4

Social Network Analysis and the Characteristics of Criminal Networks

Introduction

The focus for this chapter is on what Sparrow (1991) describes as the 'characteristics of criminal networks'. Sparrow (1991, p. 261) recognised that 'most network analysis tools have been developed within the context of retrospective social science investigations, and they are therefore designed for use on networks which are small, static, and with very few distinct types of linkages (generally only one)'. As shown in Chap. 2, however, criminal networks vary in size, are far from static and have relationships of varying types (Bakker et al. 2012; Burcher and Whelan 2015; Charette and Papachristos 2017). Because of this, Sparrow (1991) identified several unique challenges that complicate the application of existing network analysis approaches to criminal databases.

Sparrow identified four 'characteristics of criminal networks' that would make applying social network analysis (SNA) to criminal groups challenging: the *size* of criminal databases, the *incompleteness* of data, the *fuzzy boundaries* of a network and the *dynamic* nature of social networks. Since Sparrow's (1991) seminal paper, several other 'data challenges' have been recognised alongside incompleteness of criminal network data,

© The Author(s) 2020
M. Burcher, *Social Network Analysis and Law Enforcement*, Crime Prevention and Security Management, https://doi.org/10.1007/978-3-030-47771-4_4

including *incorrectness*, *inconsistences* and *data transformation*. These four characteristics, including the broader category of data challenges, have been regarded by researchers as the primary challenges of applying SNA to criminal networks (Duijn and Klerks 2014; Malm et al. 2008; Yuan et al. 2013). This chapter examines these and other challenges by drawing on the experiences of crime intelligence analysts in their attempts to apply SNA to criminal networks.

Size

Size refers to how large criminal intelligence databases can be and the fact that it may be difficult to process such datasets. Perspective on the size of the increase in global data production is evident in IBM's estimation, in 2011, that 90 per cent of the world's data had been produced in the preceding two years alone (IBM 2011). In the context of global increases in data production, it is understandable that the size of criminal intelligence databases has increased substantially (Ratcliffe 2002; Sheptycki 2017). Law enforcement agencies now collect data from 'traditional' sources, such as physical surveillance, paper trails (e.g. financial records) and telephone intercepts, as well as newer sources, including emails, text messages, computer hard drives and instant messaging (or IM)[1] (Décary-Hétu and Dupont 2012, p. 162). Interviewees highlighted this change, with one senior analyst suggesting that at an organisational level there is a strong push for analysts to obtain information from as many sources as possible:

> We teach through our program a lot in relation to collection and collation, so in that collection/collation process there are a lot of data sources that need to be considered. So, into that social networking we would be expecting them [analysts] to look at the avenues of enquiry that identify perhaps those types of social footprints online, whether it be online, through social media, whether it be through normal telecommunication type platforms.

[1] IM is a form of online chat that uses the internet for real-time text communication. IM apps, which can be used on cell phones or computers, are the end-user applications that facilitate this communication.

> So, within that space or sphere of analysing data or information, whether it be looking at crime data, whether it be looking at data in relation to persons of interest, we would be expecting them to draw out those types of information that would potentially activate other avenues of enquiry to find that type of information that we could draw into that social network analysis. (Analyst No. 17)

By collecting information from so many sources 'the age of "big data" has come to policing' (Joh 2014, p. 35). While there is no universally agreed-upon definition of big data, generally speaking it refers to 'the emergence of new datasets with massive volume that change at a rapid pace, are very complex, and exceed the reach of the analytical capabilities of commonly used hardware environments and software tools for data management' (Akhgar et al. 2015, p. 3). At present, this includes databases that contain terabytes (1000 gigabytes), petabytes (1000 terabytes) and even zettabytes (1,000,000 petabytes) of data.

Almost all of those interviewed expressed a general concern about the information technology (IT) systems available to them (explored further in Chap. 7) and several had specific concerns about the capacity of their software packages to process large quantitates of data when conducting SNA. As several analysts explained:

> From an operational point of view, probably still the biggest thing to us is that we've probably got access to a major amount of information that we're not able to fully assess properly and that's more down to the tools that we've got to assess it with. (Analyst No. 8)

> You are constrained by the power of the software. I've nearly broken Analyst Notebook. I did ring them [software developers] up and ask, how much [data] can I actually put in? (Analyst No. 2)

According to Akhgar et al. (2015, p. 102), Analyst Notebook and other analytical programs popular with law enforcement, including Palantir and SPSS, are incapable of handling even 'one-thousandth' of the information that would be considered 'big data'. Not all criminal databases would meet the definition of 'big data', but it is evident that when conducting SNA, there are restrictions on the quantity of data analysts can include due to the processing limitations of the software available to them. This would suggest there are limits on the ‘size’ of the criminal networks that analysts can examine (research participants were comfortable with broadly discussing the processing limitations of their software but were unwilling to state specifically what these limits are, citing security concerns). In this instance, size refers to both the number of actors/relationships included in the analysis and the volume of data about the network, including actor characteristics and relationship types.

Research participants also expressed several other concerns with the size of their criminal databases and how size can impact on their ability to use SNA. For example, one analyst suggested that it can be difficult to visually display large networks: ‘when you have a huge amount of data it can look like a ball of string, so at the end of the day it’s not really useful to present as a picture, you have to analyse it; when you have a lot of data it just can’t be presented in its entirety’ (Analyst No. 3). This does not mean that the link diagram component of SNA should be dismissed entirely. It is still regarded as a very useful way of presenting the results of an analysis (Strang 2014; Tayebi and Glasser 2016), for example in instances such as an analyst presenting their findings to a detective (discussed further in Chap. 6). There are also some instances where a visual inspection of a link diagram is the preferred method of analysis over mathematical computations. An example of this would be applying SNA to a small network on multiple occasions, where any changes in the network’s structure would easily be seen from a visual inspection of the link diagram (see Mullins and Dolnik 2010). Overall though, this analyst’s view of SNA would support the notion that once a network reaches a certain size the link diagram component becomes rather meaningless, in which case it becomes necessary to apply suitable mathematical computations (Bichler et al. 2016; Colladon and Remondi 2017; Johnson and Reitzal 2011; Roberts and Everton 2011).

Another concern with large criminal databases, as one analyst discussed, is that there is ultimately a trade-off because although ‘intelligence analysts will want more information, […] that can also create a great deal of distraction’ (Analyst No. 5). SNA is already regarded as a time-consuming process (Bright et al. 2015b; Kriegler 2014; Xu and

Chen 2005); as the quantity of information increases, the time it will take analysts to 'clean' the data also increases (see data transformation below). Software packages have developed considerably over the past two and a half decades, making data processing far more 'user-friendly' (Duijn and Klerks 2014, p. 150). However, in this time the rate of data collection and retention by law enforcement agencies has also rapidly increased (Brodeur and Dupont 2006; Décary-Hétu and Dupont 2012; Hutchins and Benham-Hutchins 2010; Sheptycki 2000). The impact of this massive push within law enforcement agencies to continually collect more data means that they have exceeded their ability to fully analyse this information. Analysts are unable to analyse networks that are above a certain size and thus cannot use SNA to its full potential.

Data Challenges

The second characteristic of criminal networks identified by Sparrow (1991) is the *incompleteness* of criminal databases. Since then several other issues that can be placed under the umbrella term 'data challenges' have been identified as impediments to the accuracy and efficiency of SNA. They are *incorrectness*, *inconsistency* and *data transformation* (Morris and Deckro 2013; Xu and Chen 2005).

Incompleteness

Criminal network data will inevitably be incomplete due to some actors, their relationships and the characteristics of those connections going unobserved or unrecorded by law enforcement (Malm and Bichler 2011; Medina 2014; Saidi et al. 2017). Incomplete data can distort the outputs of SNA (Décary-Hétu and Dupont 2012; Hofmann and Gallupe 2015). For example, it may be that an individual that has a low-centrality score (indicating they have few relationships with other actors) is in fact well connected to the rest of the network, but that these relationships have simply gone unobserved. Most interviewees identified incomplete data as a core challenge of applying SNA to criminal networks. As several analysts discuss:

> Social networks or the absence of social networks doesn't say that there is no connection there, it just says that we can't see a connection there. So, there's some blind spots. (Analyst No. 19)
>
> I think it's [limitations of SNA] the initial data inputs. The tools are great, the theories are great, but you don't always have access to the amount of information on a given individual or given group that would make those tools and theories effective. (Analyst No. 5)
>
> We really don't have too much to base those [SNA] assessments on. Using just the [name of this organisations primary criminal database] system which is only going to show people seen together at the same time, spoken to at the same time, somebody, some cop somewhere happened to think they were connected in some way. Sometimes the information that we have got to assess is not great. (Analyst No. 25)

Although it was noted in Chap. 2 that the mathematical computations used in SNA are relatively robust to missing data (up to 10 per cent missing), these analysts suggest that their criminal databases are often too incomplete to conduct SNA. While there are numerous factors that contribute to the incompleteness of criminal network data, one is the biases that exist within the investigative methodologies used by law enforcement (Bright et al. 2012; Burcher and Whelan 2017; Duijn and Sloot 2015; Frank 1978; Ratcliffe and Sheptycki 2004).[2] For example, mathematical computations can be distorted by the amount of data collected on particular individuals. This can occur when police collect large amounts of data about certain actors, not because they are necessarily central to the targeted network, but as a result of focusing their attention

[2] Sometimes called organisational or institutional bias (Gunnell et al. 2016; Ratcliffe and Sheptycki 2004), investigative bias should not be confused with the issue of 'cognitive bias' discussed in Chap. 3. Investigative bias is also separate from 'bias crimes' which can be defined as 'a criminal offence committed against persons, associates of persons, property or society that is motivated, in whole or in part, by an offender's bias against an individual's or group's actual or perceived; race, religion, ethnic/national origin, sex/gender, gender identity, age, disability status, sexual orientation and homeless status' (Mullane 2015, p. 5).

on these individuals more than the rest of the network (Bright et al. 2012). A further example of investigative bias, as one analyst explained, occurs when police are conducting surveillance operations and those being targeted may only be under surveillance at certain times of the day due to resource constraints and because of the inability of law enforcement to access certain locations:

> Sometimes we learn about relationships we were previously unaware of through the physical surveillance of criminal targets. That surveillance may also be not uniform [complete surveillance coverage] in that it happens at certain times of the day, more likely than others. It might be that those relationships involving people who have met in certain circumstances in public spaces versus buildings we don't have access to, might also colour that picture that's developed of the social relationships. […] But whether they form a representative picture of the most significant relationships that person has, or a complete picture. It's possible in some families that are constantly interacting with the police that we have that picture. But it's likely in many others that we have nothing like the full picture and we may not have the most significant relationships recorded at all. (Analyst No. 18)

According to this analyst, when surveillance is not 'uniform' it is likely that they are not capturing a 'complete picture' of the targeted network. This task is even more challenging when criminal groups employ counter-surveillance measures, such as having local residents or family alert the group of 'unusual police activity in the neighbourhood' (Spapens 2011, p. 30). While it is unlikely that police will ever have complete surveillance or a complete picture of the networks they are investigating, this highlights how the choices they make around who they focus their attention on will introduce a degree of bias into the information they collect. The same analyst also suggested that incomplete or 'unrepresentative data' can be the result of using information from previous investigations that was not collected with the objective of conducting SNA:

> When we are talking about unrepresentative data samples as well, we need to be aware when we are collecting data for a purpose directly related to building a picture of the social network of the target, versus data that was collected with some other objective in mind, that data collected more gen-

> erally will have an application in social network analysis but the amount of bias in it or the completeness of it will differ. Sometimes we are collecting data in one mode at one stage of an investigation and then move to another mode, but we treat the entire data pool as being accessible, as it should be, for social network analysis without taking into account that there was a bias in the collection at one phase. (Analyst No. 18)

The point here is that many of the data sources used by law enforcement agencies, including surveillance and previous investigations, are likely to be incomplete due to varying degrees of investigative bias. This reinforces a common finding of this study that decision-making, particularly with regard to target selection, should not be based solely on the findings of SNA, and instead should be used alongside other forms of intelligence analysis to influence this process. Other forms of analysis that are likely to pair well with SNA include crime script analysis (Bright 2017; Duijn and Klerks 2014; Morselli and Roy 2008), geographical information systems (or GIS) (Ratcliffe 2004) and criminal business profiles (NCIS 2000).

Despite investigative bias being a key concern for the research participants, the investigative methods employed by law enforcement agencies may not have much of an impact on SNA (Berlusconi 2013; Bright 2015; Bright et al. 2015a). For example, a study by Bright et al. (2015a), which applied SNA to an Australian drug-trafficking network, found that despite the fact that the data they had used was from the police files of only two individuals in the network, they were not always the actors that had the highest centrality scores, as might have been expected. So although the police primarily focused their attention on these two actors, the data they collected did not appear to distort the centrality measures applied. Bright et al. (2015a) concluded that this inherent bias in the data used could not account for all results. Therefore, police files focused on particular individuals may contain sufficient information about co-offenders so as not to skew findings to the point where they are unusable. While this gives a degree of confidence to analysts that the biases that exist within their investigations may not impact on the findings of SNA, further research is required to fully understand this issue.

Interviewees also noted that criminal databases are often incomplete due to poor data entry standards. For example, one analyst explained that they are often left with incomplete data internally because officers do not always fill out reports completely: 'sometimes you can get incomplete parts of the puzzle and information reports that aren't filled out completely, so you're missing bits and pieces' (Analyst No. 21). Poor data entry standards are explained, in part, by the increased collection of data discussed earlier in the chapter. For example, according to Victoria Police, 50 per cent of an officer's time during each shift is actually spent on various tasks associated with information capture and reporting (Victoria Police 2014). It is not clear, however, to what extent officers fail to fill out forms completely. Again, this is an area that requires greater examination as the extent of the problem may vary in different areas within law enforcement agencies and each jurisdiction.

Incorrectness

A further data challenge is that collected relational data can be incorrect, a factor which can, in turn, affect the outputs of SNA (Morris and Deckro 2013). Xu and Chen (2005) suggested that incorrect data, such as information about an offender's identity, physical characteristics or address, is often the result of either data entry errors by law enforcement officers or intentional deception from offenders. The deliberate supply of misinformation is arguably the most challenging source of incorrect data and the greatest difference between 'bright' and 'dark' networks. Dark networks refer to groups that 'operate covertly and illegally' (Bakker et al. 2012, p. 4). Bright or 'light' networks are those not engaged in criminal activities (Everton and Cunningham 2015; Morris and Deckro 2013), such as sports or social clubs. A common form of misinformation involves offenders using multiple aliases or nicknames to mask their identity. This issue is explained by several analysts:

> The other limitations are obviously privacy and naming conventions. For the majority of social media sites that I've seen, there doesn't appear to be any rock-solid verification procedure. I could establish a social media pro-

> file under the name of Bill Blogs and I've got a phone that I bought from Safeway [a supermarket] for $50 and used the name Bill Blogs to activate the phone and there's my whole verification procedure. There's nothing that actually verifies who someone actually is. (Analyst No. 1)

> Nine out of 10 guys will not use their correct name, they would be an idiot if they did. So, you're getting things like nicknames. (Analyst No. 10)

This will of course depend on the data source being used. As these analysts suggest, many social media sites do not have an identity verification process. However, other sources commonly used by law enforcement, such as financial records obtained from a bank, will generally have stricter verification processes. The main point here is that a problem arises for law enforcement agencies when they incorrectly record each aliases/nickname as separate actors and not as the same person. Therefore, when an analyst applies SNA, the network will appear larger than it actually is due to the counting of an actor multiple times. It will also reduce the importance or centrality of an actor due to their relationships being split across multiple identities (Malm et al. 2010).

One analyst noted that it is common practice within law enforcement agencies to apply a reliability and validity weighting in an effort to reduce the chances of incorrect data entering an analysis:

> That's where you use a weighting to describe the reliability of a piece of information. So, you'll give it an assignment of a number and a letter, A1 being the most reliable, F6 being the least reliable. That's a formal way and commonly used in intelligence analysis practice to describe the reliability of information by which you're determining a connection. Sometimes that's not available, so you would make an informal call on the weighting, if it's single source, so you've only got the connection from one particular intelligence input. It's less reliable than if you have multiple sources, preferably three sources says a connection is good and more than that the better. (Analyst No. 5)

As this analyst explained, a reliability and validity weighting is a determination on the trustworthiness and accuracy of certain information. While there are a variety of reliability and validity weightings used within the intelligence field, in this instance the letter designates the reliability of the source ('A' being most reliable and 'F' being an untested source) and the number represents a determination on the validity of the information ('1' means that the information is known to be true and '6' that the information is likely to be false). According to Carter (2009), reliability and validity weightings also play an important role in resource allocation. Using a simplified example, Carter (2009) explained that if a law enforcement agency receives information about a possible terrorist attack, but that information has a very low reliability and validity weighting, little credibility will be placed on the threat. Alternatively, as 'validity and reliability increase, the greater credibility will result in devoting more resources to corroboration and a response' (Carter 2009, p. 163). However, there currently is no convenient way to incorporate reliability and validity weightings into SNA (van der Hulst 2009). For intelligence analysts with access to Analyst Notebook it is possible to classify the relationships in a network as 'confirmed', 'unconfirmed' and 'tentative'. One analyst discussed their use of this function:

> [I] have used whether it's a confirmed or unconfirmed connection. So just because people are seen in a building doesn't necessarily mean they're there together. So, that's helped in the past and that comes into when you're doing your analysis to take into account that sort of thing. That's what the program will give you but you have to analyse it yourself. (Analyst No. 23)

This approach, however, still fails to take into account the complexity of the reliability and validity weightings commonly used by law enforcement agencies. Exploring how such weightings can be better included in SNA should be the focus of future research. Nevertheless, such weightings as they currently stand are an effective way of outlining the reliability and validity of the data used. It is for this reason that researchers may benefit from incorporating a reliability and validity weighting into their applications of SNA to criminal networks.

Researchers broadly discuss the reliability and validity of the data they use and highlight any limitations (Berlusconi 2013; Duijn and Klerks 2014; Malm et al. 2008; Varese 2013), but they mostly do not apply a formal weighting similar to those used by law enforcement agencies. For example, Duijn et al. (2014) split their data into two groups, 'dataset hard' and 'dataset soft', to indicate the relative reliability and validity of their data. However, this is still simplistic compared with the system discussed by Analyst No. 5 above (A1-F6), which involves a total of 36 possible categories regarding the reliability and validity of a piece of information. Researchers make an assumption that data that has already passed through the hands of police and subsequently through the courts has had 'its validity tested' (Leuprecht and Hall 2014, p. 95). However, researchers have gathered data from a wide variety of sources, not all of which assess the reliability and validity of the data as the courts do. According to Bright et al. (2012), the sources of data used by researchers applying SNA to criminal networks can be divided into five categories: offender databases, transcripts of physical or electronic surveillance, written summaries of police interrogations, transcripts of court proceedings, and online and print media. In addition to these sources, researchers have also used data from archival interviews (Athey and Bouchard 2013), non-government organisations (Everton and Cunningham 2012), interviews with former investigators (Koschade 2007) and think tanks (Leuprecht and Hall 2013). While all studies would benefit from using a reliability and validity weighting, it is strongly recommended that those that use these additional data sources with online and print media incorporate weightings similar to those used by law enforcement to reduce instances of incorrect data infiltrating the analysis.

Inconsistencies

A further data challenge is the inconsistencies that can occur in criminal databases (Sanders et al. 2015; Xu and Chen 2005). Several of those interviewed identified inconsistent data as a key challenge when applying SNA. For example, one analyst explained that if information is inconsistently entered into their databases it essentially becomes lost:

> So, if you wanted to find out, you might want to find out a particular person who's got a specific tattoo. So, you can run a query [search] for a tattoo, but in order to find that tattoo, someone at some point would have had to

> put in a form. […] But what you will find is that people or members, they don't put in that specific form, they will actually mention it in the dossier, which is the narrative of that particular offender. So okay, I'm trying to locate Joe Blow who could have, I don't know, a big tattoo on his head that says 'criminal', or something like that, and if I put in a query for males, 25–30, whatever, have got 'criminal' [as a tattoo]. It [the database] will never bring it back because it will be in the narrative or the dossier and then if you're going to try and search the dossier of every 25–30-year-old criminal in Victoria it will probably take six months. (Analyst No. 14)

Therefore, a situation can arise whereby analysts and detectives cannot 'see' certain information that may be relevant to their investigation. For analysts, the outcome is that there may be valuable relational data that they are not aware of and is subsequently not included in an SNA.

This issue relates to a broader point raised by many research participants concerning data standard: *what* is being stored and *how* it is being stored (Sheptycki 2004). Several analysts also noted that the data they receive from external groups, including public and private companies, is often inconsistent. Given that these are critical sources of data for law enforcement (Victoria Police 2014), this is an area that needs examination. One analyst discussed how there are inconsistencies in how long data is kept by certain private businesses:

> How long information is kept is an issue, because if somebody doesn't hang onto data then obviously, you can't use it. Locally that's seen if there is surveillance footage from a 7/11 or a service station or a shopping centre, that's great stuff, you can use that sometimes. But you have to jump onto it straight away because they only keep that thing for two weeks, sometimes a month, sometimes a week, some places only keep it for a few days. If police haven't turned up in a few days to say hang on we need such and such, it gets written over. That just comes down to their finances. (Analyst No. 7)

Many analysts were also frustrated with the lack of consistency in the information they receive from certain industries, such as telecommunications. As one analyst explained:

> What you're going to get from one telco [telecommunications] company is quite often different from the outputs you'll get from different ones and the amount of hoops you'll have to jump through to get that in the first place. It's always a balancing act of risk assessments of what you would really like to get, what's available and how difficult it is to get it. (Analyst No. 8)

In Australia, steps have been taken to address this issue with the *Telecommunications (Interception and Access) Amendment (Data Retention) Act 2015*, now requiring telecommunication providers to retain particular types of metadata for two years.[3] This law is designed to ensure that there is consistency in the telecommunications industry concerning what information is collected and how long it is held for.[4] The problem for law enforcement agencies is there are a number of other industries and potential sources of information, such as privately held CCTV footage, where no such laws exist. These other sources of relational data vary dramatically in their availability and consistency. This issue is compounded by the fact that even when analysts do have the information they require to conduct SNA, it can be an immense challenge to get that data into a usable format.

Data Transformation

Once analysts have collected the necessary relational data to conduct SNA, it will need to be converted into a usable format. A requirement of SNA is that actors are represented as nodes and that relationships are represented as ties (Xu and Chen 2005). However, many criminal databases, particularly older ones, were not designed to store data in a format

[3] Telecommunications metadata consists of information including caller identity as well as call time and place. It does not include the contents of a phone conversation. For internet activity, it includes the email address and when it was sent, but not the subject line of an email or its contents.

[4] It is important to note that there has been considerable debate surrounding this law and similar legislation from overseas due to privacy concerns (see Breyer 2005; Newell 2014). For example, in 2006 the European Union Parliament passed what has become known as the 'Data Retention Directive', requiring member countries to store telecommunications metadata for at least 6 months and for up to 24 months (Vainio and Miettinen 2015, p. 290). However, in 2014 the Court of European Justice ruled that the Directive was invalid, stating that it violated fundamental rights, namely a right to privacy (Vainio and Miettinen 2015).

that can easily be used in SNA. Analysts will often have to manually go through their databases, extract the data and then convert it. For example, evidence of a relationship between two individuals might be recorded within a criminal database simply as 'Actor A was observed meeting with Actor B'. This information must then be manually entered into an association matrix (see Chap. 2), which is often created using Microsoft Excel. Several analysts complained about the time-consuming and labour-intensive nature of this process. As one analyst explained:

> I need to spend the next day working out how this different source of information, how I can integrate this with the information I've already got in a meaningful way […]. It really comes down to people with enough energy and skills to be able to do something but to not be put off that I have to spend the next eight hours manually manipulating spreadsheets to be able to get them in a form that I can input to a chart. [However,] we should not be at a point where we are […] manually mapping networks […], it should be far more automated in the way that we do it. That's the barrier that we at least have some control over, […] but it's still only scratching the surface. (Analyst No. 8)

While it was certainly evident that the time-consuming and manual nature of converting the data into a usable format was an important consideration, it was not clear from research participants to which extent they simply choose not to use SNA because of these factors. Furthermore, the point made by this last analyst concerning the automation of the data transformation process, and more broadly the application of SNA, was only raised by the one individual. There has been some research into automating SNA. For example, Ball (2016) suggested that 'data mining' and 'text mining' techniques will be critical to this process. Data mining involves the use of various algorithms to identify patterns in very large datasets (Ye 2014). Text mining is the analysis of natural language text, including the extraction of specific information from datasets (Weiss et al. 2005). However, Ball (2016) also cautioned that a great deal of further research is needed before the automation of SNA can be realised, including the development of the algorithms necessary for the extraction of social network data from criminal databases and open sources. Overall,