# Using the Spanning Tree of a Criminal Network for Identifying Its Leaders

2017-10-16

*Presented by*
*Shailendra Rathore*
*(rathoreshailendra@seoultech.ac.kr)*

# Abstract


- Introduce a forensic analysis system called ECLfinder that identifies the influential members of a criminal organization as well as the immediate leaders of a given list of lower-level criminals.
- Criminal investigators usually seek to identify the influential members of criminal organizations, because eliminating them is most likely to hinder and disrupt the operations of these organizations and put them out of business.

# 1. Introduction

- SOCIAL groups and their relationships have long been identified using Social network analysis (SNA).
- Inspired by SNA, researchers in digital forensic investigation have been employing similar network analysis techniques for identifying criminal communities, their relationships, and their influential leaders.
- Recently, forensic investigators have shown a growing interest on using Mobile Communication Data (MCD) that belong criminal organizations to construct networks that depict the organizations and analyze these networks [12].
- Criminal forensic investigators have also shown interest on constructing networks from Crime Incident Reports that contain information about a criminal organization.
- ECLfinder (Efficient Criminal Leaders Finder) can identify the most influential members of a criminal organization.
- In the framework of ECLfinder, a network can be constructed from either Mobile Communication Data (MCD) that belongs to a criminal organization or from crime incident reports that contain information about a criminal organization.

# 2. Background and outline of the approach

*A. Background*

- For identifying the vertices that are important to query vertices, Existing methods suffer incomplete contribution and inconsistent contribution.
- Incomplete contribution occurs, if some query vertices do not contribute to the overall relative importance value of a vertex.
- The inconsistent contribution occurs, if query vertices contribute unequally to the overall relative importance value of a vertex.
- Let *v* be the current vertex under consideration.
- ECLfinder overcomes the problem of Incomplete Contribution by: (1) considering the importance of *each* query vertex to *v*, and (2) assigning a weight to each incoming edge to *v* that is outgoing from one of the query vertices (this weight represents the importance/rank of this vertex relative to all incoming edges to *v)*.
- ECLfinder overcomes the problem of Inconsistent Contribution by: (1) considering the importance of *each* query vertex to each vertex connected to *v*, and (2) accounting for the degree of relativity of *v* to all query vertices.

# 2. Background and outline of the approach

*B. Outline of the Approach*

1) Constructing a Network: constructed from either MCD or crime incident reports
2) Assigning a Weight to Each Edge in the Network: In a network constructed from MCD, the weight of an edge represents the number of phone calls/messages between two criminals. In a network constructed from crime incident reports, the weight of an edge represents the number of co-occurrences of the names of suspects and accomplices in the same reports.

3) Computing the Shortest-Path Edge Betweenness: computed by replacing edges' initial weights by their shortest-path betweenness.

4) Assigning a Score to Each Edge: Edges' shortest-path betweenness are replaced by their inverses.

5) Assigning a Score to Each Vertex in the Network Based on the Concept of Existence Dependency:

6) Identifying the Influential Members of the Criminal Organization: Criminals represented by the top ranked vertices are considered the influential members of the criminal organization.

# 3. Constructing a network

- A network can be constructed from information gathered from MCD associated with a criminal organization.
- A vertex in such a network represents a criminal caller and/or receiver. An edge represents the flow of communications between two criminals, through phone calls or messages.
- The weight of an edge represents the number of phone calls/messages between the two criminals represented by the two vertices connected by the edge.
- A network can also be created from crime incident reports that contain information about the members of a criminal organization.
- A vertex represents a criminal.
- An edge represents the relationship between two criminals, determined based on the co-occurrences of the criminals' names in the same crime incident report.
- Employs the concept of space approach [5] to construct networks automatically from crime incident reports [6].
- Employs the techniques of Stanford Named Entity Recognition [17] to determine the names of people in reports. It uses a tokenizer and stemmer to match a sequence of words against persons' names.
- Shortest-path edge betweenness is computed by using the method of Girvan–Newman [11].
- Replace edges' initial weights by their shortest-path betweenness.
- A *score* is assigned to each edge. The score of an edge is the inverse of the edge's shortest-path betweenness weight.

# 4. Identifying the influential members of a criminal organization

*A. Assigning a Score to Each Vertex in the Network Based on the Concept of Existence Dependency:*

- Construct the Minimum Spanning Tree (MST) of the network based on the edges' scores

***Algorithm CONSTRUCT-MST (NW, S, r)***

1. **for** each $u \in V[NW]$
2. **do** $key[u] \leftarrow \infty$
3. $\pi[u] \leftarrow$ NIL
4. $key[r] \leftarrow 0$
5. $Q \leftarrow V[NW]$
6. **while** $Q \neq \emptyset$
7. **do** $u \leftarrow EXTRACT - MIN(Q)$
8. **for** each $v \in Adj[u]$
9. **do if** $v \in Q$ and $S(u,v) < key[v]$
10. **then** $\pi[v] \leftarrow u$
11. $key[v] \leftarrow S(u, v)$
12. $MST \leftarrow (v, \pi[v])$

Fig. 1. Algorithm *CONSTR-MST*

# 4. Identifying the influential members of a criminal organization

*A. Assigning a Score to Each Vertex in the Network Based on the Concept of Existence Dependency:*

- ECLfinder identifies for each vertex $v$, the set $S$ of vertices, whose existence in MST is dependent on $v$.
- The removal of $v$ causes each vertex $u \in S$ to be unable to reach *each* other vertex through the paths of the MST.
- Finally, ECLfinder assigns a score to each vertex $v$ in the network. The score of the vertex $v$ is the number of other vertices, whose existence in MST is dependent on $v$.
- The score reflects the relative rank/importance of the criminal represented by the vertex $v$ in the criminal organization.

# 4. Identifying the influential members of a criminal organization

*B. Identifying the Central*

- ECLfinder can also identify the immediate leaders of a given list of lower-level criminals.
- To find immediate leader it uses the term "*query vertices*" to refer to a given list of vertices representing lower-level criminals.
- Let q1, q2, …qn denote a list of query vertices.
- A criminal represented by a vertex *v* in a network is considered an immediate leader of the criminals represented by q1, …, qn, if:
- (1) *v* has the highest score among the vertices located at the convergences of the subtrees of the MST that pass through q1, …, qn,
- (2) the existence of each of q1,…, qn in the MST is dependent on *v*.

# 5. Case studies

- A partial snapshot of Friendster social network.
- A vertex in the network represents a user.
- An edge represents a relationship between two users.
- The score of an edge is the inverse of the shortest-path betweenness of the edge
- The bold/thick edges show the path of the Minimum Spanning Tree (MST) of the network.

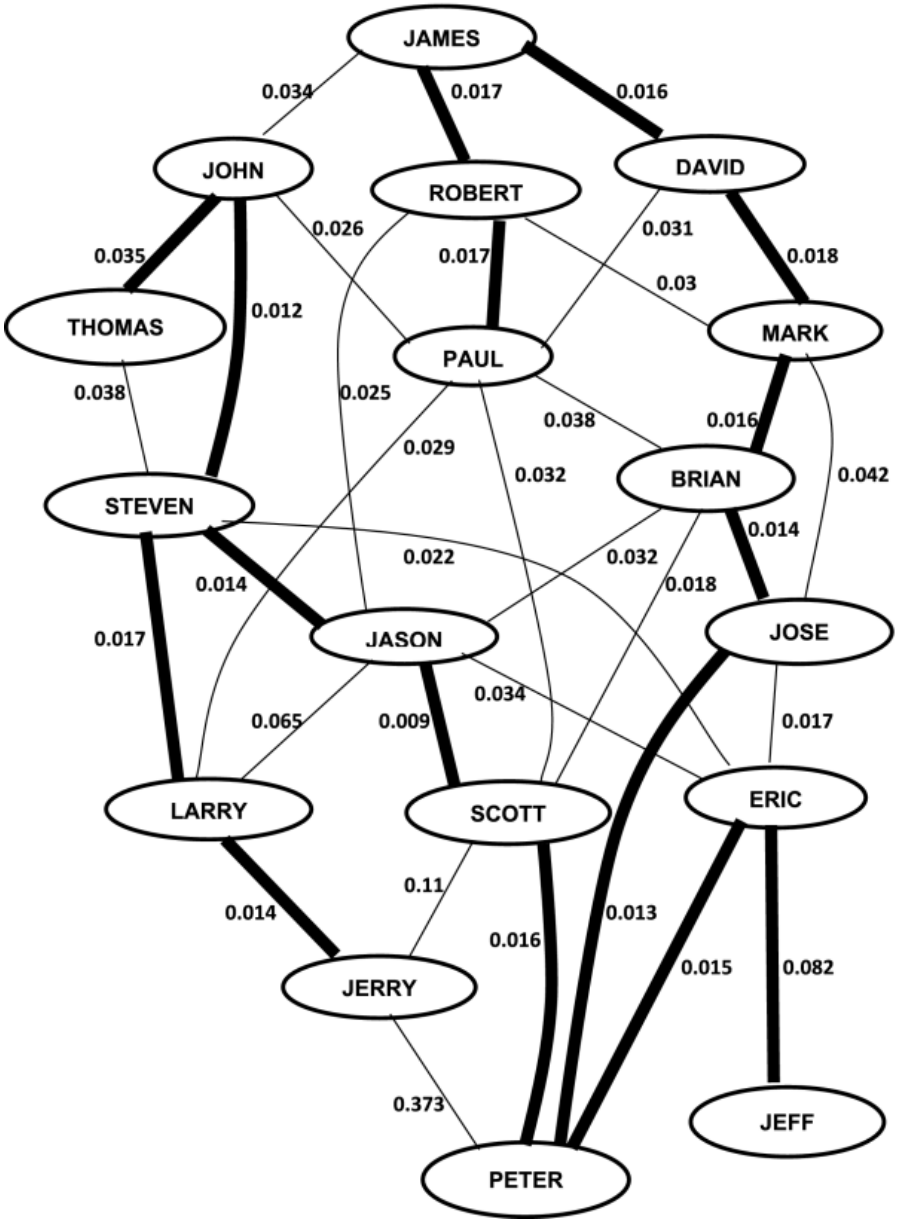


Fig. 2. A partial snapshot of Friendster social network [28]. The score of an edge is the inverse of the shortest-path betweenness of the edge. The bold/thick edges are the paths of the MST of the network.

# 5. Case studies

- Assigning a score to each vertex in the network
- How the scores of some selected vertices in the network in Fig. 3,
- The score of vertex STEVEN is 4, because the following four vertices are existence dependent on STEVEN in MST: THOMAS, JOHN, LARRY, and JERRY.
- The score of vertex PETER is 9. This is because the removal of vertex PETER will cause the following 9 vertices to be unable to reach each of the remaining vertices connected with the root vertex through the paths of the MST: ERIC, JEFF, SCOTT, JASON, STEVEN, JOHN, THOMAS, LARRY, and JERRY.

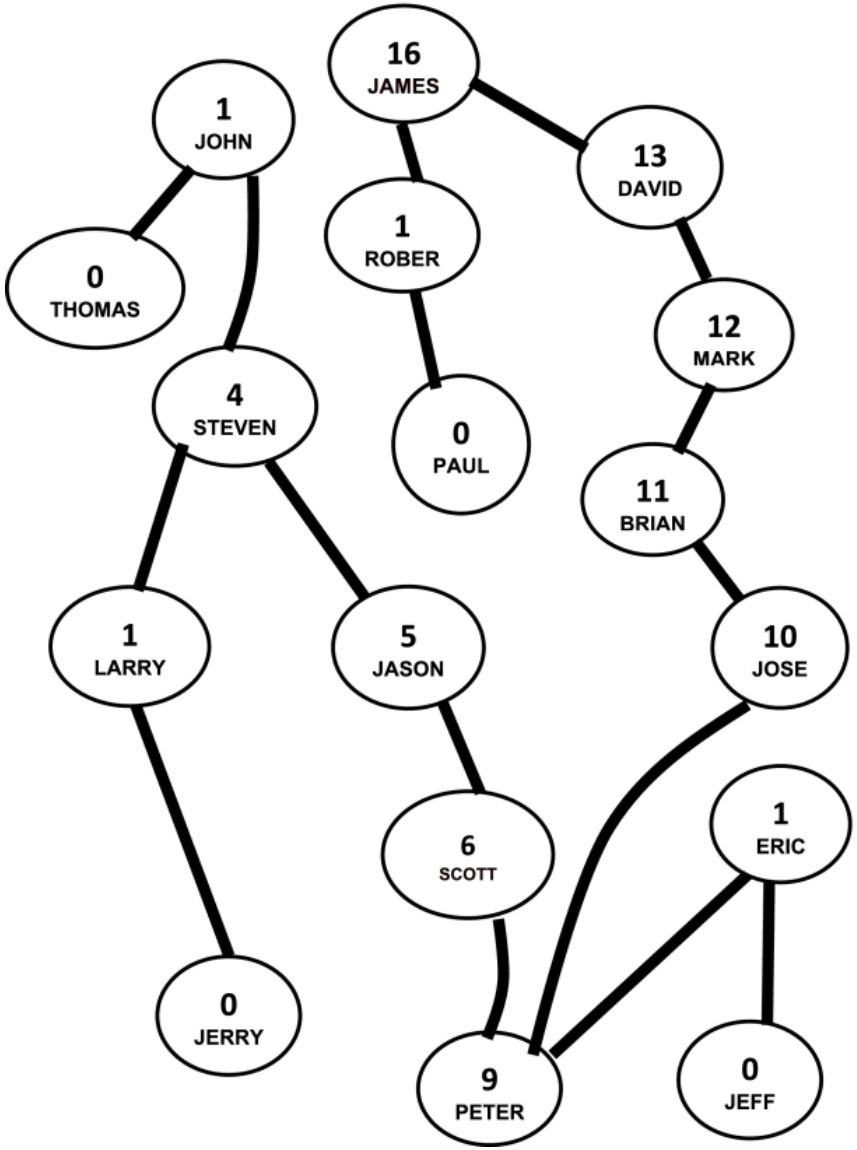


Fig. 3. The partial social network presented in Fig. 2 after assigning a score to each vertex.

# 5. Case studies

- Ranked based on their scores shown in fig. 3.
- The top ranked users in the table are the influential ones in the social network.

| Rank | Score | Criminal Name |
|---|---|---|
| 1 | 16 | JAMES |
| 2 | 13 | DAVID |
| 3 | 12 | MARK |
| 4 | 11 | BRIAN |
| 5 | 10 | JOSE |
| 6 | 9 | PETER |
| 7 | 6 | SCOTT |
| 8 | 5 | JASON |
| 9 | 4 | STEVEN |
| 10 | 1 | JOHN, ROBERT, LARRY, ERIC |
| 14 | 0 | THOMAS, PAUL, JERRY, JEFF |

Table 1: The 17 users represented by the 17 vertices in the partial network shown in fig. 2

# 5. Case studies

- Consider Fig. 3 and the following query: *Q*(“THOMAS”, “LARRY”).
- The query asks for the immediate leader of THOMAS and LARRY.
- As Fig. 4 shows, STEVEN is the immediate leader, because of the following:
- (1) vertex STEVEN is located at the convergence of the subtrees of the MST that passes through vertices THOMAS and LARRY
- (2) the existence of vertices THOMAS and LARRY in the MST is dependent on vertex STEVEN *(the removal of vertex STEVEN will cause the two vertices to be unable to reach each of the vertices in the other subtree containing the root vertex).*

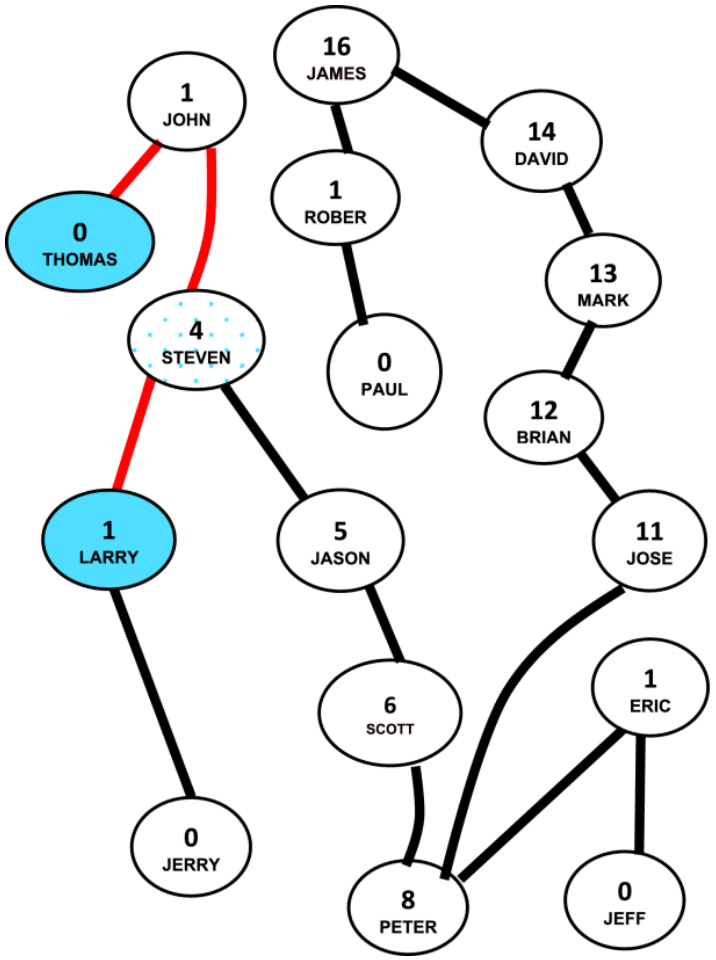


Fig. 4. The red paths show that vertex STEVEN is located at the convergence of the subtree of the MST that passes through vertices THOMAS and LARRY.

# 5. Case studies

- Consider Fig. 3 and the query: *Q*(“JERRY”, “ERIC”).
- The query asks for the immediate leader of JERRY and ERIC.
- As Fig. 5 shows, PETER is the immediate leader.

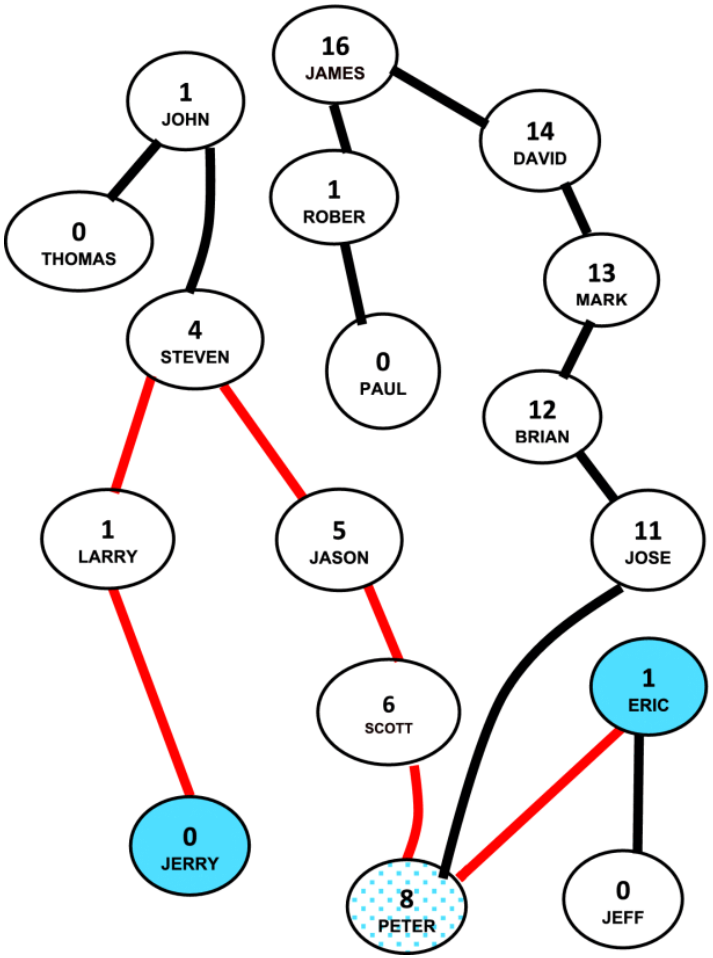


Fig. 5. The red paths show that vertex PETER is located at the convergence of the subtree of the MST that passes through vertices JERRY and ERIC.

# 6. Experimental results

- ECLfinder in Java, run on Intel(R) Core(TM) i7 processor, with a CPU of 2.70 GHz and 16 GB of RAM, under Windows 10.
- *LogAnalysis [8]:* employs Girvan & Newman [11] algorithms to identify the degree of relationships between vertices representing criminals in a criminal network.

- *CrimeNet Explorer [12]:* It uses hierarchical clustering techniques to construct network. It employs the Closeness, Degree, and Betweenness centrality metrics to determine the important vertices in a subnetwork.

- *SIIMCO [19]*: Uses formulas that quantify the degree of influences of a vertex.

- One of the key differences between ECLfinder and SIIMCO is that SIIMCO adopts vertex-centric approach while ECLfinder adopts edgecentric approach.
- In SIIMCO, the importance of a vertex $v$ is determined based on the importance of the vertices connecting $v$ with the network. In ECLfinder, the importance of a vertex $v$ is determined based on the importance of the edges connecting $v$ with the network, using the concept of *existence dependency*.

# 6. Experimental results

*A. Compiling Datasets for the Evaluation*

- ***Krebs's 9/11 dataset [26], [27]:*** Dataset of the 9/11 incident. The 9/11 were a series of four coordinated terrorist attacks on the United States on the morning of September 11, 2001.
- The network consists of 62 nodes representing all individuals involved in the incident. The network contains 153 edges.
- The average node degree in the network is 4.9.
- ***Enron email dataset [9]:*** *A* criminal scandal involved top Enron employees.
- Dataset includes 200,136 emails from 151 Enron employees.

*B. Evaluating the Accuracy of Identifying the Influential Members of a Criminal Organization*

$$Recall = \frac{N_s^c}{N_m^{top}}, \qquad F - value = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

$$Precision = \frac{N_s^c}{N_s^{top}}$$

where $N_s^c$ *is* is the number of *correct* vertices returned by a system, $N_m^{top}$ is the number of *actual correct* vertices, and $N_s^{top}$ is the number of vertices returned by a system. Let $L_{top}$ be the list of top vertices returned by a standard network metric and let $L_s$ be the list of correct vertices returned by a system.

# 6. Experimental results

- Degree is the number of ties that a vertex has.
- Vertices with high degree centralities are central in the network.
- The betweenness centrality of a vertex *v* is the number of shortest paths between other vertices that pass through *v*.
- Closeness centrality is the length of the shortest path to all other vertices. It measures how a vertex is close to other vertices.

# 6. Experimental results

TABLE 2 PERFORMANCE OF THE SYSTEMS USING THE 9/11 DATASET COMPUTED BASED ON THE TOP VERTICES RETURNED BY THE STANDARD NETWORK METRICS

| | | Recall | Precision | F-value |
|---|---|---|---|---|
| ECLfinder | Closeness Centrality | 0.66 | 0.61 | 0.63 |
| SIIMCO | | 0.62 | 0.55 | 0.58 |
| CrimeNet Explorer | | 0.54 | 0.58 | 0.56 |
| LogAnalysis | | 0.51 | 0.49 | 0.50 |
| ECLfinder | Betweenness Centrality | 0.59 | 0.57 | 0.58 |
| SIIMCO | | 0.55 | 0.50 | 0.52 |
| CrimeNet Explorer | | 0.49 | 0.43 | 0.46 |
| LogAnalysis | | 0.39 | 0.43 | 0.41 |
| ECLfinder | In Degree Centrality | 0.66 | 0.68 | 0.67 |
| SIIMCO | | 0.64 | 0.59 | 0.61 |
| CrimeNet Explorer | | 0.52 | 0.46 | 0.49 |
| LogAnalysis | | 0.52 | 0.54 | 0.53 |
| ECLfinder | Out Degree Centrality | 0.71 | 0.67 | 0.69 |
| SIIMCO | | 0.69 | 0.55 | 0.61 |
| CrimeNet Explorer | | 0.57 | 0.51 | 0.54 |
| LogAnalysis | | 0.66 | 0.61 | 0.63 |

# 6. Experimental results

TABLE 3 PERFORMANCE OF THE SYSTEMS USING ENRON DATASET COMPUTED BASED ON THE TOP VERTICES RETURNED BY THE STANDARD NETWORK METRICS

| | | Recall | Precision | F-value |
|---|---|---|---|---|
| ECLfinder | Closeness Centrality | 0.58 | 0.50 | 0.54 |
| SIIMCO | | 0.52 | 0.46 | 0.49 |
| CrimeNet Explorer | | 0.37 | 0.30 | 0.33 |
| LogAnalysis | | 0.40 | 0.34 | 0.37 |
| ECLfinder | Betweenness Centrality | 0.44 | 0.37 | 0.40 |
| SIIMCO | | 0.46 | 0.39 | 0.42 |
| CrimeNet Explorer | | 0.34 | 0.26 | 0.29 |
| LogAnalysis | | 0.44 | 0.39 | 0.41 |
| ECLfinder | In Degree Centrality | 0.69 | 0.67 | 0.68 |
| SIIMCO | | 0.64 | 0.61 | 0.62 |
| CrimeNet Explorer | | 0.40 | 0.34 | 0.37 |
| LogAnalysis | | 0.58 | 0.56 | 0.57 |
| ECLfinder | Out Degree Centrality | 0.65 | 0.59 | 0.62 |
| SIIMCO | | 0.61 | 0.52 | 0.56 |
| CrimeNet Explorer | | 0.49 | 0.44 | 0.46 |
| LogAnalysis | | 0.45 | 0.38 | 0.41 |

# 6. Experimental results

*2) Calculating the Euclidean Distances Between the Results of Each System and the Results of the Network Metrics:*

$$d(\sigma_m, \sigma_s) = \sum_{x \in |N_m^{top}} |\sigma_m(v) - \sigma_s(v)|$$

- $N_m^{top}$ are the top *n* vertices returned by network metric *m*.
- Considered *n* equals 5, 10, and 15.
- $\sigma_m \in [0,1]^{|N_m^{top}|}$ and $\sigma_s \in [0,1]^{|N_m^{top}|}$ are the top *ranked n* vertices returned by metric *m* and system *s*, respectively.
- $\sigma_m(v), \sigma_s(v)$ are the position of vertex $v \in N_m^{top}$ in the lists $\sigma_m, and\ \sigma_s$ respectively.
- Fig. 6 shows the average Euclidean Distances using the Krebs's 9/11 and Enron datasets.

# 6. Experimental results

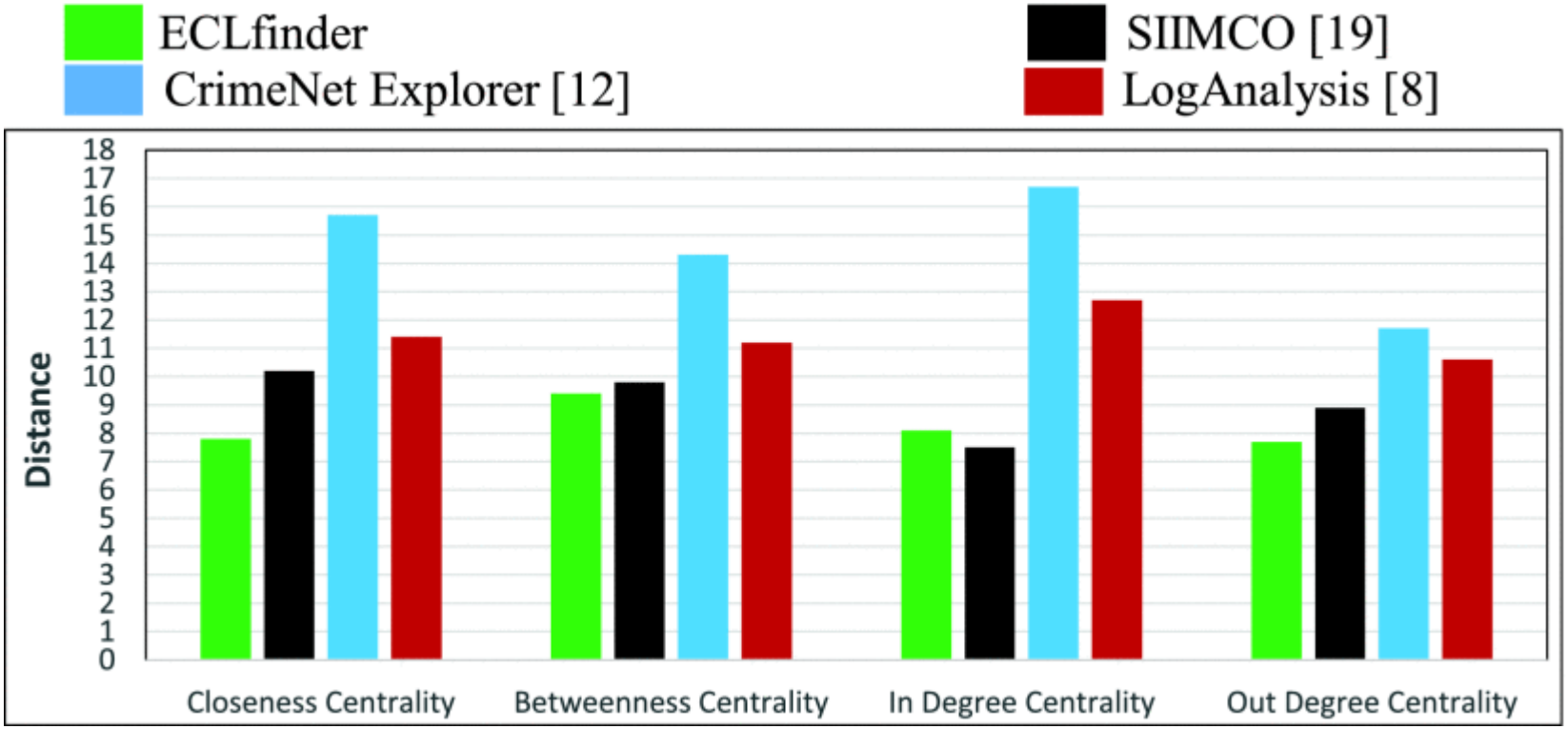


Fig. 6. Average Euclidean distances between the results returned by each of the four systems and the results returned by the standard network metrics using Krebs's 9/11 and Enron datasets.

# 6. Experimental results

*3) Evaluating the Accuracy of Identifying the Immediate Leaders of Lower Level Criminals in a Criminal Organization.*

- Randomly selected 50 lists of 2-query vertices, 50 lists of 3-query vertices, and 50 lists of 4-query vertices from each of the two networks.

# 6. Experimental results

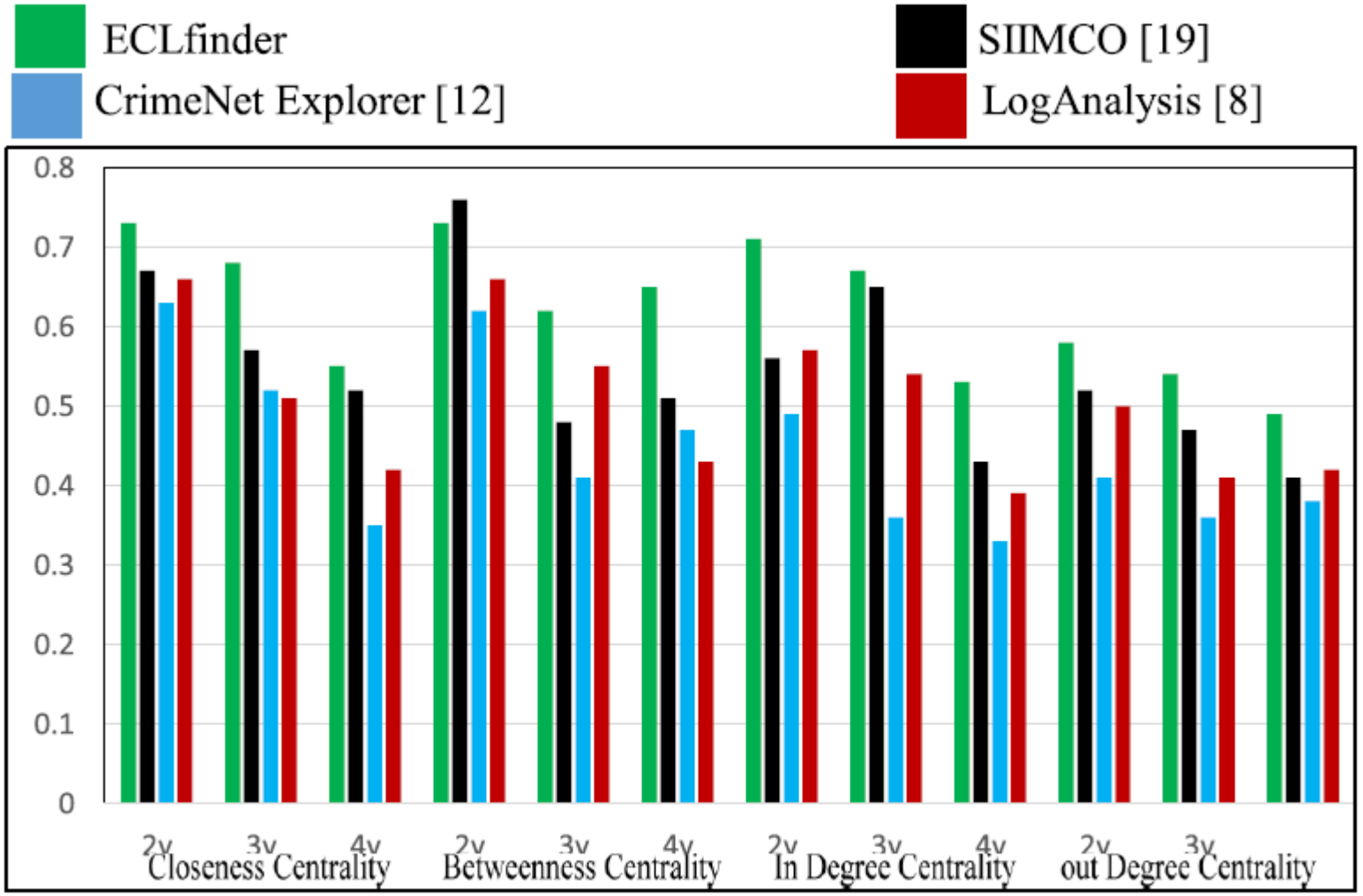


Fig. 7. (a) Recall of the four systems for identifying the important vertices to a given list of query vertices using the Krebs's 9/11 dataset. In the figure, 2v, 3v, and 4v denote the following: 2 query vertices, 3 query vertices, and 4 query vertices respectively.

# 6. Experimental results

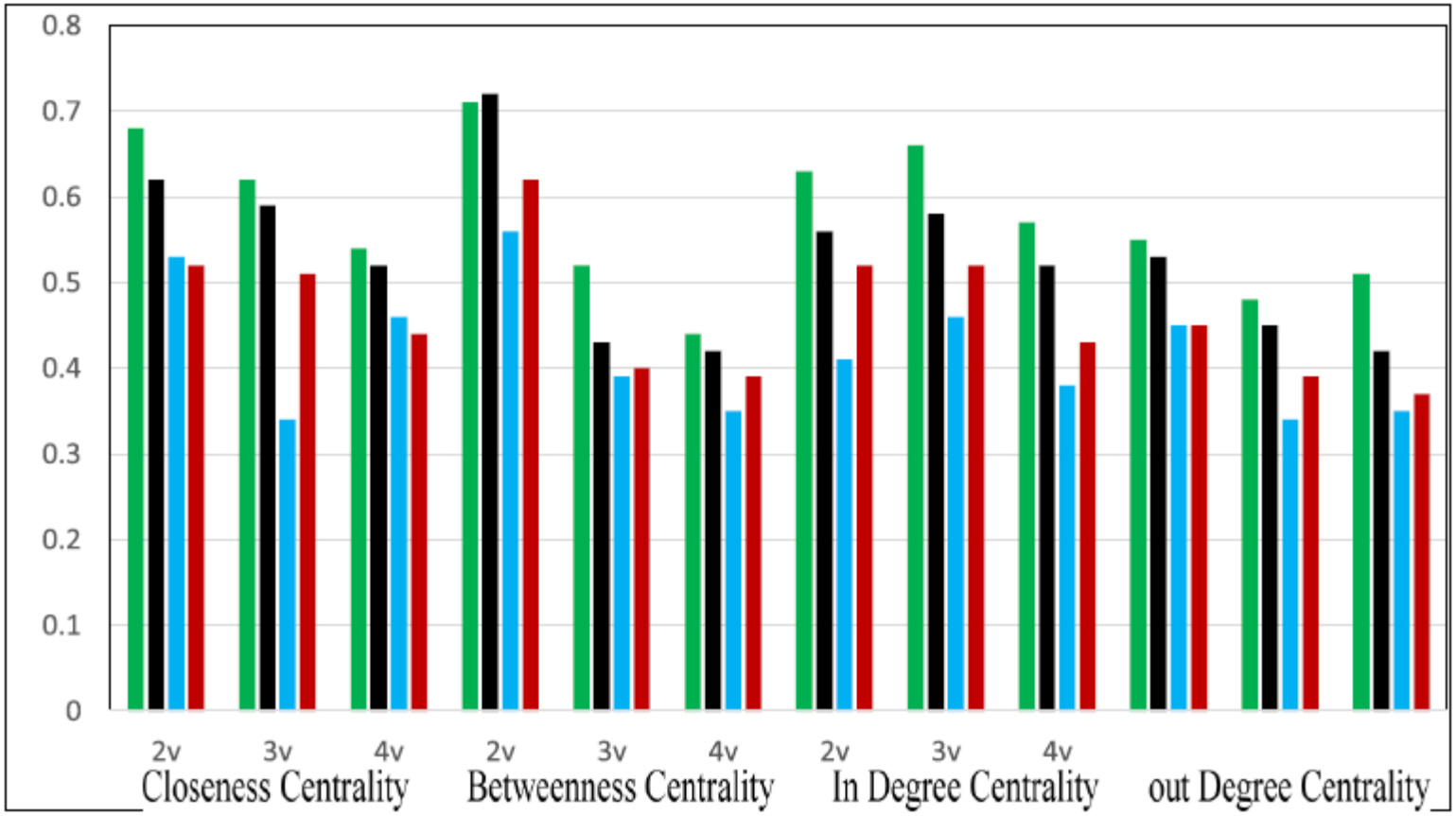


Fig. 7. (b) Precision of the four systems for identifying the important vertices to a given list of query vertices using the Krebs's 9/11 dataset. In the figure, 2v, 3v, and 4v denote the following: 2 query vertices, 3 query vertices, and 4 query vertices respectively.

# 6. Experimental results

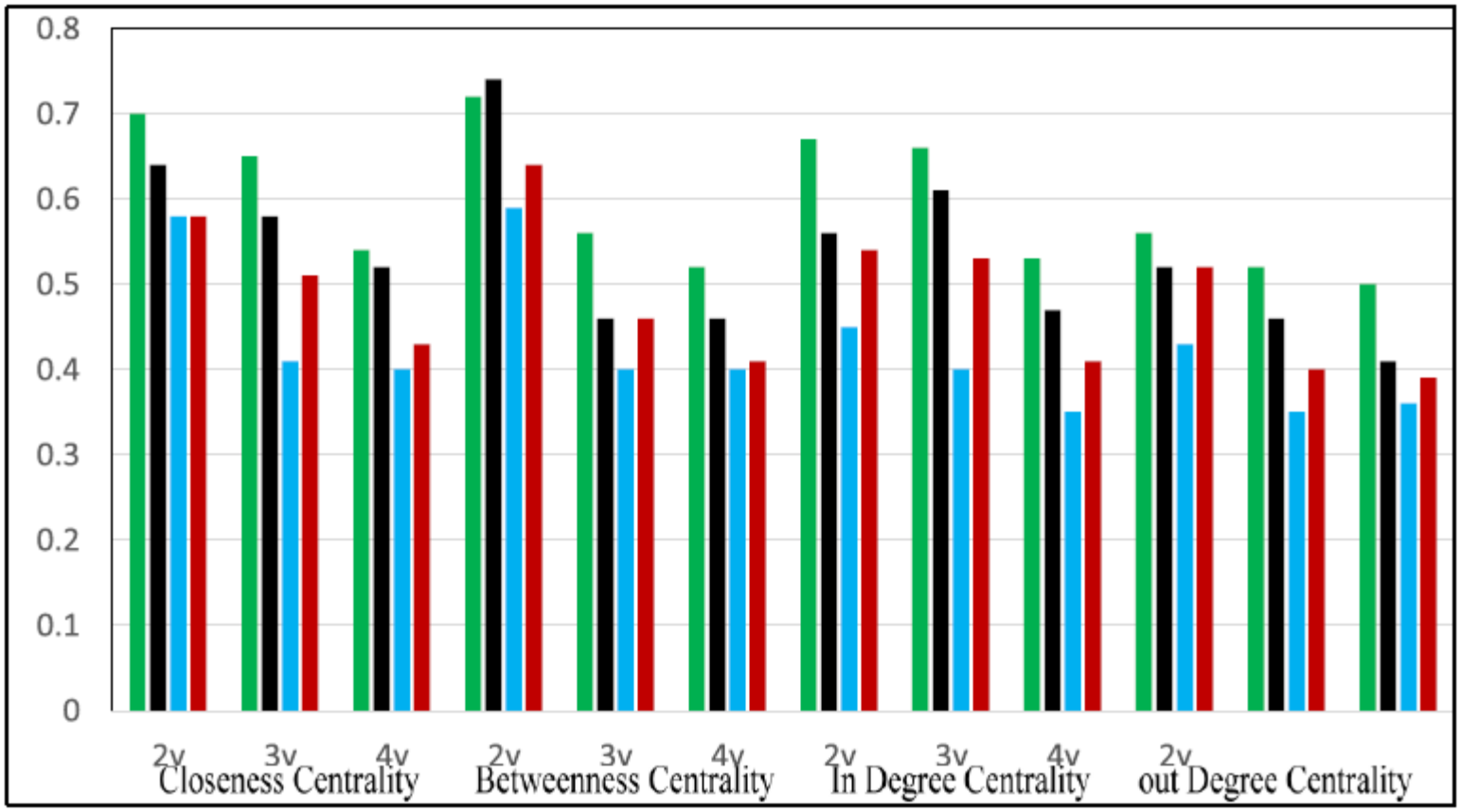


Fig. 7. (c) F-value of the four systems for identifying the important vertices to a given list of query vertices using the Krebs's 9/11 dataset. In the figure, 2v, 3v, and 4v denote the following: 2 query vertices, 3 query vertices, and 4 query vertices respectively.

# 6. Experimental results

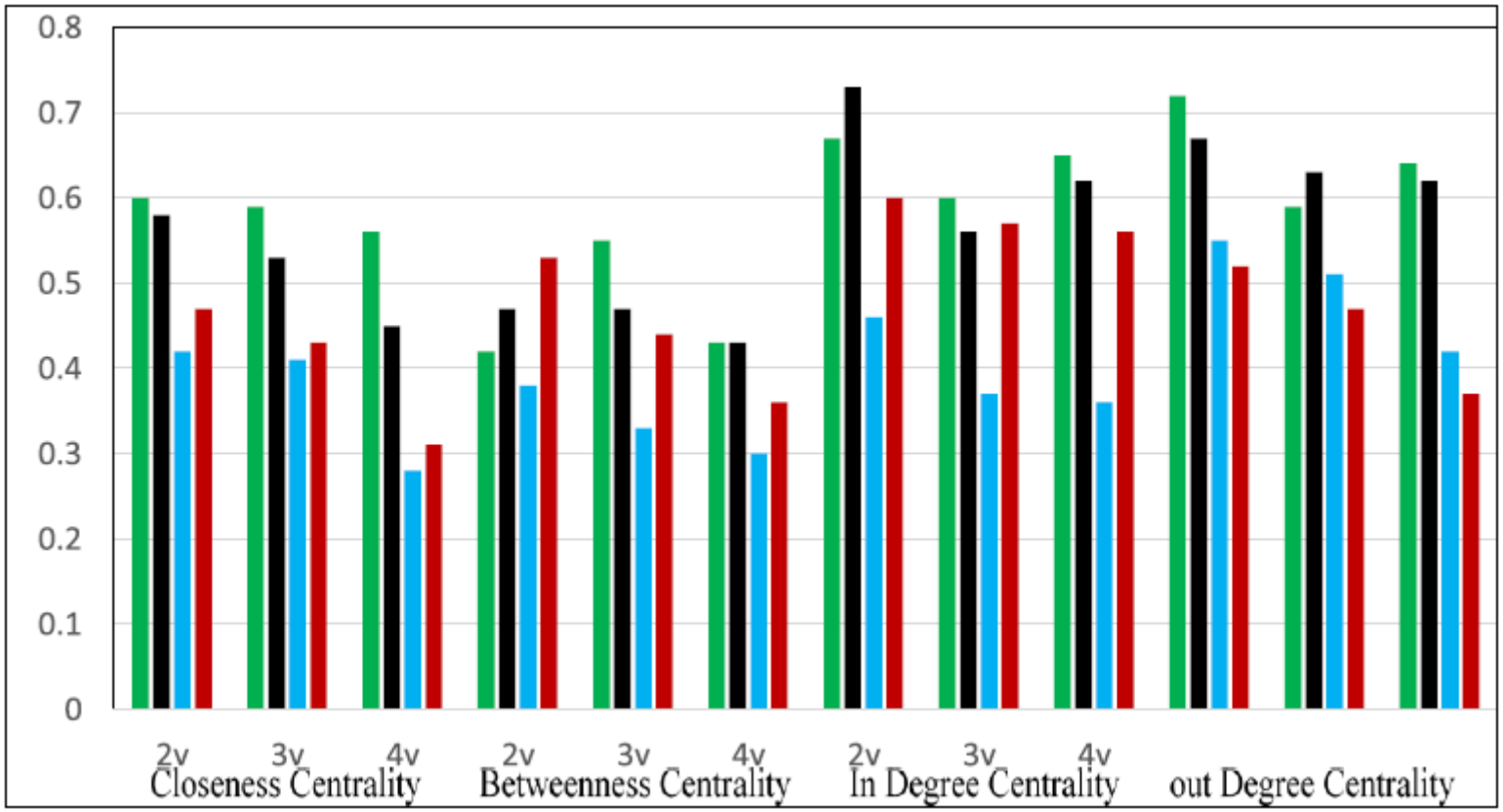


Fig. 8. (a) Recall of the four systems for identifying the important vertices to a given list of query vertices using the Enron dataset. In the figure, 2v, 3v, and 4v denote the following: 2 query vertices, 3 query vertices, and 4 query vertices respectively.

# 6. Experimental results

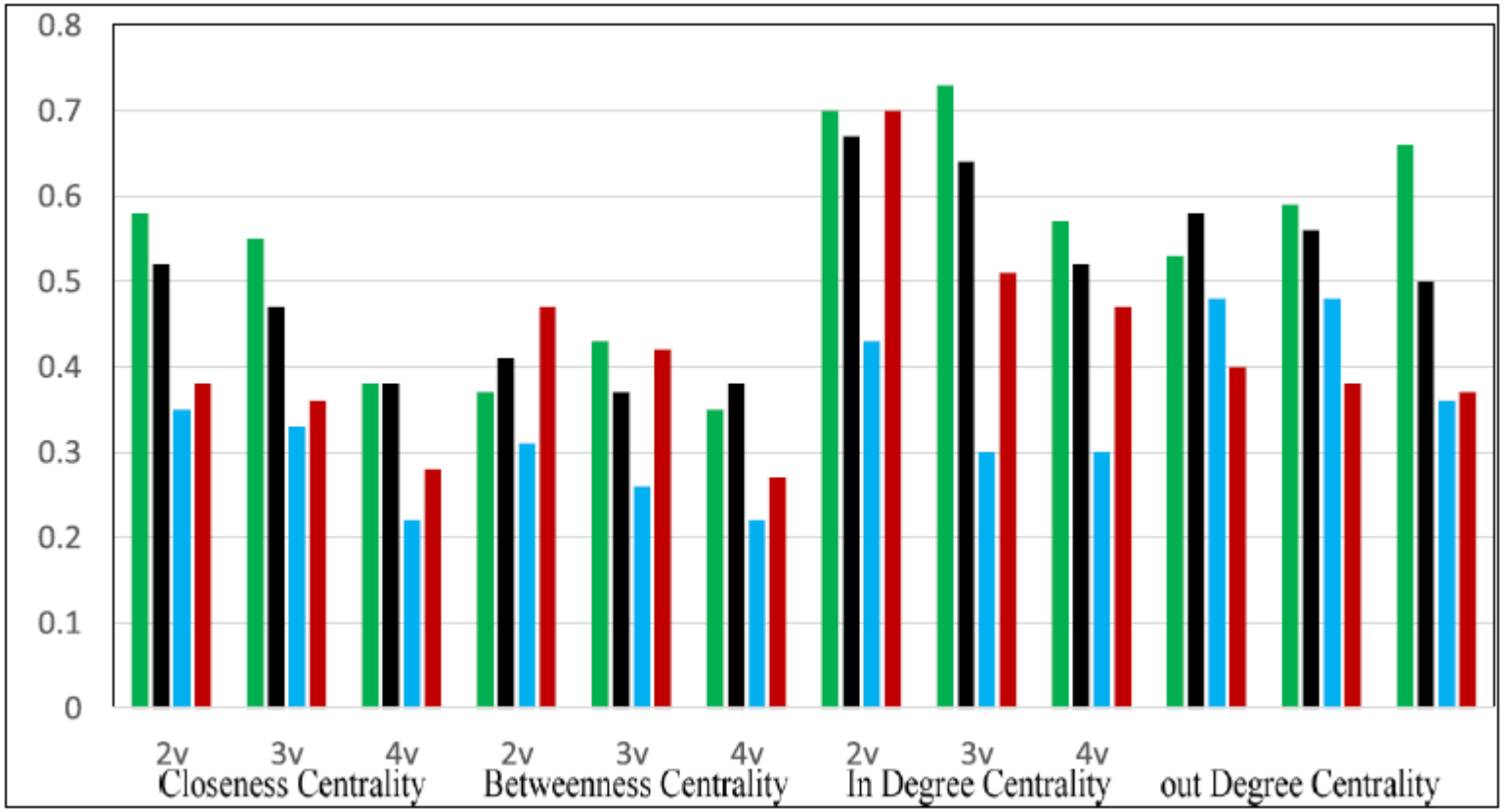


Fig. 8. (b) Precision of the four systems for identifying the important vertices to a given list of query vertices using the Enron dataset. In the figure, 2v, 3v, and 4v denote the following: 2 query vertices, 3 query vertices, and 4 query vertices respectively.

# 6. Experimental results

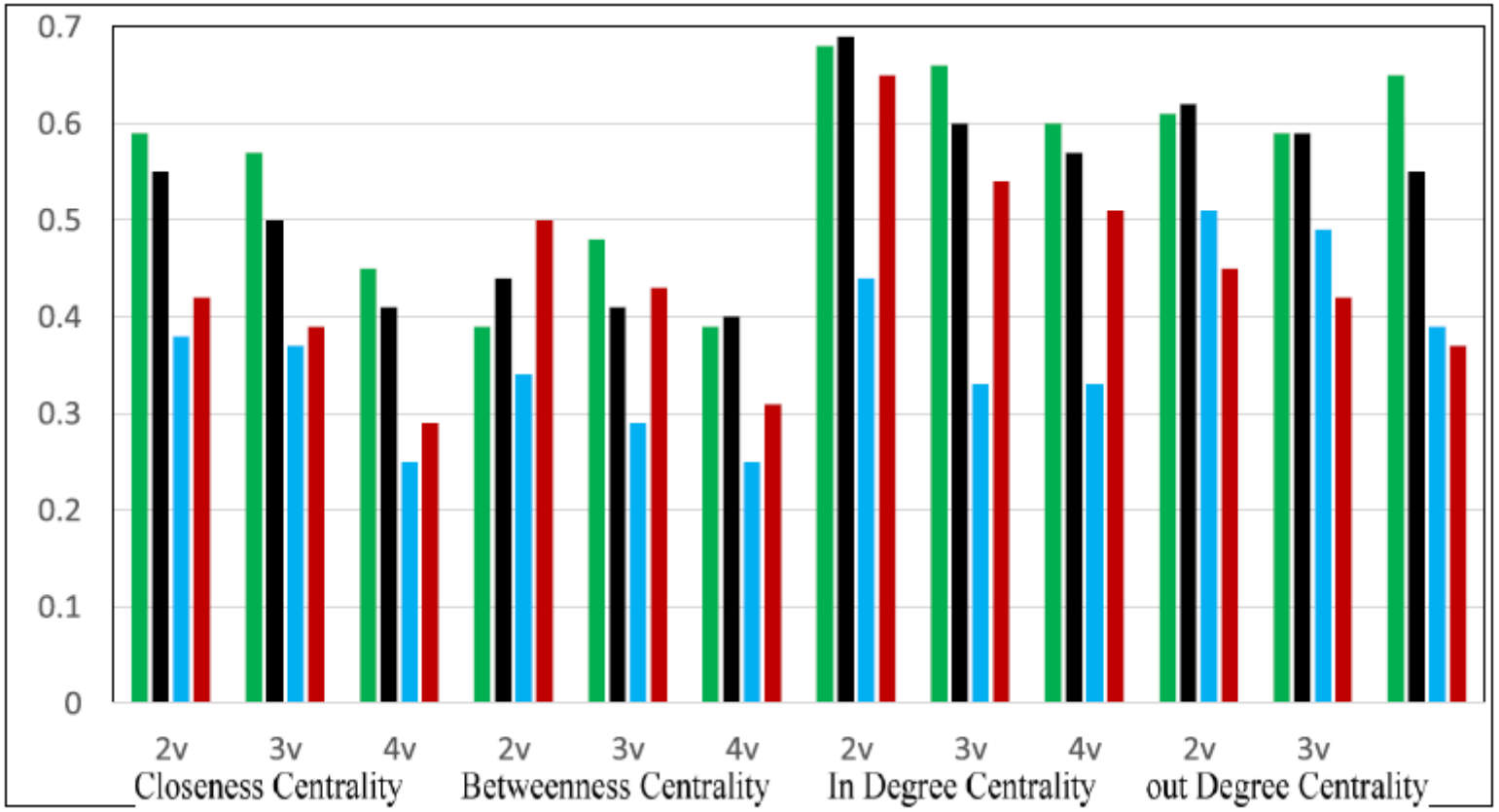


Fig. 8. (c) F-value of the four systems for identifying the important vertices to a given list of query vertices using the Enron dataset. In the figure, 2v, 3v, and 4v denote the following: 2 query vertices, 3 query vertices, and 4 query vertices respectively.

# 6. Experimental results

*4) Discussion of the Results:*

- Observations of the experimental results using the Krebs's 9/11 dataset:

1. Able to identify the key nodes in the network
2. Able to identify the nodes in the network representing the following most influential (i.e., central) actors in the incident: Atta, Al-Shehi, Jarrah, Khemais, Moussaoui, Hanjour, Al-Hazmi, Al-Shibh, and Essabar.
3. Able to identify the node representing Atta, the ringleader of the hijackers, as the most central node in the network.
4. The top four nodes identified by ECLfinder represents one of the hijackers on one of the four planes.
5. ECLfinder ranked the nodes representing Khemais, Moussaoui, and Jarrah very high.

- Observations of the experimental results using the Enron network dataset:

1. The top five nodes returned by ECLfinder in the Enron network represent the following actors in the Enron scandal:

- Arthur Andersen (auditor).
- Kenneth Lay (CEO).
- Sheila Kahanek (accountant).
- Andrew Fastow (financial officer).
- Jeffrey Skilling (COO).

2. Three of these five individuals have been charged and found guilty of various conspiracy and accounting frauds.

# 6. Experimental results

*4) Discussion of the Results:*

*1. LogAnalysis Limitations:*

- It does not work well for clustering large-size networks.
- The results showed that it clusters small-size networks more accurately than large-size ones.
- It is biased to globular clusters.
- It cannot detect and undo incorrect clustering that was done at an early stage.
- If clusters have different sizes, it may not work well.
- Due to the nature of its techniques, some vertices may not contribute to the overall importance value of a vertex (Incomplete Contribution).

*2. CrimeNet Explorer Limitations:*

- Let (*u*, *v)* be the most important incoming edge to vertex *v*. CrimeNet Explorer determines the weight of vertex *v* based *solely* on the weights of edge (*u*, *v)* and vertex *u*.

*3. SIIMCO Limitations:*

- It does not work well when the network consists of a large number of vertices and edges.

# 6. Conclusion

- Introduced  a forensic analysis system called ECLfinder.
- The system can determine the influential members of a criminal organization as well as the immediate leaders of a given list of lower-level criminals associated with the organization.
- Experimentally compared ECLfinder with SIIMCO [19], CrimeNet Explorer [12], and LogAnalysis [8] for identifying the important vertices in networks. Results revealed that ECLfinder outperforms the three systems.

# REFERENCE

Kamal Taha and Paul D. Yoo, Using the Spanning Tree of a Criminal Network for Identifying Its Leaders, *IEEE Transactions on Information Forensics and Security*, Vol. 12, No. 2, 2017, pp. 445-453