International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Data Mining in Social Networks and its Application in Counterterrorism

Krishna Ganeriwal, Gayathri P, G. Gopichand, H. Santhi

Abstract*: Social Networks are best represented as complex interconnected graphs. Graph theory analysis can hence be used for insight into various aspects of these complex social networks. Privacy of such networks lately has been challenged and a detailed analysis of such networks is required. This paper applies key graph theory concepts to analyze such social networks. Moreover, it also discusses applications and proposal of a novel algorithm to analyze and gather key information from terrorist social networks. Investigative Data Mining is used for this which is defined as when Social Network Analysis (SNA) is applied to Terrorist Networks to gather useful insights about the network..

***Index Terms*: Graph Theory, Graph Mining, Investigative Data Mining, Social Network Analysis.**

I. INTRODUCTION

A graph typically is a collection of non-linear nodes connected using edges. A social network can typically be defined using a graph, where the members of the network comprise of the nodes and their mutual connections are represented using the edges. Social network analysis is a key technique in present day human communication analysis. It has also gained a significant popularity in anthropology, biology, communication studies, economics, geography, information studies, organizational studies and human psychology and has become a popular topic of speculation and study. Humans have used the thought of social systems freely for over a century to mean complex arrangements of connections between individuals from social frameworks by any stretch of the imagination scales, from interpersonal to worldwide. Graph theory involves analysis of graphs and extracting useful information and represents it using standard metrics such as degree, etc.

Graph representation of SNA topologies are used for various purposes such as community detection, network structures, random walks and temporal networks. Social network analysis as mentioned has been useful in numerous fields and much of its impact is shared in [1] and [2]. There have been numerous issues with these ever so important social networks. A good compilation of these issues can be found [5] here.

In this paper, we discuss various metrics that can be applied to social network graphs such as centrality, closeness, betweenness, bridge, clustering coefficient, page-rank centrality, etc. Furthermore, the existing algorithms and their disadvantages are also analyzed and finally, a novel algorithm is proposed to counter some of these disadvantages. In the closing section of the paper potential areas for future research is discussed.

Manuscript published on 30 September 2019
* Correspondence Author
Krishna Ganeriwal*, SCOPE, VIT University, Vellore, India. Email: ganeriwalk@gmail.com
Gayathri P, SCOPE, VIT University, Vellore, India. Email: pgayathri@vit.ac.in
G. Gopichand, SCOPE, VIT University, Vellore, India. Email: gopichand.g@vit.ac.in
H Santhi, SCOPE, VIT University, Vellore, India. Email: hsanthi@vit.ac.in

II. GRAPH METRICS IN SNA

Typically social networks can be considered as unweighted and undirected graphs for analysis and investigative purposes. In these graphs the users represent nodes and the connection between these users is illustrated using edges between the concerned nodes. The resulting graph can be mathematically formulated as an adjacency matrix ($\mathbf{M_{ij}}$) in the following format:

$$M_{ij} = \begin{cases} 1, if\ nodes\ i\ and\ j\ are\ connected \\ 0, otherwise \end{cases}$$

This representation of the social network enables easy calculation and representation of metrics and network structure, mainly clustering coefficient, mixing time, efficiency of the network, Eigen vectors and degree.

Following is an insight into these metrics and their use in SNA:

A. Metrics

SNA Metrics is regarded as among one of the broadest researches in graphical representation of Social network analysis. By knowing relations between various nodes inside a network, we can compute useful information about networks and essentially they can reveal network structures. Having metrics additionally implies we can gather essential information about the system and utilize the information for a few purposes.

1. Centrality:

Centrality gives key insight into the importance of a node in a particular network and has been well explained by the work of Freeman [3]. There are a few basic types of centrality measures used, which are:

a. Degree centrality: Provides information on how well connected a node is in a given network and is represented using:

$$DC_i = \sum_{j=1, j\neq i}^{n} M_{ij}$$

b. Betweenness: Provides numerical data on the extent to which a node intermediates the connection of two other nodes. Gatekeepers and other important mediators in a network are exposed using betweenness:

$$BC_i = \sum_{j}^{n} \sum_{k}^{n} G_{jk}$$

In the above mentioned equation is G_{jk} is an indication if the shortest path between nodes **j** and **k** that passes through node **i**.

c. **Eigen vector centrality:** Extending the idea of centrality to the nodes which are themselves influential (well connected) or are connected to nodes that are influential, makes some nodes important and well connected (indirectly) in the network. The measure of this is done using Eigen vector centrality which is defined to be proportional to the sum of the centrality of the node's neighbors. In the formula, below, n is the total number of nodes and λ is the Eigen constant.

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} G_{jk}\, x_j$$

2. **Clustering coefficient:**
This metric gives an extent to which a node tend to form or be a part of cluster(s). In social networks, there are a lot of such clusters present and hence, clustering coefficient helps in SNA and Investigative Data Mining. It is defined mathematically for a node **V**, as the proportion of the number of links between neighbors of V (represented below as N_V) to the degree of node V.

$$CC_V = \frac{2(N_V)}{DC_V(DC_V - 1)}$$

3. **Page Rank Centrality:**
It is an approach to quantify graph/network centrality in a way similar to degree centrality. It is considered as an improved variant of in-degree centrality, used to measure the impact on different nodes in the system.
PageRank is Google's novel algorithm for inspecting the whole connection structure of the web and figure out which pages are more significant/ relevant for a particular user query.
One of the issues with Katz centrality (another node relation inspection metric) is that just by being connected to nodes with high centrality; some nodes are assigned high centrality too in a network. This issue was handled utilizing the PageRank centrality measure. The PageRank algorithm instead of allotting the entire centrality of a highly central hub just doles a fraction of this centrality to its neighboring nodes, hence, overcoming the drawback of Katz centrality.
The PageRank centrality is represented using the following equation:

$$PR(x_i) = \alpha \sum_k M_{ik} \frac{x_k}{DC_k^{out}} + \beta$$

In the above mentioned equation, **α** is the normal Eigen vector centrality, whereas **β** is the free centrality constant. In addition to this, $\mathbf{x}_i$ and $\mathbf{x}_k$ is the actual node pair for which the PR centrality is being calculated. Note that, for our calculation, α is set to 0.85 as Google also uses the same value (practicality reasons), although other values can also be assumed but it should be ensured that the inverse of the matrix MD^{-1}, should be greater than the chosen value for **α.**

Here, **D** is:

D: a diagonal matrix, for which:
$\mathbf{D}_{ii}$ = max (DC_i^{out}, 1)

β is set to 1 as a constant for positive penalty.
DC_k^{out} (Out degree of node k) is set to 1, when $DC_k^{out} = 0$, since, $M_{ik} = 0$, which makes the influence of non out-degree to zero. It makes it clear, that the relation between the out-degree and centrality is the proportion of centrality derived from neighbor nodes to the centrality divided by their out-degree.

4. **Katz Centrality:**
Katz centrality estimates the degree of impact of a node (user in a social network) in a system. Katz can be viewed as a spinoff of eigenvector centrality. For instance, it tallies the total number of "walks" ending on or beginning from a node. A penalty is charged in case the walks are lengthy. It finds the impact accounting the collective total of walks among a couple of nodes. Katz assesses this relative impact of a hub by estimating the quantity of the adjacent neighboring vertices and the vertices that are associated with the vertex through these neighboring vertices.

As the association between vertices is assessed through the walk or length for the pair of vertex, subsequently length is contrarily identified with the strength of association (strong or on the other hand frail) in the below mentioned equation:

$$KC(x_i) = \alpha \sum_k M_{ik}\, x_k + \beta$$

Here, $\mathbf{x}_i$ and $\mathbf{x}_k$ is the node pair in consideration. **α** is the normal Eigen vector centrality which in this case to converge needs to be less than the reciprocal value of the maximum Eigen value (**λ**) for adjacency matrix **M**. **β** on the other hand helps provide a positive penalty constant, when its value is set to 1, especially in the cases when **α** is zero.

5. **Mixing Time:**
Random walks [6] have a significant property: at the point when the irregular walk approximates its unfaltering state conveyance after an adequate number of jumps, the start-point and end-point of the walk are uncorrelated. This number of jumps is termed as mixing time. The property mentioned above is met quicker if the mixing time is smaller.
The steady state distribution **Ω** illustrates the probability of this random walk reaching **Ω**, after sufficient hops without having been dependent on the origin of the walk. Mathematically,

$SSD(\Omega) = \frac{deg(\Omega)}{2||E||}$; E => the total number of edges in the graph

The mixing time can hence, be enumerated as:

$$T_i(\varepsilon) = \min\{h: \phi_i(h) \leq \phi\}$$

Here, $\phi_i(h)$ is the variation distance between SSD (i) and between random walk distributions R^h (i) after **h** hops.

6. HITS: Hyperlink-Induced Topic Search (HITS) [11] and [12], also known as the Hubs and Authority algorithm is a link analysis algorithm that rates web pages.

This algorithm evaluates a graph and generates two scores for all the nodes in that graph:

a) Authority: Indicates the value of the node (page).
b) **Hubs:** Estimates the value of the links (connections) going out from a node (page).

Hits is an iterative calculation and at every iteration:

a) Authority value of each node is to be updated to the sum of the hub values for every connected node that points to it.

b) The hub values of each node are to be updated to the sum of the authority values of each connected node that it point to. The inference of this is that nodes with high hub score are those connected to authorities in a network.

III. METRICS TO DESTABILIZE SOCIAL NETWORKS

The social networks as a whole can be destabilized using the following metrics:

1. Network Efficiency:

The proficiency **E (G)** of a graph is a measure to evaluate how effectively the hubs of the system communicate and exchange data [7]. To characterize proficiency of a system G, first we compute the shortest path $\mathbf{d_{ij}}$ among $\mathbf{i^{th}}$ and $\mathbf{j^{th}}$ hubs. It is fair to assume that each hub sends data along the system, through its connections. The proficiency in the correspondence between i^{th} hub and j^{th} hub is conversely corresponding to the shortest distance: when there is no way in the graph between i^{th} and j^{th} hubs, we get $d_{ij} = +\infty$ and effectiveness ends up being zero. Let **N** is total nodes in a graph, the efficiency of G can be characterized as:

$$E(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}$$

2. Position Role Index (PRI):

It is a reality that the proficiency of a system in nearness of adherents is low in contrast with their nonappearance in the system. This is on the grounds that they are generally less associated nodes and their essence expands the quantity of less connected nodes in a system, in this manner diminishing its effectiveness.

$$PRI(G) = E(G) - E(G - v_i);\ i = 1....N$$

Here, $(\mathbf{G - v_i})$ demonstrates the network without node $\mathbf{v_i}$. On the off chance that we plot the qualities on the chart, the hubs which are plotted beneath x-pivot are supporters, while the hubs higher than remaining hubs with higher qualities on positive y hub are the watchmen.

IV. APPLICATION

Investigative Data Mining and Privacy Analysis are extensively used for various purposes, one of which is destabilizing terrorist networks. There are already existing methodologies for destabilizing such networks. In the following passages, a very widely used hierarchy algorithm [8] is explained, its issues are also stated along with it and finally, a novel algorithm to tackle the issues of the hierarchy algorithm is proposed using a terrorist network dataset.

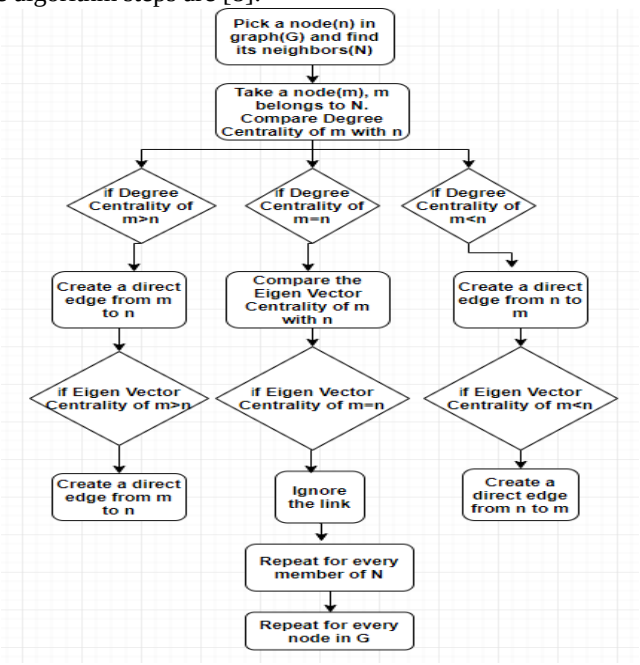
Dependency centrality gives a degree of influence a node has in a graph or in this context, a user has in a social network. This helps in creating a tree relationship between nodes based on their influence in a network.

Although this metric is helpful but there are two very clear issues when we try this for creating such an influence tree:

1. When nodes have the same centrality value, it is a problem to decide which should be parent and which the child node.
2. When more than one node have similar centrality measures, it is difficult to decide which of these nodes will be a direct parent to a node which has less influence on the network compared to these nodes.

The hierarchy algorithm solves the above mentioned issues and for this it follows two sets of algorithms to achieve this, which are:

1. The first part of the hierarchy algorithm derives a directed graph out of an undirected network graph utilizing Eigen vector centralities and degree metrics. The vertex with higher degree initiates connect to a vertex with less degree influence. On the off chance that the influence for two nodes is same then the equivalent judgment is pursued for Eigen vector value for those nodes. Furthermore, and still, after that entire if the influence of nodes is same and the Eigen vector is also same, the connection among these nodes is overlooked. Thus, the algorithm steps are [8]:

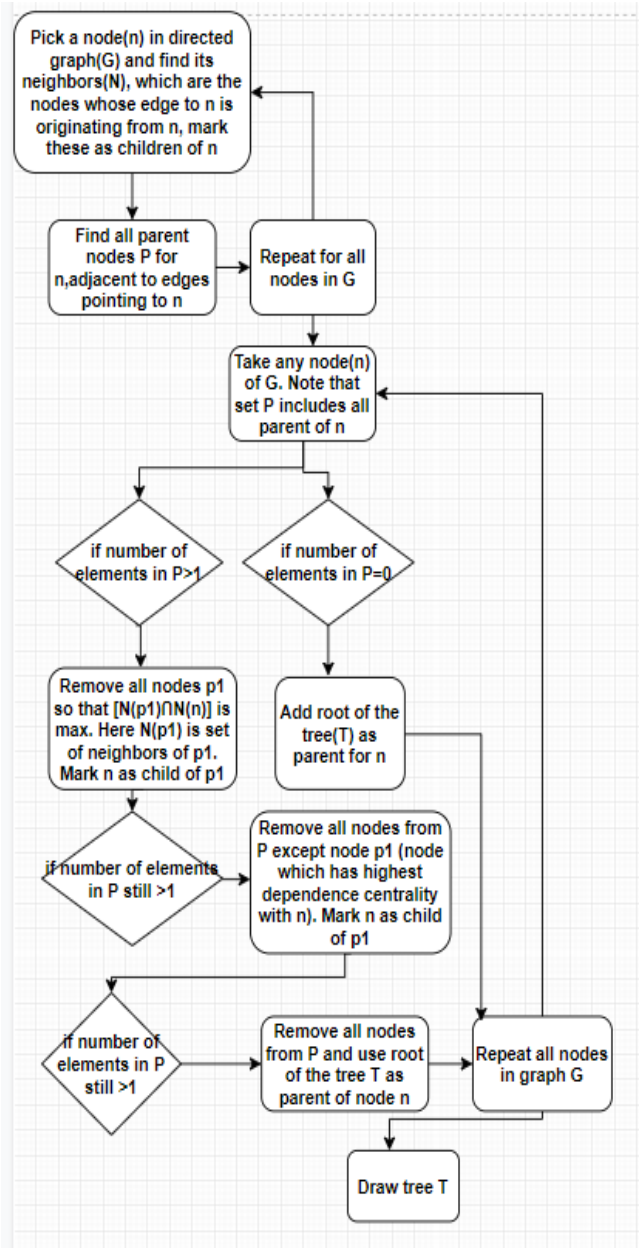


Hierarchy algorithm One

2. The input to the second algorithm is the directed graph constructed using the first hierarchy algorithm and creates the tree structure.

3. In the event that multiple nodes qualify to be the parent of a node, at that point the one with maximum recognized neighbors is considered as a parent for that specific node. This clearly depicts that a genuine influential leader, as for a node, is increasingly compelling on its neighborhood [9].

Subsequent to building the chain of parent-child tree, at last, the most influential node is distinguished from other potential nodes utilizing dependence centrality.



Hierarchy algorithm Two

Drawbacks of the hierarchy algorithm:

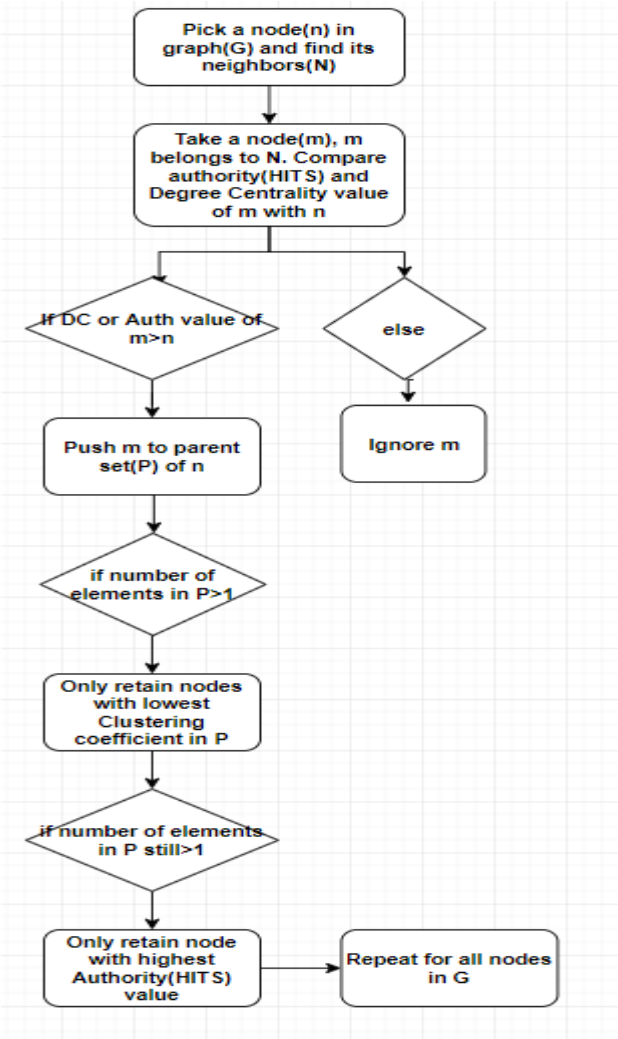
The hierarchy algorithm in itself is capable in detecting the most influential nodes/people in a social network, but the efficiency of these algorithms go for a toss in some cases and this definitely stands as a drawback of this algorithm. Following are the shortcomings of the hierarchy algorithm:

1. The nodes which shouldn't be considered because of their low degree centrality values are always taken into account.
2. Two algorithms need to be run, that too sequentially which makes the hierarchy method inefficient.
3. The transformation of an undirected to a directed graph (the end result of algorithm one) is only utilized to create the parent-child set (input to algorithm two).
4. The number of iterations and calculations to derive the neighbor, parent and children set is all sequential, whereas, it could be parallelized.

Proposed Algorithm

The algorithm proposed in this paper overcomes all the above mentioned drawbacks of the hierarchy algorithm while still delivering the required result of investigative data mining by unfolding the most influential nodes/people in a social network.

The proposed algorithm is as follows:



Proposed Algorithm

V. EXPERIMENT

The application of the novel algorithm proposed in comparison with the existing hierarchy algorithm is worked up from the same publicly available data that Krebs [Krebs 2002] initially used from the Sydney Morning Herald [Sydney 2001]

for the 9/11 terrorist attack in the U.S.A [10], which was extracted just fifteen days after the attack. Krebs uses this data as the start of a social network analysis and the work by Krebs is considered remarkable in terms of Social Network Analysis.
The social network for the 9/11 attacks is:

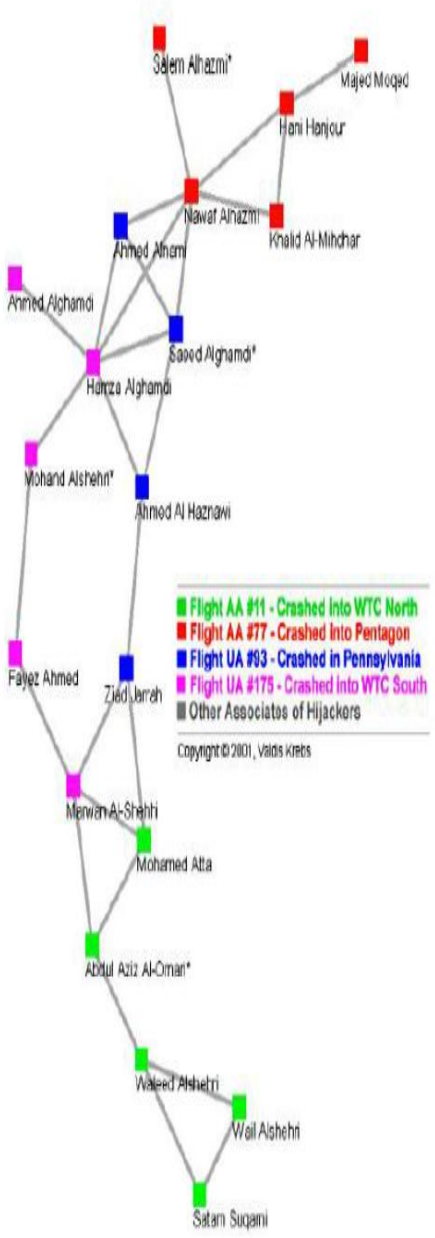


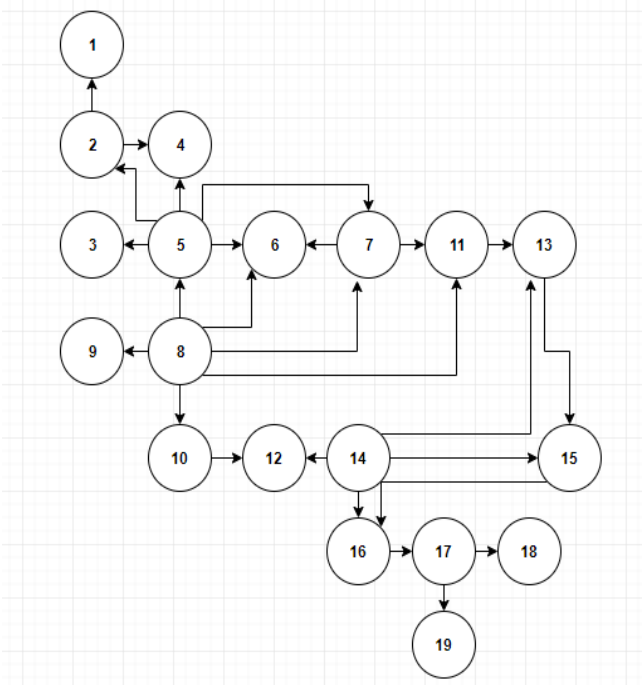
Fig1: Valdis Krebs Social Network of the 9/11 Hijacker Network

The software used to create and analyze this social network is Gephi. All the experimental analysis is performed using Gephi.
The metrics and other information observed and collected for the above mentioned graph using Gephi is as follows:

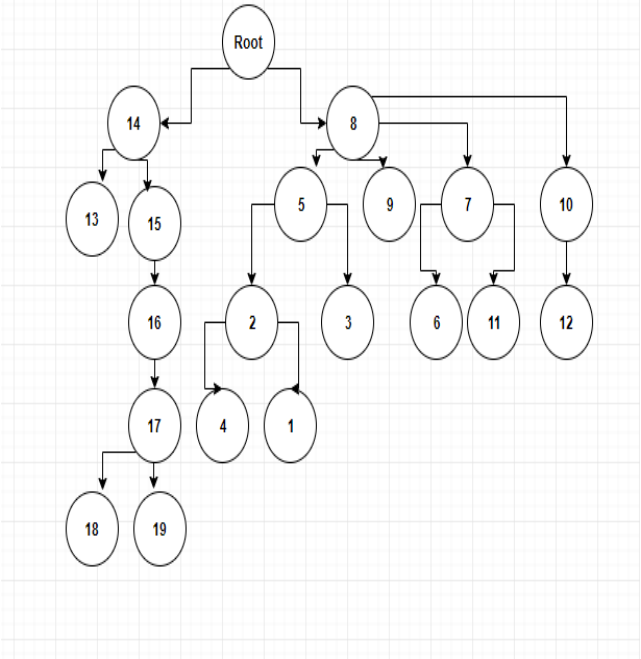
Table 1: Social Network metrics

Sno	Terrorists	DC	Eigen Centrality	PR	CC	Authority/ Hub(HITS)
1	Majed Moqed	1	0.108128	0.025182	0	0.049898
2	Hani Hanjour	3	0.388651	0.061043	0.333333	0.186855
3	Salem Alhazmi	1	0.25581	0.022644	0	0.12644
4	Khalid Al-Mihdhar	2	0.363938	0.039931	1	0.176339
5	Nawaf Alhazmi	6	0.964169	0.10409	0.266667	0.473483
6	Ahmed Alhami	3	0.748116	0.050907	1	0.370287
7	Saeed Algham di	4	0.874512	0.065795	0.666667	0.427641
8	Hamza Alghamdi	6	1	0.100876	0.266667	0.485493
9	Ahmed Algham di	1	0.263756	0.022179	0	0.129648
10	Mohand Alshehri	2	0.325431	0.038874	0	0.144843
11	Ahmed Al Haznawi	3	0.599483	0.050967	0.333333	0.27213
12	Fayez Ahmed	2	0.199852	0.039312	0	0.056903
13	Ziad Jarrah	3	0.364399	0.052293	0.333333	0.105914
14	Marwan Al-Shehhi	4	0.349954	0.070048	0.333333	0.06824
15	Mohamed Atta	3	0.303218	0.053175	0.666667	0.056247
16	Abdul Aziz Al-Omari	3	0.264189	0.055005	0.333333	0.036473
17	Waleed Alshehri	3	0.163446	0.060527	0.333333	0.012093
18	Wail Alshehri	2	0.098231	0.043576	1	0.004406
19	Satam Suqami	2	0.098231	0.043576	1	0.004406

On applying the conventional hierarchy algorithm to the dataset mentioned above fetches the following directed graph using the first part of the hierarchy algorithm (Here the node number represents the terrorist using the same serial number as mentioned above):

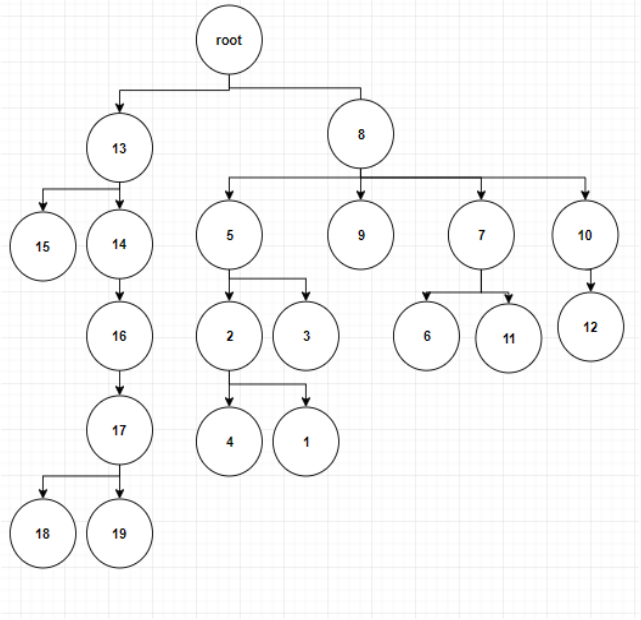


Directed graph derived from undirected graph
After the second part of the hierarchy algorithm is implemented, the following tree is derived:



Hierarchical tree derived from the directed graph

Whereas, on applying the proposed algorithm to the dataset mentioned above fetches the following directed graph. (Here the node number represents the terrorist using the same serial number as mentioned above):



Hierarchical tree derived using the proposed algorithm

RESULTS

From the experiment conducted as part of this paper it can be inferred that the same results as the traditional hierarchy algorithm can be derived from the proposed algorithm. It is clear from both the algorithms that Hamza Alghamdi and Marwan Alshehhi was the most influential terrorists in the tragic terrorist hijack activity of 9/11. The proposed algorithm could clearly achieve what the conventional algorithm achieves in far less steps. The spatial and time complexity is also less compared to the traditional hierarchy algorithm. It is needless to say that Investigative Data Mining when done effectively drastically enables a nation to prevent potential terrorist attacks by identifying malicious people connected.

VI. CONCLUSION AND FUTURE WORK

Social Network Analysis has always been helpful in providing inferences and insights into social groups and has been applied to various uses. The actors and their connections directly form a part of the social graph representing nodes and edges between the nodes, opening an array of helpful information as to how these actors behave in social settings. In this paper, a traditional social network analysis algorithm (hierarchy) is discussed and its drawbacks are also pointed out. A novel algorithm is discussed to overcome the shortcomings of the traditional hierarchy algorithm which is also used to gauge the influence of nodes in complex social network setups. This algorithm is an improvement in both spatial and time complexity. The algorithm proposed is applied in destabilizing a terrorist network applying the principles of Investigative Data Mining. Furthermore, the paper also discussed major graph theory metrics that are applied to social networks for analyzing its complexity and classifying nodes based on its strengths and weakness (based on connections).

There is still a scope for further reducing the spatial complexity of the proposed algorithm by eliminating least influential nodes in the initial phase of tree creation. In addition to this, more advanced metrics could be discovered which further convey the influence of nodes which have similar connection strengths and weaknesses to better categorize into child/parent buckets and resolve conflicts in the parent buckets.

REFERENCES

1. Scott, J.: Social Network Analysis: A Handbook, 2 edn. Sage Publications, London 2000
2. Christian Hirschi: “Introduction: Applications of Social Network Analysis” in 6th Conference on Applications of Social Network Analysis, ETH Zurich, Institute for Environmental Decisions, CHN K 76.2, Universitätstrasse 22, 8092 Zurich, Switzerland.
3. Freeman, L.C. Centrality in Social Networks: I. Conceptual clarification. Social Networks, 1:215-39 (1978).
4. Page Rank. [Online] Available:
a. https://en.wikipedia.org/wiki/PageRank
5. Privacy concerns with social networking services [Online] Available: https://en.wikipedia.org/wiki/Privacy_concerns_with_social_networking_services.
6. M. Mitzenmacher and E. Upfal, Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press, January 2005.
7. Latora, V., Massimo Marchiori, How Science of Complex Networks can help in developing Strategy against Terrorism, Chaos, Solitons and Fractals 20, 69-75 (2004).
8. Nasrullah Memon, Henrik Legind Larsen: Practical Approaches for Analysis, Visualization and Destabilizing Terrorist Networks, In: Proceedings of the First International Conference on Availability, Reliability and Security, ARES (2006).
9. Nasrullah Memon, Henrik Legind Larsen, David L. Hicks, and Nicholas Harkiolakis: 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).
10. Uncloaking Terrorist Networks [Online] Available: https://firstmonday.org/ojs/index.php/fm/article/view/941/863
11. HITS Algorithm [Online] Available:
a. http://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html.
12. HITS Algorithm, Wikipedia information [Online] Available: https://en.wikipedia.org/wiki/HITS_algorithm
13. S. Buchegger, D. Schi¨oberg, L. H. Vu, and A. Datta, “PeerSoN: P2P Social Networking,” in Social Network Systems, 2009.

14. Watts, D.J.: Six Degrees – The Science of a Connected Age. W.W. Norton & Company, New York, 2003.
15. A. Clauset, M. Newman, C. Moore. Finding Community Structure in Very Large Networks, Physical Review E70, 066111 (2004).
16. U Kang, Spiros Papadimitriou, Jimeng Sun, Hanghang Tong: Centralities in Large Networks: Algorithms and Observations, In: SIAM International Conference on Data Mining (SDM'2011), Phoenix, U.S.A. (2011).
17. Memon, N., Larsen H.L.: Investigative Data Mining Toolkit: A Software Prototype for Visualizing, Analyzing and Destabilizing Terrorist Networks. In: Visualizing Network Information, pp. 14-1 – 14-24 (2006).
18. Sarita Azad and Arvind Gupta: A Quantitative Assessment on 26/11 Mumbai Attack using Social Network Analysis, Journal of Terrorism Research, Volume 2, Issue 2 (2011).
19. M. Burgess, G. Canright, and K. Engø. 2003. A graph theoretical model of computer security: from file access to social engineering. International Journal of Information Security.
20. Carley M. Kathleen, Lee Ju-Sung, David Krackhardt. 2002. Destabilizing Networks. Connections 24 (3) 79-92.
21. U Kang, Spiros Papadimitriou, Jimeng Sun, Hanghang Tong: Centralities in Large Networks: Algorithms and Observations, In: SIAM International Conference on Data Mining (SDM'2011), Phoenix, U.S.A. (2011).
22. Sparrow, M. K. 1991. The application of network analysis to Criminal intelligence: An assessment of the prospects. Social Networks. 13, 251-274.