Social Networks xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

A graph database framework for covert network analysis: An application to the Islamic State network in Europe

Alexander Gutfraind [a,*], Michael Genkin [b]

[a] University of Illinois at Chicago, Loyola University Medical Center, Uptake Technologies, Inc., United States
[b] School of Social Sciences, Singapore Management University, Singapore

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Covert network
Terrorist network
Graph database
Terrorism
Islamic State

1. Introduction

How could a covert social network be measured with reasonable confidence? Unlike data collection for overt social networks, it is rarely possible, in covert networks, to collect respondent-generated ties via name generators and name interpreters or even to measure ties in a uniform manner. Instead researchers are forced to infer and code their network structure from incomplete, indirect, and uncertain information (Gerdes, 2015). Multiple coding decisions need to be made regarding how the ties between nodes should be measured in addition to which nodes to include. This introduces subjectivity and affects reproducibility. To some extent this is the bane of all network studies, but covert networks pose a special challenge given the secondary nature of the data.

To address this problem, the paper proposes and demonstrates a framework to encoding and analyzing covert networks, which has the advantage of systematically accounting for the uncertainty and ambiguity of covert network data. The framework relies on newly-developed methods from graph database theory (see Angles and Gutierrez 2008; Robinson et al., 2015). The approach allows the researcher to encode raw data in multi-modal form and then use the powerful tools of graph databases to project the data into a social network. Once the network has been projected from the graph database, it is analyzed using existing methods of social network analysis. The framework allows the user to conduct a sensitivity analysis to examine whether the network structure depends upon the way the researcher chose to generate their network in terms of how ties were recorded, how nodes were counted, and how sources for the information were weighted.

This paper is organized as follows. First, we begin by describing the problem of using covert network data. Next, we introduce the graph database as a tool for studying covert social networks and compare it to more familiar methods. Third, we introduce sensitivity analysis based on the graph database framework. Fourth, we apply this method to the Islamic State (IS) network in Europe that was responsible for the November 2015 Paris Attack and the March 2016 Brussels Attack and compare it to other covert networks. Finally, we conclude by discussing the findings on characterizing this network, the use of graph databases for covert network analysis, as well as the broader implications of graph databases for developing social network data standards and its potential for big data analytics.

1.1. The Challenges of Covert Network Data

Social network analysis faces a variety of unique data challenges, especially when it comes to sociocentric data (see Borgatti et al., 2013; Robins 2015).These range from specifying boundaries (Laumann et al., 1989), to adequate sample size (Costenbader and Valente, 2003; Frank, 2011) to issues with various forms of missing data (Kossinets, 2006). The problems are amplified with covert networks because the actors are consciously attempting to conceal or falsify their identities, their attributes, or their connections (Gerdes, 2015). As a result, the instruments of collecting network data directly from respondents such as name generators or name interpreters are usually not available. Data on covert networks typically come from secondary sources. In the case of terrorist or

* Corresponding author at: University of Illinois at Chicago, 1603 W. Taylor St, Chicago, IL 60612, United States.
E-mail address: agutfraind.research@gmail.com (A. Gutfraind).

http://dx.doi.org/10.1016/j.socnet.2016.10.004

organized crime networks the data are inferred retroactively from sources such as physical and electronic surveillance documents (Morselli, 2009); court records (Baker and Faulkner, 1993); or news reports (Krebs, 2002). As with any data, the analyst is forced to make coding decisions before social network analysis can even begin. There are two issues that arise, which should be kept separate. The first is at the level of analysis and the second is at the level of measurement.

First, there is the issue of what counts as a tie and to what extent should different types of ties be distinguished from one another. Some researchers have criticized the tendency in covert network analysis to amalgamate multiplex ties such as kinship, friendship, and organizational roles into a uniplex relation (see Gerdes 2015, chapter 2). For example, if actor A has a family tie to actor B as well as an operational tie with him, the combined uniplex relation is often given a double score. Doing so involves throwing out information and assigning weights arbitrarily, thus introducing the analyst's own biases (ibid). Lumping diverse relationships between actors as equivalent "ties" without theoretical justification introduces bias at the level of coding analysis.

The second issue is whether the nodes included and their relationships are in fact correctly measured. There is always the problem of missing data and whether certain nodes or relations are simply not known – the false negatives. But there is also the problem of how accurate are the relationships that are "known" – the false positives. Measurement error is especially pronounced for covert networks due to the secondary nature of the data collection. This paper introduces a framework that seeks to reduce measurement error, though it address both measurement and analysis errors.

For the purposes of this paper we refer to the problem of measuring and representing relationships in a covert network as the problem of covert network forensics. Much like a detective trying to piece together the details of a crime while sifting through a multitude of clues, covert networks forensics involves the systematic piecing together of information about relationships between nodes. Data about a covert network are usually obtained after the network has already carried out its mission. In the case of assembling information from disparate sources, there is a great deal of uncertainty as to which ties and nodes to include. Thus, what is needed is a framework that has the following four properties.

Documentation

The ability to link specific network elements (nodes or ties) to the raw source data as metadata. In this way the data are precisely documented and can be reproduced and easily re-examined.

Complex representation

The ability to store diverse entities and relations from multiple modes along with data on multiple types of edges. This is useful as the available sources often describe multi-modal relations.[1]

Reproducibility

The ability to efficiently reproduce coding decisions by other researchers. Ideally all the coding decisions are documented in a command script file.

Multiple projection

The ability to use raw data in multiple ways when deriving the network's nodes and edges. This allows the analyst to rerun her analysis and to test the robustness of one's network to different network-generating assumptions. This is especially important for covert networks because there are many arbitrary decisions that are made regarding tie measurement.

Traditional packages for social network analysis are not well-suited for these tasks because it is cumbersome to store so many different kinds of data: multiple edge attributes, multiple modes, and non-network data. Indeed many covert network datasets that are publicly available, even those derived from open sources, are poorly documented and very difficult to reproduce.

1.2. Graph databases

To respond to these challenges, this paper uses the theory of graph databases, a relatively new methodology from computer science, to perform covert network forensics. Generally, a graph database is a knowledge representation system which codes knowledge using nodes and edges, rather than storing tables of rows and columns as found in conventional databases (Angles and Gutierrez 2008; Robinson et al., 2015). Graph databases have recently matured as a technology and are capable of exceeding, in some respects, methods of conventional databases, even when working with data such as tables, text, and images. Because the internal structure of graph databases is in the form of a network, they are very well-suited to represent the information about covert networks, including members, activities, events and the relationships among them, as well as the attributes of the entities and relationships (Fig. 1). Graph databases can also contain meta-information *about* the nodes and edges, including the full text, image or movie containing the evidence, and this information can be considered when reconstructing the covert network.

Software implementations of graph databases are also equipped with a powerful and versatile query language for extracting information, analogous to how relational databases are queried using the SQL language. The query language of the graph database allows systematic extraction of information from the available raw data – a method not possible using social network analysis packages. Using a query, it is possible to determine the members of a covert network, and the relationships between them by listing only members and relationships that satisfy a particular condition. Furthermore, as new data are added to the database, the query can instantly (i.e., with one command) report the updated structure of the covert network based on any new information. The query also facilitates easy sensitivity analysis to determine whether the covert network is dependent upon the way the data are being coded (Kossinets, 2006). The social network resulting from the database query is then visualized and analyzed using existing social network tools and methods.

While the most advanced currently available network software packages have the ability to store multiple types of nodes and edges as well as their attributes, it is not generally possible in these systems to perform automated extraction of network information based on complex logical criteria. By contrast, the graph database is well-suited to store and query detailed temporal, spatial and other attributes about the nodes and relations. For example, a query can list all pairs of persons who were present in the same location within 7 days of a terrorist attack and who also visited a specific country in the past 5 years. In general, a graph database query returns any logical path (a set of one or more linked entities) which satisfies any specified logical conditions (Zouzias et al., 2014), while existing frameworks are designed for filtering only single nodes and relationships (e.g., all nodes of a particular kind).

To summarize, the methodology underpinning this paper provides the following advantages over existing frameworks. First, the data, including raw documents and extracted attributes is represented in a graph database, minimizing the need to make most

[1] For example, suppose node A and node B shared an apartment in Molenbeek while node C and node D lived in the same area of Paris. If both pairs are coded as "co-presence", one throws away the data that indicates co-presence with respect to specific locations (e.g. Molenbeek, Paris). This information might be relevant for subsequent analysis or by other researchers. While it is possible for the analyst to go back to her raw data and re-code or for other researchers to contact the original coder for their raw data, it is not efficient to do so. Ideally, the analyst should input as much as possible of the relevant raw data into machine-readable format while throwing away the least possible amount of information, at the stage of data coding.

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004

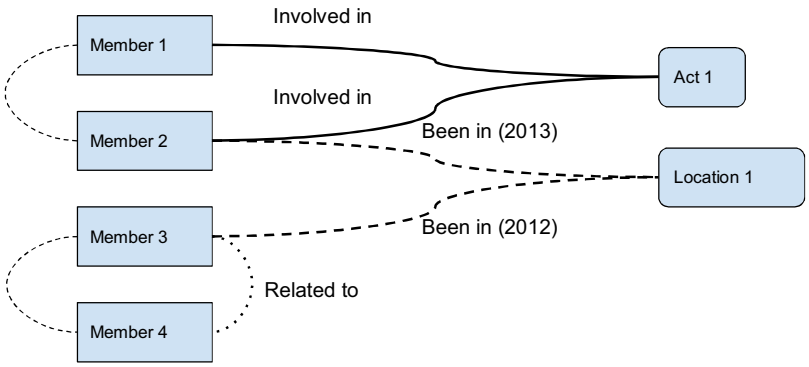

Fig. 1. Example of a graph database representation of a four-node covert network, with inferred ties (arcs on the left side). Tie 1–2 (from member 1 to member 2) is inferred because of their involvement in the same act. Members 2 and 3 were in the same location, but at a different year, and are thus not tied. Tie 3–4 is inferred because members 3 and 4 have a kinship tie to each other. The graph database often stores multiple attributes which cannot be visualized here, but which may be important for covert network forensics.

coding decisions. Second, the database query language provides a way of extracting complex structural information in a reproducible way, whereas existing methods are limited to simpler forms of filtering. Finally, the combination of richer data and better queries also enables a more comprehensive set of projections and sensitivity analyses. Although current network analysis packages, such as *statnet* (Handcock et al., 2003) can, in many cases, support similar analysis, they lack the ability to store and efficiently query large heterogeneous datasets (>10 GB) in raw form, making it difficult to work with raw event data.

Our survey of the social network and covert network literature has found that the graph database method is substantially novel, with only a handful of papers using the graph database framework for network analysis, largely by engineers and computer scientists, and no published studies have applied it to covert networks. The graph database method could be viewed as a generalization of multi-modal network analysis, as discussed in Breiger et al. (2014) and Lindelauf et al. (2012), who argued that terrorist networks could be represented as a multi-modal networks linking individuals based on their affiliation to missions or targets. It also extends the method of multi-relational (also known as multiplex) network analysis, common in covert network analysis (Everton, 2012; Xu and Chen, 2008), in which the network allows for different types of relationships (family, friendship, recruiter, etc.) but all the nodes in the network are of one type, such as persons. This is distinguished from a graph database framework advocated here in which the nodes are of several types, both human actors and non-human nodes (locations, events, attacks) and are connected with multiple types of relationships. The frameworks are compared in Table 1. Once data are stored in the graph database, it is usually projected into the multi-relational or the simple graph framework and then analyzed as a social network. Therefore, the graph database framework is not a new method for *analyzing* multiplex multimodal networks, but rather, its key contribution is to enable the encoding of raw data and then deriving the network from it.

We apply the graph database framework to two major recent terrorist events. On November 13, 2015 a set of coordinated attacks in Paris, France took the lives of 130 people and caused more than 350 injuries. Just months later, on March 22, 2016 the same network struck transportation hubs in Brussels killing 32 and injuring over 300. The attacks are considered one of the worst terrorist atrocities in Europe, and one of the deadliest attacks in an OECD country since the 9/11 attack in the US (START, 2013). The attacks were organized by the Islamic State group (IS), and carried out by a network we refer to as IS in Europe (or IS-E). IS-E has been operating in Western Europe for several years, and could be linked to several smaller operations: the attack on the Jewish Museum of Belgium (May 2014), the interdicted plot in Verviers (January 2015)(Crilly, 2015), and the high-speed train attack on the Brussels-Paris line (August 2015) (Parlapiano et al., 2015). However, IS-E does not include the Charlie Hebdo attacks, which are attributed to an unrelated group affiliated with al-Qaida (Schmitt et al., 2015) nor does it include acts by self-starter Islamists who are not directed by or affiliated with any known terrorist organization.

Because of the extensive interest in the Paris attacks and related plots, considerable information is available about the IS-E network in the open media. Based on this data, a graph database was built and it includes IS-E members, attacks, terrorist activities, sites, locations, and countries (see Methods). Following the creation of the database, we inferentially obtained a social network between the human actors and analyzed this IS-E network. Specifically, we aimed to provide a visual summary of the overall plot, and to understand the structure of the social network that made the attack possible.

We evaluated the IS-E network using measures like density, clustering, diameter, and several specialized measures such as secrecy. We then performed a sensitivity analysis to evaluate whether the measurements depend on how the ties are inferred. Finally, we compared the IS-E network to other covert organizational networks. We found that IS-E has lower mean degree compared to al-Qaida, as well as, intriguingly – lower secrecy (as defined in Lindelauf et al., 2009).

To sum, our main contribution is to propose a new methodology for systematic representation and analysis of covert networks, and secondarily, to map the Islamic Sate network in Europe. A practical outcome of this work for counter-terrorism is to better understand the structure and any vulnerabilities of terrorist networks, assisting future identification of leaders and other counter-terrorism interdiction activities.

2. Methods

2.1. Data

Data from media sources were collected from Nov. 17, 2015 to May 9, 2016 by coding large-circulation media sources, particularly French, Dutch, and English-language daily newspapers online as well as the Lexis (formerly known as Lexis-Nexis) database. The

[7] See Morselli et al. (2007) for a good discussion of the security-efficiency tradeoff.

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004

Table 1
Comparison of frameworks for covert network forensics.

	Simple Graph	Multi-relational	Multi-modal	Graph Database
Nodes Types	1 (Persons)	1 (Persons)	1 or more (Persons and other entities)	1 or more (Persons and other entities)
Edges Types	1	1 or more	1	1 or more
Raw Source Information	External	External	External	Internal
Storage Capacity[a]	$<10^6$	$<10^6$	$<10^6$	$>10^9$
Representative Software[7]	NetworkX, Igraph	Ucinet, Statnet, PNet		Neo4j, Titan, DSE Graph

[a] Capacity of nodes and edges is based on leading implementations: Neo4 j v3.0.1, UciNet 6.605, Statnet 2015.11.0, Titan 1.0.0.

full text of sources of the articles are included in an online website complementary to this paper.[2]

2.2. Graph database analysis

3.2.1. Nodes

The following classes of nodes were used:

1. Persons: Human actors with the attributes age, gender, role, citizenship and status (free, wanted, or dead). They were also coded with a role attribute: weapons, logistics (other than weapons), leadership (recruiter, planner).
2. Attacks: Attack events including "successful" operations as well as interdicted plots.
3. Activities: Missions related to covert activities or terrorism (e.g. in/exfiltrations, weapon preparation, etc.) and activities not related to covert missions or attacks (e.g. family circle, sports club).
4. Sites: Geographical places with addresses where all present likely know each other (e.g. private residences, safe houses, and staging sites).
5. Localities: Larger locations such as communities and cities commonly reported as places of residence.
6. Countries: Sovereign states visited for covert activities.

To build the IS-E network, we started from attacks, which were only included if they were claimed by IS (or IS-inspired actors) and were also orchestrated by IS-E leadership. Although IS inspired and orchestrated other attacks in Europe, they were not included in IS-E if there was little evidence of any operational connections to IS-E. Only attacks that had formal IS members with links to IS were included. We next sought out news reports about the attacks, and extracted names and relationships using our coding scheme above.

The next step used graph database queries to define the network boundaries for the IS-E network. Namely, individuals were included based on the following criteria: (1) Attackers (i.e. used/planned to use a weapon on a particular target); (2) Known to have been involved with an attacker/actor in a covert/terrorism-related task (as organizers or in logistical roles); (3) Wanted or arrested in relation to the attacks for unspecified reasons. Thus, for IS-E, any individual that assisted a wanted or dead member of IS-E is included, but not individuals who played no known role (such as innocent family members). Detailed query logic, the analysis program, the database, and raw media data are linked to this paper and are also made available publicly through the GitHub collaborative repository: http://github.com/sashagutfraind/nov13/.

2.2.2. Relationships

The database contains person to entity relations, as well as a relatively small number of person to person relations. Covert network forensics is often reliant on information, which belongs to several fairly well-characterized types: (1) Person Involved-In Attack: a person that actively participated in an attack on a particular site; (2) Person Involved-In Activity: a person who participated in a covert mission or a non-covert group or activity; (3) Person Linked-To Person: a pair of linked persons due to kinship, affinity, operational work, or unspecified reasons; (4) Person Present-In Site, Locality or Country: an individual connected to a location, including dates if available, where he or she is known to have resided in, temporarily stayed in, or countries visited (if relevant to the covert mission). When recording the relationship in the graph database, we also recorded the specific source that reported it. In the case of IS-E, most of the sources were large circulation media sources in English and French. The web link to the source was also stored in the database.

For the IS-E network, two individuals were tied to each other in the social network on one of several grounds, where one is sufficient for a tie: if they were involved in an attack on the same target, jointly involved in a covert-related activity, were present in the same location and are both suspected of a covert activity. A pre-existing strong tie (kinship, friendship or other) was also grounds for a tie, as long as both persons were included as members of IS-E (see above criteria). Ties projected based on location were examined closely as part of the sensitivity analysis (see below). While such ties are expected to be genuine operational ties particularly when the location is small (e.g. two persons living in the same a safe house), in some cases are spurious (e.g. two persons living in the same area of Brussels) and so including them would have increased the false positive rate.

2.2.3. Representation of covert network data in a graph database

In our analysis we used the free and open-source Neo4j database and its Cypher query language (Miller 2013; Robinson et al., 2015; Webber 2012), but the analysis does not depend on this specific tool and it could be implemented using any graph database software. Neo4j can be accessed either with Cypher or from common languages such as Java, R and Python. The syntax of Cypher is illustrated in Box 1 and 2.

Our data processing was based on the R language. As a first step, we used the open-source *RNeo4j* library for R to load the raw data into the *Neo4j* database. Next, we used *RNeo4j* to query the database and obtain lists of matching nodes and relationships. Finally, we

[2] Neo4j: Neo Technology, Inc. San Mateo, CA. 2016. Available from: https://neo4j.com/product/ : UciNet: Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies. Available from: https://sites.google.com/site/ucinetsoftware/downloadshttps://sites.google.com/site/ucinetsoftware/downloads ; Networkx: networkx team. 2014. Available from: https://networkx.github.iohttps://networkx.github.io ; Statnet: Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2003. Available from: http://statnetproject.org; Igraph: igraph team. 2003. Available from: http://igraph.org ; Pnet: P Wang, G Robins, P Pattison 2006. "PNet A program for the simulation and estimation of exponential random graph models" University of Melbourne. Available from: http://www.swinburne.edu.au/fbl/research/transformative-innovation/our-research/MelNet-social-network-group/PNet-software/index.htmlTitan: Titan community. 2015. Available from: https://github.com/thinkaurelius/titan ;DSE Graph: DataStax Analytics, Inc, Santa Clara, CA. 2016. Available from: http://www.datastax.com/products/datastax-enterprise-graph.

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004

Box 1: Sample code in the Cypher language to create a pair of nodes and an edge. Actual database contains more attributes (see above). The "ref1" attribute stores a reference to the media source.

```
CREATE (t1:Person {name:"Abdelhamid Abaaoud", age:27, ref1:"DM1"})
CREATE (s1:Site {name:"Charleroi apartment", ref1:"NYT10"})
CREATE (t1)-[r1s21:PRESENT_IN{ref1:"NYT10"}]->(s1)
```

Box 2: Sample code in the Cypher language for querying the graph database. The code produces tables of matches which are loaded into igraph for network analysis.

```
sameAttackPairs = '
MATCH (p1:Person)-[:INVOLVED_IN]->(l:AttackSite)<-[:INVOLVED_IN]-(p2:Person)
RETURN p1.name, p2.name'

sameActivityPairs = '
MATCH (p1:Person)-[:INVOLVED_IN]->(l:Activity)<-[:INVOLVED_IN]-(p2:Person)
RETURN p1.name, p2.name'

sameSitePairs = '
MATCH (p1:Person)-[:PRESENT_IN]->(l:Site)<-[:PRESENT_IN]-(p2:Person)
WHERE (p1.status <> "free") AND (p2.status <> "free")
RETURN p1.name, p2.name'
```

performed social network analysis using the open-source *igraph* software.[3] for R. All statistical analysis used the R *stats* package.[4]

2.2.4. Sensitivity analysis

The greatest source of uncertainty in the IS-E network has to do with using co-location to infer ties. For this reason we considered several variants of the IS-E network: (1) IS-E restricted: this network only has ties for persons who were jointly involved in attacks or covert activities or had a pre-existing link. Sharing of space is not considered, thus it excludes ties of members who e.g. lived in the same house. (2) IS-E extended: expanded the IS-E network by tying two members whenever they were at the same locality (for IS-E members, it was often Molenbeek) or affiliated with a non-covert activity. We obtained these three networks by modifying the queries to the graph database.

2.2.5. Social network metrics

The resulting networks were studied structurally, based on several metrics that were previously used to characterize covert networks, particularly terrorist networks (Helfstein and Wright 2011; Krebs 2002; Lindelauf et al., 2009). The basic metrics were: density (number of edges divided by the number of edges in a complete network), clustering (the number of actual triads divided by the possible) and the diameter (the length of the shortest path between the most distant pair of individuals) – a measure of the communication time across the network. We also considered efficiency, a measure of the network's ability to perform terrorist operations as computed from the harmonic mean of person-to-person distances (Gutfraind, 2010) and secrecy S_1, the fraction of the network that remains unexposed if a single person is detected (Lindelauf et al., 2009). Both efficiency and secrecy range in [0,1] with 1 considered the maximal possible for a network. Terrorist networks tend to be highly modular (organized into cells) (Sageman, 2004), which enables them to operate covertly and avoid accidental exposure. To measure modularity, we used a modularity score originally proposed by (Newman and Girvan, 2004) and calculated it using the Louvain approach of Blondel et al. (2008). This modularity measure identifies the optimal separation of the network into communities (including their number) and computes a modularity score in [0,1] that reflects how strongly is the network separable (1 being the highest).

2.2.6. Media coverage sensitivity data

In order to examine a possible relationship between media coverage and the structure of the network, information about media coverage was sought for all the individuals in the network. Using Lexis (formerly Lexis-Nexis), we determined (1) total number of English-language articles about the individual (up to May 27, 2016), (2) the date of first coverage, and (3) the date of last coverage. From the dates we calculated the duration of coverage in days for each actor. The results were correlated with network attributes such as degree, betweenness centrality, and the local clustering coefficient.

2.2.7. Comparison of terrorist networks

We compared the IS-E network to several datasets from past covert networks (Table 2). The motivation for doing so was to see if there are some underlying similarities among covert networks that organize political violence. A particularly significant empirical reference are networks affiliated with al-Qaida, which competes with IS in the radical Islamist milieu (Watts, 2016). We also included networks from other terrorist organizations and a covert non-terrorist network (FTP), in order to determine whether IS-E has distinct properties when compared to other covert networks. Next, we extracted from the IS-E network the operational subnetwork responsible for the 2015–2016 attacks in Paris and Brussels (termed, IS-E supercell) and compared it to several operational networks under al-Qaida's control, including the 9/11 attacks in the US and the March 11 attack in Madrid. Operational networks are organized to carry out a single operation (or related operations) and, unlike organizational networks, do not include the leadership or the auxiliary echelons of the organization.

[3] https://github.com/sashagutfraind/nov13/
[4] Version 1.0.1.

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004

Table 2
Network datasets which were compared to IS-E.

Network	Dates Active	Type	Movement	Source
9/11 attacker network	2001	Terrorist, Operational	Islamist radicals	(Krebs, 2002)
Strasbourg Cathedral plot	2000	Terrorist, Operational	Islamist radicals	(Xu and Chen, 2008)
Bali Bombings network	2002	Terrorist, Operational	Islamist radicals	(JJATT, 2009)
Madrid bombers (M11)	2003	Terrorist, Operational	Islamist radicals	(Rodriguez, 2004)
Jemaah Islamiyah (JI)	2000–2005	Terrorist, Operational	Islamist radicals	(Everton, 2012)
Jakarta hotel bombings	2009	Terrorist, Operational	Islamist radicals	(JJATT, 2009)
al-Qaida global	2001	Terrorist, Organizational	Islamist radicals	(Xu and Chen, 2008)
Francs-tireurs et Partisans (FTP)	1941–1945	Urban insurgency, Organizational	World-war II Resistance	(Gutfraind, 2010)
Revolutionary Org. November 17 (17N)	1975–2002	Terrorist, Organizational	Marxist	(JJATT, 2009)

3. Results

3.1. The structure of the Islamic State in Europe network (IS-E)

Applying the graph database framework to the IS-E network resulted in a database of 119 nodes and 232 relationships. The core of the database consists of nodes of type Attack, Person, Activity, and Site; as well as the relationships between them, represented in Fig. 2. Further, the network contains 71 persons who have the following demographic characteristics.[5] The mean age is 28 (Range: 13–40) and the gender distribution is 99% male (69 males and 1 female). The most common countries of citizenships are: Belgium – 29 (41%), France – 18 (25%) and Algeria – 5 (7%).

A striking feature of the network is the high degree centrality of the A. Abaaoud node, i.e., his high number of connections to persons and entities. He is one of the field leaders of the Nov. 13 attacks, as well as the organizer of previous attacks by IS-E. The node of S. Abdeslam, another leader, is also high in degree. Moreover, many of the actors have multiple paths to each other, suggesting that the network is relatively robust to the arrest of bridging individuals.

Applying the projection procedure to the database (see Methods), we obtained a covert social network with 71 individuals and 146 undirected relationships. This reveals the cells involved in the attacks, which appear as cliques (Fig. 3). Each of the attacks cells (e.g. Stade de France, Café Bonne Bière, the Bataclan Theater) are distinguished. Other cells include the smuggling cell of S. Abdeslam, the cell responsible for exfiltration of Ahmet Dahmani, the group around St. Denis and finally the IS leadership grouping, which is connected to A. Abaaoud. Abaaoud and S. Abdeslam are clearly vital for connecting the network.

The degree of a node can highlight its relative importance to the network (Krebs 2002). Examining the degrees of the nodes in IS-E (Fig. 4A) it could be seen that A. Abaaoud has the highest degree followed by S. Abdeslam. The same nodes also have the highest betweenness centrality (Fig. 4B). Indeed, Abdelhamid Abaaoud is believed to be a leader of IS-E and he was responsible for multiple terror attacks in Europe while S. Abdeslam is known as one of the co-leaders, possibly with an additional person (Rubin, 2015). People loosely tied to the plot have the lowest degree. Thus, in IS-E just as in other Islamist groups, the degree of the node appears to be a reliable indicator of leadership role (Medina, 2014) – a finding that may assist future counter-terrorism investigations.

3.2. Dependence of network structure on media coverage

We obtained media coverage information for 66 individuals out of a total of 71 in the network. The remaining five persons never had their names identified in any open sources, making it difficult to estimate the true coverage. The median number of English-language articles was 96.5 (range: 3–16052). The median duration of coverage was 136.5 days (range: 1–1715). Univariate regression model of duration on the number of articles was not significant (F-test: $p > 0.45$).

We tested for a possible relationship between coverage variables and three outcomes: the degree, betweenness centrality, and the person's clustering coefficient. We found that the duration in the news was not significant in any of the three outcomes, when controlling for number of articles (F-test: $p > 0.17$). In this multivariate regression analysis, a positive linear relationship was found between degree and the number of articles (estimate: 0.00178, standard deviation (SD): 0.000173), the betweenness (estimate: 0.0665, SD: 0.0072) and clustering (estimate: −6.70e-05 SD: 1.50e-05) with high significance in all cases (F-test: $p < 0.0001$). Thus, an increase of 1.78 in the degree of a person in IS-E is associated with additional 1000 media articles. The adjusted R-squared values in the three regressions are 0.66, 0.60 and 0.30, respectively. When removing two outliers (>5000 articles), the adjusted R-squared values are 0.43, 0.35, 0.37, respectively.

3.3. Structural network analysis and comparison to other covert networks

Table 3 shows the structural properties of the IS-E network and compares IS-E to other organizational networks in our dataset.

Table 3 shows that IS-E is one of the smallest networks, and it has a relatively low mean degree as compared to all other organizations. It also has a low secrecy score, as compared to al-Qaida and FTP, but higher than 17N and JI. The clustering of IS-E is lower than FTP and JI, but higher than al-Qaida. The modularity is intermediate and comparable to al-Qaida.

Although IS-E is the best estimate of the true network of IS behind these attacks, the graph database allowed us to evaluate the sensitivity of the results by considering alternative assumptions. We considered two such alternative assumptions with corresponding variant networks: (1) the only real ties are those around terrorist activities or known links (IS-E restricted), and (2) many ties were unreported and individuals should be tied when they shared a site, location or a non-covert activity (IS-E extended). Under this evaluation, the relatively low degree, higher clustering and low secrecy of the IS-E network as compared to al-Qaida is robust to how the network was inferred. However, the clustering increases substantially in IS-E extended, which implies that clustering is dependent on the assumptions used to infer the network.

Focusing on the supercell responsible for the Paris and Brussels attacks, we found that this subnetwork contains 69% of the nodes of the IS-E network. We also compared the Paris/Brussels supercell to operational networks from past attacks by al-Qaida (Table 2). In this comparison, the supercell was found to have a relatively low mean degree (but not the lowest) and low clustering. From

[5] Version 3.3.1.

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004

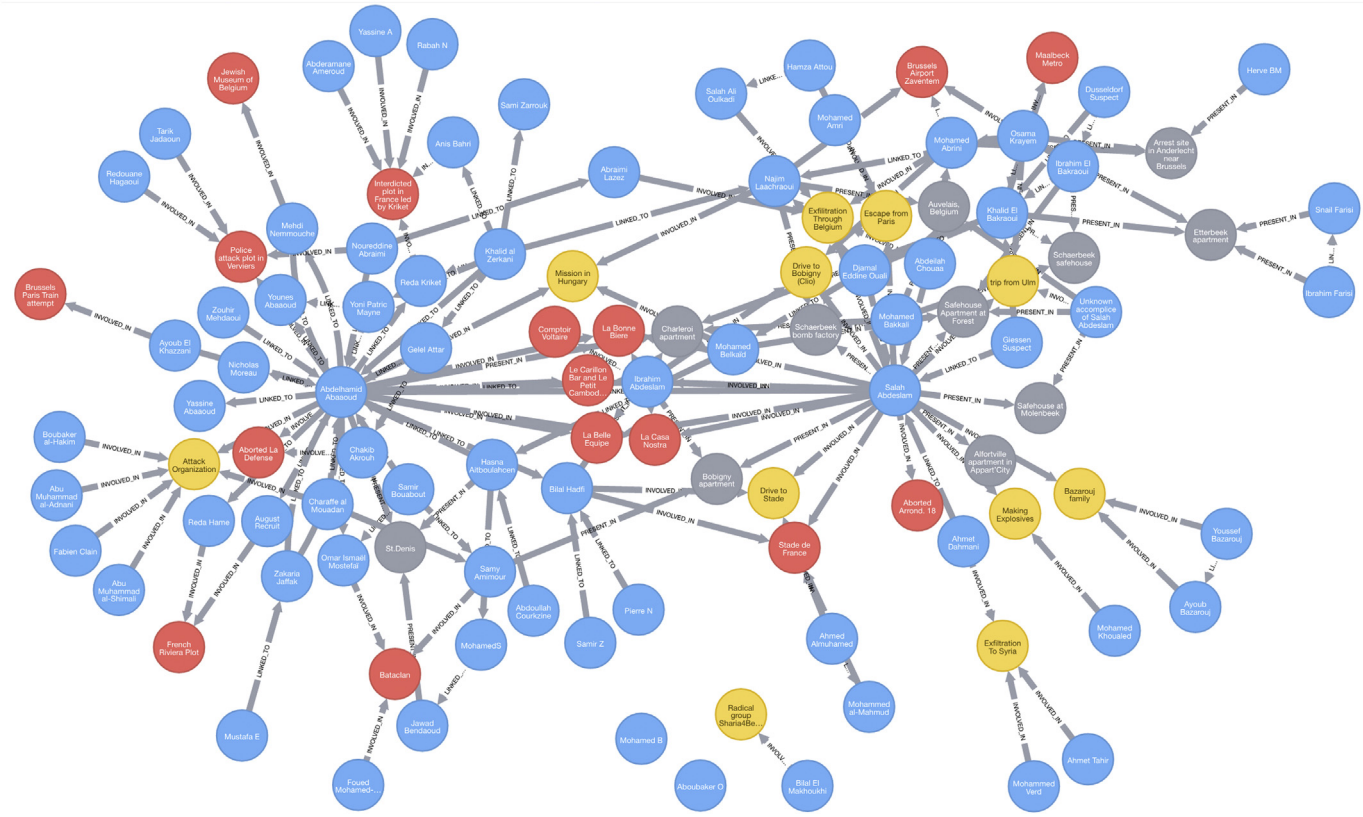

Fig. 2. The nodes and relationship in the IS-E network in the graph database representation: attacks (red), activities (yellow), persons (blue) and sites (gray). To reduce clutter, localities and countries are not shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Network properties of IS-E and of other covert organizational networks.

Organization	Nodes	Edges	Components	Diameter	Density	Mean Degree	Clustering	Efficiency	Secrecy	Modularity
IS-E	71	146	7	5	0.059	4.11	0.27	0.36	0.93	0.55
IS-E extended[a]	71	222	7	5	0.089	6.25	0.53	0.40	0.90	0.50
IS-E restricted[b]	71	117	12	5	0.047	3.30	0.25	0.30	0.94	0.59
17N	18	46	1	2	0.30	4.11	0.50	0.65	0.66	0.21
al-Qaida	368	1481	5	7	0.02	8.05	0.16	0.33	0.98	0.56
FTP	174	264	1	9	0.02	3.03	0.41	0.20	0.98	0.81
JI	79	623	2	14	0.20	15.77	0.56	0.56	0.79	0.33

[a] IS-E extended: extends IS-E by including ties based on sites and locality of residence. [b]IS-E restricted: includes only ties based on known terrorist activity, attacks, or personal links.

Table 4
Comparison of covert operational networks.

Operations	Nodes	Edges	Components	Diameter	Density	Mean Degree	Clustering	Efficiency	Secrecy	Modularity
IS-E supercell[a]	49	109	3	5	0.093	4.449	0.305	0.427	0.889	0.470
9/11	62	152	1	5	0.08	4.90	0.36	0.40	0.90	0.53
Bali 2002	17	158	1	3	0.45	11.7	0.72	0.72	0.53	0.22
Jakarta 2009	28	38	3	4	0.10	2.71	0.36	0.22	0.87	0.61
Madrid M11	70	240	7	6	0.10	6.86	0.57	0.37	0.89	0.46
Strasbourg	18	26	2	5	0.17	2.89	0.13	0.45	0.78	0.38

[a] The subnetwork of IS-E responsible for the 2015–2016 attacks in Paris and Brussels.

the lower mean degree figure, we hypothesize that IS-E might use smaller cells as compared to al-Qaida (Table 4).

4. Discussion

Our study proposed a framework based on graph databases to document and infer covert networks relying on data drawn from large-circulation open media sources. The framework has several advantages over existing approaches in the areas of reproducibility, documentation, complexity of representation, and multiple projection of the data. The framework was applied to represent the network of the Islamic State group in Western Europe (IS-E).

For the Islamic State in Europe (IS-E), it was found that IS-E leaders at the operational level could be identified by their much higher

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004

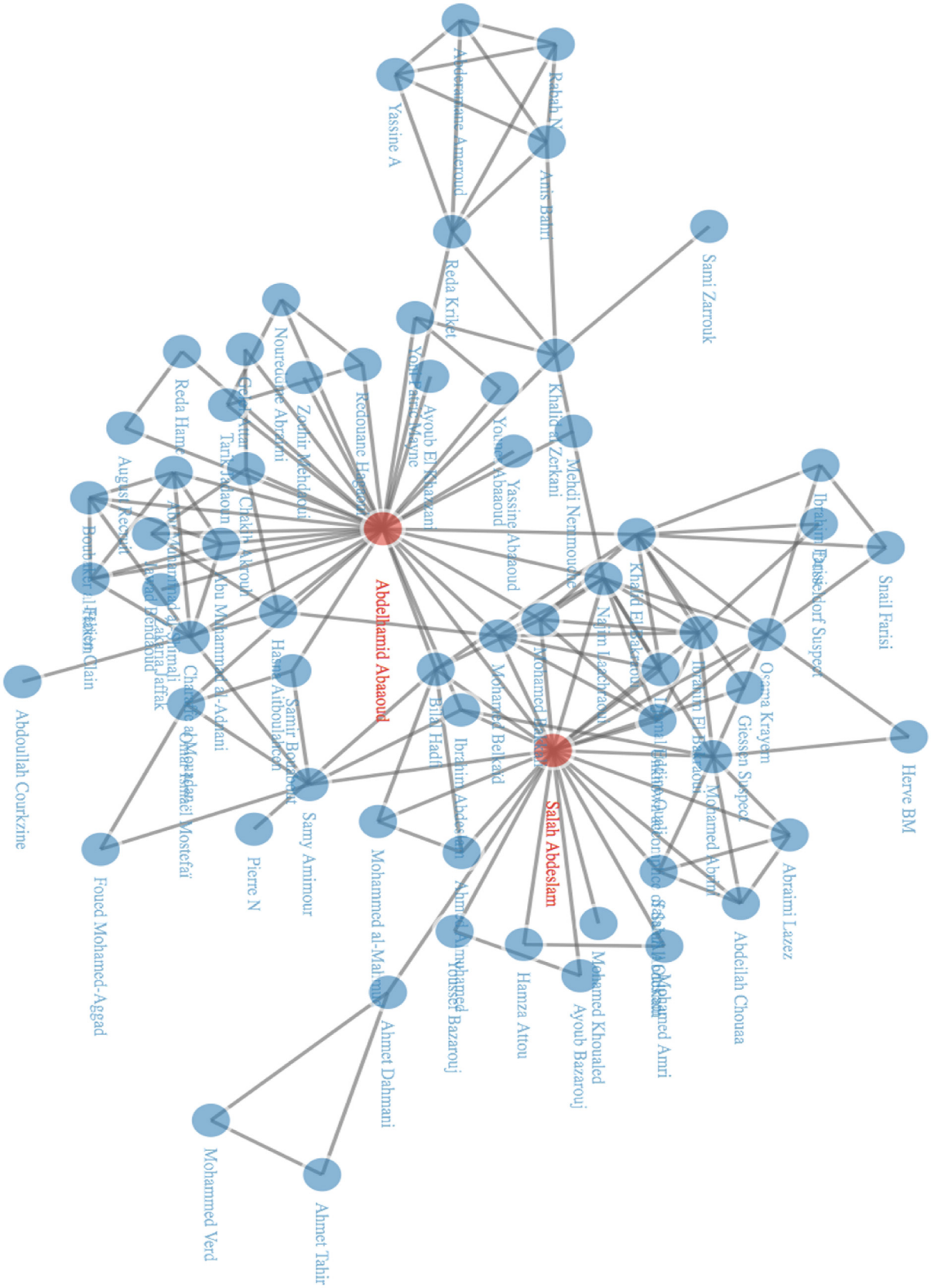

Fig. 3. Plot of the IS-E actor network. The network nodes represent all the individuals involved in the covert organization, and the edges link individuals that probably interacted with each other. Additional singleton nodes (6) are not shown (see degrees in Fig. 4A).

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004

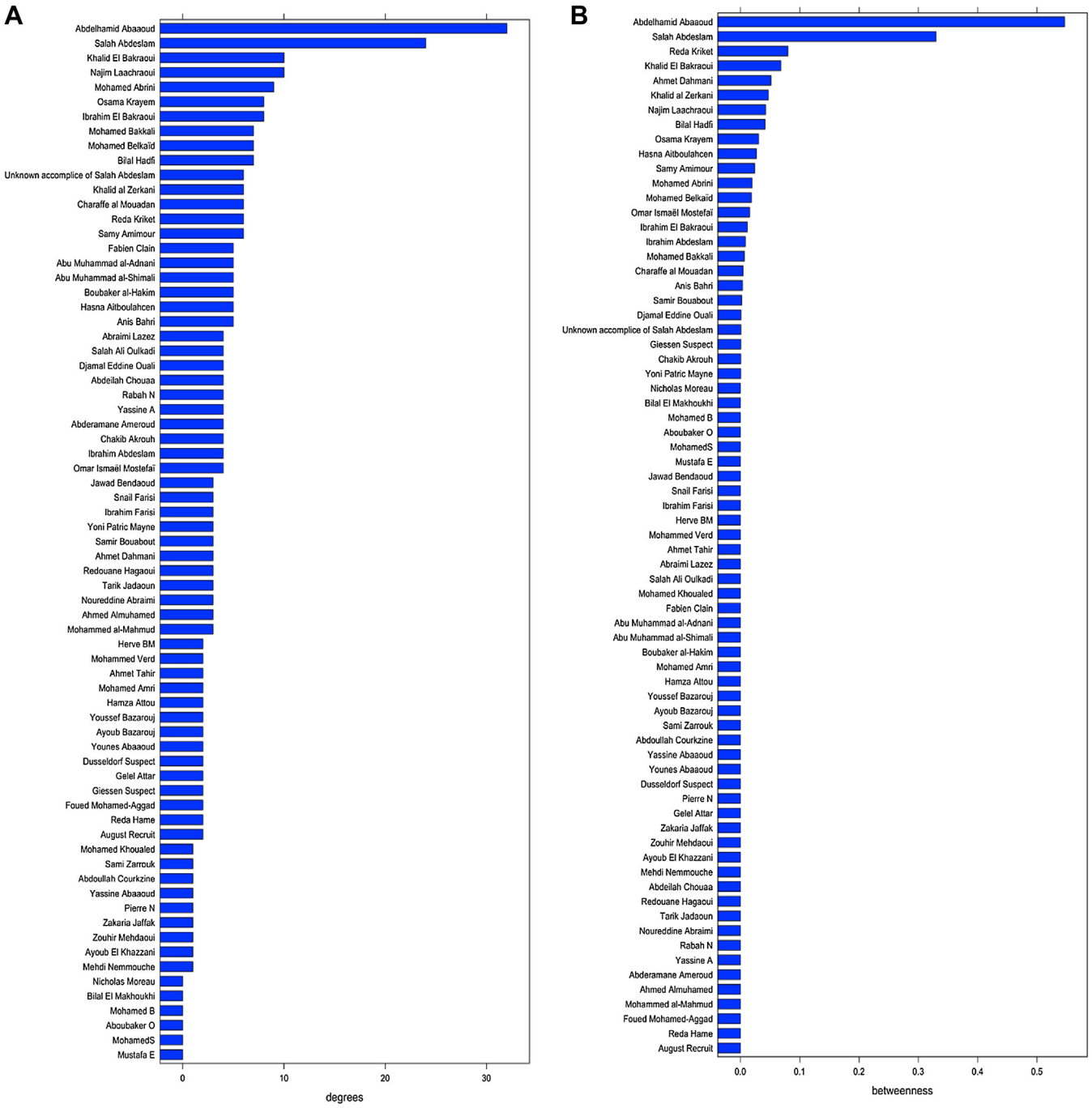

Fig. 4. Node centrality of the IS-E actor network. (A) Node degree centrality and (B) Betweenness centrality.

degree and betweenness centrality, which is consistent with past studies of al-Qaida (Krebs, 2002; Medina, 2014). Because the leaders of the IS-E network have relatively high degree, the IS network might be organized to maximize the leaders' efficiency, i.e., ability to recruit and carry out attacks. This efficiency comes at the cost of lowered security by subjecting the leaders to interception and arrest (Gutfraind, 2010). The efficiency-security tradeoff[6] may have been buffered by the documented adoption of new cryptographic communication technologies in planning and coordinating the IS-E attacks (see Rotella, 2016). Examination of the IS-E network suggested that its network could be characterized by relatively low mean degree and low secrecy, as compared to the al-Qaida network. Taken together, IS-E could be characterized by relatively small cells of about four persons, which are somewhat loosely knit together, as indicated by the lower secrecy score. The finding undermines the hypothesis that all covert networks, or at least, all covert networks in a particular operational milieu would have a similar network structure (Gutfraind, 2010; Lindelauf et al., 2009).

Even with more time and more data, there are basic limitations to the process of reconstructing covert network structure when relying on exclusively public media sources. It is likely that

[6] One person has no available attributes.

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004

some ties between the individuals might not be revealed for security reasons until many years after the attack. Fortunately, because information about arrests and arrest warrants is public and widely reported, the public record has good coverage of IS-E operations surrounding the recent attacks. Indeed, those ties that were consistently reported by multiple independent media sources are likely to be real. There are also particular difficulties in the case of the IS-E because its commanders are based in Syria, and hence some of its higher echelons are not characterized in detail, although some were reported and included in our database.

We found evidence of a positive correlation between a suspect's network properties, particularly, degree and betweenness, and the number of articles on that person. This finding is intriguing but of course does not indicate causality. We hypothesize that there is a mutually reinforcing effect whereby greater terrorist activity leads to greater media coverage, as well as media coverage giving greater network information. The number of articles is positively related to degree (R-squared = 0.43), a finding that sets an upper bound on the number of missing links in our data – approximately 2 additional links per 1000 articles. Because the median number of articles per person is just 96, which is much less than a 1000, the effect is small. It implies that a typical node, which has a degree of 4.1, has an error of at most 0.2 links in its degree, or just 5%.

The comparison of terrorist networks requires that they are coded using the same framework, though this is not current practice. Indeed, there are systematic differences in how various authors define membership and relationships in the covert networks they study, which are often not possible to reconstruct and reanalyze. We have attempted to address this problem, and indeed this is our main contribution, by including raw data in the graph database, automating the process of building the social network, and testing our findings using sensitivity analysis. We publicly share our code and data, and encourage other researchers to examine them and perhaps add to the analysis.

Indeed, part of the great advantage of graph databases is that they allow for the cumulative development of social network knowledge. It may be desirable for the social network community to adopt standards for coding secondary network data using graph databases such that all the decisions of the analyst are carefully documented for future researchers to replicate and build upon. This is in contrast to current practice where coding decisions are largely hidden and are not easily reproducible by subsequent researchers.

Given the power of graph databases, we argue that these methods could considerably advance covert network forensics, and social network analysis in general, by enabling more consistent data aggregation and automatic sensitivity analysis of the results. As the field is increasingly adopting Big Data methods such as machine learning and natural language processing, it is likely that information extraction including network information extraction would become increasingly automated. It would be possible, in the near future, to build an automated system that would continuously read public media sources about a particular event, such as a terror attack, represent the data in a graph database, and produce a robust estimate of the underlying covert network.

5. Conclusion

This study proposes the use of graph databases to obtain a high-resolution map of covert networks based on public media sources and perform structural inference. The analysis method was applied to the Islamic State network responsible for the November 2015 Paris Attack and the March 2016 Brussels Attack as well as related plots. Our study is one of the first to examine the structure of a network operated by the Islamic State, and to compare it to other covert networks. We found that the network, as compared to al-Qaida operations, has a lower secrecy, which may make it relatively more vulnerable to interdiction.

An important direction for future research is to standardize the coding of clandestine and terrorist networks to enable a more consistent comparison of their structure. This promises to highlight universal patterns as well as systematic differences in the organization of these networks. The graph database framework offers an efficient way to do so.

Acknowledgement

The authors wish to thank Roy Lindelauf for detailed feedback. AG was partly supported by Uptake Technologies, Inc. and National Institutes of Health grant R01AI101229.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.socnet.2016.10.004.

References

Angles, Renzo, Gutierrez, Claudio, 2008. Survey of graph database models. ACM Comput. Surv. (CSUR) 40 (1), 1–39.
Baker, Wayne E., Faulkner, Robert R., 1993. The social organization of conspiracy: illegal networks in the heavy electrical equipment industry. Am. Sociol. Rev. 58 (6), 837–860.
Blondel, D.Vincent, Guillaume, Jean-Loup, Lambiotte Renaud, Lambiotte, Renaud, Lefebvre, Etienne, 2008. Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. 2008 (10) (P10008).
Borgatti, S.P., Everett, M.G., Johnson, J.C., 2013. Analyzing Social Networks. SAGE Publications, London, UK.
Breiger, Ronald L., Schoon, Eric, Melamed, David, Asal, Victor, Karl Rethemeyer, R., 2014. Comparative configurational analysis as a two-mode network problem: a study of terrorist group engagement in the drug trade. Soc. Netw. 36, 23–39.
Costenbader, E., Valente, T.W., 2003. The stability of centrality measures when networks are sampled. Soc. Netw. 25 (2), 283–307.
Crilly, Rob, 2015. Police Detain Suspected Ringleader of Belgian Terror Cell, Says Source. in (Retrieved: 04.04.16 http://www.telegraph.co.uk/news/worldnews/europe/greece/11353214/Greek-police-detain-suspected-ringleader-of-Belgian-terror-cell-says-source.html).
Everton, Sean F., 2012. Disrupting Dark Networks. Cambridge University Press, New York, NY.
Frank, Ove, 2011. Survey sampling in networks. In: Scott, J.C., Carrington, Peter J. (Eds.), The Sage Handbook of Social Network Analysis. SAGE Publications, London, UK, pp. 390–402.
Gerdes, Luke M., 2015. Illuminating Dark Networks: The Study of Clandestine Groups and Organizations. Cambridge University Press, New York.
Gutfraind, Alexander, 2010. Optimizing topological cascade resilience based on the structure of terrorist networks. PLoS One 5 (11), e13448.
Handcock, Mark S., Hunter, David R., Butts, Carter T., Goodreau, Steven M., Morris, Martina, 2003. Statnet: Software Tools for the Statistical Modeling of Network Data, Available from: http://statnetproject.org.
Helfstein, Scott, Wright, Dominick, 2011. Covert or convenient? evolution of terror attack networks. J. Confl. Resol. 55 (5), 785–813.
JJATT, 2009. John Jay & Artis Transnational Terrorism Database. John Jay College, City University of New York (Available from: http://doitapps.jjay.cuny.edu/jjatt/index.php).
Kossinets, Gueorgi., 2006. Effects of missing data in social networks. Soc. Netw. 28 (3), 247–268.
Krebs, Valdis E., 2002. Mapping networks of terrorist cells. Connections 24 (3), 43–52.
Laumann, Edward O, Marsden, Peter V, Prensky, David, 1989. The boundary specification problem in network analysis. In: Freeman, L.C., White, D.R., Romney, A.K. (Eds.), Research Methods in Social Network Analysis. Transaction Publishers, New Brunswick, NJ, pp. 61–68.
Lindelauf, Roy, Borm, Peter, Hamers, Herbert, 2009. The influence of secrecy on the communication structure of covert networks. Soc. Netw. 31 (2), 126–137.
Lindelauf, Roy, Borm, Peter, Hamers, Herbert, 2012. One-Mode projection analysis and design of covert affiliation networks. Soc. Netw. 34 (4), 614–622.
Medina, Richard., 2014. Social network analysis: a case study of the islamist terrorist network. Sec. J. 27, 97–121.
Miller, Justin J., 2013. Graph database applications and concepts with neo4j. Atlanta, GA, USA In: Proceedings of the Southern Association for Information Systems Conference, 2324.
Morselli, Carlo, Giguère, Cynthia, Petit, Katia, 2007. The Efficiency/Security trade-Off in criminal networks. Soc. Netw. 29 (1), 143–153.
Morselli, Carlo., 2009. Inside Criminal Networks. Springer, New York, NY.

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69 (2), 026113.
Parlapiano, Alicia Andrews, Haeyoun, Wilson Park, Larry, Buchanan, Sarah, Almukhtar, 2015. Unraveling the Connections Among the Paris Attackers. New York Times 2016, Retrieved: 04.04.16 (http://www.nytimes.com/interactive/2015/11/15/world/europe/manhunt-for-paris-attackers.html?_r=1).
Robins, Garry L., 2015. Doing Social Network Research: Network-Based Research Design for Social Scientists. Sage Publications Ltd, Los Angeles, CA.
Robinson, Ian, Webber, Jim, Eifrem, Emil, 2015. Graph Databases: New Opportunities for Connected Data: Sebastopol. O'Reilly Media, CA: Sebastopol, CA.
Rodriguez, A. Jose, 2004. The March 11th Terrorist Network: In Its Weakness Lies Its Strength. Working Papers EPP-LEA. University of Barcelona, Spain.
Rotella, Sebastian, 2016. The Dark Side of Privacy: How ISIS Communications Go Undetected Pacific Standard, Retrieved: 30.07.16 https://psmag.com/the-dark-side-of-privacy-how-isis-communications-go-undetected-890aec4e86c3#.6pa3fargo.
Rubin, J. Alissa, 2015. Cellphone Contacts in Paris Attacks Suggest Foreign Coordination. New York Times (Retrieved: 04.04.16) http://www.nytimes.com/2015/12/31/world/europe/cellphone-contacts-in-paris-attacks-suggest-foreign-coordination.html?_r=0.
START, 2013. Global Terrorism Database [Data File], Retrieved from http://www.start.umd.edu/gtd.
Sageman, Marc, 2004. Understanding Terror Networks. University of Pennsylvania Press, Philadelphia, PA.
Schmitt, Eric, Schmidt, S. Michael, Higgins, Andrew, 2015. Al Qaeda Trained Suspect in Paris Terror Attack, Official Says. New York Times, Retrieved: 04.04.16 http://www.nytimes.com/2015/01/09/world/europe/paris-terror-attack-suspects.htm.
Watts, Clint, 2016. Deciphering Competition Between Al-Qa'ida and the Islamic State, 9., 7 CTC Sentinel., West Point, NY, pp. 1–6.
Webber, Jim, 2012. A programmatic introduction to neo4j. In: Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity, ACM, pp. 217–218.
Xu, Jennifer, Chen, Hsinchun, 2008. The topology of dark networks. Commun. ACM 51 (10), 58–65.
Zouzias, Anastasios, Michail, Vlachos, Vagelis, Hristidis, 2014. Templated search over relational databases. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, pp. 21–30.

Please cite this article in press as: Gutfraind, A., Genkin, M., A graph database framework for covert network analysis: An application to the Islamic State network in Europe. Soc. Netw. (2016), http://dx.doi.org/10.1016/j.socnet.2016.10.004