Yoo, Paul D. (2019) Shortlisting the influential members of criminal organizations and identifying their important communication channels. IEEE Transactions on Information Forensics and Security , ISSN 1556-6013.

Shortlisting the Influential Members of Criminal Organizations and Identifying their Important Communication Channels

Kamal Taha, *Senior Member, IEEE* and Paul D. Yoo, *Senior Member, IEEE*

Abstract— **Low-level criminals, who do the legwork in a criminal organization are the most likely to be arrested, whereas the high-level ones tend to avoid attention. But crippling the work of a criminal organizations is not possible unless investigators can identify the most influential, high-level members and monitor their communication channels. Investigators often approach this task by requesting the mobile phone service records of the arrested low-level criminals to identify contacts, and then they build a network model of the organization where each node denotes a criminal and the edges represent communications. Network analysis can be used to infer the most influential criminals and most important communication channels within the network but screening all the nodes and links in a network is laborious and time consuming. Here we propose a new forensic analysis system called IICCC (Identifying Influential Criminals and their Communication Channels) that can effectively and efficiently infer the high-level criminals and short-list the important communication channels in a criminal organization, based on the mobile phone communications of its members. IICCC can also be used to build a network from crime incident reports. We evaluated IICCC experimentally and compared it with five other systems, confirming its superior prediction performance.**

***Index Terms*— Forensic investigation tool, criminal network, mobile communications data, influential criminals, low-level criminals.**

I. INTRODUCTION

Sanctioning and removing the leaders of a criminal organization can lead to a significant reduction in the organization's criminal activities [11, 32]. These criminals play crucial and decisive roles in controlling the flow of information within the organization. Digital forensics tools are widely used by criminal investigators to identify the leaders of criminal organizations. Some of these tools are inspired by social network analysis (SNA) [39] and others employ the *k*-clique technique [21, 41, 45]. For example, the *k*-clique technique was used to identify a group of hackers called Shadowcrew [29], and to identify a community of criminals from a large dataset of Canadian offenders [16].

Hierarchically structured criminal organizations mainly use mobile phones and emails to communicate, so access to mobile communications data (MCD) and email records allows the relationships among the members of such organizations to be depicted as networks, often allowing the identification of leaders and the most important communication channels [16, 18, 22, 28, 29, 30, 41, 45]. For example, the email addresses of Nigeria-based scammers were linked to their Facebook profiles and the *k*-clique was used to identify the scammers and their leaders, based on a social network with 40,000 nodes [18]. More recently, crime incident reports that involve criminal organizations have been used to depict the relationships among its members [8]. Most current approaches compute the *relative importance* of each node in a criminal network to infer the primary nodes representing the leaders, a strategy known as the r*elative importance* problem [19]. These approaches are usually inspired by the *k*-clique technique [18, 20, 29, 48], network metrics [26], and semantic similarities [3, 25, 44]. However, such strategies generally cannot short-list the important communication channels in a network and the channels need to be investigated directly to provide insight into the criminal organization and its influential members. There are numerous communication channels in criminal networks, and investigating all the channels is time-consuming and distracting, when the investigators would prefer to focus on the key channels that reveal the most information.

In order to address this challenge, we have developed the IICCC system (**I**dentifying **I**nfluential **C**riminals and their **C**ommunication **C**hannels) which exploits the hierarchical structure of typical criminal organizations to infer not only the high-ranking members but also the vital communication channels, based on the propagation of information among nodes. The key contribution of this paper lies in identifying the vital communication channels in a criminal network. These vital channels involve a sequence of nodes that form part of one or more communication paths and depict influential criminals, who propagate important information diffused by a top-ranking member. These communication channels are likely to hold vital information about the criminal organization [49]. The ultimate goal of our proposed system IICCC is to identify these important communication channels. Since the information propagated through these channels originate from influential members in an organization, we also describe a methodology in the paper for identifying these influential members.

The authors of [23] observed that the change in the flow rate of information diffused by higher-level criminals is slower than that of lower-level criminals. This has led us to hypothesize that the steadiness of information flow rate increases as the chain of command gets higher. That is, we hypothesize that influential leaders diffuse information in more steady rate than less influential ones. IICCC identifies the top leaders of a criminal organization based on the above-mentioned hypothesis by computing the chi-squared (χ^2) values [17, 34] of the network's nodes. Criminals represented by nodes with high χ^2 are considered top leaders. Finally, IICCC identifies the important communication channels originating from these top leaders. These channels are likely to hold vital information about the criminal organization. An important

---

K. Taha is with the Electrical and Computer Engineering Department, Khalifa University, UAE (e-mail: kamal.taha@kustar.ac.ae).

P. Yoo is with the CSIS within Birkbeck College at the University of London, United Kingdom (email: p.d.yoo@ieee.org)

communication channel is a sequence of influential criminals who propagate information diffused by a top leader. In the framework of IICCC, these important channels are identified by computing the χ^2 values of the communication channels connecting each two adjacent nodes that are part of a communication path originating from a top leader. The channels with large χ^2 values are considered critical communication channels.

II. Motivation and Outline of the Approach

Low-level criminals who do the leg work are usually the easier to be caught and taken into custody. Criminal investigators, usually, take advantage of these detained criminals as a bridge for identifying their top leaders [1, 55]. To do so, the investigators would ask mobile service providers for the MCD that belongs to these arrested low-level criminals, their callers, the callers of their callers, and so on. Then, they would build a communication network depicting the criminal organization. This is an effective approach, because the vast majority of criminal organizations, nowadays, contemplate their plots using mobile phones [21]. Many techniques have been proposed in literature for identifying the relative importance of nodes in such networks [18, 25, 29, 48]. However, most of these techniques may not work well for short-listing the critical communication channels, which provide insight into criminal organizations and their influential members. This is because these techniques have been designed to identify the influential nodes in a network and not the critical communication channels that pass high and steady rate of information diffused by these influential nodes. We introduce in this paper the forensic system IICCC, which can infer the high-level criminals and short-list the important communication channels in a criminal organization, based on the mobile phone communications of its members.

Identifying high-level criminals and short-listing the important communication channels originating from them can help investigators monitor the criminal organization. This monitoring can give an insight into the organization's work and the identity of a large number of the insiders and outsiders who deal with it. This will eventually lead to the arrest of these people, which will most likely result in at least crippling the organization's work. Thus, for the monitoring procedure to be effective, the influential members of a criminal organization should be identified. Sanctioning and removing (e.g., arresting) these influential leaders can lead to significant reduction in the organization's criminal activities [11, 32].

We use the term "significant communication path" to denote a sequence of nodes in a criminal network representing *influential criminals* who propagate information diffused by a top-ranked leader in the criminal organization. We use the term "critical communication channel" to denote a portion of a "significant communication path" that has a high and steady rate of information flow relative to the other paths in the network.

Let p be a "significant communication path" originating from a top-ranked leader r in a criminal organization. A "critical communication channel" denotes a sub-path $p' \subseteq p$ that has a high and steady flow rate of information diffused by r and propagated by influential criminals to lower-level criminals. Some of this information may be passed to p' via other significant communication paths originating from r.

A "critical communication channel" is likely to hold vital information about a criminal organization. The ultimate goal of IICCC is to identify the "critical communication channels" in a network. To identify the "critical communication channels", IICCC identifies the influential nodes representing the top-ranked members as a precursor. To identify the top-ranked members, IICCC identifies the "significant communication paths" as a precursor.

IICCC infers the high-level criminals and the critical communication channels in a criminal organization by going through the following sequence of steps:

1) Construct a network based on either MCD or crime incident reports pertaining to the organization. Section III describes this process in detail.
2) Identify the *significant communication paths* originating from each node in the network. Subsection IV describes this process in detail.
3) Compute the χ^2 value of each node in a significant communication path, based on the actual (observed) and expected betweenness centralities of the nodes. Subsection V-A describes this process in detail.
4) Identify the influential criminals by ranking the nodes based on their χ^2 values. Subsection V-B describes this process in detail.
5) Identify the *critical communication channels* in the network by: (1) computing the actual (observed) betweenness centralities, expected betweenness centralities, and χ^2 values of the communication channels (edges) connecting each two adjacent nodes m and m' that are part of a significant communication path originating from an influential node in the network; (2) computing the summation of the χ^2 values of the communication channels, which are part of the different significant communication paths that pass through m and m'; and (3) identifying the channels with large χ^2 values, which will be considered the critical communication channels. Subsection VI describes this process in detail.

Fig. 1 shows a process legend that visualizes the sequential processing steps taken by IICCC to identify influential leaders and "critical communication channels".

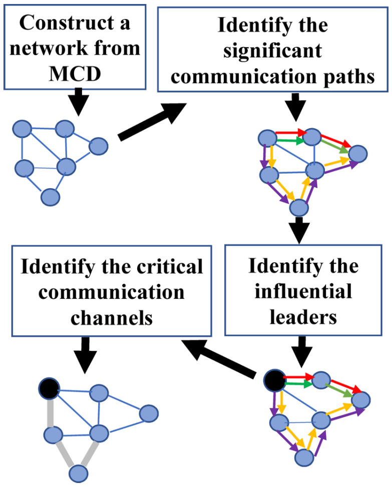
Construct a network from MCD
Identify the significant communication paths
Identify the critical communication channels
Identify the influential leaders

Fig. 1: A process legend that visualizes the sequential processing steps taken by IICCC to identify influential leaders and "critical communication channels".

III. COMPUTING NODE BETWEENNESS CENTRALITY

IICCC can construct networks from either MCD or crime incident reports. In a MCD network, a node denotes a caller/receiver criminal and an edge denotes the flow of information (phone calls, text messages or emails) between two criminals. IICCC adopts the space approach concept [7] to automatically build a network from crime incident reports [8]. We assume that criminals who appear in the same crime incident reports collaborate in committing crimes. Let *e* be an edge linking two nodes n_1 and n_2 in a network built from crime incident reports. Then *e* denotes the co-occurrence of two criminals represented by n_1 and n_2 in the same crime incident reports.

Node betweenness has been used extensively to indicate the centralities of nodes in a network [45]. A node with a relatively large betweenness centrality acts as a bridge that controls the flow of information between the nodes of the network. Thus, the betweenness centralities of nodes are reflective of the relative influences of these nodes in the network. IICCC therefore computes the betweenness centralities of all nodes in a network to capture their relative influence over the information flow in the network.

Several variations that capture the notion of node betweenness have been proposed. In one widely adopted measure [14], the betweenness $B(v)$ of a node *v* is the fraction of the number of shortest paths from all nodes to all other nodes that pass-through *v*, as defined in Equation 1:

$$B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where $\sigma_{st}(v)$ is the number of shortest paths from node $s \in V$ to node $t \in V$ that passes through v, and σ_{st} is the overall number of shortest paths from node s to node v.

Calculating the betweenness centralities of nodes in a network that has *n* nodes and *k* links (edges) using a breadth-first search (BFS) algorithm takes $O(kn^2)$ [42]. This is because, computing the shortest path between two nodes takes $O(k)$ and there are $O(n^2)$ pairs of nodes. A faster algorithm that calculates the betweenness centrality is based on the dependency accumulation technique [47]. Nodes are accessed in the reverse order compared to BFS. The algorithm takes $O(kn)$ on an unweighted network, where *k* is the number of links. In the framework of IICCC, we use this technique to compute node betweenness, as defined in Equation 2:

$$B(v) = \sum_{s \neq v} \delta_{s\bullet}(v) \tag{2}$$

where $B(v)$ is the betweenness of a node *v*, and $\delta_{s\bullet}(v)$ is the dependency of a node *s* on node *v*, as defined in Equation 3.

$$\delta_{s\bullet}(v) = \sum_{w: v \in P_s(w)} \frac{\sigma_{sv}}{\sigma_{sw}} (1 + \delta_{s\bullet}(w)) \tag{3}$$

where $P_s(w)$ is the set of predecessors of a node *w* located on the shortest paths from node *s*, and σ_{sv} is the overall number of shortest paths from node s to node v.

To illustrate the concepts proposed in this paper, we use a running example based on the network shown in Fig. 2. The network depicts communication attempts within a criminal organization based on its MCD, where a node denotes a criminal and an edge denotes the flow of information (phone calls and text messages) between two criminals. The network consists of 45 nodes and 103 edges (communication channels).

Example 1: Table 1 shows the betweenness centralities of the 45 nodes in the running network presented in Fig. 2.

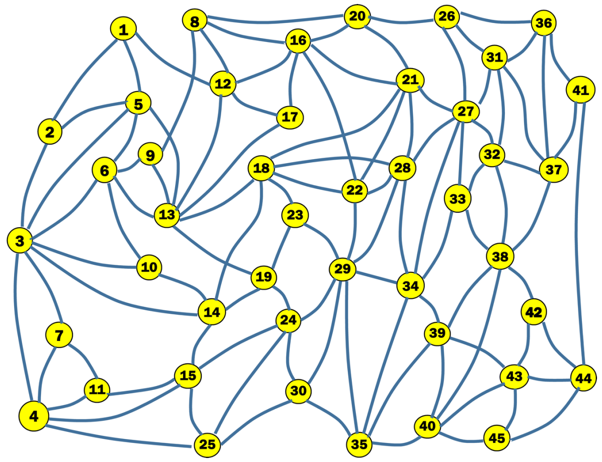

Fig. 2: A network depicting the communication attempts within a criminal organization based on its MCD. The network contains 45 nodes representing 45 criminals in the organization. The network contains 103 edges representing 103 communication channels among the 45 criminals.

Table 1: The betweenness centrality of each of the 45 nodes in our running network presented in Fig. 2.

Node	Betweenness Centrality	Node	Betweenness Centrality	Node	Betweenness Centrality	Node	Betweenness Centrality	Node	Betweenness Centrality	Node	Betweenness Centrality	Node	Betweenness Centrality	Node	Betweenness Centrality	Node	Betweenness Centrality
1	0	**6**	0.050	**11**	0.003	**16**	0.049	**21**	0.099	**26**	0.120	**31**	0.048	**36**	0.029	**41**	0.022
2	0.005	**7**	0.005	**12**	0.056	**17**	0	**22**	0.039	**27**	0.178	**32**	0.036	**37**	0.024	**42**	0.005
3	0.075	**8**	0.015	**13**	0.265	**18**	0.193	**23**	0.013	**28**	0.118	**33**	0.031	**38**	0.068	**43**	0.044
4	0.064	**9**	0	**14**	0.026	**19**	0.059	**24**	0.235	**29**	0.118	**34**	0.187	**39**	0.093	**44**	0.020
5	0.085	**10**	0.001	**15**	0.061	**20**	0.047	**25**	0.08	**30**	0.121	**35**	0.119	**40**	0.066	**45**	0.007

IV. IDENTIFYING THE SIGNIFICANT COMMUNICATION PATHS ORIGINATING FROM EACH NODE IN A NETWORK

IICCC identifies the significant communication paths originating from a node *n* as follows. Let *S* be the set of nodes that are directly connected to *n* by edges. For each node *m* in *S*, IICCC determines the significant communication path *p* originating from *n* and passing through *m* as follows. Let m' be a node at hierarchical level *l* in *p*. Let S' be the set of nodes at hierarchical level $(l+1)$ that are directly connected to m' by edges. From among the nodes in S', IICCC selects the node m'' that has the highest betweenness centrality. Thus, m'' becomes part of the path *p*. This process continues until the path *p* starts to converge to itself. That is, path *p* starts from *n* and ends at n', which is the node in the path where the path starts to converge to itself.

Example 2: Let us identify the significant communication paths originating from node 6 in our running network shown in Fig. 2. The set of nodes that is directly connected to node 6 by edges is {5, 9, 13, 10, 3} indicating five significant communication paths.

For easy reference, we refer to these paths using the following notations: 6→5, 6→9, 6→13, 6→10 and 6→3. Fig. 3 shows these five channels. For example, path 6→13 is identified as follows:

- The set of nodes directly connected to node 13 by edges is {9, 5, 12, 17, 18, 19}. Node 18 has the largest betweenness centrality among the nodes in this set (Table 1).
- At level 2 of channel 6→13, the information therefore passes through node 18. The set of nodes directly connected to node 18 by edges is {21, 28, 22, 23, 14}. Node 21 has the largest betweenness centrality among the nodes in this set (Table 1).
- At level 3 of channel 6→13, the information therefore passes through node 21.
- Ultimately, the sequence (6→13→18→21→27→34→28→27) starts to converge to itself at node 27. This happens because (1) the information passes through node 28 at level 6; (2) node 27 has the largest betweenness centrality among the nodes that are directly connected to node 28; and (3) the information has already passed through node 27 at level 4.

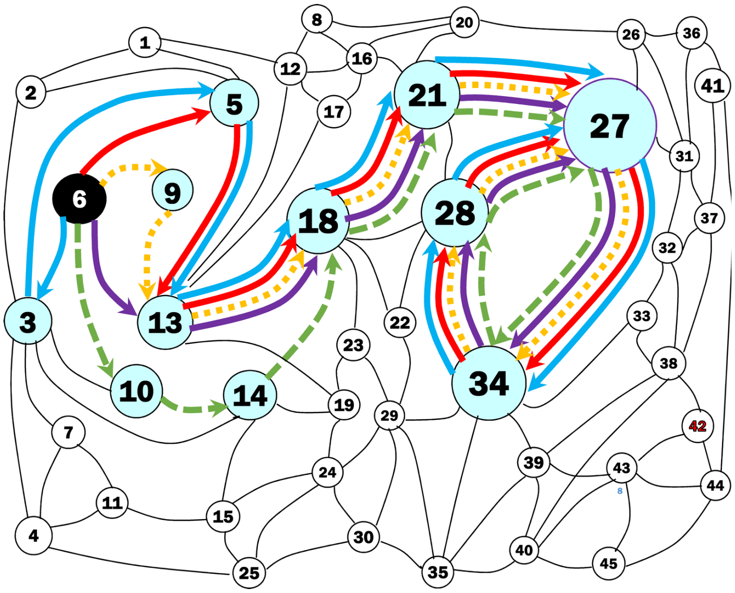

Fig. 3: The five significant communication paths originating from node 6 in our running example network. Each path is assigned a different color.

V. IDENTIFYING THE INFLUENTIAL CRIMINALS IN THE NETWORK

Arresting the influential leaders of a criminal organization can lead to the reduction of the rate of information flow, which in turn destabilizes the criminal network [36]. The information diffused by these leaders is propagated by criminals in the chain of command [37]. As [35] indicates, the directives from the leaders of a criminal organization are transmitted from higher-level criminals to lower-level criminals in the chain of hierarchy. The authors of [23] observed that the change in the rate of information flow decreases as the number of the receivers of this information increases. As a result, these authors assume that the change in the flow rate of information diffused by higher-level criminals is slower than that of lower-level criminals. This finding conforms to several studies, which observed that as the influence of a criminal gets higher, the centrality score of the node representing this criminal increases [36] (i.e., a higher score of a node could be an indicative of a larger number of receivers of the information diffused by this node). All the above revelations have led us to hypothesize that the steadiness of information flow rate increases as the chain of command gets higher. That is, we hypothesize that influential leaders diffuse information in more steady rate than less influential ones.

IICCC identifies the top leaders of a criminal organization based on the above-mentioned hypothesis. More specifically, if the difference between the actual and expected betweenness centralities of the nodes located in the significant communication paths originating from a node *n* is small, IICCC considers *n* as influential. To achieve this, IICCC uses χ^2 analysis [17, 34] to compute the goodness of fit of the observed and expected betweenness centralities of the nodes in a significant communication path. This is because χ^2 analysis can effectively determine whether an observed distribution conforms to any particular expected distribution. IICCC considers criminals represented by nodes with large χ^2 values as influential.

A. Computing the χ^2 Value of a Node

In IICCC, the degree of influence of a node *n* is represented by its χ^2 value [17, 34]. The degree of influence of *n* is the summed degrees of influence of the nodes in the significant communication paths originating from *n*. Thus, the χ^2 value of *n* is the overall χ^2 value of the nodes in the significant communication paths originating from *n*. That is, the χ^2 value of *n* is the summation of the χ^2 values of the paths originating from *n*. The χ^2 value of each of these paths is the summation of the χ^2 values of the nodes in the path. The χ^2 value of each of these nodes is determined as follows. Let *p* be a significant communication path originating from *n*. Let *m* be a node located at hierarchical level *l* of *p*. The χ^2 value of *m* is determined based on the betweenness centralities of (1) the nodes in *p*; and (2) the nodes at level *l* of the other paths originating from *n*, as defined in Equation 4 [17, 34]:

$$\chi^2 = (O-E)^2/E \qquad (4)$$

where *O* is the observed betweenness centrality of the node, and *E* is the expected betweenness centrality of the node.

The Expected betweenness centrality (E_n) of a node *n* located at level *l* of a path *p* is computed using Equation 5:

$$E_n = \frac{B_{p_n} \times B_{l_n}}{B_{all}} \qquad (5)$$

where B_{p_n} is the betweenness centrality of path *p* originating from node *n*; B_{l_n} is the summation of the betweenness centralities of the nodes at level *l* of the paths originating from *n*; and B_{all} is the summation of the betweenness centralities of the paths originating from *n*.

Let p_u and p_v be two significant communication paths originating from a node *n* under consideration. Let p_u contain *u* levels and p_v contain *v* levels. Let p_u be the longest significant communication path originating from *n*. Thus, $v < u$. IICCC would assign a zero observed betweenness centrality (*O*) for each hypothetical node located at levels $v+1$, $v+2$, …., *u*.

Example 3: Let us compute the χ^2 value for node 6 in our running network as shown in Fig. 3. The left-hand portion of Table 2 shows the observed betweenness centralities (*O*) of the nodes comprising the five significant communication paths originating from node 6 (refer to Example 2). As Fig. 3 shows, path 6→3 is the longest path and contains nine levels. Given that each of the paths 6→5, 6→9 and 6→10 contains only eight levels, the hypothetical nodes at level 9 of these three paths are assigned zero observed betweenness centralities (Table 2). Because path 6→13 contains only seven levels, the hypothetical nodes at levels 8 and 9 of this path are assigned zero observed betweenness centralities (Table 2). The middle portion of Table 2 shows the expected betweenness centralities (E) of the nodes comprising the five paths. For example, the expected betweenness centrality of node 5 at level 1 of path 6→5 is calculated as follows: $E = \frac{1.303 \times 0.426}{6.022} = 0.092$. The right-hand portion of Table 2 shows the χ^2 values of the nodes comprising the five paths, the summed χ^2 values of the five paths (bottom row), and the χ^2 value of node 6 (the bottom right-hand corner). For example, the χ^2 value of node 5 at level 1 of path 6→5 is calculated as follows: $x^2 = \frac{(0.085 - 0.092)^2}{0.092} = 0.001$. The χ^2 value of a path is the summation of the χ^2 values of the nodes comprising it. For example, the χ^2 value of path 6→5 is 0.227 (Table 2). The χ^2 value of a node from which significant communication paths originate is the summation of the χ^2 values of these paths. For example, the χ^2 value of node 6 is the summation of the χ^2 values of the five paths (3.323; bottom right-hand corner of Table 2).

Table 2: Observed betweenness centralities, expected betweenness centralities, and χ^2 values of the nodes comprising the significant communication paths originating from node 6 as well as the overall χ^2 value of node 6 in our running network as described in Example 3.

Level	Observed Betweenness (O)						Expected Betweenness (E)					Chi-squared (X^2)					
	Path 6 →5	Path 6 → 9	Path 6→13	Path 6→10	Path 6→ 3	Σ	Path 6 →5	Path 6 → 9	Path 6→13	Path 6→10	Path 6→ 3	Path 6 →5	Path 6→ 9	Path 6→13	Path 6→10	Path 6→ 3	Σ
1	0.085	0	0.265	0.001	0.075	0.426	0.092	0.086	0.086	0.064	0.097	0.001	0.086	0.373	0.062	0.005	0.527
2	0.265	0.265	0.193	0.026	0.085	0.834	0.180	0.169	0.169	0.125	0.191	0.04	0.054	0.003	0.078	0.059	0.234
3	0.193	0.193	0.099	0.099	0.265	0.849	0.184	0.172	0.172	0.128	0.194	0	0.003	0.031	0.007	0.026	0.067
4	0.099	0.099	0.178	0.178	0.193	0.747	0.162	0.151	0.151	0.112	0.171	0.025	0.018	0.005	0.039	0.003	0.09
5	0.178	0.178	0.187	0.187	0.099	0.829	0.179	0.168	0.168	0.125	0.19	0	0.001	0.002	0.031	0.044	0.078
6	0.187	0.187	0.118	0.118	0.178	0.788	0.171	0.159	0.168	0.118	0.180	0.001	0.005	0.011	0	0	0.017
7	0.118	0.118	0.178	0.118	0.187	0.719	0.301	0.145	0.145	0.108	0.165	0.111	0.005	0.008	0.001	0.003	0.128
8	0.178	0.178	0	0.178	0.118	0.652	0.141	0.132	0.132	0.018	0.149	0.01	0.016	0.132	1.422	0.006	1.586
9	0	0	0	0	0.178	0.178	0.039	0.036	0.036	0.027	0.041	0.039	0.036	0.039	0.027	0.458	0.596
Σ	1.303	1.218	1.218	0.905	1.378	6.022	1.449	1.218	1.218	0.825	1.378	0.227	0.224	0.601	1.667	0.604	3.323

B. *Identifying the Influential Criminals*

Nowadays, criminal organizations can be classified as either hierarchically well-structured [27] or hierarchically loosely structured [52]. Usually, a hierarchically well-structured organization operates under the directions of leaders of varying degrees of influences [27]. Hierarchically loosely structured organizations are usually composed of loosely connected groups of criminals [52]. Each group includes some influential criminals, who provide some type of directions to the group. The framework of IICCC can be applied to both, the hierarchically well-structured and loosely structured organizations, as long as their communication networks can be depicted. IICCC can help investigators to identify the influential criminals in these organizations by generating short-lists comprising small and tightly-defined groups of their most influential individuals. It does so by ranking the nodes in a criminal network based on their χ^2 values (Subsection V-A). Criminals represented by the top-ranked nodes are considered by IICCC as the most influential ones in the organization.

Example 4: Table 3 ranks the nodes in our running network based on their χ^2 values, which are calculated using the techniques described in Subsection V-A and Example 3. Investigators should focus on the top-ranked nodes, especially node 6.

Table 3: The nodes in our running network ranked based on their χ^2 values.

Rank	Node	Ch. Sq.	Rank	Node	Ch. Sq.	Rank	Node	Ch. Sq.
1	6	3.323	**16**	32	2.037	**31**	11	1.236
2	34	2.916	**17**	35	1.909	**32**	17	1.216
3	29	2.908	**18**	37	1.897	**33**	44	1.145
4	3	2.894	**19**	14	1.831	**34**	33	1.123
5	38	2.822	**20**	20	1.803	**35**	8	1.1
6	21	2.815	**21**	15	1.772	**36**	41	0.926
7	12	2.753	**22**	4	1.699	**37**	9	0.921
8	43	2.677	**23**	19	1.688	**38**	28	0.858
9	13	2.562	**24**	16	1.678	**39**	23	0.852
10	18	2.523	**25**	30	1.527	**40**	40	0.848
11	22	2.401	**26**	25	1.431	**41**	7	0.592
12	39	2.274	**27**	31	1.344	**42**	2	0.568
13	36	2.261	**28**	26	1.289	**43**	1	0.38
14	27	2.136	**29**	5	1.25	**44**	42	0.363
15	10	2.113	**30**	11	1.236	**45**	45	0.254

VI. IDENTIFYING THE CRITICAL COMMUNICATION CHANNELS

Each member of a criminal organization O is controlled either directly or indirectly by a core influential criminal(s) i. The information diffused by i is propagated to the members of O through communication channels. The order in which the criminals receive and propagate the information within these channels corresponds to their hierarchical influence in O [27]. Therefore, the portions of these channels that involve influential members of O are likely to hold vital information about O, hence their designation as critical communication channels. In IICCC, a critical communication channel is a sequence of influential criminals who form part of one or more significant communication paths that pass information diffused by a top-ranking member. In other words, let p be a significant communication path originating from a top influential criminal i. A critical communication channel is a sub-path (channel) $p' \subseteq p$ that encompasses a sequence of influential criminals, who propagate information through either (1) p alone, or (2) p and other significant communication paths originating from i. Based on the betweenness centralities of the edges in a network [15], IICCC calculates the expected betweenness centralities and χ^2 values of the communication channels connecting each two adjacent nodes m and m' that are part of a significant communication path originating from the node under consideration. Then IICCC computes the summation of the χ^2 values of the communication channels that are part of the different significant communication paths that pass through m and m'. For example, if there are k significant communication paths that pass through m and m', we sum the χ^2 values of the channels of these paths between m and m'. That is, we sum the χ^2 values of the portion of these paths (channels) at their hierarchical levels between m and m'. This summation is considered the χ^2 value of the communication channel connecting m and m'. The channels with large χ^2 values are considered critical communication channels.

Example 5: Let us identify the critical communication channels originating from the most core influential node in our running network, which is node 6 (Table 3). Based on the betweenness centralities of the edges in the network [15] (the observed betweenness centralities of the edges), we calculate the expected betweenness centralities and χ^2 values of the communication channels connecting each two adjacent nodes that are part of the significant communication paths originating from node 6. As shown in Table 4, the χ^2 value of each communication channel connecting each two adjacent nodes m and m' that are part of a significant communication path p is computed based on the hierarchical level of m and m' within p. For example, the communication channel at level 3 of path 6→5 (i.e., the portion of the path between nodes 13 and 18) has an observed betweenness centrality of 364.1, an expected betweenness centrality of 250.1, and a χ^2 value of 52. The overall χ^2 value of a communication channel connecting two adjacent nodes m and m' is the summation of the χ^2 values of the communication channels connecting m and m' that are part of the significant communication paths passing through m and m'. Fig. 4 shows the overall χ^2 value of each communication channel. For example, Table 5 shows the overall χ^2 value of the communication channel connecting nodes 13 and 18, which is computed as follows. As Fig. 3 shows, paths 6→5, 6→9, 6→13, and 6→3 pass through nodes 13 and 18 at different hierarchical levels: paths 6→5 and 6→9 pass through the two nodes at the third hierarchical level, path 6→13 passes through the two nodes at the second hierarchical level, and path 6→3 passes through the two nodes at the fourth hierarchical level. Therefore, the χ^2 values of the different communication channels connecting nodes 13 and 18 are computed based on the above hierarchical levels. Table 5 shows the corresponding hierarchical levels and χ^2 values. Accordingly, the overall χ^2 value of the channel connecting nodes 13 and 18 is 525.2, which was computed by summing the χ^2 values of the different channels. The critical communication channels are therefore those connecting the following nodes: (6, 13), (13, 18), and (18, 21). Although Fig. 4 shows there are 103 edges (communication channels) in the network, IICCC was able to short-list three critical communication channels allowing investigators to focus on these as key targets.

Table 4: Observed betweenness centralities, expected betweenness centralities, and χ^2 values of each communication channel connecting each two adjacent nodes m and m' that are part of a significant communication path p originating from node 6 in our running network. The values were computed based on the hierarchical level of m and m' within p.

Level	Observed Betweenness (*O*)						Expected Betweenness (*E*)					Chi-squared (X^2)				
	Path 6→5	Path 6→9	Path 6→13	Path 6→10	Path 6→3	∑	Path 6→5	Path 6→9	Path 6→13	Path 6→10	Path 6→3	Path 6→5	Path 6→9	Path 6→13	Path 6→10	Path 6→3
1	9.33	18.9	137.5	46.7	64.8	277.2	60.4	54.3	56.6	39.9	66.1	43.2	23.1	115.6	1.16	0.03
2	209.2	69.1	364.1	30.6	64.9	737.9	160.8	144.4	150.6	106.2	175.9	14.6	39.3	302.7	53.8	70
3	364.1	364.1	140.8	69.4	209.2	1147.6	250.1	224.6	234.3	165.1	273.6	52	86.6	37.3	55.5	15.2
4	140.8	140.8	162.9	140.8	364.1	949.4	206.9	185.9	193.8	136.6	226.3	21.1	10.9	4.9	0.13	83.9
5	162.9	162.9	133.8	162.9	140.8	763.3	166.3	149.4	155.8	109.8	181.9	0.1	1.22	3.1	25.7	9.3
6	133.8	133.8	158.7	133.8	162.9	723	157.6	141.5	147.6	104	172.3	3.6	0.42	0.83	8.5	0.51
7	158.7	158.7	104.7	158.7	133.9	714.7	155.7	139.9	145.9	102.8	170.4	0.1	2.5	11.6	30.4	7.8
8	104.7	104.7	0	104.7	158.7	472.8	103	92.6	96.5	68	112.7	0.03	1.58	96.5	19.8	18.8
9	0	0	0	0	104.7	104.7	22.8	20.5	21.4	15.1	25	22.8	20.5	21.4	15.1	254.1
∑	1283.5	1153	1202.5	847.6	1404	5890.6	1283.6	1153.1	1202.5	847.5	1404.2	157.5	186.1	593.9	210.1	459.6

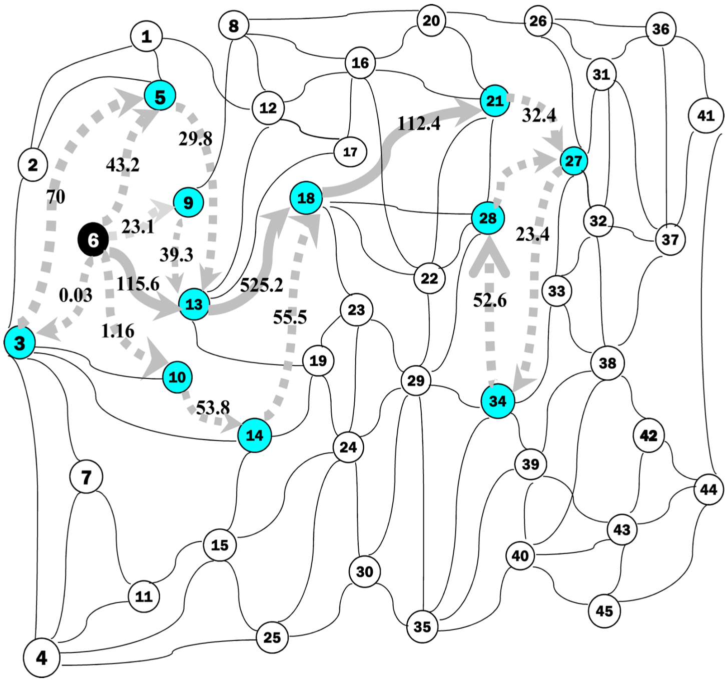

Fig. 4: The overall χ^2 value of each communication channel connecting each two adjacent nodes that are part of a significant communication path originating from node 6 (the most influential node) in our running network. The critical communication channels originating from node 6 are marked with thick solid arrows and the rest are marked with thick dotted arrows. The χ^2 values of each communication channel originating from node 6 are also shown.

Table 5: The overall χ^2 value of the communication channel connecting nodes 13 and 18 computed by summing the χ^2 values of the different channels connecting these two nodes based on the hierarchical level of the two nodes.

Path	Hierarchical level of path between nodes 13 and 18	χ^2 value
Path 6 →5	3	52
Path 6 → 9	3	86.6
Path 6→13	2	302.7
Path 6→ 3	4	83.9
∑ (overall χ^2 value of channel 13-18)		**525.2**

VII. EXPERIMENTAL RESULTS

We implemented IICCC in Java and ran it on a computer featuring an Intel Core i7 processor (2.70 GHz) and 16 GB of RAM running Windows 10 Pro. We compared IICCC experimentally with five competing systems: Locality Sensitive Hashing (LSH) [40], SIIMCO [45], ECLfinder [41], CrimeNet Explorer [21], and LogAnalysis [13]. Each of these systems is briefly described below.

LSH [40] employs locality sensitive hashing techniques for identifying the important edges in a network. [40] used this technique for identifying energy-efficient paths. It applied energy-efficient for low importance edges and full fidelity for high importance edges. SIIMCO [45] and ECLfinder [41] are systems that we previously proposed as tools to identify the leaders of a criminal organization based on MCD. The key difference between IICCC and these systems is that IICCC adopts the *significant communication path* and *critical communication channel* concepts, whereas ECLfinder applies the existence dependency concept, and SIIMCO uses Fisher's exact test to assign a score to each node reflecting its relative importance in the network. CrimeNet Explorer [21] uses Blockmodeling [5] and the shortest path algorithm to compute the relationships between nodes. It uses the closeness, degree and betweenness centrality metrics to identify the leaders of a criminal organization represented by a MCD network. Finally, LogAnalysis [13] computes the edge betweenness for a network using the Girvan & Newman algorithm [15], and then uses the greedy algorithm [33] to cluster the network hierarchically. The top clusters confine the important nodes in the network.

A. Compiling Datasets for Evaluation

The algorithms were evaluated using three real-world communications datasets, namely the Caviar dataset [6, 31], the Enron email dataset [12], and the Krebs's 9/11 dataset [50, 51]. We converted each dataset to a network depicting the communication attempts between the individuals incriminated in the incidents. Each dataset is briefly described below.

- Caviar dataset [6, 31] is drug-trafficking operation's communications among a Canadian gang called Caviar. The Caviar gang operated in Montreal, Canada, and dealt with importing and distributing hashish and cocaine. A network was created based on the phone calls among the drug traffickers between the years 1994 and 1996. The network consists of 110 nodes representing 110 gang members. Since the identities of the gang members have been kept confidential, the members are represented in the network by node designations (e.g., N1, N2, ….). The dataset is available at [9].
- Enron email dataset [12, 24] is a corpus of email messages exchanged between top Enron employees and associates. The corpus came to light in 2001 following a criminal investigation about alleged white-collar crime within the Enron Corporation. Most of these emails revolve around this. The dataset consists of 619,446 messages exchanged between 158 Enron employees. After cleaning the data, we obtained 200,136 messages exchanged between 151 employees. The investigation of Enron wrongdoing incriminated 28 Enron employees and associates. The names and identities of these 28 employees have been released to the public. In our evaluations, we considered the identities of these 28 employees as ground-truth data. The raw corpus is currently available online at [12].
- Krebs's 9/11 dataset [50, 51] is a corpus depicting the interactions between the terrorists involved in 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 Krebs's 9/11 dataset includes a network depicting interactions between the individuals incriminated in the terrorist attacks. The network contains 62 nodes depicting the terrorists implicated in the plot. It contains 153 links (edges) depicting the interactions between the terrorists. The average degree of a node is 4.9. We considered the lists of the terrorists ranked by Krebs [50] based on the Degree, Betweenness, and Closeness centralities of the nodes representing them in the network as a ground-truth dataset.

B. Evaluating the Accuracy of Detecting Influential Nodes

We compared the influential nodes returned by each system with the corresponding ones returned by the standard Betweenness, Closeness, Out Degree, and In Degree centrality metrics [4, 54]. We compared the results in terms of the following standard quality metrics:

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

where N_s^c is the number of *correct* nodes predicted by a system, N_m^{top} is the *actual* number of *correct* nodes, and N_s^{top} is the number of nodes predicted by a system. Let L_{top} and L_s be the lists predicted by a network metric and a system, respectively, ranked based on their influences. $N_s^c \subseteq L_{top}$ and $N_m^{top} = |L_{top}|$.

Accordingly, we computed the Recalls, Precisions, and F-values of each system with regards to each network centrality metric and each quality metric. Table 6 shows the results using the Caviar dataset [6, 31], Table 7 shows the results using the Enron email dataset [12], and Table 8 shows the results using Krebs's 9/11 dataset [50, 51]. Fig. 5 shows the overall average Recalls, Precisions, and F-values of the five systems.

Table 6: The quality of the results predicted by each system computed by comparing the system's top-ranked nodes with the corresponding ones predicted by the network metrics using the Caviar dataset [50, 51].

		Recall	Precision	F-value
IICCC	Closeness Centrality	0.62	0.66	0.64
ECLfinder		0.56	0.55	0.55
SIIMCO		0.57	0.60	0.58
CrimeNet Explorer		0.55	0.52	0.53
LogAnalysis		0.46	0.52	0.49
IICCC	Betweenness Centrality	0.64	0.53	0.58
ECLfinder		0.46	0.41	0.43
SIIMCO		0.59	0.47	0.52
CrimeNet Explorer		0.52	0.40	0.45
LogAnalysis		0.48	0.42	0.44
IICCC	In Degree Centrality	0.68	0.74	0.76
ECLfinder		0.66	0.71	0.67
SIIMCO		0.68	0.62	0.65
CrimeNet Explorer		0.55	0.51	0.53
LogAnalysis		0.61	0.58	0.59
IICCC	Out Degree Centrality	0.73	0.69	0.71
ECLfinder		0.68	0.63	0.65
SIIMCO		0.62	0.57	0.59
CrimeNet Explorer		0.54	0.47	0.50
LogAnalysis		0.54	0.51	0.52

Table 7: The quality of the results predicted by each system computed by comparing the system's top-ranked nodes with the corresponding ones predicted by the network metrics using the Enron dataset [12].

		Recall	Precision	F-value
IICCC	Closeness Centrality	0.57	0.54	0.55
ECLfinder		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
IICCC	Betweenness Centrality	0.55	0.48	0.51
ECLfinder		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
IICCC	In Degree Centrality	0.72	0.67	0.69
ECLfinder		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
IICCC	Out Degree Centrality	0.69	0.73	0.71
ECLfinder		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

Table 8: The quality of the results predicted by each system computed by comparing the system's top-ranked nodes with the corresponding ones predicted by the network metrics using the Krebs's 9/11 dataset [50, 51].

		Recall	Precision	F-value
IICCC	Closeness Centrality	0.77	0.64	0.70
ECLfinder		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
IICCC	Betweenness Centrality	0.68	0.73	0.70
ECLfinder		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
IICCC	In Degree Centrality	0.62	0.53	0.57
ECLfinder		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
IICCC	Out Degree Centrality	0.73	0.66	0.69
ECLfinder		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

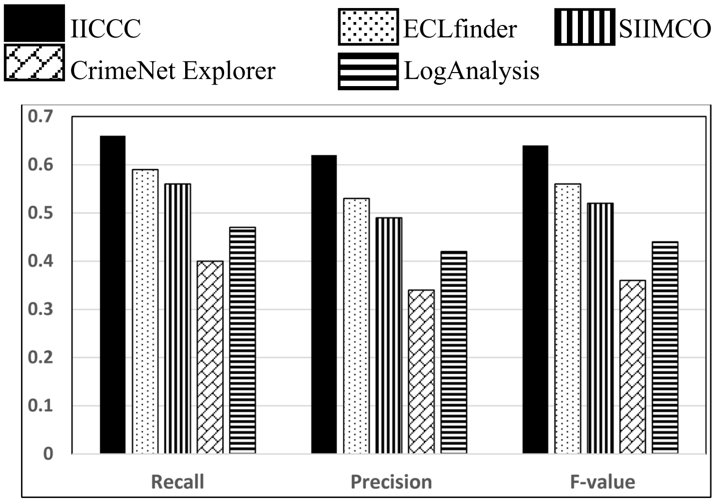

Fig. 5: The overall average Recall, Precision, and F-value of the five systems.

We computed the overall average execution time of IICCC and the other four methods for identifying the influential nodes in the thee ground-truth datasets described in section VII-A. Fig. 6 shows the results. As the figure shows, the computation time of IICCC is acceptable, and it is outperformed by only five methods.

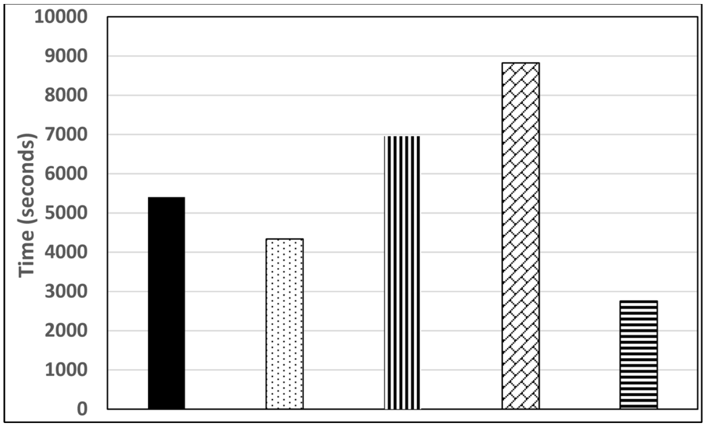

Fig. 6: Overall average execution time for identifying the influential nodes in the three ground-truth datasets described in section VII-A

C. Evaluating the Accuracy of Detecting Critical Communication Channels

We compared IICCC with the technique proposed in [40] for identifying the important communication paths (e.g., the critical communication channels) in the Caviar dataset [6, 31]. The authors of [40] proposed a framework for identifying the important edges originating from influential nodes in a graph. To the best of our knowledge, the technique proposed in [40] is the closest to our work for identifying the important paths originating from the influential nodes in a network. ECLfinder, SIIMCO, CrimeNet, and LogAnalysis are not designed for identifying important paths. [40] employs locality sensitive hashing techniques for identifying important edges. We refer to this technique by LSH (Locality Sensitive Hashing) for easy reference. [40] used the LSH technique for computing energy-efficient by applying energy-efficient for low importance edges and full fidelity for high importance edges.

We used the Caviar dataset [6, 31] described in section VII-A for the evaluation. The most influential nodes in the Caviar dataset have been designated by nodes N1, N12, and N3 [10]. The gang member represented by node N1 was heading the hashish drug trafficking. The gang member represented by node N12 was heading the cocaine drug trafficking. The gang member representing by node N3 was the intermediary between the N1 and N12 as well as between the two and non-traffickers [10]. *Therefore, we considered the paths originating from nodes N1, N12, and N3 as ground-truth critical communication channels.*

We used the detection accuracy (Acc) formula shown in Equation 6 as a metric for the evaluation. Fig. 7 shows the results.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

- TP (True positive): Number of paths correctly predicted as non-critical.
- FP (False positive): Number of paths incorrectly predicted as non-critical.
- TN (True negative): Number of paths correctly predicted as critical
- FN (False negative): Number of paths incorrectly predicted as critical.

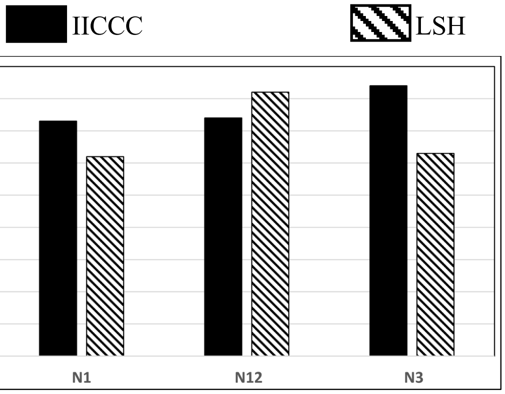

Fig. 7: The *detection accuracy* of IICCC and LSH [40] for identifying the critical communication channels in the Caviar dataset [6, 31].

D. Comparing the Systems in terms of Euclidean Distances

In this test we aim at measuring the *degree of conformation* between the predictions made by IICCC and the corresponding predictions made by the standard network metrics. That is, we evaluated the accuracy of IICCC in terms of the distances between: (1) the position of each node *m* in a list ranked by IICCC according to the influences of nodes in the network, and (2) the position of the same node m in a list ranked by a standard network metric according to the influences of nodes in the network. A ranking of nodes is a permutation of the integers 1, 2, ... *Intuitively, the smaller the distance the better IICCC*. Especially, we are interested in measuring the distances between the positions of IICCC's *top-ranked* nodes and the positions of the same nodes in the lists ranked by the standard network metrics. Towards this, we computed the average Euclidean distances between the positions of the top n nodes in the lists ranked by IICCC, and the corresponding positions of the same nodes in the lists ranked by the three-network metrics, where n is considered to be 5, 10, or 15. We employed the Euclidean measure shown in Equation 7. Fig. 8 plots the *average* Euclidean distance for each system using the three datasets described in subsection VII-A.

$$d(\sigma_m,\sigma_s) = \sum_{x \in N_m^{top}} |\sigma_m(v) - \sigma_s(v)| \tag{7}$$

where N_m^{top} is the list of the top n nodes predicted by network metric m, $\sigma_m \in [0,1]^{|N_m^{top}|}$ is the list of the ranked top n nodes predicted by network metric m, $\sigma_s \in [0,1]^{|N_m^{top}|}$ is the list of the ranked top n nodes predicted by a system S, and $\sigma_m(v)$ and $\sigma_s(v)$ are the positions in the lists σ_m and σ_s of node $v \in N_m^{top}$.

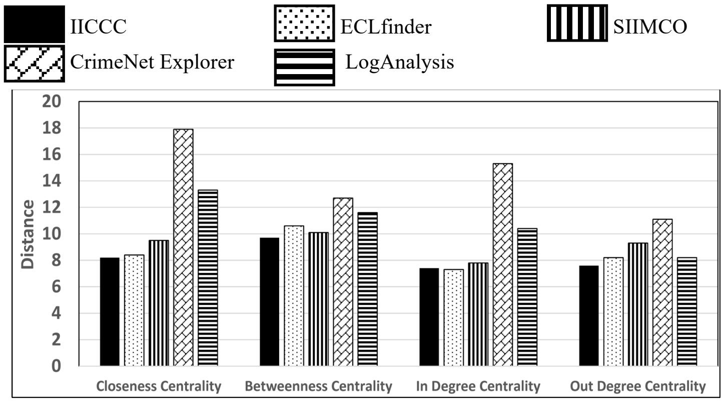

Fig. 8: The average Euclidean distance for each system using the three datasets described in subsection VII-A.

E. Discussion of the Results

As shown in Fig. 7 that IICCC outperformed LSH [40] in terms of Detection Accuracy of the critical communication channels in the Caviar dataset [6, 31]. Based on our observation of the experimental results, we attribute this outperformance to the combination of the "significant communication paths" and "critical communication channels" concepts employed by IICCC. The important paths identified by the "locality sensitive hashing technique" employed by [40] is equivalent to the "significant communication paths" identified by IICCC. However, the performance of IICCC over [40] stems from IICCC's employment of the "critical communication channels" concept, which shorts list the "significant communication paths" based on their importance in networks.

As Tables 6-8 and Figs. 5 and 8 show, IICCC outperformed SIIMCO, ECLfinder, CrimeNet Explorer, and LogAnalysis for identifying the influential nodes in the datasets. After studying the experimental results, we attributed the superior performance of IICCC to the following factors:

a) IICCC returned eight Enron employees as the most influential ones in the organization. Five out of these eight employees are publicly known to be the most involved ones in the crime, according to the investigation and the sentencing records that have been released to public domain. They were charged and found guilty of various conspiracy and accounting frauds. These five employees are: Arthur Andersen (auditor), Andrew Fastow (financial officer), Kenneth Lay (CEO), Rick Causey (Chief Accounting Officer), and Jeffrey Skilling (COO). On the other hand, ECLfinder and SIIMCO identified correctly four of them, LogAnalysis identified correctly three of them, and CrimeNet Explorer identified correctly only one of them.

b) IICCC determined the influence of each node in the evaluation networks by considering not only the node's number of links, but also the relative influences of the nodes connected to the node (using the concept of *significant communication paths*). IICCC also considered the relative influences of the edges connected to the node (using the concept of *critical communication channels*).

c) CrimeNet Explorer assigned a weight for each node based on its topology in the evaluation networks with regard to the node *n* under consideration. Therefore, these nodes did not contribute equally to the influence of *n*, a phenomenon described as incomplete contribution. This is one of the key limitations of CrimeNet Explorer.

d) When identifying the influences of the nodes in the evaluation networks, LogAnalysis considered only the weights of the edges connected to these nodes. This is one of the key limitations of LogAnalysis.

e) SIIMCO and generated a much larger number of correctly predicted influences of nodes in the parts of the evaluation networks that have less dense connections. That is, they did not perform as well in dense parts. This is one of the key limitations of SIIMCO and ECLfinder.

Comparing IICCC with the real-world drug trafficking operation that identified the Caviar Canadian gang [6, 31] can give a supporting evidence of the possibility of applying the IICCC method to solve real-world practical problems. The similarities between IICCC and the Caviar operation can be summarized as follows:

- The Caviar operation constructed the network that depicts the drug traffickers based on the phone calls exchanged between them. Similarly, IICCC constructs a network depicting a criminal organization based on the phone calls exchanged between the criminals in the organization.
- The Caviar operation identified the influential criminals in the network (i.e., the criminals represented by nodes N1, N12, and N3) in order to monitor them and to gain insight into the traffickers' work. Similarly, IICCC identifies the influential criminals in a criminal organization in order to monitor them and to short-list the important communication channels originating from them. This monitoring can give an insight into the organization's work and the identity of a large number of the insiders and outsiders who deal with the organization. This will eventually lead to the arrest of these people, which will most likely result in at least crippling the organization's work.

VIII. CONCLUSION

Many techniques have been proposed to determine the relative importance of nodes. However, most these techniques do not generate useful short-lists of critical communication channels in a criminal organization. We have developed a forensics system called IICCC that can (1) infer the top influential criminals in a criminal organization, and (2) short-list the vital communication channels in the criminal organization. IICCC can create a network from either MCD or crime incident reports. IICCC employs the following concepts: (1) "significant communication paths", denoting sequences of influential criminals who propagate information diffused by the criminal under investigation to other criminals, and (2) "critical communication channels", which are sub-paths linking two adjacent influential criminals in the network and passing vital information from the top influential criminals to others lower in the hierarchy. IICCC identifies the influential criminals and significant communication paths by computing the χ^2 value of each node in a significant communication path.

We compared IICCC experimentally with LSH [40], SIIMCO [45], ECLfinder [41], CrimeNet Explorer [21], and LogAnalysis [13]. For the evaluations, we used the following datasets: Caviar dataset [6, 31], Enron email dataset [12, 24], and Krebs's 9/11 dataset [50, 51]. The experimental results showed that IICCC has a superior prediction performance.

The key contribution of the paper lies in identifying the vital communication channels in a criminal network. That is, the key contribution of IICCC is to generate useful short-lists of the critical communication channels in a criminal organization. This is because there are numerous communication channels in criminal networks, and investigating all the channels is time-consuming and distracting, when the investigators would prefer to focus on the key channels that reveal the most information.

The experimental results confirmed the usefulness of shortlisting critical communication channels. For example, in the Caviar dataset used in our experiments there are 86, 138, and 227 communication channels connected directly and indirectly to the most influential nodes designated by nodes N1, N12, and N3 respectively. Out of these channels, IICCC identified only 9, 5, and 14 as critical communication channels connected to N1, N12, and N3 respectively.

Even though the paper focusses on identifying the important communications paths in criminal networks, the framework of IICCC can also be used to solve many other practical real-world problems depicted in the form of networks, such as the following:

- In social networks, it can be used for detecting communities. A number of studies successfully identified the boundaries of communities based on the strength of the information flow in the paths/edges connecting their nodes [43].
- In metabolic networks, it can be used for identifying functionally-related units. A fully described path between two units represents the dynamics and dependencies among them [46].
- In energy conservation computation, it can be used for identifying energy-efficient paths. A path in an energy network that passes energy with high flow rate is likely to utilize energy effectively [38].
- In health care, it can be used for quantifying the impact of human mobility in spreading infectious diseases [53] (e.g., an influential path depicts a route with high human mobility).
- In urban planning, it can be used for identifying congested roads [53] (e.g., an influential path depicts a congested road).

REFERENCES

[1] A. Milani Fard, M. Ester, "Collaborative Mining in Multiple Social Networks Data for Criminal Group Discovery", *IEEE International Conference on Social Computing (SocialCom)*, 2009.

[2] Baker, W. E. and Faulkner, R. R. 1993. "The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry". *Amer. Soc. Rev. 58*, 837–860.

[3] Baldi, P. and Hatfield, W. (2002), "DNA Microarrays and Gene Expression", *Cambridge University Press, Cambridge, UK*.

[4] Borgatti, Stephen P. (2005). *"Centrality and Network Flow". Social Networks. Elsevier. 27: 55–71.*

[5] Breiger, R. L., Boorman, S. A., and Arabie, P. 1975. "An algorithm for clustering relational data, with applications to social network analysis and comparison with multidimensional scaling". *J. Math. Psych. 12*, 328–383.

[6] Bahulkar, A., Szymanski, B., Baycik, N., and Sharkey, T. "Community Detection with Edge Augmentation in Criminal Networks". 2018 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Spain, 2018.

[7] Chen, H. and Lynch, K. J. 1992. "Automatic construction of networks of concepts characterizing document databases". *IEEE Trans. Syst. Man Cybernet. 22*, 885–902.

[8] Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., and Schroeder, J. 2003. "Coplink: Managing law enforcement data and knowledge". *Commun. ACM 46*, 28–34.

[9] Caviar dataset. Available at: https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/caviar

[10] C. Morselli and K. Petit "Law-Enforcement Disruption of a Drug Importation Network", *Global Crime, vol. 8, Issue 2,* 2007.

[11] Díaz, C., Patacchini, E., Verdier, T., Zenou, Y. "The influence of leaders on criminal decisions". *VOX, CEPR's Policy Portal*, 2018.

[12] Enron Email Dataset. Available at: http://www-2.cs.cmu.edu/~enron/.

[13] E. Ferrara, P. De Meo, S. Catanese, and G. Fiumara, "Detecting criminal organizations in mobile phone networks," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5733–5750, 2014.

[14] Freeman, L. C. (1977). "A set of measures of centrality based on betweenness". *Sociometry,* 40(1), 35-41.

[15] Girvan, M., Newman, M. (2002). "Community structure in social and biological networks". *Proceedings of the National Academy of Sciences*, 99(12), 7821.

[16] Glsser, "Estimating Possible Criminal Organizations from Co-offending Data". *Public Safety Canada*, 2012.

[17] Greenwood, P. E., Nikulin, M.S. (1996), "A guide to chi-squared testing". *Wiley*, New York. ISBN 0-471-55779-X

[18] H. Sarvari, E. Abozinadah, A. Mbaziira, and D. McCoy, "Constructing and analyzing criminal networks", Proc. IWCC, USA, 2014, pp. 84–91.

[19] H. Wang, C. K. Chang, H.-I. Yang, and Y. Chen, "Estimating the relative importance of nodes in social networks," *Journal of Information Processing*, vol. 21, no. 3, pp. 414–422, 2013.

[20] J. Pattillo, N. Youssef, and S. Butenko, "Clique relaxation models in social network analysis," in *Handbook of Optimization in Complex Networks*. Springer, 2012, pp. 143–162.

[21] J. J. Xu and H. Chen, "CrimeNet explorer: A framework for criminal network knowledge discovery," *ACM Trans. Inf. Syst.*, vol. 23, no. 2, pp. 201–226, Apr. 2005.

[22] Klerks P., "The Network Paradigm Applied to Criminal Organisations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands", *Connections* 24(3): 53-65, 2001.

[23] Kathleen C., Matthew D., Max T., Jeffrey R., Natasha K. "Destabilizing Dynamic Covert Networks". *Proc. ICCRTS 2003*, Washington DC, USA.

[24] Keila, S. and D. Skillicorn (2005), "Structure in the Enron email dataset", *Computational & Mathematical Organization Theory*, 11(3), 183–99.

[25] L. Langohr, "Methods for finding interesting vertices in weighted

graphs," *Ph.D. dissertation*, 2014.

[26] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

[27] Leeson, Peter T., and Douglas B. Rogers. 2012. "Organizing Crime." *Supreme Court Economic Review* 20: 89-123.

[28] Mcandrew, D. 1999. The structural analysis of criminal networks. In *The Social Psychology of Crime: Groups, Teams, and Networks*. D. Canter and L. Alison, Eds. Dartmouth Publishing, Aldershot, UK, 53–94.

[29] M. Akbas, R. Avula, M. Bassiouni, and D. Turgut, "Social network generation and friend ranking based on mobile phone data" *IEEE International Conference on Communications, Budapest*, Hungary, 2013.

[30] Memon, Bisharat, "Identifying Important Nodes in Weighted Covert Networks Using Generalized Centrality Measures". 2012 *European Intelligence and Security Informatics Conference,* Odense, Denmark.

[31] Morselli, Carlo, & Cynthia Giguere, "Legitimate strengths in criminal networks", *Crime, Law and Social Change* 45(3), 2006, 185-200.

[32] Mansour, Abdala, Nicolas Marceau, and Steve Mongrain, "Gangs and Crime Deterrence," *Journal of Law, Economics & Organization*, Oct 2006, Vol. 22 Issue 2, p 315-339.

[33] Newman, M. (2004). "Fast algorithm for detecting community structure in networks". *Physical Review E*, 69(6), 066133.

[34] Nikulin, M.S. (1973). "Chi-squared test for normality". In: *Proceedings of the International Vilnius Conference on Probability Theory and Mathematical Statistics*, v.2, pp. 119–122.

[35] Phil W., "Transnational Criminal Networks". In *Networks and Netwars: The Future of Terror, Crime, and Militancy*, ed. Arquilla and Ronfeldt.

[36] P. A. C. Duijn, V. Kashirin, and P. M. A. Sloot, "The relative ineffectiveness of criminal network disruption," *Scientific Reports*, vol. 4, article 4238, 2014.

[37] P Dey, S Roy. "Centrality based information blocking and influence minimization in online social network". *2017 IEEE ANTS*, India.

[38] P. Plonski, P. Tokekar, V. Isler, "Energy-efficient path planning for solar-powered mobile robots". *J. of Field Robotics*," Vol.30, pp. 583-601, 2013.

[39] Shang, X., Yuan, Y. *Social Network Analysis in Multiple Social Networks Data for Criminal Group Discovery*. Proc. *2012 CyberC*, Sanya, China.

[40] S. M. Faisal, G. Tziantzioulis, A. M. Gok, N. Hardavellas, S. Ogrenci-Memik, and S. Parthasarathy, "Edge importance identification for energy efficient graph processing," in 2015 *IEEE Big Data*, 2015, pp. 347-354.

[41] Taha, K. and Yoo, P. "Using the Spanning Tree of a Criminal Network for Identifying its Leaders". *IEEE Transactions on Information Forensics & Security*, 2016, 12 (2), pp. 445 - 453.

[42] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "Introduction to Algorithms". *MIT Press, Cambridge*, MA, 2nd edition (2001).

[43] Taha, K. "Disjoint Community Detection in Networks based on the Relative Association of Members". *IEEE Transactions on Computational Social Systems*, 2018, Vol. 5, issue 2.

[44] Tversky *A: Features of Similarity*. *Psycholog. Rev* 1977, 84:327-352

[45] Taha, K. and Yoo, P. "SIIMCO: A Forensic Investigation Tool for Identifying the Influential Members of a Criminal Organization". *IEEE Transactions on Information Forensics & Security*, 2015, Vol. 11, issue 4, pp. 811 – 822.

[46] Taha, K. "Inferring the Functions of Proteins from the Interrelationships between Functional Categories". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016, Vol. 15, issue 1, pp. 157-167.

[47] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[48] U. Glsser, "Estimating Possible Criminal Organizations from Co-offending Data". *Public Safety Canada*, 2012.

[49] U. K. Wiil, J. Gniadek, N. Memon; "Measuring Link Importance in Terrorist Networks. Social Network Analysis", *International Conference On Advances in Social Networks Analysis and Mining, ASONAM 2010.*

[50] V. E. Krebs, "Uncloaking terrorist networks," *First Monday*, vol. 7, pp. 4–11, 2002.

[51] V. E. Krebs, "Mapping networks of terrorist cells," *Connections*, vol. 24 (3), pp. 43–52, 2002.

[52] [Victor E. Kappeler and Gary W. Potter. "The Mythology of Crime and Criminal Justice", *5th edition, Waveland Press*, Inc., 2017.

[53] Vazquez-Prokopec G et al. *"*Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment*". PLoS ONE 8, 2013, e58802.*

[54] Wasserman, S. and Faust, K. (1994). "Social Network Analysis: Methods and Applications". *Cambridge University Press*.

[55] Yang, L. "Based on social network crime organization relation mining and central figure determining". *2012 IEEE International Conference on Computer Science and Automation Engineering*, South Korea, June 2012.

Kamal Taha is an Associate Professor in the Department of Electrical and Computer Engineering at Khalifa University, UAE, since 2010. He received his Ph.D. in Computer Science from the University of Texas at Arlington, USA. He has over 80 refereed publications that have appeared in prestigious top ranked journals, conference proceedings, and book chapters. Over 20 of his publications have appeared in IEEE Transactions journals. He was as an Instructor of Computer Science at the University of Texas at Arlington, USA, from August 2008 to August 2010. He worked as Engineering Specialist for Seagate Technology, USA, from 1996 to 2005 *(Seagate is a leading computer disc drive manufacturer in the US)*. His research interests span bioinformatics, *defect* characterization of *semiconductor wafers*, Information Forensics & Security, information retrieval, data mining, and databases, with an emphasis on making data retrieval and exploration in emerging applications more effective, efficient, and robust. He serves as a member of the Program Committee, editorial board, and review panel for a number of international conferences and journals, some of which are IEEE and ACM journals. He is a Senior Member of the IEEE

Paul Yoo is currently with the CSIS within Birkbeck College at the University of London. Prior to this, he held academic/research posts in Cranfield (Defence Academy of the UK), Sydney (USyd), South Korea (KAIST) and the UAE (Khalifa). In his short career, he has amassed more than 60 prestigious journal and conference publications, has been awarded more than US$ 2.3 million in project funding, and a number of prestigious international and national awards for my work in advanced data analytics, machine learning and secure systems research, notably IEEE Outstanding Leadership Award, Capital Markets CRC Award, Emirates Foundation Research Award, and the ICT Fund Award. Most recently, he won the prestigious Samsung award for research to protect IoT devices. He serves as an Editor of IEEE COMML and Journal of Big Data Research (Elsevier). He is also affiliated with the University of Sydney and Korea Advanced Institute of Science and Technology (KAIST) as a Visiting Professor. He is a Senior Member of the IEEE.