeshkumar Pugalendhi[1,*] *and Shanmugapriya Kumaresan*[2]

[1] Department of Information Technology, Anna University, Regional Campus, Coimbatore, India
[2] Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, India

Received: 2 Jan. 2019, Revised: 22 Mar. 2019, Accepted: 29 Mar. 2019
Published online: 1 Jul. 2019

Abstract: In this paper, we propose a modified Cluster Walk Trap (CWT) approach combined with Closed Cycle Approach(CCA). The proposed method uses probability measure in a multimode network to evaluate group of persons involved in common crime and group of crime done by a common person thereby avoiding the difficulty of adjacency matrix. The usage of transitivity based on closed cycle in node identification overcomes the problem of computational complexity with the increased number of nodes and edges. Based on the clustering coefficient value, the influential nodes are classified as victim, suspect and witness. Experiments are conducted using the publicly available KONECT dataset and the simulation results show that the proposed approach is good in identification of communities and influential nodes with improved degree of accuracy than the other approaches reported in the literature.

Keywords: Community Detection, Influential Nodes, Graph Analysis, Multi-mode Network, Criminal Network

1 Introduction

Crime [1] is the behavior of an individual or group of people which is illegal and considered harmful and offensive to society. In recent days, criminal behaviorism is highly publicized in media (online and offline). People in society are well-informed about the crimes and the law is quite complicated. In general, there are many different crimes happening in the society that can be broadly classified into four major types viz., personal crimes, property crimes, inchoate crimes and statutory crimes. In general, the crime analysis [2] is performed manually using the crime data collected in form of record by an investigative officer for identifying types of crime, responsible individuals, etc. Recently the crime incidents are recorded in the form of a network and computer-based analysis methods [3] are emerging as a helping tool for the investigating officer to get an overall picture of the crime, crime scene and the people involved.

Identifying the communities and the influential person in a crime has received much attention in criminal network analysis as they are covert and the information is not open. Most of the criminal network analysis focuses on the one-mode network whereas analysis of multi-mode network has received only less attention. Due to the incognito nature of multimode network, the identification of link between the persons and crime is inherently difficult. There are large numbers of community detection algorithms existing for multimode network. Most of the algorithms handle the problem as a graph-partitioning approach [4] that splits the input graph into number of groups while minimizing the cost of edge cut, using the number of communities and the size of the community as parameter. But for analyzing the social crime network, it is not possible to identify number of community and its size in advance using partitioning of graph communities can be formed based on tie existence [5].

A divisive method of graph clustering based on the removal of edges with largest edge betweenness value splits the graph into communities in a hierarchical manner. This method was proposed by Girvan and Newman [6] that considers the modularity as a quality function whose complexity is $O(n_3)$ and it is limited to networks of size 103 nodes. Radicchi et al. [7] upgraded the version of the Girvan and Newman algorithm to enhance the computational complexity and the limit of network size. This approach removes the edges in the graph based on the high clustering coefficient rather than considering the betweenness value. The computational complexity of this approach is $O(n_2)$.

* Corresponding author e-mail: ganesh_p2154@yahoo.com

Instead of splitting the graph based on some parameters, merging the vertices arises for community detection. Newman [8] proposed a greedy algorithm which starts with n communities and merges the vertices by optimizing the quality function called modularity as a quality of partition. Computational complexity of this approach is $O(m_n)$. The eigen vectors of the laplacian matrix of graph is to measure similarity of vertices. The complexity determined for eigen values computations $O(n_3)$ time for sparse matrices [9]. The walktrap approach uses the random walks to define a distance which measures the structural similarity between the vertices and between the communities. Random walktraps the dense community at certain point of time in the graph. On transversing through the graphs, the walks created are repeated within the set of nodes as they are linked to one another. The relationship between the nodes which makes the walk to repeat within set of nodes results in the cluster of nodes or community. As the random walk is repeated for n-number of steps, there is only some set of nodes left outside the community. The results of each step of random walk are merged to generate different set of communities in bottom up fashion. In general, the random walk follows the divide and conquers approach which helps to reduce the time complexity in community detection.

Focuss et al. [10] carried out the Euclidean commute time distance based on which merging of the vertices to form communities. The dissimilarity index based on the same quantity, named this hierarchical algorithm as net walk. Markov cluster algorithm uses two matrix operations iteratively to form clusters in the limit state [11, 12]. Unfortunately the computational complexity of this approach is $O(n_3)$. The walktrap approach starts initially with each vertex as its own community and starts to merge the vertices and communities based on their minimum distance value computed and the process is iterated. The quality function associated with walktrap approach maximizes the modularity [13].

Louvain algorithm [14] is the fast modularity optimization algorithm which repeats the merging of communities formed by Pons method in second phase. This is advanced by re-merging the super-nodes in the new network to achieve high modularity [15]. But iteratively merging the communities in the already formed community structure sometimes lead to omitting some important vertices in the graph [15]. So far, the implementation of above discussed approaches were carried out on Zacharys karate club network [16], college football network [17], protein interaction network [18], scientist collaboration network [19], internet map and web graph [20]. All these networks contain a binary valued matrix.

The community detection in bipartite graph is difficult when compared to normal graph. Since the bipartite graph has two set of vertices for analysis, one cannot omit any vertices in the graph. Clusters of the two distinct vertex set and the edges connect to the vertices of different sets by maximizing the modularity based on the probability of existence of edge between the vertices [21]. A score which indicates the node which has the clustering behaviour of complete network. They considered 4-closed paths and 6-closed paths for measuring the score [22].

In this paper, a novel method combining modified Cluster Walk trap (CWT) and Closed Cycle Approach (CCA) is used for forming the communities and identifying the influential nodes in the multi-mode crime network. The proposed method calculates the membership and modularity values based on which the communities are formed. In general, membership is defined as the degree to which the nodes are connected in the network. In crime network, membership is the measure of active participation of the individuals in the crime. As membership is calculated for each node, it shows the strength of the node in the network which is important for clustering without convincing the node role. Similarly modularity is the structural measure which identifies the strength of splitting the network into clusters. Maximum of the time modularity is calculated with the help of membership value of nodes, eigen vector or other similarity measures. Here, the modularity measure is calculated by taking the probability of the degree of the nodes in network to form the cluster. Further, the clustering coefficient value is calculated for individual nodes, clusters and entire network for predicting the most influential nodes in the crime network. The Performance of the proposed CWT-CCA is tested using publicly available KONECT crime dataset. Simulations have been conducted to evaluate the efficiency and accuracy of the proposed method in terms of community detection and influential person identification.

The rest of the paper is organized as follows. Section 2 provides an overview of the traditional approach for crime detection. Section 3 describes the architecture, and steps in proposed computer-based approach for crime detection. Section 4 describes the modified CWT and CCA for identification of influential persons in the crime network. Section 5 represents the results obtained from applying the proposed algorithm to real crime data sets and its performance measures. Section 6 provides the conclusion.

2 Crime Detection: a Traditional Approach

The crime detection is defined as a function that involves identification and pattern analysis of crimes and disorders. The crime detection involves both professional and technical analysis. The overall picture of detecting crime is shown in Fig. 1.

There are different crimes occurring in different part of city which is shown in the left side of the crime analysis figure. The crimes happen in presence of the people in society who are the witness. The crime involves

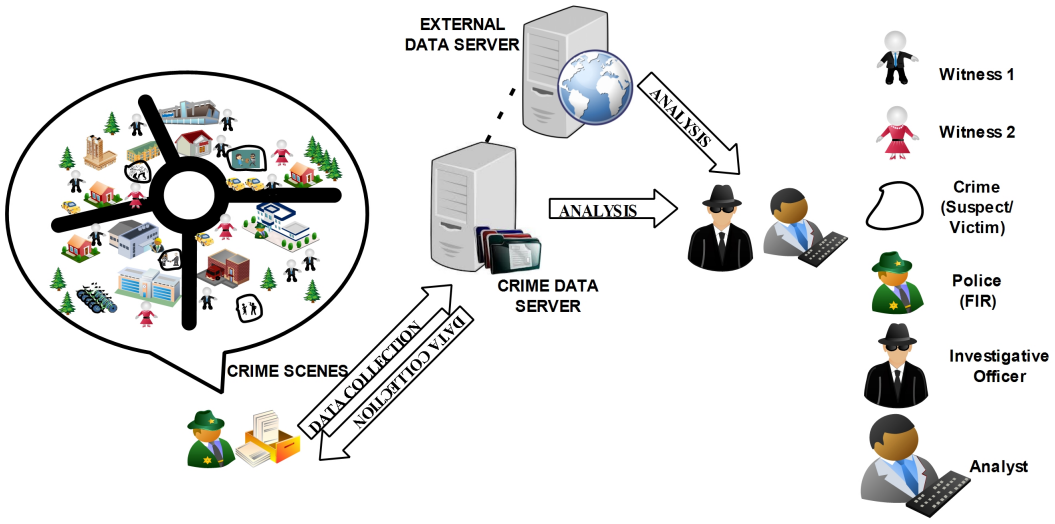


Fig. 1: Traditional way of detecting crimes

victim and the suspect. Victim is responsible for crime and Suspect is the one who may or may not be responsible for crime. Once the crime occurred in the place, the police officer collects all the data related to the crime scene and prepares a First Information Report (FIR). The collected data in form of FIR is then fed into the Crime Data Server (CDS) for future analysis and reference. This process is done in various regions of state. The data related to various parts of the state are aggregated from each CDS and stored in External Data Server (EDS) which helps in the comparative analysis of crime scene. The investigative officer is the professional who takes the data from the CDS and EDS for investigation and identifies the criminal involved and provides him/her the punishment based on law. The investigative officer analysis is manual process. The analyst is a non-professional who takes the data for developing different computational-based approaches for identifying types of crime, individuals responsible, etc.

In this paper, the role of non-professional analyst for analyzing the saved crime data is taken into account and a novel computational method using modified cluster walktrap and closed cycle is developed for detecting communities and the influential nodes in the crime network.

3 Proposed Computer-Based Approach for Crime Detection

The proposed method considers the crime data as a network in form of a bipartite graph $BG = \{P,C,R\}$, where P represents set of people, C represents set of crime and R denotes relation between P and C. Fig. 2 shows the schematic diagram of the proposed work. The analyst gathers data from the crime scene and generates the data set in required format for analysis.

The input to the proposed work is the crime event matrix which consists of rows representing the persons and the columns representing the crimes generated by the analyst. The values in the matrix represent the relation between the persons in the crime. The first step is to convert the crime event matrix into a bipartite graph. The second step is to identify the community existing in the graph. The constructed bipartite graph is fed as input to the proposed cluster walktrap approach for community detection. The result of the algorithm is set of clusters. The clusters show the set of people associated to a common crime and vice versa.

The third step is to identify the influential nodes in the graph using the proposed closed cycle approach. The constructed bipartite graph is projected into two different one-mode graphs. The first projection is based on the person and the second projection is based on the crime. The projection 1 forms the person wise adjacency matrix and projection 2 forms crime wise adjacency matrix. For identifying the influential nodes, projection 1 is taken for analysis. The result of this gives the list of nodes which are highly influential in the crime.

Finally, from the set of influential nodes identified along with the detected communities based on the suspect and witness are identified.

4 Implementation of Cluster-Walk-Trap (CWT) Closed Cycle Approach (CCA) for Crime Detection

In this section, detailed steps involved in detecting communities and the influential persons of a crime network using our proposed CWT and CCA are discussed. Fig. 3 depicts the implementation procedure of the proposed CWT and CCA for crime detection.

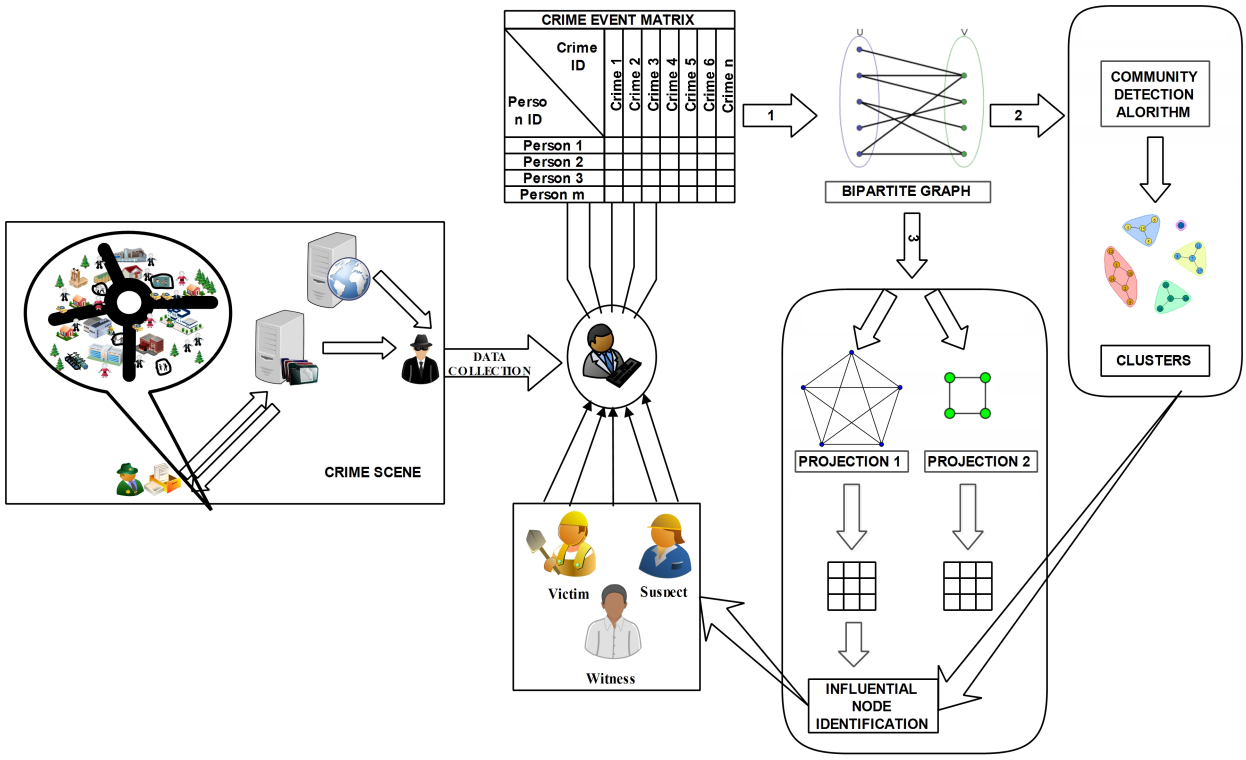


Fig. 2: Schematic diagram of proposed approach for crime detection

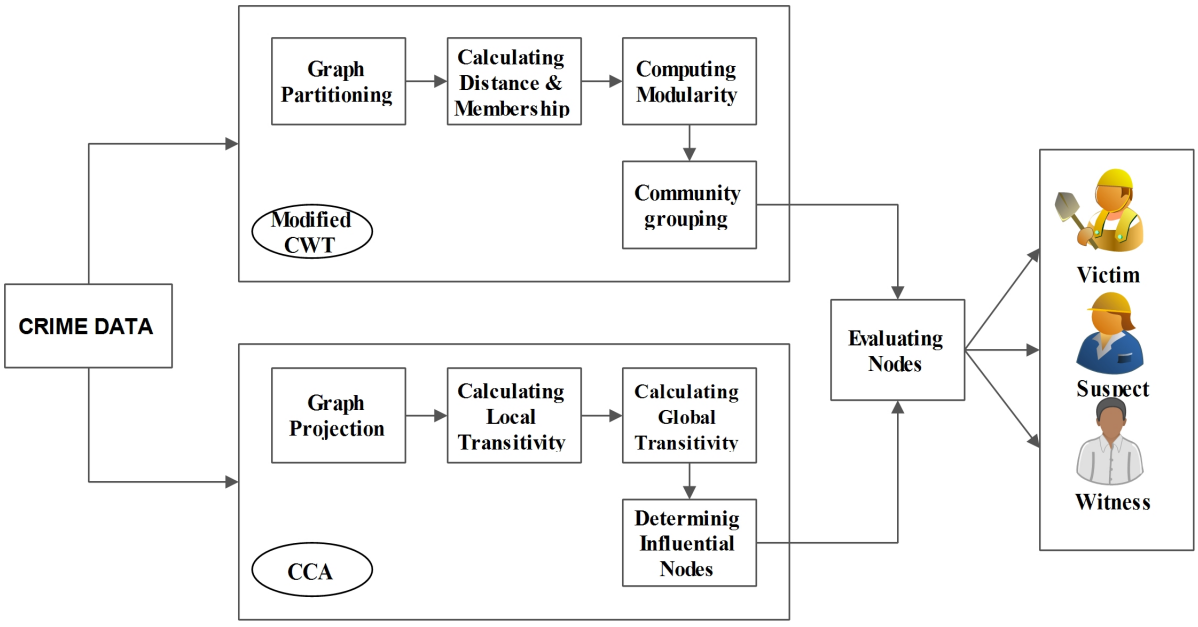


Fig. 3: Implementation procedure of proposed CWT-CCA for crime detection

4.1 Community Detection using Modified CWT approach

The main focus of the proposed work is to determine the community i.e. group of people linked to the particular crime. The community can be determined using the modified cluster walktrap algorithm. There are so many measures which can be used to carry on the walktrap approach, here the probability-based modularity is considered as the key measure for clustering.

The modified cluster walktrap measures the structural similarity between the vertices and between the communities in the one-mode network. It traps the dense part of community in the network by random walks. The communities at each step are merged with the minimum distance and the process is repeated till maximum modularity is obtained. The membership is calculated based on the partitioning the graph at each step. Based on the membership value the modularity is computed. Here, the modified walktrap algorithm is implemented to the two-mode network. For implementing the algorithm in two-mode network some suitable modifications are carried out in the walktrap approach. In modified cluster walktrap approach, the similarities between the vertices and communities are measured by probability distance value.

Algorithm 1: Modified Cluster Walktrap Approach

Input: bg–Bipartite graph
Output: C-Clusters formed
Method:

1. procedure commDetec(bg){
2. for each vertex in bg {
3. if($i \neq 1$) then
4. Partition into community (C_i)
5. Else
6. Stop partition
7. for each vertex in bg{
8. $PM_i = \frac{n}{C}$
9. }
10. if($PM_i < PM_{i-1}$) then {
11. $dist_i = \frac{n(n-1)}{2}$
12. $PD_i = \frac{dist_i}{\sum_{i=1}^{n} dist_{i_i}}$
13. }
14. for each k {
15. if($PD_i < \Gamma$) then {
16. $PT_k = \{C_1, C_2$
17. $C_3 = C_1 \cup C_2$
18. $PT_{k+1} = C_3$
19. } }
20. for each vertex in PT_k\{
21. $memb_i = \frac{n}{C}$
22. }
23. $pm_{ij} = \frac{m_i m_j}{\sum_{i,j=1}^{n} m_i m_j}$
24. $PM = \frac{1}{2n} \sum_{i,j=1}^{n} \left(IM_{ij} - \frac{Pm_{ij}}{2n} \right) \delta(C_i C_j)$
25. } }

The Algorithm 1 modified-CWT approach starts with the input bipartite graph. It starts partitioning the graph into n communities which reduces to single vertex in each community where $P1 = \{\{v\}, v \in V\}$. Then it computes distance between 3 all adjacent vertices $dist_i$. The probability-based distance measure is computed as PD_i for all vertices. Algorithm selects communities, merges the selected communities based on the PD_i and updates the distance value based on the distance limit Γ. Also the measures namely membership $(memb_i)$ division of node to communities and probability-based modularity (PM) are calculated.

4.2 Identification of Influential Person in Community

The communities formed will have set of nodes based on the similarity. All nodes are not important but some nodes play a major role in the community. So, the next task in the proposed work is to identify the important nodes within the community, i.e. to identify the important person involved in crime and within community. Identifying important nodes in the one-mode projection network is direct and simple. This can be done by computing a measure called centrality within the communities. But here we consider a two-mode projection network where identifying important nodes is difficult because the communities vary in size and also have elements from both the sets. The measure used for identifying important nodes is clustering coefficient.

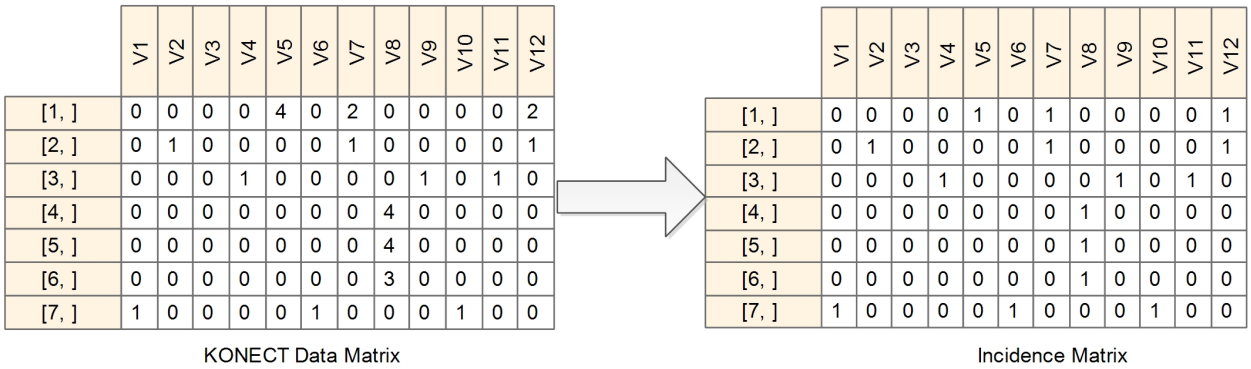
The proposed method identifies the important person involved in crime and also important person within the community. First step is to project the bipartite graph into two one-mode projections. Next, consider the projection on person as input and the metric of transitivity measure of clustering coefficient is calculated. Based on the value calculated, the influential node in the network is identified.

Algorithm 2: Closed Cycle Approach

Input: bg–Bipartite graph
Output: infl-Set of Influential Nodes
noinfl-Set of Non-Influential Nodes
Method:

1. procedure inflNode(g){
2. for each vertex in g {
3. proj1 $= \{v, v \in V(persons)\}$
4. proj2 $= \{v, v \in V(crimes)\}$
5. }
6. for proj1 and proj2
7. $cc_L = \frac{p_{(i,c)}}{p_c}$
8. $cc_G = \frac{p_{(i,g)}}{p_g}$
9. }
10. if($cc_L(i) >= cc_G$) then
11. infl $= i$
12. else
13. no_infl $= i$
14. }

Algorithm 2 CCA starts with projecting the graph as proj1 is person wise projection and proj2 is crime wise



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
[1,]	0	0	0	0	4	0	2	0	0	0	0	2
[2,]	0	1	0	0	0	0	1	0	0	0	0	1
[3,]	0	0	0	1	0	0	0	0	1	0	1	0
[4,]	0	0	0	0	0	0	0	4	0	0	0	0
[5,]	0	0	0	0	0	0	0	4	0	0	0	0
[6,]	0	0	0	0	0	0	0	3	0	0	0	0
[7,]	1	0	0	0	0	1	0	0	0	1	0	0

KONECT Data Matrix

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
[1,]	0	0	0	0	1	0	1	0	0	0	0	1
[2,]	0	1	0	0	0	0	1	0	0	0	0	1
[3,]	0	0	0	1	0	0	0	0	1	0	1	0
[4,]	0	0	0	0	0	0	0	1	0	0	0	0
[5,]	0	0	0	0	0	0	0	1	0	0	0	0
[6,]	0	0	0	0	0	0	0	1	0	0	0	0
[7,]	1	0	0	0	0	1	0	0	0	1	0	0

Incidence Matrix

Fig. 4: Construction of incidence matrix from KONECT data

projection. For each projection, the local clustering coefficient values are calculated as cc_L and global clustering coefficient values are calculated as cc_G based on the probability of connected nodes. In this algorithm the limit is based on cc_G. The nodes with cc_L less than the limit are non-influential nodes and others are influential. The algorithm on implementation with proj1 gives the list of persons who are more influential in the crime events in the cluster formed. The same on proj2 gives the list of crimes involved by common people. From set of influential nodes determined, set of victims, suspects and witness are determined based on the same probability measure.

5 Simulation Result

In this section, simulation carried out using *R* [23] for the proposed work is discussed and the results are reported. In *R*, two packages namely igraph (Network Analysis and Visualization) and sna (Social Network Analysis) are used for the proposed experiment.

5.1 Dataset

5.1.1 Training Data

The proposed method of modified CWT–CCA approach is trained using the sample two-mode valued crime matrix of individuals involved in the crime event. The dataset itself cannot be divided into training and testing data because it is a two-mode matrix. Each value corresponding to persons (row-wise) and crime (column-wise) are interrelated and plays vital role in the community detection and the influential node identification. For instance, if the dataset is divided into training and testing row-wise it omits the interrelation of persons among the crimes which in turn would reflect deviation in influential person identification. In other hand, if the dataset is divided column-wise it omits the interrelation of crimes among the people which in turn would reflect deviation in community detection of crimes. To overcome this the sample training data is generated using the concept of data wrangling in *R* tool keeping the KONECT crime data as reference. The sample data is generated using the grammar for data manipulation package.

Table 1: Properties of the KONECT crime dataset

Properties	Details
Number of actors	1380 (Persons + Crimes)
Number of links	1476 (Involvements)
Average degree (overall)	2.1391 edges/vertex
Average person degree	1.7805 edges/vertex
Average crime degree	2.6788 edges/vertex
Diameter	32 edges
Mean shortest path length	13.37 edges

5.1.2 Testing Data

In this work, KONECT Crime data set [24] is used. This data set is a 870×557 two-mode valued matrix of individuals involved in crime events. The crime event matrix is coded as 1-victims, 2-suspects, 3-witnesses and 4-duals (victims + suspects). This data forms a bipartite network which contains persons who appeared in crime case as given in Table 1.

5.2 Data Pre-processing

The KONECT crime data has 870 persons and 557 crimes. At first, this data set is constructed as a graph using *R*, whose adjacency matrix is considered for pre-processing. This data set is not a square matrix and hence it is not suited to construct the graph. So the first step is to make the data suitable for constructing a bipartite or a two-mode network. For that, the data set's matrix is converted as an incidence matrix or bi-adjacency matrix. The incidence matrix is the one which has two set

of vertex: one for persons and other for crimes. The value in the matrix shows the tie between the two classes of vertices involved in the relation. The incidence matrix is a non-square matrix from which the bipartite network can be constructed. The incidence matrix shows the existence of tie between the vertices in the graph.

Fig. 4 shows the conversion of matrix to incidence matrix. The given matrix is a weighted matrix that has 7 rows (1 to 7) and 12 columns (V1 to V12). The matrix is then converted into an incidence matrix which has 7 vertices as rows indicating the person denoted as 1 to 7 and 12 vertices as column that indicates the crimes denoted as 8 to 19. The weight is not considered for incidence matrix. Incidence matrix thus forms holds the values 0 for no tie and 1 for existence of tie in the network. Once the incidence matrix is formed, the graph is constructed using *R* in a bipartite network structure.

5.3 Graph Construction

The KONECT crime data set is in the form of matrix. For the analysis of network of crimes, first the data set is mapped as a graph using igraph package in *R*. Since the network is bipartite in structure, the matrix is considered as a bi-adjacency (incidence) matrix and an undirected graph is formed with 1427 vertex and 1487 edges. The vertex is identified as person that counts to 870 vertices and crimes that counts to 557 vertices.

For clear visualization, a simplified form of this undirected graph with 19 vertex and 15 edges that has 7 persons and 12 crimes formed from sample bi-adjacency (incidence) matrix in R_i graph is shown in Fig. 5.

In Fig. 6, the bipartite structure of the graph shown in Fig. 5 is given with 7 vertices (1 to 7) on top that denotes persons and 12 vertices (8 to 19) on bottom that denotes crimes.

5.4 Modified Cluster Walktrap Approach

The modified walktrap algorithm uses random walks to find the distance based on the similarity measure between the vertices and between the communities. At a particular stage the algorithm traps into the denser part of the network by merging the communities based on the distance measure and the process is iterated. The algorithm stops when maximum modularity value is reached. On implementing the algorithm there are 5 groups with modularity value of 0.668.

Fig. 7 shows the groups formed after performing the clustering. Person 1 is involved in the crimes 5,7 and 12, person 2 is involved in crimes 2,7 and 12, person 3 involved in the crimes 4,9 and 11, persons 4,5 and 6 are involved in crime 8, person 7 is involved in crimes 1,6 and 10. Crime 3 is excluded for further analysis because there is no link to person. On visualizing Figs. 5 to 7, there is no direct link between person 2 and crime 12 as the data set is viewed as two mode graphs. So, the two-mode data is projected for identifying the hidden links with the help of incidence matrix generation shown in Fig. 4. Since person 2 is not directed linked to crime 12 but on analyzing the network it is seen that there is a connection between person 2 and 1 with common crimes 9 and 14 where in turn the person 1 is linked to crime 12, there exists a hidden link between person 2 and crime 12. Hence such type of hidden links is identified in the proposed approach using the incidence matrix generation as step of data pre-processing that is making the data ready for implementing the proposed approach of community detection.

The partitioning of graph into communities is shown as hierarchical structure of communities using dendrogram in Fig. 8.

Fig. 8 shows the tree structure of partitioning made by the algorithm. In this tree structure leaves correspond to the vertices and the internal nodes correspond to the merging of communities in the algorithm i.e. union of communities corresponding to its children. Here there are 19 vertices in the *x*-axis which are grouped into five groups and the *y*-axis shows the height of the tree.

5.5 Identifying Important Persons

The influential nodes in the cluster are identified by projecting the bipartite graph into two graphs projected in view of person and crime. As the input data is two-mode valued it represents the persons and crime involvement. On projecting the graph down into two single mode valued representation, projection 1 gives person wise graph and projection 2 gives crime wise graph. The projection 1 graph has person as nodes and link shows the relationship among person based on their involvement in the crime. The projection 2 graph has crime as nodes and link shows the crimes done by common people.

Fig. 9 shows the projection of persons in the input graph. The graph is projected with 7 nodes and its connections. The 7 nodes are the persons 1 to 7. The link between them shows their involvement in the crime. Persons 1 and 2 are involved in more than one crime so it is denoted in dark line. Persons 4, 5 and 6 are connected in single crime. Persons 3 and 7 are not involved in any common crime.

Table 2 depicts the adjacency matrix of the projection 1 shown in Fig. 9.

Fig. 10 shows the projection of crimes in the input graph. The graph is projected with 12 nodes and its connections. The 12 nodes are the crimes 1 to 12. The link between them shows the persons involved in the crime. Crimes 2, 5, 7 and 12 are carried by single person. Also Crimes 7 and 12 are carried by more than one person so it is denoted in dark line. Crimes 4, 9 and 11 are carried by single person. Crimes 1, 6 and 10 are carried

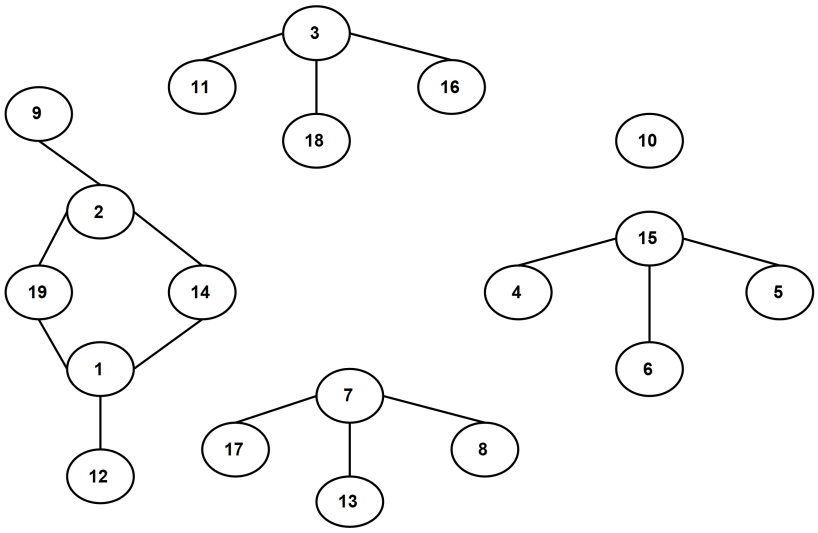


Fig. 5: Visualization of KONECT crime dataset using selected vertex and edges

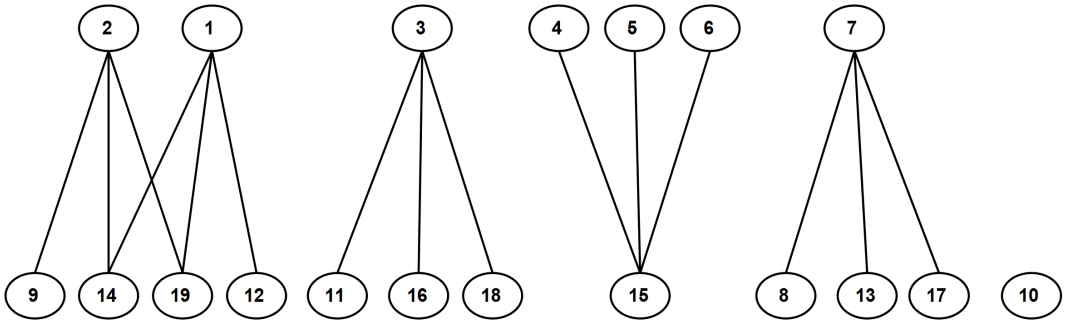


Fig. 6: Bipartite structure of KONECT crime data set

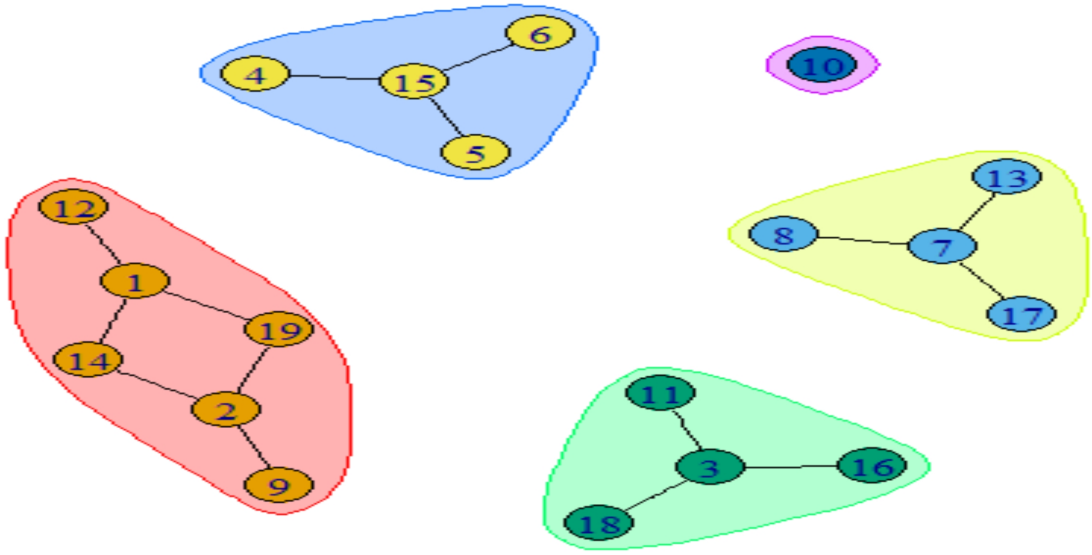


Fig. 7: Visualization of groups

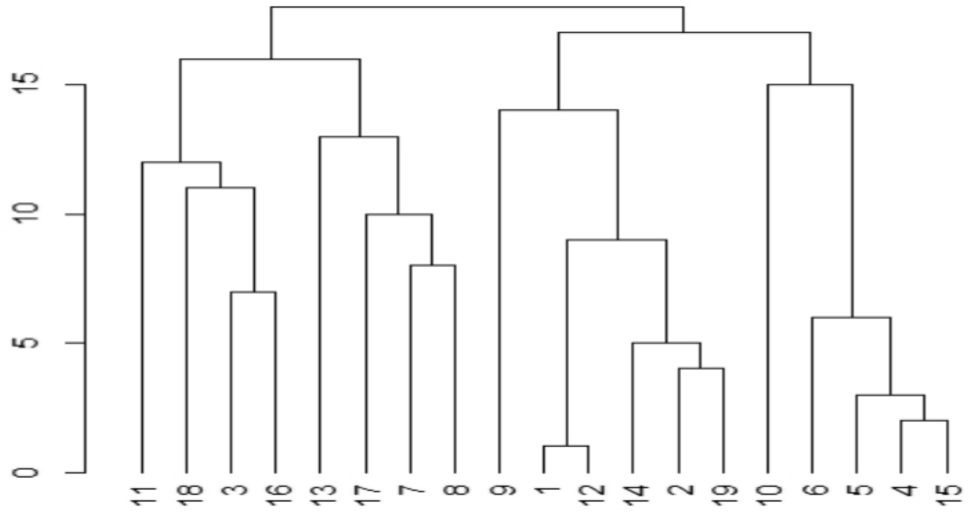


Fig. 8: Dendrogram representation of clusters formed using Modified CWT

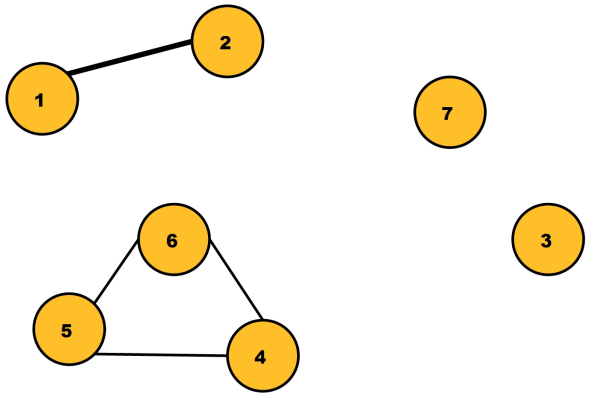


Fig. 9: Visualization of projection1 (persons)

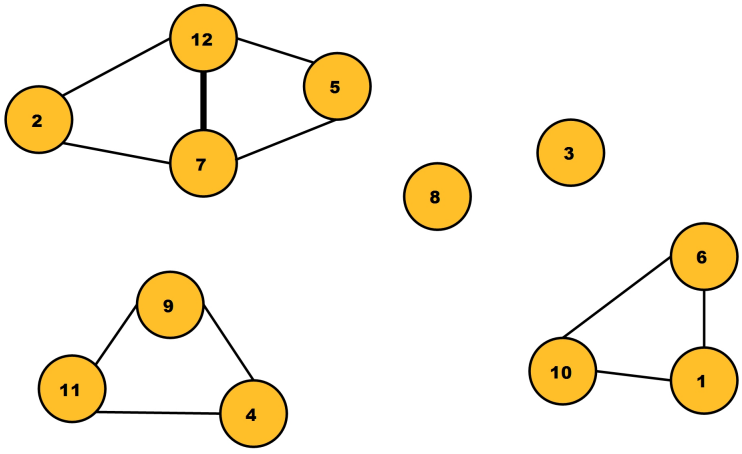


Fig. 10: Visualization of projection2 (crimes)

Table 2: Adjacency Matrix of projection1 (persons)

		Person ID						
		[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
Person ID	[1,]	0	2	0	0	0	0	0
	[2,]	2	0	0	0	0	0	0
	[3,]	0	0	0	0	0	0	0
	[4,]	0	0	0	0	1	1	0
	[5,]	0	0	0	1	0	1	0
	[6,]	0	0	0	1	1	0	0
	[7,]	0	0	0	0	0	0	0

Table 3: Adjacency matrix of projection2 (crimes)

		Crime ID											
		[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
Crime ID	[1,]	0	0	0	0	0	1	0	0	0	1	0	0
	[2,]	0	0	0	0	0	0	1	0	0	0	0	1
	[3,]	0	0	0	0	0	0	0	0	0	0	0	0
	[4,]	0	0	0	0	0	0	0	0	1	0	1	0
	[5,]	0	0	0	0	0	0	1	0	0	0	0	1
	[6,]	1	0	0	0	0	0	0	0	0	1	0	0
	[7,]	0	1	0	0	1	0	0	0	0	0	0	2
	[8,]	0	0	0	0	0	0	0	0	0	0	0	0
	[9,]	0	0	0	1	0	0	0	0	0	0	1	0
	[10,]	1	0	0	0	0	1	0	0	0	0	0	0
	[11,]	0	0	0	1	0	0	0	0	1	0	0	0
	[12,]	0	1	0	0	1	0	2	0	0	0	0	0

by single person and Crimes 8 and 3 not carried by common person.

Table 3 depicts the adjacency matrix of the projection 2 shown in Fig. 10.

6 Performance Measure

For the KONECT crime data set there are 443 groups with modularity value of 0.715. The clustering is performed over several times to obtain the optimal groups. The result of each step of clustering along with the modularity value is shown in Fig. 11 where X-axis represents the number of steps and the Y-axis in first graph represents the number of clusters and in second graph represents Modularity value.

The clustering starts with first level of grouping and continues till the optimal group is reached. At step 14, 146 groups with 0.908 modularity is obtained. Later on continuing at step 15, 146 groups with 0.909 modularity is obtained. At step 16, also 146 groups with 0.909 modularity is obtained. In between the steps 1 and 16, the number of groups decreases suddenly at step 4 which in turn is clustered at later steps. The sudden decrease in modularity is due to unlinked vertex in the data which is omitted in later steps. The community of people involved in same crime and vice versa are grouped which shows

Table 4: Local clustering coefficient value for person wise projected KONECT crime data set

cc_L	Number of persons	Inference
0	180	No Influence in the crime
Below cc_G	116	No major influence in crime
Above cc_G	49	Some influence in crime
1	526	Important Person in the Crime

the involvement of persons in the crime case based on their weight attribute.

The degree of each vertex is calculated and is shown in Fig. 12 for the sample crime data. The X-axis shows the vertex and Y-axis shows the degree associated with each vertex. The maximum degree in person vertex shows that they are involved in more crimes and that in crime vertex shows that more persons are involved in same crime. For the sample crime event graph there are around 2 cliques obtained which results in the 16 count in the maximum clique. There are 5 articulation points among 19 vertices in the data. For the entire KONECT crime event graph there are around 2 cliques obtained which result in the 1531 count in the maximum clique. There are 488 articulation points among 1427 vertices in the data.

The centrality closeness values of each vertex are calculated and are depicted as chart in Fig. 13. The X-axis shows the vertices and Y-axis shows the closeness value computed. The maximum value shows the central person in the community detected. Person 2 holds the value 0.00384 in group 1 community who acts as centre person and associated to person 1 in the crimes 7 and 12.

On analysis, it is found that the hub score and authority score are same for the sample crime event graph and the scores are shown in Fig. 14. The X-axis shows the vertex and the Y-axis shows the authority score. The maximum value associated with person vertex shows that the persons involved in similar crime and that associated with crime vertex shows that the common peoples of that crime. Here person vertex 4 and 5 have maximum value which shows that they are involved in the crime 8 which holds maximum value in crime vertex.

The analysis of calculated local clustering coefficients (cc_L) of the projection 1 (persons) are listed in Table 4. The global clustering coefficient (cc_G) of projection 1 is 0.5815762. 180 persons have no connection with any crime whose cc_L value is 0. 116 persons have no major connection with any crime whose cc_L value is below the cc_G value. 49 persons have some influence in the crime whose cc_L value is above the cc_G value. 526 persons are major actors in the crime whose cc_L value is 1.

The local clustering coefficient value of nodes above the global clustering coefficient value treated as influential persons in the crime event is shown in Fig. 15. The X-axis shows the Person_ID and Y-axis shows the local clustering coefficient value (ccL). The IDs of persons falling under

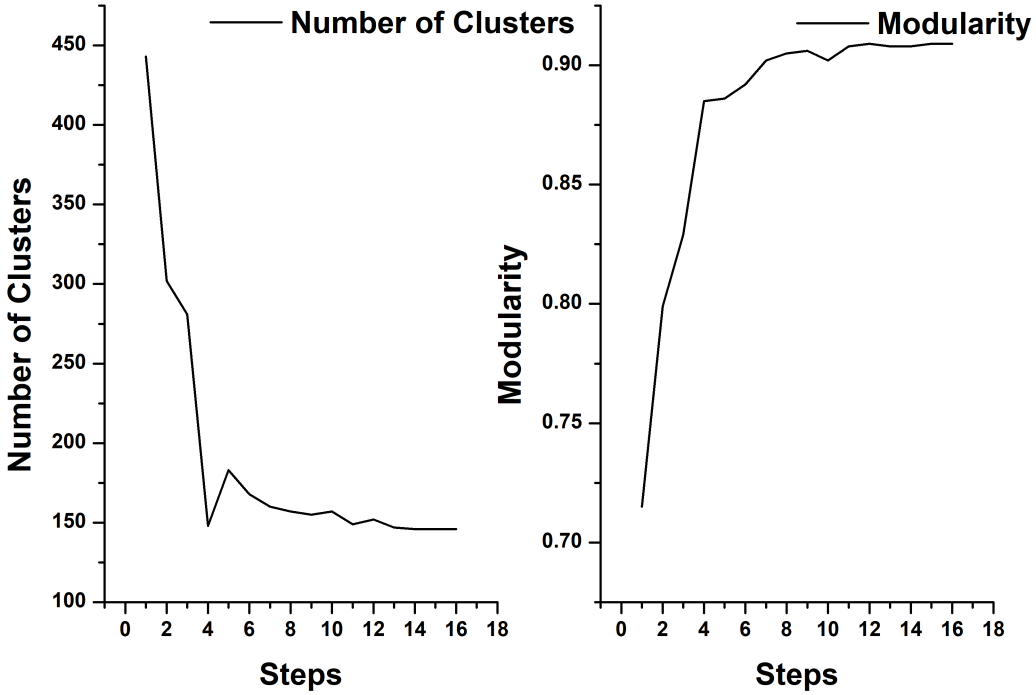


Fig. 11: Details of cluster and modularity value for KONECT crime data set

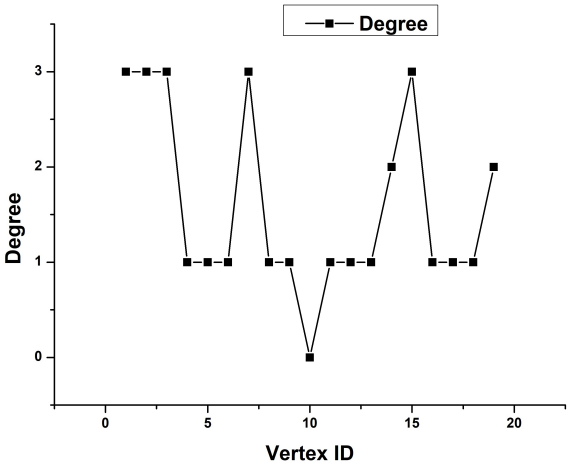


Fig. 12: Degree distribution of KONECT crime data

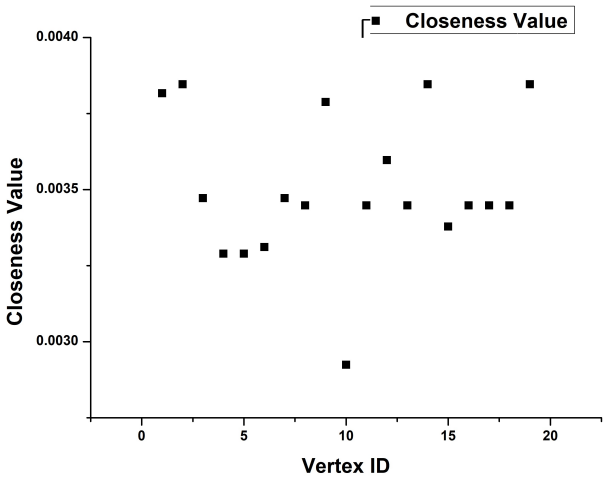


Fig. 13: Closeness distribution of KONECT crime data

this category are 46, 61, 76, 101, 119, 137, 140, 154, 161, 176, 237, 299, 319, 321, 329, 343, 349, 381, 386, 417, 421, 427, 435, 440, 458, 460, 473, 498, 502, 516, 539, 560, 562, 593, 603, 667, 670, 671, 684, 701, 767, 775, 780, 811, 813, 816, 820, 830, 851.

The local clustering coefficient value of nodes below the global clustering coefficient value treated as influential persons in the crime event is also shown in Fig. 16. The X-axis shows the Person_ID and Y-axis shows the local clustering coefficient value (ccL). The IDs of persons falling under this category are 2, 10, 12,

Table 5: Comparison of community detection using KONECT crime dataset with other approaches

Algorithm	Number of Clusters		Iterations	Effectiveness	CPU execution time
	NR	R			
Modified_Cluster_Walktrap	146	13	14	10.43	3.81 s
Cluster_Edge_Betweenness	126	20	13	9.69	3.25 s
Cluster_Walktrap	130	10	18	7.22	4.86 s

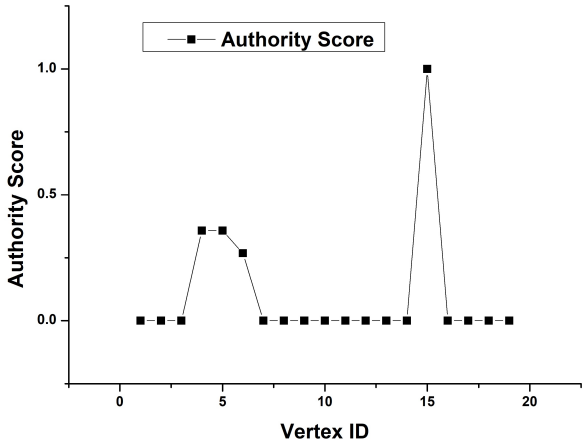


Fig. 14: Score distribution of KONECT crime data

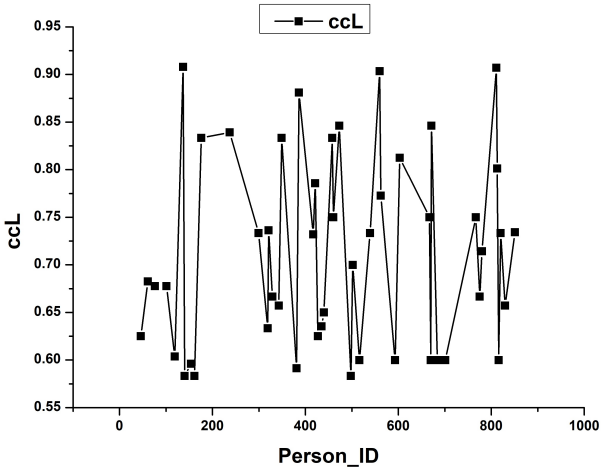


Fig. 15: Influential nodes in KONECT crime data set

17, 21, 37, 47, 54, 57, 59, 60, 65, 77, 82, 85, 92, 105, 110, 112, 116, 120, 124, 138, 155, 164, 168, 180, 194, 200, 212, 225, 230, 231, 232, 254, 256, 257, 261, 267, 279, 284, 293, 307, 310, 313, 315, 320, 351, 353, 354, 355, 362, 363, 371, 376, 382, 383, 391, 402, 405, 413, 418, 426, 428, 432, 433, 436, 446, 447, 462, 464, 476, 478, 490, 496, 500, 507, 524, 530, 542, 565, 572, 578, 597, 608, 615, 624, 647, 668, 676, 687, 694, 712, 716, 717, 722, 726, 731, 732, 751, 755, 759, 765, 772, 776, 782, 800, 801, 804, 825, 835, 849, 852, 855, 859, 862.

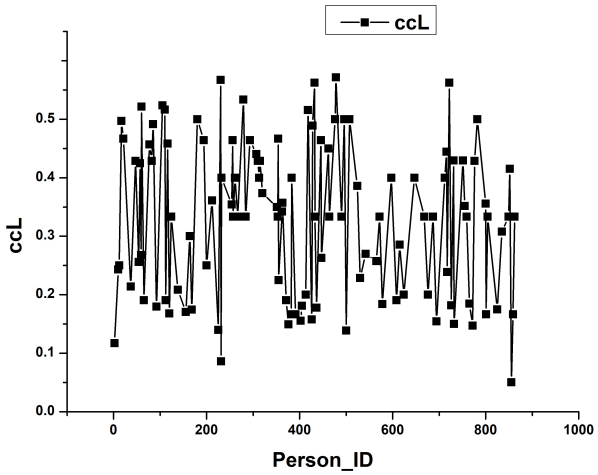


Fig. 16: Non-influential nodes in KONECT crime data set

For clear visualization, the sample network with 7 persons, 12 crimes and 15 links is shown. Fig. 17 depicts the sample network where persons are represented in the circle and the crimes are represented in the box. The structure shows the bipartite representation of the sample crime event data. The dashed circles represent the individuals who are influential in the entire network. The dark circles represent the individuals who are influential in the community.

Later the nodes with the local clustering coefficient values are considered for categorization into victim, suspect and witness. The cc_L is used as the limit for categorizing the nodes along with the consideration of nodes probability measure within the cluster formed. There are overlapping of nodes that exist in the victim and suspect. The witness is a direct categorization which holds the nodes with cc_L value as 0.

6.1 Performance Analysis

The performance analysis of the proposed approach is carried out using *R* tool. For the KONECT crime data set considered for implementation, the performance analysis of clustering the nodes with the proposed method of walktrap and with other methods like cluster_edge_betweenness and cluster_walktrap is shown in Table 5.

The methods other than modified cluster walktrap need to project the two-mode crime network into two

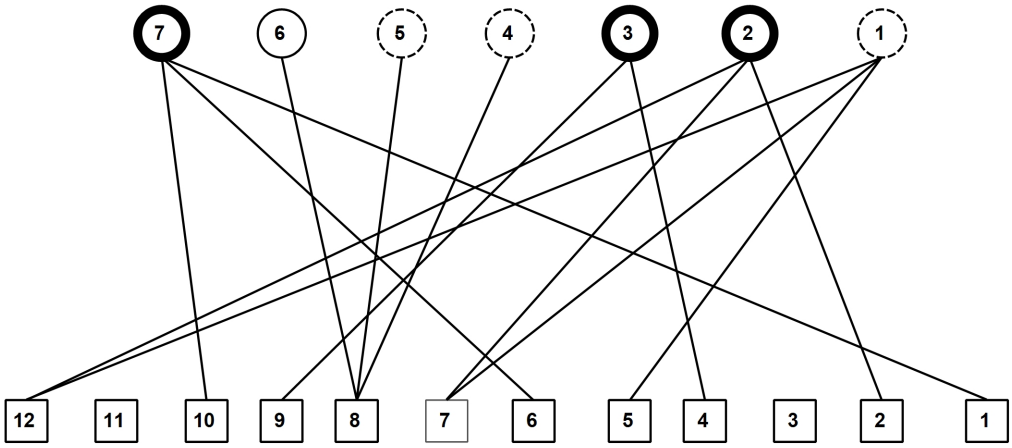


Fig. 17: Influential nodes in sample data

Table 6: Performance of categorizing the persons of KONECT crime data

		Implementation result			Sensitivity	Specificity	Positive predicted value	Negative predicted value	F1 score
		V	S	W					
Dataset	V	525	23	21	0.9227	0.9345	0.9099	0.9494	0.9162
	S	47	574	61	0.8417	0.9568	0.9456	0.8713	0.8906
	W	5	10	180	0.9231	0.9345	0.6871	0.9873	0.7878

one-mode projections for identifying the clusters. The performance is measured by means of number of clusters formed with Non-Repetitive (NR) set of nodes and also Repetitive set (R) of nodes in the network and the number of iteration for tracking the entire network. The effectiveness of the approaches is measured as the ratio between NR and number of iterations. Even though, the number of iterations of edge betweenness is low compared to modified walktrap and number of non-repetitive groups of nodes formed is high.

On the whole the persons of KONECT crime data play 1446 roles in the constructed crime network. Among the roles played, our proposed method paves way to categorize the persons into three types namely victim-V, suspect- S and witness-W. The identified number of victims is 525 out of 569, a suspect is 574 out of 682 and a witness is 180 out of 195. Sensitivity, specificity, positive predicted value, negative predicted value and F1 score based on recall and precision are important key measures for evaluating the performance of categorization which are shown in Table 6.

7 Conclusion

In this paper, a combined cluster walktrap and closed cycle approach for effective identification of influential persons in crime network is discussed. Using the membership and modularity measures, a community detection model is developed for clustering the nodes in the crime network. The probability-based modularity measure is considered as a threshold value for the modified cluster walktrap approach. As the communities are formed, the multi-mode crime network is projected into different one-mode networks for identifying most influential nodes in the network with the help of the hidden attribute information. The influential nodes are identified by means of transitivity value calculated by closed cycle approach. The proposed CWT-CCA method uses the degree-based probability measure which improves the accuracy of the clusters formed. Also the method supports the linking of the clustering coefficient values of the nodes in the projected one-mode network with the clusters formed. On analysing the result with the existing approaches, the proposed method shows better clustering of nodes and the accurate set of prominent nodes. The proposed algorithm can also be implemented for network analysis in any domain. Furthermore, the proposed approach is tailored to a two-mode crime network, which can also be employed over the two-mode networks of cyber crime. The variation can be made in the proposed work by employing a machine learning method to the two-mode crime network after projection. As an extension to the present work, it is planned to implement fuzzy-based domain driven approach with multiple constraint behaviour for identification of influential nodes in the crime network in an optimal way.

References

[1] Farmer and Lindsay, Crime, definitions of in Cane and Conoghan, *The New Oxford Companion to Law*, 263 (2008).

[2] R. Boba, Introductory guide to crime analysis and mapping, *Report to the Office of Community Oriented Policing Services Cooperative Agreement #97-CK- WXK-004*, 15–16 (2001).

[3] Y. Xu, L. Mingyang, A. Ningning and Z. Xinchao, Criminal detection based on social network analysis, Proceedings of the Eighth International Conference on Semantics, *Knowledge and Grids*, 24–31 (2012).

[4] M. Fiedler, Algebraic connectivity of graphs. *Czechoslovak Mathematics Journal*, **23**(2), 298–305 (1973).

[5] B.W. Kernighan, and S. Lin, An efficient heuristic procedure for partitioning graphs, *Bell System Technical Journal*, **49**(2), 209–308 (1970).

[6] M.E.J. Newman and M. Girvan, Finding and evaluating community structure in networks. *Physical Review E*, **69**(2) (2004).

[7] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, Defining and identifying communities in networks. *PNAS*, **101**(9), 2658–2663 (2004).

[8] M.E.J. Newman, Fast algorithm for detecting community structure in networks, *Physical Review E*, **69**(6) (2004).

[9] L. Donetti and M.A. Munoz, Detecting network communities: a new systematic and efficient algorithm, *Journal of Statistical Mechanics*, **2004**(10), 10012–10019 (2004).

[10] A. Focuss, F. Pirotte and M. Saerens, A novel way of computing dissimilarities between nodes of a graph, with application to collaborative filtering, *Workshop on Statistical Approaches for Web Mining (SAWM)*, 26–37 (2004).

[11] H. Zhou and R. Lipowsky, Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and sub communities, *Proceedings of International Conference on Computational Science*, 1062–1069 (2004).

[12] S.V. Dongen, *Graph clustering by flow simulation*, PhD thesis, University of Utrecht. (2000).

[13] P. Pons and M. Latapy, Computing communities in large networks using random walks, *Journal of Graph Algorithms Application*, **10**(2), 191–218 (2006).

[14] V.D. Blondel, J.L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory & Experiment*, **2008**(7) (2008).

[15] A. Lancichinetti and S. Fortunato, Community detection algorithms: a comparative analysis, *Physical Review E*, **80**(5), 117–128 (2009).

[16] W.W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research*, **33**(4), 452–473 (1977).

[17] M. Girvan and M.E.J. Newman, Community structure in social and biological networks. *PNAS*, **99**(12), 7821–7826 (2002).

[18] H. Jeong, S. Mason, A.L. Barabasi and Z.N. Oltvai, Centrality and lethality of protein networks, *Nature International Journal of Science*, **411**(5), 41–42 (2001).

[19] J. Gehrke, P. Ginsparg and J. Kleinberg, Overview of Kdd cup, *Proceedings of Ninth Annual ACM SIGKDD Conference*, **5**(2), 149–151 (2003).

[20] M. Hoerdt and D. Magoni, Completeness of the internet core topology collected by a fast mapping software, *Proceedings of the 11th International Conference on Software*, Telecommunications and Computer Networks, 257261–257269 (2003).

[21] T. Alzahrani and K.J Horadam, Community detection in bipartite networks: Algorithms and case studies, *Complex Systems and Networks Dynamics, Controls and Applications*, Springer, 25–50 (2016).

[22] J. Liebig and A. Rao, Identifying influential nodes in bipartite networks using the clustering coefficient, *Proceedings of Tenth IEEE International Conference on Signal-Image Technology and Internet-Based Systems*, Morracco, 23–27 (2014).

[23] D. Murdoch, The comprehensive R archive network, *CRAN-r-base* (2017).

[24] L.C. Freeman, *The development of social network analysis: A study in the sociology of science*, Vancouver (2004).

Ganeshkumar Pugalendhi received his B.Tech, MS (by Research), and Ph.D. degrees all in Information Technology in 2003, 2008, and 2012 from the University of Madras, Anna University, Chennai, and Anna University, Coimbatore, respectively. He completed his Post Doctorate from the School of Computer Science and Engineering, Kyungpook National University, South Korea in 2016. His research interest includes Social Network Analysis, Machine Learning, Big Data Analytics and IoT.

Shanmugapriya Kumaresan received her B.E., M.Tech, and MBA degrees in Computer Science and Engineering, Information Technology and Human Resource in 2009, 2011, and 2014 from the Anna University, Chennai, Anna University, Coimbatore and Bharathiyar University, Coimbatore respectively. Her research interest includes Social Network Analysis, Machine Learning.