7

Open-Source Exploitation for Understanding Covert Networks

Kathleen M. Carley

Most people take for granted that the Web is the go-to place for information. The amount of Internet-available information is huge and growing at unprecedented rates. In 2012, Domo estimated that every minute users created 571 new Web sites, 100,000 tweets, and 684,478 pieces of Facebook content.[1] Those numbers have skyrocketed; for example, in 2013, nearly 230,000 fans per minute tweeted about the Red Hot Chili Peppers during their half-time performance at the Super Bowl.[2] Because people and organizations turn to the open sources of the Internet to obtain and provide information, contemporary analysts require methods of open source exploitation to extract meaningful information from this digital data forest. Open source exploitation includes the gathering, coding, analyzing, and visualizing of open-source data to provide significant insights into the activities of individuals and groups, the responses provoked by those activities, and the ways those activities have changed.

Open source exploitation can be used to study terrorists, dark networks, and the population's responses to their activity. By 2000, most if not all terror groups had established a Web presence (Weiman, 2004). For example, al-Shabaab uses Twitter to communicate, and al-Qaida used the Internet to plan 9/11 (Thomas, 2003). However, information about dark networks also appears on the Web in online newspapers and blogs

This work was supported in part by the Air Force Office of Sponsored Research, FA9550-11-1-0179, the Office of Naval Research Minerva N000141310835, and the Office of Naval Research Muri N000140811186. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems and the Institute for Software Research at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the AFOSR, the ONR, or the U.S. government.

[1] http://www.visualnews.com/2012/06/19/how-much-data-created-every-minute/

[2] http://www.statista.com/statistics/252221/super-bowl-moments-with-the-most-tweets-per-minute/

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

dedicated to studying the activities and membership of terror groups, criminal organizations, narcotics traffickers, and pirates. Such data can be exploited to understand dark networks (Krebs, 2002a, 2002b); however, it is often incomplete, filled with errors, out of date, and spread across numerous, difficult-to-find sources. Indeed, dark network sites are often hidden and "move about" as they are shut down and then reemerge.

This chapter discusses the open source exploitation process and pays particular attention to dark networks. Key topics include: the techniques needed for making sense of open-source data, the limitations of the data sources and the techniques, and the challenges open source exploitation has yet to meet. This chapter uses real-world illustrations to shed light on the complex process of open source exploitation and its potential pitfalls. Major challenges are scalability, data multidimensionality, data collection, analysis of big data, data integration, data layering, data fusion, and the applicability of methods in field settings (Roberts, 2011). The basic story is both simple and quite nuanced: open source exploitation holds great promise for the understanding of dark networks, and great strides have been made in the technologies available for exploiting this information. However, the current state of the art is still in its infancy.

I. The Data: From Text to Networks

Open-source data includes text-based data such as online news articles, tweets, and blogs. In addition to raw, content-rich messages, such sources also include aspects of structured meta-data. For example, Twitter messages contain "sender" and "receiver" fields, as well as numeric data on the date and time of the message and, for users who have chosen not to opt out of Twitter's geo-locating services, information on the user's latitude and longitude at the time the message was sent. Open source exploitation leverages both structured and unstructured data to identify topics, assess sentiments, identify groups, determine key actors within those groups, and measure lines of influence among them.

It is possible to extract information about dark networks from open source documents (Carley, Bigrigg, & Diallo, 2012; Diesner & Carley, 2012). The information extracted can include the dark network and the roles of actors; overall organizational structure (Kenney et al., 2013), linkages to subgroups (Gerdes, 2012), instances of adaptation (Horgan et al., 2014), and the topics associated with the dark network. Network analytics can be applied to assess organizational vulnerabilities and strengths (Sparrow, 1991; Baker & Faulkner, 1993; Klerks, 2001; Ronfeldt & Arquilla, 2001a) and to determine the role of age (Sageman, 2004), ethnic, religious, and family affiliations in the development of dark networks (Ronfeldt & Arquilla, 2001b). Such information allows analysts

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

Table 7.1. *Data sets*

Set	Foci	Sources
Enron	What is found about the conspirators?	E-mail, corporate information on organizational position
Sudan	Who are the key actors and did they change as South Sudan separated?	News stories, Twitter, climate, subject matter expertise, demographics
Kenya	What is the role of al-Shabaab?	News stories, Twitter
Benghazi	Did the movie *Innocence of Muslims* incite the embassy and consulate attacks?	News stories, Twitter
Arab Spring (includes Syria)	Can revolution be predicted?	News stories Twitter
Myanmar	Is there a basis for collaboration?	News stories, subject matter expertise

to determine strategies to disrupt dark networks (Everton, 2012b) and to forecast potential impacts of their implementation (Carley, Lee, & Krackhardt, 2001).

Given its integrative nature, it is unsurprising that open source exploitation requires many tools. Examples include the TweetTracker developed by Arizona State University to capture Twitter data (Kumar et al., 2011); REA developed by Carnegie Mellon University (CMU) to capture news (Carley & Pfeffer, 2012); network analysis and visualization packages, like the ORA software also developed at CMU (Carley, 2014); text-mining tools like CMU's AutoMap software (Carley, Columbus, & Landwehr, 2013b); and tools to ensure data privacy, like the social media de-identifier Netanomics developed to remove personally identifiable information (PII) from social media data. Although these tools are all powerful in their own right, they are most effective when linked into tool chains that enable raw data to be collected and cleaned before extracting networks and conducting preliminary analysis, and then re-cleaning data, reanalyzing networks, and forecasting future outcomes. This process is a "mixed-initiative" approach that requires a human-in-the-loop, especially for data-cleaning tasks that require topical expertise, like identifying aliases, concepts of relevance, and bots. Simply put, not all of the process involved in open source exploitation can be done automatically.

To illustrate the state of this field, this chapter employs data and analyses from a large number of studies, as summarized by Table 7.1. Each of these data sets was created using a text-to-network workflow that employs the tools just described (for details, see Carley et al., 2002).

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

II. Lessons Learned: The Current Reality of Open Source Exploitation

The text-to-network process results in attributed, high-dimensional dynamic-network data sets, referred to as meta-networks (Carley, 2002). A meta-network is an ecology of networks that is multimode (it contains multiple types of nodes), multilink (it contains multiple types of relations between nodes), and multilevel (nodes and meta-nodes coexist). Meta-networks can describe the "who," the "what," the "where," the "how," and the "why" among entities and can also summarize their relationships at multiple points of time. Across numerous studies, we find that meta-networks are more valuable in assessing dark networks than are social networks that only cover who is connected to whom.

A. Who Is Critical?

A common network analysis tasks is key player identification; that is, determining which nodes stand out from other network nodes by virtue of their structural position. Multiple metrics for identifying node criticality exist. These centrality metrics are derived from theories of power. An example is total degree centrality, which measures the number of nodes connected to the node of interest.

Centrality measures are deceptively simple and frequently misapplied by analysts. Metrics like degree, betweenness, eigenvector, and closeness centrality were originally developed as one-mode metrics; consequently, the common interpretation assumes one-mode data containing only one class of nodes, such as humans. When these metrics are applied to two-mode or multimode data, analysts need to carefully consider the interpretation. The meaning of degree centrality is relatively straightforward in a multimodal context; it measures how many things the node of interest is connected to. Interpretation becomes more difficult when applying other metrics to two-mode data (Bonacich, 1991). Nevertheless, unless the software prevents it, analysts will often apply these metrics to multimode networks resulting in misidentifying critical nodes and misinterpretation of the results.

Similarly, the meaning of metrics also changes with weighted data. For example, in a binary network, degree measures the number of alters with which nodes interact; however, when analysts apply the standard formulation of degree to weighted networks, they measure the number of relationships in which nodes participate. Thus, a node with a single relationship with a weight of 100 would have higher degree than a node with ninety-nine relationships each with a weight of one (Opsahl, Agneessens, &

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

Skvoretz, 2010). Analysts often commit errors because they neglect to consider the implications of weighting.

If analysts avoid such errors when measuring centrality, the general wisdom is that actors with power, such as leaders, will be highly central. This notion derives in part from the idea that access to information is power, and it assumes all lines of communication are accurately reflected in the network data. However, as Bienenstock and Salwen highlight in their discussion of exchange theory in this volume's first chapter, these assumptions often go unmet in covert networks. Consequently, centrality metrics applied to dark networks hidden in other networks may not find critical actors.

There are several reasons. First, the critical actors may be "hiding" and so rarely appear in the data. Consider the Enron case (Diesner & Carley, 2005; Diesner, Frantz, & Carley, 2005). The data consists of requisitioned corporate e-mails. The art of operating a dark network involves segmenting communication across multiple media; consequently, networks derived from any single form of media may not accurately reflect the entire network. Conspirators in the Enron case could have accomplished this information segmentation by communicating certain information in face-to-face meetings and/or by regularly deleting potentially incriminating e-mails. Indeed, as a matter of policy, many corporations regularly flush the e-mail of top executives. It is, therefore, unsurprising that the Enron corpus contains few e-mails by top executives. Given that administrative assistants typically broadcast information to large groups, the critical actors in the conspiracy were not high in most classical centrality metrics; secretaries were.

Despite such obfuscations, key leaders are still identifiable in the Enron data. Not because they are high in the traditional centrality metrics, but because they are key interstitial actors (top of Figure 7.1) and linked into densely connected working cells (bottom of Figure 7.1). Special patterns and motifs, such as butterflies (a.k.a. checkerboards), near-cliques, and chains, are often more important than traditional metrics when it comes to identifying key actors in dark networks.

Second, critical actors may act only intermittently, such as when a sleeper cell suddenly appears. In the case of the Benghazi consulate attack, the group that conducted the attack hid so well that it was not discussed in the Libyan news or Twitter stream prior to the event (Carley et al., 2013).

Third, the data publisher may alter the data in ways that make traditional metrics invalid. For example, Twitter takes chains of re-tweets and attaches them to the original tweeter, not the user who forwards content created by someone else. So, if *A* re-tweeted *B*, who had re-tweeted *C*, then it would appear that *A* re-tweeted *C*. New metrics are needed,

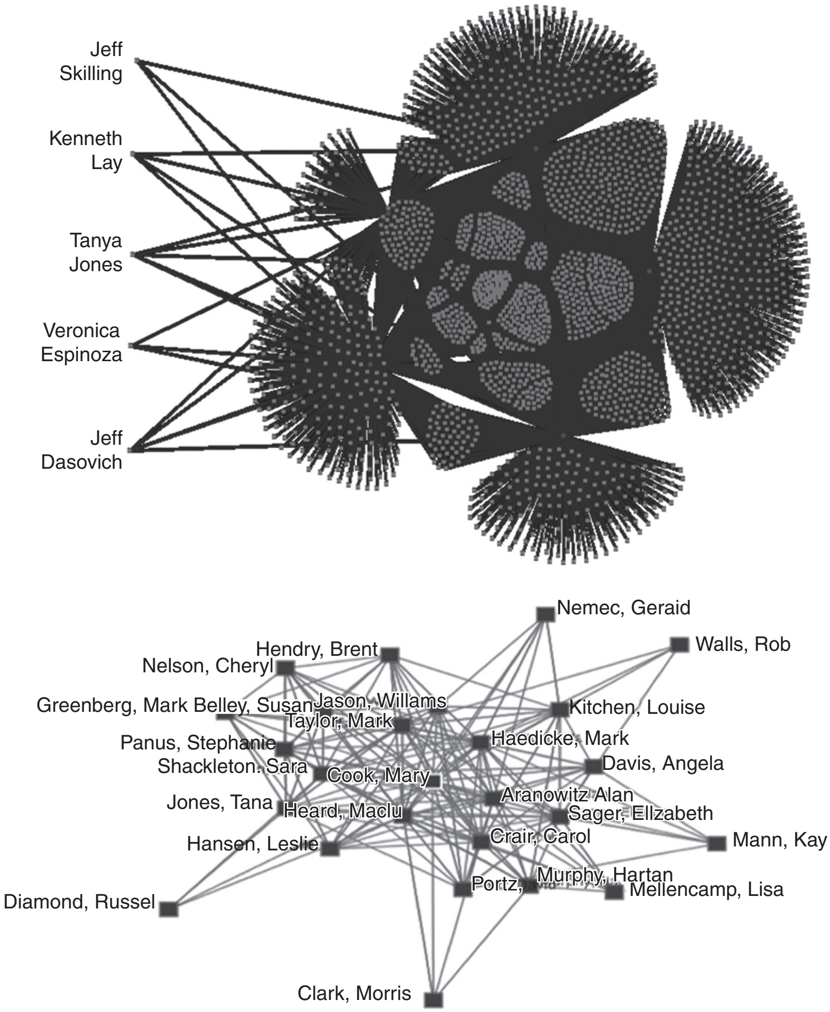

Figure 7.1. Patterns for identifying key actors in dark networks. *Top*: key conspirators are interstitial actors in a fuzzy group mapping. *Bottom*: the cell containing many conspirators.

such as a measure showing which actors show trust in other actors, while taking into account reciprocal re-tweeting and verification. Logic suggests that two tweeters who re-tweet each other regularly follow, attend to, and consider relevant the information provided by each other. The users act in a consistent and mutually friendly way; they exhibit reciprocal trust. Verified actors are those that Twitter determined are who they

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

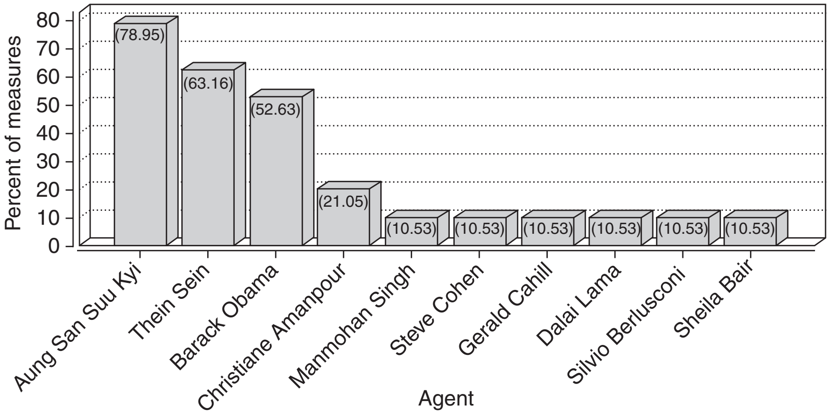

Figure 7.2. Recurring top-ranked actors in Myanmar.

say they are and so can be "trusted" to speak for themselves. Measures that combine this notion of trust with Twitter's concept of verification could offer domain-specific conceptions of centrality that outperform standard centrality metrics in the analysis of data obtained from Twitter and other similar "micro-blogging" services.

Fourth, biases in who uses social media (and how they use it) will impact what stands out as key. Myanmar provides a useful illustration. A CMU team collected news data on Myanmar for about two months, and then analyzed it with ORA. The team also formed meta-networks from co-occurring Lexis-Nexis news tags. News, in general, focuses on celebrities or the occasional human interest story. Consequently, the key actors tend to include celebrities such as politicians, major industrial leaders, Hollywood stars, radio personalities, or sports stars. Without data cleaning, celebrities such as Justin Timberlake and Justin Bieber and major politicians such as Barack Obama and Burmese President Thin Sein dominate the analysis. Removing athletes and other members of the entertainment industry resulted in the key agent list shown in Figure 7.2; Myanmar's political elite came to the fore, while the role of American politicians, who were often tied to celebrities, decreased in importance.

Benghazi provides an additional example (Carley et al., 2013). A CMU-ASU-Netanomics research team was collecting data at the United States European Command (EUCOM) when the Benghazi consulate was attacked and a riot occurred at the U.S. embassy in Cairo, Egypt. The team identified key actors by analyzing tweets posted in the first seventy-two hours following the Libyan attack. Table 7.2 shows top tweeters based on the re-tweet network (outdegree). However, the Twitter handles listed in the figure's first column require some interpretation; for example, "StateDept" is the handle of the U.S. Department of State. Top

Table 7.2. *Critical tweeters and hashtags*

Most Re-tweeted (outdegree)	Most Connected (outdegree)
AlArabiya_Brk	#Benghazi
StateDept	#egypt
Hadeelalsh	#secclinton
BBCBreaking	#gnc
Cnnbrk	#usa
ShababLibya	#us
Reuters_MLGum	#ليبيا
ahlalsunna2	#syria
JedediahBila	#cairo
JomanaCNN	#tripoli
Gatewaypundit	
Eljarh	

tweeters were largely news agencies reporting breaking news. At one level, this ranking simply demonstrates the role of Twitter as a source of breaking news and the prevalence of news agencies and governmental agencies as providers of that news.

At another level, the strong and almost exclusive presence of news agencies suggests that the Benghazi attack was not a ground-up event, but a planned event. When Twitter is used to plan an event, or at least to coordinate it, there are often tweets about where and when the participants should meet. In contrast to Benghazi, the Twitter stream surrounding the Egyptian event contains messages about the reason for the gathering (the inflammation), as well as coordination messages. Moreover, this activity occurred both prior to and during the Cairo protest. Because similar messages from the general public were absent from the Libyan Twitter stream, news agencies dominated Twitter at the time of the Benghazi attack, suggesting that the attack was a planned event, not a protest that morphed into a riot. The research question becomes: What "patterns" need to be extracted from social media to facilitate similar assessments in the future?

Additional forensic evidence can be gained by simultaneously mining both social media and traditional media. Consider the question: In Benghazi, was there an insider in the press who was informed by the terror group so that the press could be on the scene? Potentially relevant digital forensics would include the relative timing of the news articles and the tweets by news agencies, Twitter content, and the timestamps on images. In the tweets collected, there was no evidence of an insider press agent. Rather, many tweets by news agencies referred readers to stories in traditional media. The news stories were rife with speculations that

Table 7.3. *Identification of opinion leaders' roles using topics*

Agent	Unique Topics
THEIN SEIN	SEAFOOD PROCESSING, FIBER OPTICS, COMPUTER AND INTERNET LAW, HARBOR AND PORT SECURITY, LAND SUBDIVISION, INFANT MORTALITY, PRODUCT DEVELOPMENT, FERTILIZERS, REGULATORY COMPLIANCE, BANKING LAW, CRIMES AGAINST PERSONS, ...(39 total)
AUNG SAN SUU KYI	SURGERY AND TRANSPLANTATION, INFANTS AND TODDLERS, SUITS AND CLAIMS, MYANMAR CYCLONE NARGIS, INDIAN OCEAN TSUNAMI, CORPORATE GOVERNANCE, MEDICAL CHARITIES, LIVING STANDARDS, HEALTH CARE POLICY, NATIONAL HEALTH INSURANCE, HEALTH CARE REFORM, ...(22 total)

the event might be a ground-up protest due to the movie *The Innocence of Muslims* – a film that offers a deliberately offensive portrait of the prophet Mohammed. Having multiple open-source media and baseline data enabled rapid digital forensics to refute this speculation.

B. How Can Roles Be Assessed?

It is often vital to know both who is important and what role that person plays. Role identification requires extracting both the social network and the knowledge network because roles can be characterized by a combination of the actor's structural position in the social network (e.g., the re-tweet network and co-mentions in news) and their uniqueness or similarity to others in the knowledge network. Consider Myanmar. Table 7.3 shows topics that are unique to the top two key actors, Thein Sein and Aung San Suu Kyi – the Nobel Peace Prize laureate and prodemocracy advocate who now serves as a member of Myanmar's parliament. Both actors serve as opinion leaders; Thein Sein is tied to "economy_first" because of his concern over business interests, while Aung San Suu Kyi is tied to "quality_of_life," owing to her strong focus on health and natural disasters.

C. What Data Should Be Used?

A standing problem in network analysis is group boundary definition, that is determining who is in and who is out of "the network." For dark networks, this question is particularly critical when the list of actors is not known a priori. When collecting open-source data, the analyst typically provides a set of search terms that guides data capture. Examples

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

include names of groups, countries of interest, names of individuals, lists of key events, and geographical bounding boxes. The search result is a set of documents, many of which contain unwanted information. The dark network is then inferred from connections to known members, but biases in the data and data cleaning can impact who is construed as being in the network.

Again consider Benghazi. Data were collected using geographic bounding boxes for Libya and Egypt, as well as key terms such as "embassy," "consulate," "riot," and for the movie *The Innocence of Muslims*. The resulting data set provided extensive networks; however, as occurred with the Myanmar collection, the data included many actors focusing only on sporting events.

Kenya provides a second example. A CMU-Netanomics team was collecting data on the country in September 2013, when al-Shabaab terrorists overran Nairobi's Westgate Shopping Center and killed more than sixty people in a four-day-long assault. Collection terms included: "Kenya," "al-Shabaab," and the names of the country's political leaders, such as "Kenyatta." The resulting data were again filled with sports news, spam, and random re-tweets by disruptive groups not necessarily associated with the dark networks. The top tweeters were "Reganyare" and "Abdirizak2327," both of whom utilized special software to randomly re-tweet a wide range of information. In both cases, it was necessary to remove unwanted tweets and topics so that the analysis could focus on the core issues surrounding the armed assault and gain insight into the activities of the dark network responsible for the attack.

Two points follow from this experience. First, raw open-source data needs to be cleaned before it is useful. Second, if an individual generates an excessive number of tweets in a short time period, moves rapidly between locations, generates tweets on an excessive number of topics, re-tweets vast quantities of unrelated information, and uses nonsense phrases containing current high-visibility concepts in the body of the tweet, then it is likely that the tweeter is a "bot" masquerading as a human. Analysts should consider removing known bots and overly frequent re-tweeters, and may also wish to filter results to exclude content dealing with celebrity athletes and film starts. Our experience also suggests that analysts should remove any re-tweets originally posted more than a year prior to the period of inquiry.

Topic cleaning is also critical. In the associated knowledge network, such as the Twitter hashtag co-occurrence network or the news topic co-occurrence network, analysts need to pay attention to what nodes to include. Node removal and node merger are necessary so that results are not misleading. For example, as seen in Table 7.2, both the hashtags "#usa" and "#us" occur. They respectively rank fifth and sixth, but, when merged, their rank improves to third place.

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

Analysts should reduce the set of topics/hashtags to between 200 and 500, which is generally sufficient for describing community interests. It is also vital to delete topics/hashtags to which the vast majority of nodes connect; such topics are nondiscriminatory. As a rule of thumb, rank topics/hashtags by frequency of use and remove all that are not used more than twice or by more than two actors. Then, for the top 1,000 nodes, merge those referring to the same topic. While the exact numbers can vary, this type of procedure tends to be sufficient to create a reasonably clean data set for operational use. To merge nodes, techniques based on topic modeling such as latent Dirchlet allocation (LDA), latent semantic analysis (LSA), or simple string comparison can be used. As Kenney and Coulthart argue in this volume's fourth chapter, subject matter expertise may also serve as a viable means to consolidate topics.

D. What Scale Should Be Used?

How should the data be segmented for analysis? Temporally, should the data be broken down into six-hour increments, days, or months? Geographically, should the data be collapsed at the block, city, region, or state level? Two examples highlight that scaling decisions often determine analytic outputs.

- *Syria* – Data were collected using both TweetTracker and REA. When we examined the data by day, we found no relation between the number of daily tweets and the number of daily news articles, but geographic scaling to the city of Homs revealed a different story. The Twitter data signaled a prominent instance of self-immolation, which a refugee conducted to protest reductions in international aid, by exactly one day, but news sources provided more detailed secondary information than Twitter.
- *Sudan* – Data were captured by REA for a period covering the separation of South Sudan from the north, between 2003 and 2008. Because the data exist on a daily level, they can be analyzed by day, week, or month. Figure 7.3 displays a "by-year" analysis; the top panel shows an analysis of topics, and the bottom panel shows an analysis of key actors. These graphs tell very different stories. The topical analysis suggests Sudan was fairly stable; the same basic topics are prominent at each time period. The key actor analysis, however, exhibits volatility; for example, the prominence of Omar al-Bashir, the president of Sudan, increased. Similarly, Sulima Arcua "Minni" Minnawi, who led a large faction of the Sudanese Liberation Army until South

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

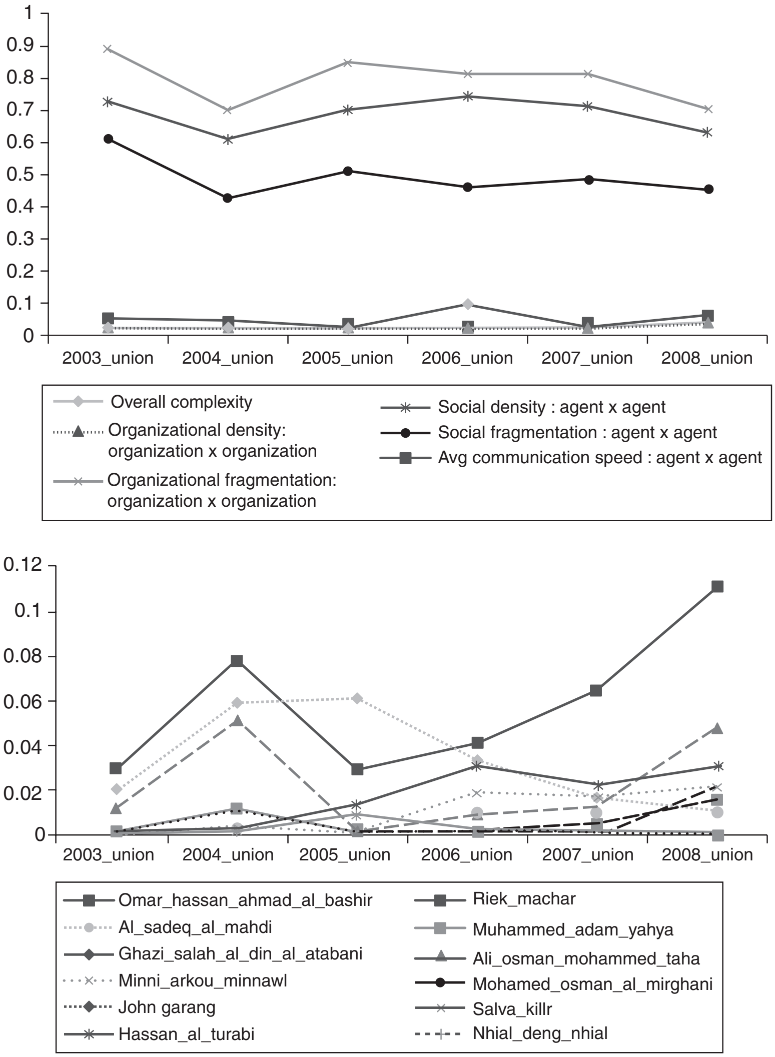

Figure 7.3. Trends in the Sudan.
Top: At the country level, the network is fairly stable; the degree of interconnectivity among the nodes, the relative speed with which information is expected to spread, etc., remain fairly consistent over time.
Bottom: the position of specific individuals within that network is changing, often quite dramatically; Minnawi gains in importance, and Garang, who dies, decreases in importance.

Sudan's secession, emerged as a leader to watch. Monthly scaling shows great fluctuations in topics surrounding the elections. By-day scaling reveals weekly patterns in the ebb and flow of news – especially news outlets' tendency to release more news

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

on Sunday. Human interactions appear random at this scale (Diesner, Tambayong, & Carley, 2012; van Holt et al., 2012).

In practice, we have found that appropriate levels of granularity are dependent on the context of inquiry. Specifically, our experience suggests that analysts studying disaster response, unfolding crises like the Benghazi event, and cyber forensics should use a granularity of four to six hours. By contrast, analysts studying typical news events should use a granularity of one day. Shorter time periods might appear better, but current technology and current organizational practices cannot handle higher temporal frequencies. In general, it is better to first look at chunked data, before moving to fine-grained analysis.

E. When Is an Event Special?

Baseline data enables the critical task of rapid situation assessment. Consider the Benghazi case, which occurred at roughly the same time as violence near the U.S. embassy in Egypt. In order to determine that the two events differed both from one another and from typical events in their respective countries, we used previously collected data from "normal" time periods in the Middle East. This information allowed us to quickly discern that the movie *The Innocence of Muslims* had nothing to do with the Benghazi consulate attack, that the riot in Egypt and news coverage of it were similar to other events in the country, and that the media coverage of Benghazi was atypically expansive for Libyan events. Thus, baseline data allowed us to place these attacks within the larger context far faster than would have been possible if collection efforts began only after the respective events.

F. How Many Sources Should Be Used?

We repeatedly found that multiple sources greatly facilitated analysis. Multiple sources allowed us to identify errors in one source, clarify facts in others, and confirm the timing of interactions. If we had relied on only a single data source, we would have been unable to assess key variables, such as geospatial factors. Individual sources often omit detailed information about where interactions occur, but when sources are viewed in aggregate, they typically reveal crucial details about locations. That said, a cross-media ontology needs to be created and used when combining multiple sources in order to account for the inevitable variation in place names, aliases, and the like. As Kenney and Coulthart's chapter on al-Muhajiroun discusses in detail, such data cleaning tasks can be quite time consuming.

The difficulty of disambiguating public sentiment and speculation from fact represents an additional challenge of media sources. For Benghazi,

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

a key issue was determining if anger over the movie *The Innocence of Muslims* led to the attack. If the only data available to the analyst were newspaper reports, then the movie would have appeared critical, especially during the hours immediately following the attack. This is because the movie is highly central in the topic network and has links to both “consulate” and “attack.” However, in the Libyan tweets, the movie is not relevant; it is absent from the rhetoric. By contrast, the Egyptian tweets suggest that the movie is linked to other riot issues, making *The Innocence of Muslims* an important factor in the protests. Analysts who had access to both Twitter and standard news sources could see within a few hours that the two events had very different media profiles and that the movie was not relevant in Benghazi.

Determining the number of sources to use is difficult. There are a number of issues to consider: Do the sources provide different information, or information at different scales, or at different tempos? How much information is available, and can it be processed in the needed time frame? Does additional data improve understanding; that is, when do analysts face diminishing returns? Can the different data sources be fused? Unfortunately, hard-and-fast rules do not exist to answer these questions. While it is clear that two data sources are better than one, the relative value of the third, fourth, and fifth sources is not well understood. It is clear that the amount of data available can lead to different results, as might occur when analysts base assessments on 1 percent of tweets on a given topic instead of 10 percent of relevant tweets, but the specific biases caused by data availability are just beginning to be understood. It is not clear when more information is advantageous; contrary to popular belief, big data is not necessarily better data.

The answers to these questions depend on the domain of study, the means of analysis, and the type of media included in the assessment. Consider the differences between Twitter and standard news sources. Twitter can, but does not always, provide more up-to-date information, including alerts about what to find in other media and information from the “on-line” public, whereas news sources provide more detailed information, broader speculation, and historical coverage. Some groups use Twitter to mobilize for events, such as announcements in the Egyptian Twitter stream to meet at a specific time and place. However, Twitter is generally not a site for open plan coordination. Rather, bystanders who observe a plan in action and notice the anomaly may post about it.

Twitter can provide information on the pulse of a population. Using Twitter, analysts can assess public interest and examine how various individuals, groups, and organizations try to influence public opinion. However, the observed opinion may not be truly “public” because

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

observations are limited to the subsection of the population who uses Twitter. In many countries, this subgroup disproportionately represents younger, more educated members of society living in the major cities. Consequently, the majority opinion in Twitter may not reflect general public sentiment. For example, in the Kenyan elections, attending to Twitter alone would have led to the conclusion that violence was likely to ensue following Kenyatta's election because there was little support for him. This speculation proved untrue; the majority of Kenyans did not share Twitter users' opinions, and widespread violence did not follow Kenyatta's election. Thus, analysts must understand the demographics of Twitter in a given country before determining that users' opinions accurately represent broader societal consensus.

G. How Can Forecasts Be Performed?

Analysts often need to forecast changes in the behavior of dark networks. Agent-based dynamic-network simulation systems, such as Construct (Carley, 1990; Carley, Martin, & Hirshman, 2009), are particularly well suited to reasoning about the evolution and impact of interventions on networks. For example, simulation facilitates the assessment of questions about the impact of incarcerating the leader of a dark network by looking at the likely behavior of his followers.

A problem with simulation models is that they are generally one-use systems. Model reuse is made problematic by the use of specialized data and difficulties in acquiring appropriate data. However, if simulations use social and knowledge networks that evolve through time, open source exploitation can be used to auto or semi-auto instantiate the model, thus promoting reuse (Joseph et al., 2014). This enables analysts to quickly move from data collection to analysis to what-if reasoning.

Consider the comparison of HAMAS and al-Qaida. Using subject matter literature and news data, we constructed meta-networks representing each organization. We then extracted both the social and the knowledge networks. Figure 7.4 show the resulting social networks of organizational structure, with HAMAS on the top and al-Qaida on the bottom. We next analyzed these networks in ORA and used the results to instantiate the Construct simulator. Finally, we conducted a simple virtual experiment, in which we examined the expected performance of the dark networks with and without their respective top leaders.

Figure 7.5 shows the expected impact of these interventions. The simulations predict that the removal of Osama bin Laden was expected to degrade al-Qaida's ability to operate, while the removal of Ahmed Yasin, the now deceased founder of HAMAS, was expected to actually *improve* the Palestinian group's ability to operate, albeit only slightly.

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

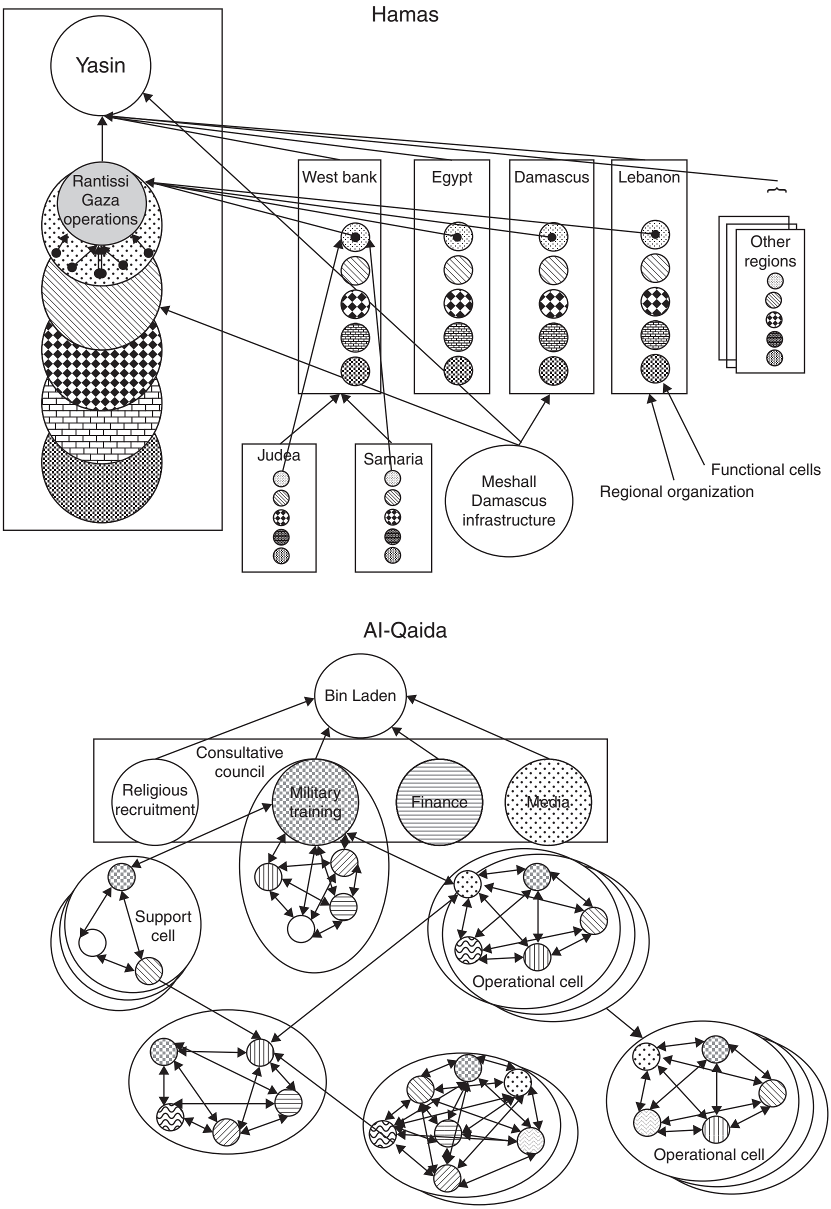

Figure 7.4. Organizational structures of HAMAS and al-Qaida.

H. Reuse: How Could These Technologies Be Used on a New Case?

Open source exploitation is facilitated by linking together many tools into simple interoperable tool chains. This is already the case for

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

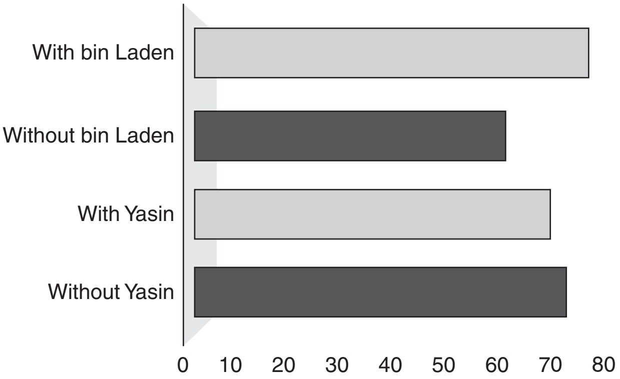

Figure 7.5. Expected impacts of removing the top leader in HAMAS and al-Qaida.

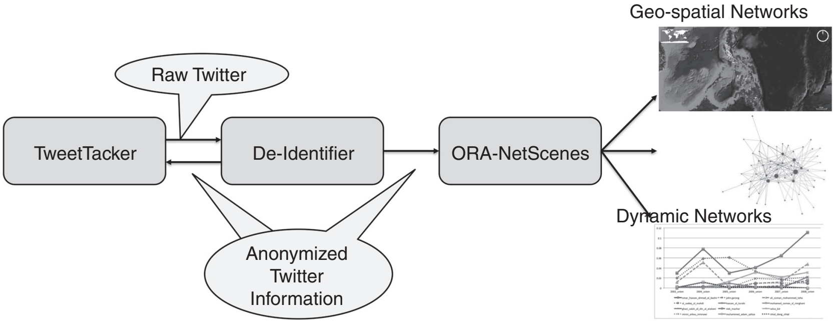

Figure 7.6. Open-source exploitation tool chain.

TweetTracker, the social media de-identifier, AutoMap, NetMapper, and ORA. As Figure 7.6 shows, these tools have been linked into a confederated system supporting reuse. They can be run live or in batch mode, thus supporting regular data collection and analysis. Text-mining tools that support key entity extraction can be run from within ORA to extract information from the body of tweets and news articles.

Other tools can be linked into such tool chains using the confederation process. The advantages of confederation are that no one tool need dominate, multiple tools can be added as they are developed, and existing tools can be improved and reintegrated with almost no cost, as long as they maintain a constant interchange format.

Analysts working in any new topical domain can deploy this tool chain in four steps:

1. *TweetTracker* – Analysts specify the topics, location, and/or specific tweeters on which they want to collect data. The software exports the collected tweets.

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

2. *Social Media De-Identifier* – The software removes personally identifiable information from the tweets before exporting the anonymizied data.
3. *ORA-NetScenes* – Analysts use text mining to identify hot topics, key actors, lines of trust and influence, and actors with special concerns from the de-identified tweets. ORA serves to clean the data by removing apparent bots and uninteresting nodes, as well as merging nodes that represent the same actor or construct. ORA also serves to visualize the data as topic maps and networks, and when the data contains multiple time periods, the software automatically identifies and visualizes trends. Finally, the tool chain produces data exports in a variety of formats (KML, XML, PNG, GIF, TSV, PDF, etc.) used by other systems.
4. *Automation* – Analysts script loops to regularly conduct steps 1–3, thereby ensuring periodic data collection and assessment. If data refinement continues, this human-in-the-loop function necessitates rerunning the ORA script.

III. Technical Challenges

The foregoing examples demonstrate that although many technical challenges have been met, key issues remain:

- *Challenge 1* – Although key entity extractors typically correctly extract 70 to 90 percent of the actors and organizations, results vary by language. Therefore, translation remains a concern. The extracted entities often contain spelling errors or typo variants, such as “Barach Obama” in lieu of “Barack Obama.” Erroneous networks can result, but improved extractors will help to reconcile such variance (Carley et al., 2012).
- *Challenge 2* – Defining the initial search strings used to collect data is difficult, and search strings often need to be modified to improve data collection. Semi-automated filter generation, using directed learning algorithms, would support this costly analytic process. Research on effective data search is needed.
- *Challenge 3* – Reporting biases and “features” of on-line social media tools impact what part of the dark network is observed by affecting the availability of raw data. One bias is celebrity-focused attention, which causes the political elite, entertainment stars, and sports personalities to dominate resulting social networks. As highlighted by the simulations Arney, Bell, Coronges, & Merkl present in this volume’s eighth chapter, ego-focused searches cause additional biases. When a single

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

individual serves as the starting point for a snowball sample, this person is typically among the most central individuals in the resulting social network, irrespective of his or her status in the real world. Analysts can mitigate this bias by selecting multiple starting points and limiting their snowball sample to three steps beyond the starting nodes. However, these recommendations are based on first-person experience; additional research is needed to identify additional biases and to determine empirically valid solutions.

- *Challenge 4* – Representing and accounting for error and uncertainty with network analytics and visualization is difficult in most current systems. Additional research is necessary to identify appropriate methods to model such ambiguity.
- *Challenge 5* – Big data. Many algorithms scale quite well and have been used on networks up to 10^6 nodes; however, those based on calculating the shortest path are less scalable. Incremental approaches appear promising (Kas, Carley, & Carley, 2013; Kas et al., 2013), as do approximation methods (Pfeffer & Carley, 2012b), but more research is again necessary.
- *Challenge 6* – Although open source exploitation supports rapid situation assessment, the process still takes too long. Time-limiting factors include slow data collection due to data access limitations imposed by the Twitter API, low volume of tweets from within country, and low bandwidth and intermittent Internet access. These problems are more pronounced in the field than in a facility with good technical infrastructure.
- *Challenge 7* – Although auto-instantiation of agent-based simulations using the data extracted from texts is possible, it is not clear if non-diffusion models can be auto-instantiated. Future research is needed to replicate this approach for other types of models and to extend it to address issues of timing, levels of violence, and finer-grained analysis.

IV. Conclusions

Open source exploitation is valuable for understanding dark networks. The techniques involved are improving at a rapid pace. More messages, data, documents, and other textual information can be processed faster and with higher fidelity than ever before. However, such techniques must be used cautiously. There is no substitute for understanding the domain and the data.

In general, it is critical to link back to the raw text for a sample of the data. By moving between the coded data and the raw text, analysts

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press

can determine whether the extracted concepts have been over-collapsed or inappropriately merged for that domain. This approach also allows analysts to identify instances when the automated approach is creating false relations, and interaction with raw texts also provides an additional opportunity to gain qualitative insights that support better interpretation. That said, automated techniques are critical because it is simply impossible to process the amount of data currently available without automated support.

To be especially valuable, open source exploitation should not be limited to the extraction of the dark network itself. It is necessary to move beyond connections among actors to embrace the attributes of the nodes and the connections to other factors, such as who is talking about what. Extracting both the social network and the knowledge network facilitates role identification, change-assessment, and impact forecasting of interventions. As it relates to activity in dark networks, analysts need to examine the relative temporality and locality of topics, as well as the groups that discuss them in various types of open-source data. Automated techniques are particularly valuable for extracting, cleaning, and fusing high-dimensional networks from data. In the near future, it is reasonable to expect improved open source exploitation technologies to be leveraged to improve our understanding of dark networks, as well as our ability to predict their behavior.

https://doi.org/10.1017/CBO9781316212639.008 Published online by Cambridge University Press