Novel Algorithm for Terrorist Network Mining

[1]Nisha Chaurasia, [2]Akhilesh Tiwari
Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior (M.P.), India
[1]chaurasianisha21@gmail.com, [2]atiwari.mits@gmail.com

***Abstract*— The evolutionary effect of social network analysis has feasibly enabled the security agencies to include the approach for estimating terrorist networks. The approach in this sense named as Terrorist Network Mining, has helped in successfully achieving the terrorist network behaviour on web along with destabilization of network using hierarchy of algorithms. However these algorithm results in fruitful outcomes, there is a need to propose a typical algorithm for destabilization. In consideration to this, the present paper proposes a novel algorithm for destabilization of terrorist network revealing the hidden hierarchy followed by the network.**

Keywords*— *Social Network Analysis, Data Mining, Terrorist Network

I. INTRODUCTION

Although much of works have been formalized to tackle the problem of terrorism, still there is a need to diminish their existence. From the initial creation of the network of terrorists to the visualization and tracking of their activities, terrorists future attacks prevention is made possible. Though social network analysis (SNA) is a broad research field with much of appreciable work been performed, SNA under data mining has evolved as an emerging field of research. The challenges faced and the prominent need to overcome the terrorism has made the analysts learn about handling and contributing to nation security by preventing terrorist future attacks.

The SNA because of its ability to uncover network behaviour has led the law enforcement agencies to use the concept for studying the various hidden terrorist groups on the web. This is generally done utilizing the graph theory where the network is considered as a graph and its nodes as edges and link representing relationship shared. Thus the study of terrorist network utilizing SNA and Graph Theory is termed as Investigative Data Mining or Terrorist Network Mining. The IDM can be defined as "The technique which is used for determining associations and predicting criminal behavior by analyzing network structure in order to identify key nodes for the purpose of destabilizing criminal/terrorist networks" [11].

The present paper adds an advantage to the detection process of terrorist networks by introducing a novel algorithm for terrorist network destabilization. The algorithm makes use of the two centralities i.e. PageRank Centrality and Katz Centrality enabling an effective estimation of hierarchy followed by the terrorist networks. The paper in consideration to this, includes the following sections further: related work of the field, a brief discussion about social network analysis along with the description about terrorist network mining, available centrality measures, detection approach of terrorist network, destabilization concept, proposed methodology consisting recommended centrality measures for the proposed algorithm and description about the algorithm, finally experimental results are shown which were obtained applying the proposed algorithm.

II. RELATED WORK

The strength of the terrorists groups on web came into shed by 9/11 attacks which left the everlasting memories of their inhumanity. The attack made the intelligence and the law enforcement agencies to concern more strongly about the security as the success of the attack involved the intensive use of the internet. This is one reason for the major effort made by law enforcement agencies around the world in gathering information from the Web about terror-related activities [5]. The goal of gathering information was to prevent the future attacks possibilities.

The very first attempt in order to analyze the terrorist networks was made by Valdis Krebs in 2002 after September 2001 attacks. He used network analysis to provide an extensive analysis of the 9/11 Hijackers network, explains three problems he encountered very early on drawing on the work of Malcolm Sparrow [4]. He mentioned three problems that a social network analyst would encounter while constructing the terrorist organizations network graph. These were [4]:

- Incompleteness - the inevitability of missing nodes and links that the investigators will not uncover.
- Fuzzy boundaries - the difficulty in deciding who to include and who not to include.
- Dynamic - these networks are not static, they are always changing.

Krebs on the basis of the knowledge he gained about the September, 2001 framed a network consisting of terrorist nodes and evaluated the importance and the contribution of each node in the attacks. The major contributors for his study were the priori study done by:

- Malcolm Sparrow (1991), examining the application of SNA to criminal activity.
- Wayne Baker and Robert Faulker (1993), suggests looking at previously stored or known data to find the relationship data.

- Bonnie Erickson (1981), explaining the importance of trusted prior contacts (that came in touch before) for the effective

- Functioning of a secret society (such as terrorist groups).

Krebs during his study of terrorist networks evaluated the links in the network on the basis of their strength. The strength of the tie depends on the time spent by user together. He categorized the tie or strength on three scales:

- Strongest tie link reveals cluster of network players or leaders of the group. The node pair with the strongest tie would largely be governing the group.

- Moderate strength or medium thickness links reveals the nodes through which maximum transactions are done or information is forwarded.

- Weak tie or the thinnest links reveals the nodes having a single transaction, or an occasional meeting and no other ties.

Using the SNA centrality measures, Krebs evaluated the involvement of each node or user in the attacks. This beneficial step helped law enforcement agencies to detect terrorist networks more effectively. However the centrality measures are significant for the analysis but these are very sensitive to minor changes in network connectivity [1]. After the successful analysis of the criminal network, Kreb discussed the shortcomings that were to be considered. He discussed about the difficulty faced in discovering of links that may be the stronger ties but because of their low frequency of activation they may appear to be weak ties. The second consideration was about the network detection as the less active the network, the more difficult it is to discover [1].

The next beneficial and novel effort for terrorist group detection on the web was made by Y.Elovici, A.Kandel, M.Last, B.Shapira and O. Zaafrany in 2004. They as a group analyzed the behavior of terrorists on the web using data mining techniques. They tried to solve the problem still faced about the dynamic switching of IP addresses and URLs by terrorist users. Hence in place of tracking terrorists on the basis of their IP addresses, they proposed a methodology by monitoring on all the ISPs traffic to detect the users accessing the terrorists' related information, keeping in mind the privacy issues. They defined three essential design goals to be fulfilled for the methodology. These are [5].

- Training the detection algorithm should be based on the content of existing terrorist sites and known terrorist traffic on the Web.

- Detection should be carried out in real-time. This goal can be achieved only if terrorist information interests are presented in a compact manner for efficient processing.

- The detection sensitivity should be controlled by user-defined parameters to enable calibration of the desired detection performance.

Y.Elovici et al. methodology also includes integration of three research fields to detect as well to evaluate typical terrorists behavior. The three topics were the following:

- Computer Security: It was performed using Intrusion Detection System (IDS). An IDS purpose is to detect abnormal actions if any, in the constantly monitored environment. It could an individual system, or number of computers in the network or the network itself. It tracks the activity of users on the web and evaluates measures viz. accuracy, completeness, performance, efficiency, fault tolerance, timeliness, and adaptivity. Along with these, the interest measures are true positive rate (TP, the percentage of intrusive or abnormal activities, such as terror-related pages viewed, detected by the system), false positive rate (FP, the percentage of normal actions, such as the pages viewed by normal users, incorrectly detected as intrusive by the system. Accuracy which is the percentage of alarms found to represent abnormal behavior out of the total number of alarms [5].

- Information Retrieval: For information retrieval the vector space model is used. The content of the web page is represented as a document in form of an n-dimensional vector. The similarity among two documents is calculated using any of the distance measuring methods such as Euclidean distance or Cosine. The cosine distance evaluated the similarity between an accessed web page and a given set of terrorists' topic of interest. The terrorists' interests are represented by several vectors where each vector relates to a different topic of interest [5].

- Data Mining: This research field involved uses unsupervised clustering to cluster or partition the web pages as collection belonging to terrorists' with same topic of interest. Cluster analysis is the process of partitioning data objects (records, documents, etc.) into meaningful groups or clusters so that objects within a cluster have similar characteristics but are dissimilar to objects in other clusters [7]. For each collection of web pages or a cluster, a centroid is determined and represented by the vector space model.

After combining these three fields as the steps, learning of Typical-Terrorist-Behavior is obtained. Typical-Terrorist-Behavior depends on the number of clusters. When the number of clusters is higher, the Typical-Terrorist-Behavior includes more topics of interest by terrorists where each topic is based on fewer pages [5]. Next step to this is to monitor the ISP traffic and detect the terrorists present on the web. For this purpose, each accessed web page by user is converted into vector named access vector using a vector generator. The access vector is then compared for similarity against all centroid vectors of the Typical –Terrorist-Behavior using cosine measure.

If the similarity among the two is higher than a predefined threshold then an alarm is generated by the detector indicating user who accessed those web pages as illicit user. The sensitivity of the detection process depends on this predefined threshold value. Higher value of threshold will decrease the sensitivity of the detection process, decrease the number of alarms, increase the accuracy and decrease the number of false alarms. Lower value of threshold will increase the detection process sensitivity, increase the number of false alarms and decrease the accuracy. The optimal value of threshold, T depends on the preferences of the system user [5].

The similarity may be calculated using the inequality:

$$\max\left[\frac{\sum_{i=1}^{m}(tC_{i1}.tA_i)}{\sqrt{\sum_{i=1}^{m}tC_{i1}^2.\sum_{i=1}^{m}tA_i^2}},\dots,\frac{\sum_{i=1}^{m}(tC_{in}.tA_i)}{\sqrt{\sum_{i=1}^{m}tC_{in}^2.\sum_{i=1}^{m}tA_i^2}}\right] > T$$

Next step to this research was made by Nasrullah Memon and Abdul Rasool Qureshi in 2005 by introducing a data mining tool named Investigative Data Mining (IDM) for the terrorist network analysis.

As opposed to traditional data mining aiming at extracting knowledge form data, mining for investigative analysis, called Investigative Data Mining (IDM), aims at discovering hidden instances of patterns of interest, such as patterns indicating an organized crime activity [6]. An important problem for identifying of terror/crime networks was resolved by IDM, based on available intelligence and other information. It is also capable to determine the subgroups if present in the network.

In contrast to traditional data mining where mining is performed on large databases, the IDM analyses and predicts the risks from the access behavior and the association existing among the terrorists network. With this, it has also been a helping hand in the destabilizing of the network.

Nasrullah Memon and Henrik Legind Larsen in 2006 proposed the two highly beneficial algorithms for the destabilizing of the terrorist networks. N. Memon along with his fellow mates, after the deep study about the criminal networks proved various possible solutions to enhance the strategies for the efficient detection of terrorist networks on web.

He proposed the hierarchy of algorithms such that destabilization is achieve in a promising manner. The hierarchy of algorithms was based on the SNA approach along with the measures defined previously. The first algorithm objective was to convert the network undirected graph obtained by SNA approach into a directed graph. While the second algorithm works specifically for destabilization by constructing a tree from the graph, by calculating the dependency of each node to other nodes in the network. They also introduced a new centrality measure called dependence centrality. The dependence centrality (DC) of a node is defined as how much that node is dependent on any other node in the network [8**]**. The nodes with less DC are predicted as the key player (leader or the gateway) nodes as they are the nodes with highest number of direct links to other nodes and do not depend on any other nodes in the network for communication.

Same year, Muhammad Akram Shaikh, Wang Jiaxin presented the effectiveness of IDM along with its working in identification of key nodes in the network. According to the study, IDM offers the capability to track the criminal network more effectively and also to analyze the interaction patterns within the network. The IDM generally uses the graph theory and the SNA techniques for the estimation of the network.

III. SOCIAL NETWORK ANALYSIS

Social Network Analysis (SNA) is a data mining technique which usually analyses the various social networks present on the web. The technique is profitably used for studying the social behaviours of the networks. Thus social network analysis, from a data mining perspective, is also called link analysis or link mining [7].

The SNA uses a concept of centrality measures pointing out who is the central node(s) in the network. It is because of this that SNA is utmost utilized technique by the law-enforcement agencies for studying trends of hidden terrorist network.

In this context, when the SNA is applied for investigating of terrorist networks on the web then it is acknowledged as Investigative Data Mining (IDM), also known as Terrorist Network Mining. The ultimate goal of IDM is to investigate terrorist networks in order to find out who the suspicious people are and who is capable of carrying out terrorist activities and how to destabilize them [11]. It is a technique defined for the analysis of hidden terrorist network that uses SNA and Graph Theory for the investigation. IDM offers the capability to track the criminal network more effectively and also to analyze the interaction patterns within the network.The technique discovers the most promising node(s) within the network and the goal is to remove this node(s) from the network in order to neutralize the network activities.

IDM is recognized as a combination of data mining and subject-based automated data analysis techniques where data mining serves as an approach which uses predictive approach for discovering patterns in dataset and subject-based automated data analysis regulate models to data for predicting the behavior, access risk, determine associations, or perform other types of analysis. IDM aims to connect the dots between individuals and map and measure complex, covert, human groups, and organizations [13].

Because of the magnificent applicability of SNA for analysing network behaviour has attracted the various law-enforcement agencies to use the technique for analysing various hidden terrorist group on the web and enforce suitable remedies for their neutralization.

IV. AVAILABLE CENTRALITY MEASURES

The centrality measures are estimated using Graph Theory. The graph in mathematical form is represented as an adjacency matrix, A_{ij} such that

$$A_{ij} = \begin{cases} 1 & if\ i\ and\ j\ are\ connected \\ 0 & otherwise \end{cases}$$

following the property of symmetric matrix, i.e. $A_{ij} = A_{ji}$.

A. Degree

The Degree of a vertex in a network is the number of edges attached to it [9]. It is calculated as the sum of all directly linked nodes connected to a node for which degree is measured.

B. Betweenness

Betweenness is the number of geodesics (shortest paths between any two nodes) passing through it. An individual with high betweenness may be a gatekeeper in the network.

C. Closeness

Closeness is the sum of all geodesics between the particular node and every other node in the network [6].

D. Eigenvector

Eigenvector centrality (EC) of a node in a network is defined to be proportional to the sum of the centralities of the node's neighbours, so that a node can acquire high centrality either by being connected to a lot of others (as with simple degree centrality) or by being connected to others that themselves are highly central [12].

V. ANALYSIS OF TERRORIST NETWORK

For the analysis of the terrorist network, the network is discovered from the web by using approach such as content-based detection of terrorists on web. Whenever a terrorist network is detected, the network influential roles and the network hierarchy is uncovered using Investigative Data Mining scheme.

One way to detect terrorist activity on the Web is to eavesdrop on all traffic of Web sites associated with terrorist organizations in order to detect the accessing users based on their IP address [5]. But the solution was not much convincing as these users do not use fixed IP addresses or URLs. Hence the law enforcement agencies tried to detect the terrorists by monitoring all ISPs traffic.

After traffic analysis, the network is preliminary studied using social network analysis (SNA) approach. The detected terrorist network is then studied for estimating promising roles. The analysis of each node (user) is done in the network and the centrality measures are calculated respectively for each node. The main centrality measures are degrees (number of direct connections that a node has), betweenness (the ability of an individual to link to important constituencies) and closeness (a position's ability to monitor the information flow and to "see" what is happening in the network) [4].

VI. DESTABILIZATION OF TERRORIST NETWORK

To understand the dynamics of covert networks, and indeed any, network we need to understand the basic processes by which networks evolve [3]. Hence in consideration to this, terrorist network roles are discovered and accordingly destabilization is achieved.

The destabilization is attempted by performing role analysis within the network. This is done usually by evaluating the efficiency of the network, critical components of a network, a proposed measure "Position Role Index" (PRI) and dependence centrality.

- Efficiency of the network, to quantify how efficiently the information is exchanged among the nodes in the network.
- Critical components of a network, for finding the measure of the centrality of a node, using which the drop in the network efficiency is evaluated when that node is deactivated from the network.
- Position Role Index (PRI), highlighted a clear distinction between followers and gatekeepers (It is a fact that leaders may act as gatekeepers) [10] and depends on the efficiency of the network.
- Dependence Centrality, for finding the node dependency on other nodes of the network and finding the leader/gatekeepers.

With respect to the role analysis, hierarchy of the terrorist network is determined. Discovering hierarchy in a terrorist network is a process of comparing different centrality values of different nodes to identify which node is more powerful, influential or worthy to neutralize than others [12].

VII. NEED FOR CONSTRUCTING HIERARCHY

The need for revealing hierarchy of terrorist network was to achieve efficient destabilization. The need was found by N. Memon et al. [8] where the author explained the concept behind the hierarchy along with an introduction to dependence centrality and hierarchy of algorithms defined for destabilizing terrorist activities on web.

The purpose of proposing the hierarchy of algorithms is to solve the dilemma that occurred during the estimation of influential nodes considering the hierarchy in form of a tree. The dilemma constituted two concerns:

- The first dilemma was that sometimes there are chances where centrality measures for more than one nodes holds the same value henceforth creating difficulty in identifying which node will be parent and which will be a child.

- The second dilemma was to judge which node would be parent and which would be a child in the hierarchy, if more than two nodes qualify as powerful nodes over another particular node.

The dilemma was familiarized as ABC. After resolving the dilemma, hidden hierarchy is detecting using the two algorithms. The algorithms worked as following:

- The first algorithm was to covert an undirected network graph into a directed graph using degree and eigenvector centralities.

- The second algorithm intakes the directed graph obtained from the first algorithm constructs the hierarchy of parent and child nodes in form of a tree structure.

After constructing the hierarchy, the hierarchical relationship among the parents of a node is discovered. Finally, the most promising parent is detected from possible parents using dependence centrality.

VIII. PROPOSED METHODOLOGY

The proposed methodology includes two subsections: Centrality Measures i.e. PageRank, Katz and Dependence Centrality measures for discovering the hidden hierarchy of the terrorist network and a new proposed Algorithm for Destabilization of terrorist network.

A. Centrality Measures used in Algorithm

The centrality of a node in a network is interpreted as the importance of the node [14].

1. *Katz centrality:* Katz centrality measures the extent of influence of a node in a network i.e. it counts the number of walks starting from a node or ending on a node, providing penalties to longer walks.

2. *PageRank Centrality:* PageRank Centrality is a way to measure network centrality similar to degree centrality. It is considered as an enhance version of indegree centrality used to measure the influence on other nodes in the network.

3. *Dependence Centrality (DC):* Memon et al [9] defined the dependence centrality of a node as how much that node is depending on any other node in the network.

B. Proposed Algorithm for Destabilization

This section proposes a new-fangled algorithm for destabilization of terrorist networks.

The algorithm works well by resulting in limiting nodes in the graph, i.e. typically focusing on the influential node(s).

The algorithm takes into account all nodes and searches for the key node(s) by utilizing PageRank and Katz centralities.

The framework of the proposed algorithm is as follows:

Proposed Algorithm:

1. Find neighbors of all nodes in the graph
2. Compare PageRank Centrality of each node to its neighbor node and calculate the Parent and Children set.
3. If value of PageRank is same for two nodes, compare their Katz centrality to resolve conflict and extent the two sets.
4. Else ignore the link
5. After calculating the Parent and Children sets, find the hierarchy.
6. If a node has no parent, it is attached to root node.
7. For Parent set with one value, node is the child of that set value node.
8. For node with Parent set with more than one value, maximum $[N(P1) \cap N(n)]$ is estimated and the parent node with maximum value is set as parent of a node. For $N(P1) \cap N(n) = 0$, node is overlooked.
9. Again if Parent set has more than one value, the parent node with maximum Dc value is set as parent of a node.
10. Even then the Parent set has multiple values the node is attached to root.

IX. EXPERIMENTAL RESULTS

The section involves the analysis of the effectiveness of the proposed algorithm. The performance of the proposed algorithm is assessed through the 26/11 attacks dataset involving thirteen terrorists who were responsible for the disaster. The simulation environment for calculating the centralities is a tool defined for social network analysis, recognized as UCINET [2]. The name of the thirteen terrorists of 26/11are listed in Table I with a number assigned respectively.

Table I: Name of Terrorists Involved in 26/11

S.No.	Name of Terrorist
0	Abu Kaahfa
1	Wassi
2	Zarar
3	Hafiz Arshad
4	Javed
5	Abu Shoaib
6	Abu Umer
7	Abdul Rehman
8	Fahadullah
9	Baba Imran
10	Nasir
11	Ismail Khan
12	Ajmal Amir Kasab

The Table II consist the values of centralities calculated for the algorithm:

Table II: Value of PageRank and KATZ Cantralities

Name of Terrorist	Pagerank	Katz
Abu Kaahfa	0.074	1.087
Wassi	0.210	2.572
Zarar	0.074	1.087
Hafiz Arshad	0.146	1.467
Javed	0.074	0.992
Abu Shoaib	0.109	0.773
Abu Umer	0.146	1.467
Abdul Rehman	0.069	0.260
Fahadullah	0.069	0.260
Baba Imran	0.039	0.514
Nasir	0.039	0.514
Ismail Khan	0.039	1.000
Ajmal Amir Kasab	0.039	1.000

Among the 13 terrorists, Wassi is expected as the most crucial node in the network as it led multiple communications through it with highest number of direct connections with other nodes in the network. The output obtained after accessing the proposed algorithm could be visualized through Figure1.

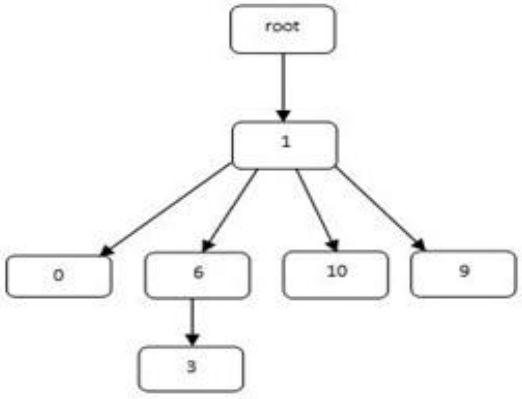

Figure 1: Hierarchy of 26/11 obtained after applying the Proposed Algorithm

X. CONCLUSION

This paper introduces a significant algorithm for destabilization of terrorist networks. The algorithm is expected as the most beneficial strategy for revealing pattern of the terrorist activities. The output obtained from experimental analysis on 26/11 dataset as already discussed uncovers Wassi as the dominating node with maximum communication flowed through it. Hence in view to this, the proposed algorithm can be considered as the most effective algorithm for uncovering the network hierarchy and neutralizing network activities.

REFERENCES

[1] V. E. Krebs, (2002), "Uncloacking Terrorist Networks". First Monday, Volume 7, 4 - 1 (2002)

[2] S. P. Borgatti, M. G. Everett, and L. C. Freeman, (2002), "UCINET 6 for Windows", Analytic Technologies, Cambridge, MA: Harvard University Press (2002).

[3] K. M. Carley, J. Reminga, and N. Kamneva, (2003), "Destabilizing Terrorist Networks", In Proceedings of the 8th International Command and Control Research and Technology Symposium, 2003.

[4] P. V. Fellman and R. Wright: Modeling, (2003), "Terrorist Networks - Complex Systems at the Mid-Range", In: Proceedings of Complexity, Ethics and Creativity Conference, LSE (2003).

[5] Y. Elovici, A. Kandel, M. Last, B. Shapira and O. Zaafrany, (2004), "Using Data Mining Techniques for Detecting Terror-Related Activities on the Web", In: Proceedings of Journal of Information Warfare (2004).

[6] N. Memon and A. R Qureshi, (2005), "Investigative Data Mining and its Application in Counterterrorism", In Proceedings of the 5th WSEAS Int. Conf. on Applied Informatics and Communications, Malta, pp. 97-403, 2005.

[7] J. Han and M. Kamber, (2006), "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, 2006.

[8] N. Memon and H. L. Larsen, (2006), "Practical Approaches for Analysis, Visualization and Destabilizing Terrorist Networks", Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06), IEEE 2006.

[9] N. Memon, D. L. Hicks, D. M. A. Hussain and H. L. Larsen, (2006), " Practical Algorithms And Mathematical Models For Destabilizing Terrorist Networks", In Sharad Mehrotra, Daniel Dajun Zeng, Hsinchun Chen, Bhavani M. Thuraisingham, Fei-Yue Wang (Eds.): ISI 2006, LNCS 3975, pp. 389. Springer-Verlag Berlin Heidelberg, 2006.

[10] N. Memon and H. L. Larsen, (2006), "Structural Analysis and Destabilizing Terrorist Networks", In Proceedings of The First International Conference on Availability, Reliability and Security, 2006. ARES 2006, IEEE 2006.

[11] M. A. Shaikh, and W. Jiaxin, (2006), "Investigative Data Mining: Identifying Key Nodes in Terrorist Networks", Multitopic Conference, INMIC '06, pp. 201-207 IEEE 2006.

[12] N. Memon, H. L. Larsen, D. L. Hicks, and N. Harkiolakis, (2008), "Detecting Hidden Hierarchy in Terrorist Networks: Some Case Studies", In Proceedings of Springer-Verlag Berlin Heidelberg 2008, ISI 2008 Workshops, LNCS 5075, pp. 477–489, 2008.

[13] U. K. Wiil, N. Memony, and P. Karampelas, (2010), "Detecting New Trends in Terrorist Networks", In Proceedings of International Conference on Advances in Social Networks Analysis and Mining, 2010.

[14] U. Kang, S. Papadimitriou, J. Sun and H. Tong, (2011), "Centralities in Large Networks: Algorithms and Observations", In Proccedings of SIAM International Conference on Data Mining (SDM'2011), Phoenix, U.S.A., 2011.