20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

A first approach to resolving ambiguity in hidden terrorist group detection in communications networks

T. Bogomolov [a] **and B. Chiera** [a]

[a] *University of South Australia, SA 5001, Australia*
Email: Timofei.Bogomolov@unisa.edu.au

Abstract: One of the most challenging problems in detecting terrorist groups in communications networks is that of identity ambiguity. Node identification mechanisms for modern communications networks can range from a mobile phone number to an email address, IP-address or VoIP account name, meaning that terrorist group members can easily assume a new network identification or possess multiple identifications simultaneously. To compound matters, terrorists are also known to employ sparse pseudorandom communication patterns while maintaining constant connectivity between group members. We are interested to address the issue of correctly identifying members of a hidden terrorist group after a change of identification has occurred.

We propose a method we call *implied connections* as a first approach to resolving such ambiguity. We begin by collecting connectivity information over an m-size neighbourhood for each node in a network over a specified observation period. This information is then converted to a weighted graph of unobserved but potential relationships between nodes we call implied connections.

The method of implied connections is tested on two real-life dynamic networks derived from mobile phone and Internet traffic data consisting of approximately 10,000 and 2,000,000 unique nodes respectively. For each network type we construct two graphs of implied connections to capture network characteristics of interest before and after a suspected change of identification has occurred. We then adapt a method of inexact subgraphs similarity and calculate β-signatures for both the subgraphs of implied connections of the network member with a new identity as well as potential candidates for the original network identity. As such, a β-signature is calculated as a column vector of probabilities of the next connection of a chosen network member to any other member of their m-size neighbourhood.

Following calculation of the β-signatures, we find the best match between subgraphs of implied connections based on Euclidean distance as defined in the multi-dimensional space of the subgraphs members. Individuals whose subgraphs are of higher similarity, that is, have shorter a distance between their β-signatures, are considered more likely to be the same person.

Our results indicate that an analysis of implied connections improves the characterisation of the relationships between nodes and substantially increases the probability of correctly identifying members of the terrorist group after an identity change has occurred.

Keywords: *Hidden network, ambiguity, implied connections, dynamic graph series*

"Tell me who's your friend and I'll tell you who you are."— Proverb, Russian

1 Introduction

The terrorist attacks of Tokyo (March 20, 1995), New York (September 11, 2001) and London (July 7, 2005) demonstrate the terrifying reality of the modern world where the actions of relatively small terrorist groups can have devastating effects. One difficulty in preventing such events is the enormous challenge of detecting the presence of a terrorist group during the planning phase of an attack. Terrorists may attempt to camouflage their connectivity by adopting a communications strategy to appear randomly connected to one another and/or the general communications network. In actuality, all terrorist group members may remain persistently connected over a number of non-homogeneous time instants in spite of the change in the flow of communication. Such behaviour gives rise to what is known as a hidden terrorist group.

A second, complementary problem in detecting hidden terrorist groups in communications networks is that of identity ambiguity. Node identifications used in modern communications networks can range from a mobile phone number to an email address, IP-address or VoIP account name, meaning that terrorist group members are able to easily assume a new network identification or possess multiple identifications simultaneously. Content analysis of communications is often infeasible or impossible and thus node identifications are the most readily available means of analysing communications. Given the limitations of this scenario, social network techniques are popularly employed to detect hidden groups [Baumes et al., 2004, 2006].

A similar problem known as router aliasing arises in studies of Internet topology [Magoni and Hoerdt, 2005; Gunes and Sarac, 2009] in which a number of unique IP-addresses might belong to the same router, however would appear as individual and potentially independent nodes. The methods developed for router alias resolution are limited in application to the ambiguity problem described here as these methods rely on cooperation with routers. For example routers will readily answer a large number of structured requests such as *traceroute* and its variants, with these responses able to be further analysed. We can not expect a similar level of co-operation from potential terrorist suspects.

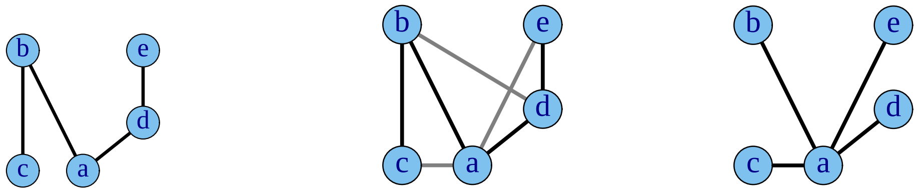
In this paper we introduce the *method of implied connections* as a means to address ambiguity resolution of hidden terrorist groups. This method expands an observed network, seen as a proxy for an unobserved relations network, by considering an extended network neighbourhood of *potential* communications. We examine the scenario in which members of a communications network are persistently connected by some prescribed shortest distance, that is the number of hops between network members, and are thus thought to share strong implied connections. As such we might suspect the existence of a hidden relationship between these individuals even if direct communication is never observed. Since connections between network members are a function of their relations, which are not readily observable in their entirety, any observed communications network will most likely be incomplete. While it can be said that a direct link between two nodes is indicative of a relationship between the two individuals, the opposite is not true. That is, the absence of a direct connection does not reflect a lack of relationship between individuals. Even though direct communication between network members may be sparse or random, it is reasonable to expect their hidden relationships are stable.

To resolve the ambiguity problem, we adapt a graph similarity technique commonly used for approximate subgraph matching [Amin et al., 2012], namely when a new individual appears in the network, a network snapshot is taken and we attempt to match a subgraph of this individual's implied connections to the subgraphs of network members derived from a previous snapshot. We calculate the Euclidean distance of the β-signatures for each pair of subgraphs in the multi-dimensional space defined by the network members, with a short distance indicating that two subgraphs are more likely to belong to the same person.

We test our approach on two dynamic real life networks comprised of mobile phone [Eagle et al., 2009] and Internet traffic [CAIDA, 2012] data. We show that any member of the network can be matched to their original identity with reasonably high probability and that there is a difference between the required approach when the person with the new identity is a terrorist or a non-terrorist.

2 The Method of Implied Connections

We begin by observing a dynamic communications network over a collection of predefined time intervals we call *observation cycles*, with cycle lengths depending on the intensity of network communications. Observation cycles can be taken as long as one day for a small mobile phone network, or as short as one minute, in the case of Internet traffic data. An assumption commonly used in the literature is that all members of the hidden terrorist group communicate with at least one other member during an observation cycle [Baumes et al., 2004].



$$A(a) = \begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1 & 0 & 1 & 0 \\ b & 1 & 0 & 1 & 0 & 0 \\ c & 0 & 1 & 0 & 0 & 0 \\ d & 1 & 0 & 0 & 0 & 1 \\ e & 0 & 0 & 0 & 1 & 0 \end{array} \quad A^2 = \begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1 & 1 & 1 & 1 \\ b & 1 & 0 & 1 & 1 & 0 \\ c & 1 & 1 & 0 & 0 & 0 \\ d & 1 & 1 & 0 & 0 & 1 \\ e & 1 & 0 & 0 & 1 & 0 \end{array} \quad A^2(a) = \begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1 & 1 & 1 & 1 \\ b & 1 & 0 & 0 & 0 & 0 \\ c & 1 & 0 & 0 & 0 & 0 \\ d & 1 & 0 & 0 & 0 & 0 \\ e & 1 & 0 & 0 & 0 & 0 \end{array}$$

(a) Two-step neighbourhood of *a* (b) Order 2 implied connections (c) Subgraph of implied connections of *a*

Figure 1: Building a subgraph of implied connections for a chosen node

To deduce a relations network from the general communications network, we introduce the concept of *implied connections* which is a set of equally important direct and indirect links. To define these connections we employ an m-reach measure, $m \geq 1$, [Borgatti, 2006] such that any pair of network members with shortest distance bounded by m are considered to be connected. Alternatively, we might say that if two nodes are in an m-size neighbourhood then there exists an implied connection between them.

For a communications network $G(t)$ we can compute the m-reach matrix $A^m(t)$ ($m \geq 1$) for each observation cycle $t = 1, ..., T,\ T \geq 1$, from which we calculate $A^m_T = \sum_{t=1}^{T} A^m(t)$ to yield a weighted adjacency matrix to represent the implied connections. The graph of implied connections, $G^m(T)$, can be determined directly from A^m_T. The weights of the graph edges of $G^m(T)$ indicates the strength of an implied connection between nodes, with weights ranging from 0 (no communication between two nodes of interest during the observation period) to T (two nodes of interest are in communication during each time cycle $t = 1, \dots, T$).

If an implied connection persists over a number of observation cycles, reflected by an edge with a high weighting, this can imply the existence of a relationship between two nodes even if there is no observed direct communication between the nodes involved. Such an implied relationship can occur legitimately in real life — for example, a wife may call both her husband and mother regularly, thereby linking the two individuals known to one another, even if the husband does not directly call his mother-in-law. An alternative explanation pertinent to the focus here is that these implied connections may indicate the presence of a hidden terrorist group with members attempting to avoid direct contact as much as possible, in order to hide their existence. The question of how to distinguish between these two types of groups is further explored in what follows.

The method of implied connections is an egocentric analysis meaning that a subgraph of implied connections for any network member will always be a star network, that is a network in which one individual forms the centre of the network, with only direct connections between this individual and each member of the network neighbourhood. Individual interaction between counterparts of the network member-of-interest can be discarded since all information relevant to this member is included in the implied connections. Figure 1 provides an example of how to build a subgraph of implied connections of order two for the node of interest, *a*, from the network members *a* – *e*. Figure 1 (a) shows the two-step neighbourhood of node *a* and the accompanying adjacency matrix $A(a)$; Figure 1 (b) depicts the implied connections for $m = 2$ from the two-step neighbourhood of *a*; and Figure 1 (c) shows only the retained edges that are connected to *a*. The edges in Figure 1 (c) are considered equally important irrespective of whether they are direct or implied and thus result in a subgraph of implied connections of order $m = 2$ for node *a* with adjacency matrix $A^2(a)$.

A subgraph of the implied connections of any network member can be characterised by a type of column-vector called a β-signature. Amin et al. [2012] introduce the β-signature as a tool to measure the similarity of subgraphs which they calculate using a random walk with restart. However for any star network, the β-signature can be calculated explicitly as the corresponding column of the column-normalised adjacency matrix A^m_T and is interpreted as containing the transition probabilities from the node of interest to any other node in the network. Here, these transitions correspond to relations between the member-of-interest and others in the communications network. The β-signature of the subgraph of implied connections of node a in Figure 1 (c) is

$$\beta(a) = \begin{array}{c} \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} a \\ \left(\begin{array}{c} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{array}\right) \end{array}$$

as all edges have equal weighting.

If, during some observation cycle t a new member appears in the network, we might suspect that this individual has already appeared in the network, albeit under a different identity. To test this suspicion we can take the subgraph G_{new} of all order m implied connections for the new member, with the vertices of G_{new} providing the set $\mathcal{P} = \{p_1, p_2, ..., p_k\}$ of 'friends' of this new member. Then, we check the network history for the observation cycles $t - n, \quad n = 1, ..., T$ before the suspected identity change and take the subgraph G_{old} of implied connections for each member of $\mathcal{P}$. The vertices of G_{old} yields the set $\mathcal{R} = \{r_1, r_2, ..., r_l\}$ which lists 'friends' of the 'friends' of the new member. Assuming that the behavioural pattern of the new member does not alter after an identity change and that they do not contact themselves, we might expect that the previous incarnation of the new member is a member of a set $\mathcal{S} = \mathcal{R} - \mathcal{R} \cap \mathcal{P}$.

We can then take the distance $d(v_1, v_2)$ between the new member v_1 and every member $v_2 \in \mathcal{S}$ in the space of members of $\mathcal{R} \cup \mathcal{P}$

$$d(v_1, v_2) = \sqrt{\sum_{i \in (\mathcal{R} \cup \mathcal{P})} (a_i - b_i)^2}$$

where v_1, the new member-of-interest in graph G_{new}, has β-signature $\beta(v_1) = (a_1, a_2, ..., a_k)$ and v_2, the member of $\mathcal{S}$ who appeared in graph G_{old}, has β-signature $\beta(v_2) = (b_1, b_2, ..., b_k), \ k = | \mathcal{R} \cup \mathcal{P} |$.

We can also calculate the degree of similarity $P(v_1, v_2) = 1 - d(v_1, v_2)$ between the subgraphs of implied connections of nodes v_1 and v_2 [Amin et al., 2012]. The similarity measure $P(v_1, v_2)$ takes values from 0 to 1 and can be interpreted as the probability that nodes v_1 and v_2 are the same individual but with different identities. Thus a member of $\mathcal{S}$ with the lowest distance $d(v_1, v_2)$ and respectively highest probability $P(v_1, v_2)$ is most likely the alter ego of the new member of the network.

3 Testing on Real World Networks

To test the method of implied connections we employ two real life dynamic data sets. The first is the mobile phone data set collected by the Reality Mining project as conducted by the MIT Media Laboratory [Eagle et al., 2009]. The project ran from September 2004 until June 2005 with 94 subjects, comprised of under- and postgraduate students and faculty, and used Nokia 6600 phones equipped with a purposely designed logging application. We take daily snapshots of the communications activity including incoming and outgoing answered and non-answered calls as well as SMS, to build communications networks for observations cycles $t = 1, \ldots, T$. The total number of unique phone numbers recorded is 10,056 while the average number of nodes in the daily network snapshots is 250 with a standard deviation of 88.8. The average number of unique edges is 239 with a standard deviation of 88.4.

The second data set is Internet traffic data provided by The Cooperative Association for Internet Data Analysis [CAIDA, 2012]. The data repository contains one minute data sets of passive traffic traces from the equinix-sanjose monitor on a high-speed Internet backbone link. In our analysis we consider a sub-graph of User Datagram Protocol (UDP) traffic, with each of these one-minute data sets consisting of, on average, 300,000 nodes and 290,000 links.

In our tests we consider two scenarios: (1) when the network contains a hidden terrorist group embedded within it; and (2) when there is no terrorist network in the larger communications network. Testing an identity change when there is no terrorist network is relatively straightforward. We randomly select a graph $G(t)$ from one of the observation cycles from each data set and from these selected graphs we randomly choose a network member who will change identity and thus becomes a person of interest. To perform matching using the distance method we take a history of T observation cycles immediately prior to $G(t)$ and try to find the best match for the individual who has adopted a new identity. An implicit assumption made in the identity change of a non-terrorist is that they would behave differently to terrorists belonging to a hidden group. We would expect a non-terrorist would have nothing to hide and as such, makes direct calls whenever needed.

We next consider the situation where there is a terrorist group embedded within the larger communications network. Previous analyses of terrorist attacks [Krebs, 2002; Koschade, 2006] suggest that hidden networks of terrorists are small in nature. In Cruickshank and Ali [2007], Mustafa Setmariam, an architect of Al Qaeda structure, is quoted as providing direct instructions to terrorist recruits to act autonomously and not exceed ten group members in size. We arbitrarily choose a terrorist group to be of size six and to fulfil the definition of a hidden terrorist group [Baumes et al., 2004] we create a communications network to form the hidden group in each time cycle. We extend the work of Lindelauf et al. [2009] which provides the optimal structure of a covert group G_H that maximises a total performance measure $\mu(G_H)$ capturing a tradeoff between secrecy and the need for communication

$$\mu(G_H) = S(G_H)I(G_H) = \frac{n^2 - n - 2e}{n^2} \cdot \frac{n(n-1)}{T(G_H)}$$

where $S(G_H)$ and $I(G_H)$ are secrecy and information measures respectively; n is the number of nodes; e is number of edges; and $T(G_H)$ is the sum of the shortest distances between all nodes in the hidden group G_H.

We are interested to use a structure potentially that is less than optimal, since the optimal structure for a covert network is known to be a star [Lindelauf et al., 2009] and it is unlikely, although not impossible, that a real life hidden group would achieve an optimal structure for their communications. Rather, we assume the hidden group structure will differ to yield a network structure in which the group members intuitively try to minimise the number of communications but at the same time still guarantee an effective flow of information. To produce such a network, we begin with a complete graph G_H of the hidden group of size n. We select an edge at random and delete the edge if its removal increases the total performance measure $\mu(G_H)$ and continue to remove edges until there is no possibility for further improvement to the total performance measure. The result will be the hidden group G'_H which is potentially less than optimal in structure.

To embed a hidden group in either the mobile phone or Internet traffic networks in observation cycle $t = 1$ we randomly select five actual members of each network and re-assign their status as members of the hidden group. We introduce a sixth, artificial member to complete the hidden group, add the connections from G'_H and assign this artificial member as the person of interest. We repeat this process for observation cycles $t = 2, \ldots, 5$ before we enter the next observation cycle $t = 6$ in which the person of interest changes identity. At this point we try to find the best match for this newly appearing member amongst all network members, including the artificial member observed during the five previous cycles $t = 1, \ldots, 5$.

It is apparent that this approach to identity matching depends on a number of behavioural factors, namely how frequently the new member and their counterparts appear in the network. If the selected member does not appear in the network during the first five observation cycles then the chance of correctly identifying the new individual is zero. The same result occurs for the other members of the hidden group if they also do not appear in the first five observation cycles, making it impossible to match a new individual to their previous identity. To overcome these potential problems and test the boundaries of the method of implied connections, we apply an extra condition for selecting members of the communications network: we ensure we select individuals from the subset of members who appear in the network in at least 2, 4 or 6 observation cycles, including the last cycle when the change of identity occurs.

4 Results

To conduct the analysis we limit our observation period to $T = 5$ cycles before a change of identity on the sixth cycle, with cycles defined as days or minutes for the Reality Mining and Internet traffic data respectively. We chose $T = 5$ as we wish to demonstrate the method in a reasonably difficult situation, whereas for T larger both before and after an identification change, we would expect ambiguity resolution to be less challenging.

For both data sets, we repeat subgraph matching of implied connections 1,000 times for both scenarios in which a terrorist network is and is not embedded in the general communications network. To allow for discontinuity in either communications behaviour or observed communications, we consider the scenarios where members of the hidden terrorist group appear over 2 or 4 observation cycles as well as the best-case scenario with each member of the hidden terrorist group appearing during each observation cycle. Figures 2 and 3 provide the averaged results for all these cases.

To determine the success rate of the method, we first consider the most stringent scenario, in which we consider ambiguity resolution successful if the individual with the highest probability $P(v_1, v_2)$, that is the highest subgraph similarity, is also the individual with the new identity. We then relax this criterion to successively

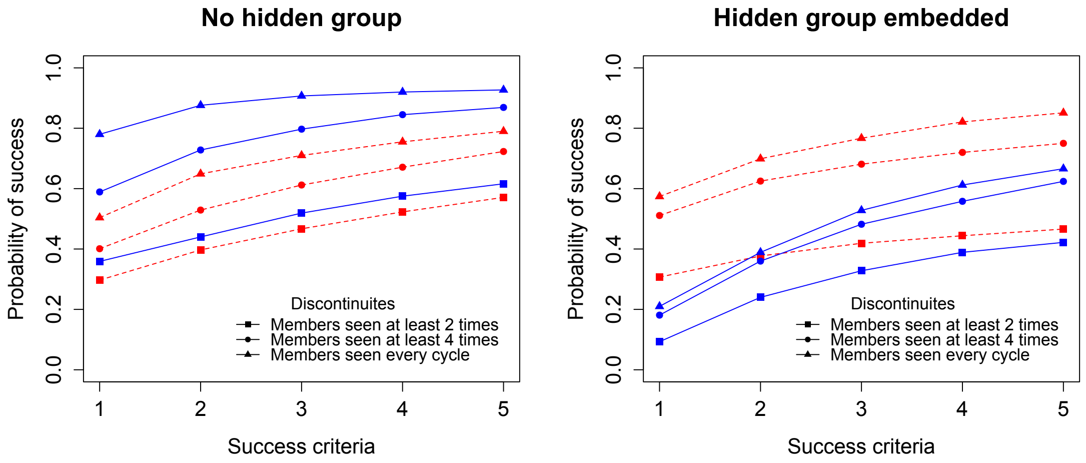


Figure 2: Reality Mining data. Implied connections of order 1 (–) and order 2 (- -) are shown.

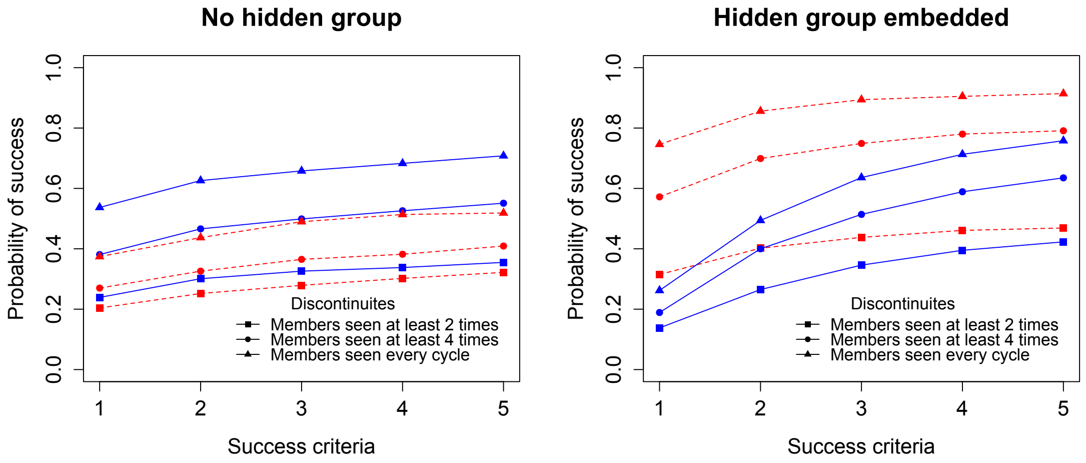


Figure 3: Internet traffic data. Implied connections of order 1 (–) and order 2 (- -) are shown.

consider the highest two, three, four and five similarities, to provide a pool of suspects of corresponding sizes. If the individual with a new identity appears in these pools, a success is counted.

From Figures 2 and 3 we see that both data sets demonstrate a similar pattern in results. In the event there is no hidden group, an order 1 connection provides the best result in resolving the ambiguity problem considered here. People who are non-suspicious members of the communications network do not hide their connectivity and, as a result, can be identified by their direct connections. In this case, using implied connections of at least order 2 does not add any further useful connectivity information. Rather, by exploring higher orders we introduce noise into the analysis that serves to reduce the effectiveness of matching individuals. It is worth noting that the analysis of mobility traces in the mobile phones network by de Montjoye et al. [2013] demonstrates similar results to those presented here, namely that it is possible to uniquely identify 95% of the individuals in a network, although in de Montjoye et al. [2013] spatio-temporal points given by the mobile carrier's antennas were used to achieve this objective.

In the case where there is a hidden group embedded and we now have members of the network trying to camouflage their connectivity, the probability of successful identification of these members through the analysis of direct connections reduces by a factor of 1.5–2. It seems the strategy of hidden group members attempting to appear randomly connected both successfully conceals their identity whilst maintaining regular contact within the group. Having said that, it is possible to instead identify members of the hidden group by analysing implied connections of an order higher than 1. From Figures 2 and 3 we can see that there is a substantial improvement in ambiguity resolution when using order 2 implied connections with the probability to pinpoint a member-of-interest trebling. Finally, in all cases we see that the success of the method depends on the quality of the network history. If there is any discontinuity in the communications pattern of the hidden group then the probability of success will decline. The more active both a network member-of-interest and their immediate connections, the more easily ambiguity in network identification can be resolved.

5 Conclusions

In this paper we consider the issue of ambiguity resolution in a communications network when an individual from a hidden terrorist group changes their observed identity. To combat this problem we introduce the method of implied connections as a proxy for unobserved relations in a communications network. We use a simplified version of a subgraph similarity measure together with Euclidean distance, to identify the most likely set of individuals from which the individual with an identity change could be chosen. We tested this approach on two real-life data sets based on mobile phone and Internet traffic data over a number of observation cycles.

Our results indicate it is possible to match, with a reasonably high probability, the new and old identities of a member of the communications network. Non-terrorist individuals in the network can be identified easily using direct connections only, however members of a hidden terrorist group use a pseudorandom communication pattern and as such appear to successfully conceal their identity if only direct connections are considered. Analysis of their implied connections of a higher order focuses on the possible relations between members of the hidden group rather than their communications structure and as such is immune to the pseudorandom character of their communications. For the type of ambiguity resolution problem considered here, we are able to demonstrate that for the two data sets under consideration, the ability to successfully match the new and old identities of the same terrorist individual trebles with increasing activity in the network.

Acknowledgement

This research was supported under DSTO contract number PG 080579.

References

Amin, M. S., R. L. Finley, Jr., and H. M. Jamil (2012, November). Top-k similar graph matching using tram in biological networks. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 9*(6), 1790–1804.

Baumes, J., M. Goldberg, M. Hayvanovych, M. Magdon-Ismail, W. Wallace, and M. Zaki (2006). Finding hidden group structure in a stream of communications. *Intelligence and Security Informatics*, 201–212.

Baumes, J., M. Goldberg, M. Magdon-Ismail, and W. A. Wallace (2004). Discovering hidden groups in communication networks. In H. Chen, R. Moore, D. Zeng, and J. Leavitt (Eds.), *Intelligence and Security Informatics*, Volume 3073 of *Lecture Notes in Computer Science*, pp. 378–389. Springer Berlin Heidelberg.

Borgatti, S. P. (2006). Identifying sets of key players in a social network. *Computational and Mathematical Organization Theory 12*(1), 21–34.

CAIDA (2012). Anonymized internet traces.

Cruickshank, P. and M. Ali (2007). Abu musab al suri: Architect of the new al qaeda. *Studies in Conflict & Terrorism 30*(1), 1–14.

de Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen, and V. D. Blondel (2013). Unique in the crowd: The privacy bounds of human mobility. *Nature. Scientific reports 3*.

Eagle, N., A. S. Pentland, and D. Lazer (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences 106*(36), 15274–15278.

Gunes, M. H. and K. Sarac (2009, December). Resolving ip aliases in building traceroute-based internet maps. *IEEE/ACM Trans. Netw. 17*(6), 1738–1751.

Koschade, S. (2006). A social network analysis of jemaah islamiyah: The applications to counterterrorism and intelligence. *Studies in Conflict and Terrorism 29*(6), 559–575.

Krebs, V. (2002). Mapping networks of terrorist cells. *Connections 24*(3), 43–52.

Lindelauf, R., P. Borm, and H. Hamers (2009). The influence of secrecy on the communication structure of covert networks. *Social Networks 31*(2), 126–137.

Magoni, D. and M. Hoerdt (2005). Internet core topology mapping and analysis. *Computer Communications 28*(5), 494–506.