Social Network Analytics in Policing and Security

Efficacy and Ethical Considerations

Derek Dempsey, Independent Consultant, Empyrric

Social Network Analysis (SNA) is the establishment of connections between entities, typically persons (nominals) and organisations and the generation of networks showing that one entity is connected to other entities in the network. In crime investigation social network analysis seeks to identify associated persons to aid and direct investigation. If suspect A is connected to person B and person B to person C then persons B and C may be of interest for an investigation. However, while Person B is directly connected in some way to A, Person C is only indirectly connected and we should require additional evidence to indicate any kind of actual association with A.

Application of SNA in Policing

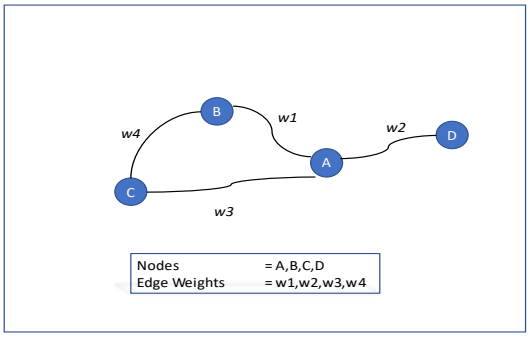
In policing, network analysis allows information and intelligence stored on the police data systems to be systematically processed to identify connections between nominals or individuals and thus potentially identify gang members and organised criminal activity in a more efficient and effective way. While the data may be in the system it is only through automated data processing and network generation that connections can be identified in a consistent and timely manner, thus generating new leads and freeing time for in depth investigation.

The process of network generation needs to closely controlled to ensure that individuals or nominals are not erroneously linked to gangs or criminal activity, or identified as suspects on the basis of spurious network connections. Ensuring that this due diligence takes place requires a governance process to be in place. This is one of the roles of the Ethics Committee but it is also encumbent on the police data science teams to ensure best practice principles are being applied, that there is transparency in the process and that guidelines for operational usage are provided.

Network Analysis in policing is primarily descriptive and aimed at providing improved intelligence for investigators. One of the main applications is through dashboards that allow users both to investigate specific cases as well as explore group dynamics for insights that can direct other types of intervention.

Graph Theory

The basis of network analysis (graph analytics) is derived from the mathematical discipline of topology. In the study of graphs or networks we can distinguish **nodes or vertices** which are the endpoints and **edges** which connect them. Networks can be **cyclic** as we see below with A, B and C or **acyclic** if here are no cyclical elements. They can be **directed** or **undirected** which would represent a direction of travel along the edges. For most practical purposes networks tend to be undirected and only these will be considered here.



.

Figure 1: Simple Network

Figure 1 shows a simple network connecting 4 nodes (or entities). The network has 4 edges each of which is associated with a weight. In practical terms this usually represents the degree of connectivity in cases where there is some variation in the edge criteria. If all edges are equal these will default to 1 and this is the most common configuration.

This network can be described by the number of nodes (4) and edges (4) and its cyclic property. Other mathematical metrics can be applied but we need to look at more complex networks for the descriptive metrics most commonly used in practical analysis.

Centrality

One of the most widely used metrics for analysing networks is Centrality. This aims to find the most important nodes in a network. There may be different notions of importance and hence there are several centrality measures. The most commonly used are:

Degree Centrality:

> The simplest Centrality definition. This is the number of edges connected to a node. In Fig 1, Node A has the highest degree centrality with 3 edges.
>
> The higher the degree centrality the more important the node.

Closeness Centrality:

> This is the average length of the shortest path from the node to all other nodes in the network. In Fig. 1 this is 1 for node A but 1.33 for B and C and 1.67 for D (average 'hops' to each other node in the network).
>
> The lower the Closeness Centrality the more central the node.

Betweenness Centrality:

> Number of times a node is present in the shortest path between 2 other nodes. Once again this will be A in the Fig 1 which appears in 2 of the shortest paths (B – D, C – D) while none of the others appear in any.
>
> The higher the Betweenness Centrality the more important the node.

These centrality measures also have variations and can be implemented using different algorithms. This means there are different definitions and algorithms used in identifying key network measures.

Network Density

Network density is a measure of how connected a graph is and usually defined as the ratio of actual edges to potential edges if every node was connected with every other. In Fig 1 we have 6 potential edges for full connectivity and 4 actual edges giving a density of 0.67 or 67%. The actual definition will vary depending on type of graph and, once again, there can be variations in this calculation.

This measure is sometimes also called 'Connectedness' and is also a measure of how 'cohesive' the network is.

Network Randomization

Network Randomizations aims to establish a probability estimate for the network by comparing this with 100s or 1000s of randomly generated networks given the data constraints. In theory this provides an indication of the significance of this particular network configuration.

In practice this is a difficult metric to calculate.

Network Analytics in Practice

In practice most networks are homogenous with respect to nodes or entities but heterogenous with respect to edges. In other words, if our entity is a person they can be connected in numerous ways to other persons – by name, address, phone number, location, workplace, social places, social networks and so on. In practice therefore edges usually need to be labelled or identified to provide visual clarity of the connection.

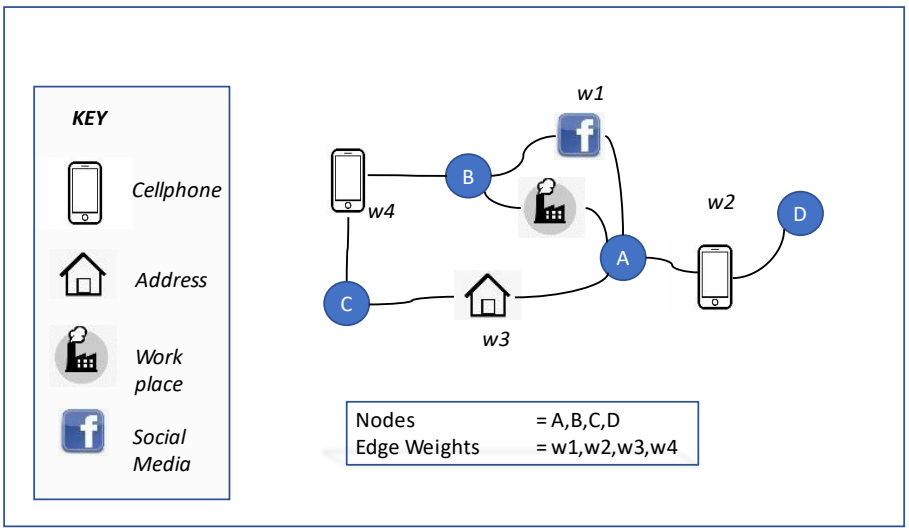


Figure 2: Network with Multiple Edge Types

Here we can see that it is possible for nodes to be connected by more than one edge. In this case, A and B share a common workplace and also a social media connection. Typically this would strengthen this connection and thus we can expand further on the network description concepts such as density if we wish. However, for the purpose of this document, it is not necessary to delve into the intricacies of network analytics in too much more detail.

Once we have heterogeneous edges then we may want to assign weights based on edge type importance, for example, the cell phone connection is weighted as stronger or more important than the social media connection. This can in turn be incorporated into additional metrics.

In other cases, links are treated purely as binary connections: there is a link or not. This has the advantage of simplicity and treats all links as of equal strength. Much of the network generation in police and security contexts takes this form. This then adds extra onus on the operational guidelines to determine the nature of the link established – is it from prosecutions, from interviews, from prison data and so on.

Common Data Science Challenges

The establishment of a connection is not an exact science and can generate spurious connections particularly as the degrees of separation or 'levels' increases. In Milgram's famous paper of 1967[1], that established the famous 'six degrees of separation' hypothesis, Milgram showed that any individual on the planet could be connected to almost any other individual through, at most, six connections, through a series of experiments[1]. The result is disputed, particularly given modern communication methods, but the principle is almost certainly correct. Once you progress beyond the direct connection, 'one hop', then the existence of a real connection becomes increasingly inferential. Therefore, while direct connections are clearly relevant any connection at the second degree or beyond needs to viewed with a degree of caution.

This is particularly relevant to law enforcement and security to ensure that individuals are not assumed to be criminal or part of a criminal gang, because they are connected in some way to a known criminal.

In policing and crime investigation this is the concern about 'peripherals' which, simplistically, means that if you have some 'social network' connection to a known criminal or suspect then you could be identified as a potential suspect regardless of any other evidence. The question of association will always be problematic in any situation. It is certainly questionable whether being identified as a network connection provides a sufficient basis for investigation or intervention activity on its own but, where networks are identifying vulnerable individuals for safeguarding activity this may be different. These may be precisely the individuals where some intervention is most useful.

A second factor to consider is that many systems use some form of fuzzy matching so that any specific connection may itself be either inferential or inexact. For example, a common address may refer to the exact same property but often the connection will be established based on living in the same block of flats for example, or indeed, living in the same postcode, same street and so on. In part this is inherent where addresses may be recorded differently in different systems and in part this is because the scope of the investigation may seek to identify individuals who live or work in the same vicinity.

Any use of fuzzy matching, inference or data enhancement needs to be explicit in the information shown by the system so that an investigator can assess the degree of match. Addresses are often 'cleaned' by reference to the Post Office address file (PAF) which provides the standard UK benchmark but this process may also require a degree of inference. These may only be the marginal cases but any inference of this type needs to be captured somewhere and made explicit. This also informs the

[1] *Milgram, Stanley (1967). "The Small World Problem". Psychology Today. **2**: 60–67.*

previous judgment about 'peripherals'. The mere establishment of a network connection should always be followed by review of the nature of the connection defined as a matter of procedure.

Social Network Analysis (SNA)

SNA is not a form of behavioural analysis of persons or nominals in any real sense but it can be extended to include elements of behavioural analysis. This capability is perhaps overstated by the proponents of the method. It connects individuals or entities to other entities within the data universe that provides its source. However, there are also differences in understanding the nature of the social network. In one guise it uses static data elements such as addresses and phone numbers and establishes connections between them. This creates a network of potential contacts or associates within the data pool. This is the common form of usage in the fraud and security domain which are primarily driven by amalgamating data from multiple sources.

In another guise SNA uses relationships of family, friendship and known associations based on information gathered in interviews and investigations or through text analysis or social contacts analysis. In this form it has a different character but all the previous metrics still apply. Here it is the reliability of the intelligence that drives the reliability of any results. Once again, any connections derived need to be transparent in terms of their nature and provenance.

Typically, multiple data sources are used and one of the challenging problems of data-based network analysis is entity disambiguation or identifying unique entities across all its data sources. It is not always straightforward to show that person A in dataset X is the same as person B in dataset Y. This is easier where a national ID is being used but, where name, address and, perhaps, date of birth are the main elements there can be challenges and false identifications can occur. If they do occur then this has a knock-on effect. Alternatively, it is not uncommon to find the same person appearing as two or more distinct entities but this is part of a continuous process of iterative enhancement that most data projects require and unlikely to be an issue within the police data.

Three Categories of SNA

McGloin & Kirk[2] proposed three categories of SNA:

- Descriptive graphs
- Network measures
- Advanced network modelling techniques.

Descriptive graphs are those that show the networks but without applying any metrics to them. This probably accounts for 90% of the real and effective usage of these networks. They visually show, and can show in summarised report format, a network connecting various entities, usually persons in various ways. A summary report will provide summary metrics on the size of the network, the number of entities, the number of connections and so on.

Network measures include size, density and centrality as outlined previously. The challenge of generating viable and meaningful network measures is very real and their interpretation remains open to debate.

[2] McGloin, Jean & Kirk, David. (2010). An Overview of Social Network Analysis. Journal of Criminal Justice Education. 21. 169-181.

Advanced modelling techniques are more problematic and not extensively used. In theory these can be layered onto the Network measures but in general they are simply an extension of the network measures rather than representing anything genuinely additional or enhancing. They can be useful in operational contexts to automatically identify networks of interest rather than have this driven by interest in specific entities or individuals.

Network Dismantling

This is a relatively new approach to network analysis that aims to model the removal of key entities or nominals, breaking down a large network into subnets. This can be useful from many perspectives, particularly in modelling the impact of interventions that target key entities. Often the analysis addresses the cost-benefit analysis, cost of intervention versus the impact, to optimise impact versus cost.

The approach is also used to measure network robustness. The principle here is that those networks dependent on one or two key entities are not robust if they fragment greatly on the removal of these key entities while networks that retain greater structure are more robust and will persist beyond the removal of the key entities.

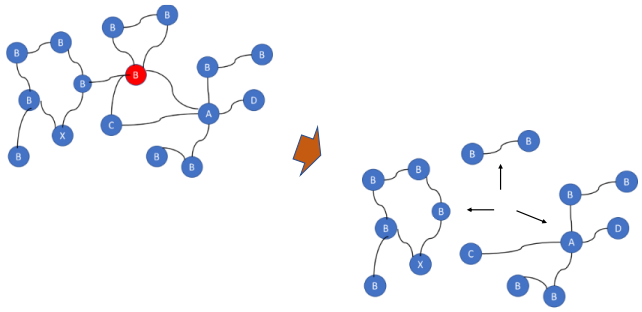


Figure 3: Network Dismantling

The metrics associated with dismantling are quite opaque. They use the concept of Giant Connected Component (GCC) which measures the number of nodes in the largest remaining network after entity removal. In effect, a measure of fragmentation.

From a law-enforcement perspective, it is clearly useful to get a measure of network robustness. In terms of interventions this can clearly influence policing strategy.

Operational Modes

There are two primary operational models for network analytics: investigation and monitoring.

Investigation: this is the dominant mode. Networks are redrawn on a periodic basis, daily or weekly for example, and updated networks made available for investigators. The analysts or investigators then explore the networks generated through searches on specific entities. Who is X connected to? Who else is connected to this address, and so on. This is the standard method of usage.

Network metrics can support this search but do not really offer guided search capabilities.

Network analytics can score or rank networks based on the various network characteristics and thus provide a ranked list of networks based on the problem criteria posed.

Monitoring: in this mode the network is used to automatically monitor all new input and generate network information from the input data. Is X part of a network, network size, density, etc. This approach is used in financial crime monitoring where quick decisions need to be taken although it can be challenging from a performance perspective.

In the law enforcement context monitoring usage is feasible but in practice, as in industry, the investigative mode dominates.

Ethical Issues

There are a number of specific ethical issues that arise in relation to social network analytics. It goes without saying that best practice data science principles must be observed in their development, that all data must be accurate and that a robust data science governance process is in place.

It is also critical that operational usage guidelines are defined and that there is sufficient transparency to ensure that monitoring can take place.

Data Protection Issues

SNA relies on establishing links between people and this cannot be done with fully anonymized data. There are ways that individual data fields can be anonymised but preserve any structural relationships within the data fields. A simple case would be surnames being consistently encoded so this feature was preserved. However, structure also need to be preserved across fields and it is very difficult to completely transform and codify a dataset and preserve all the important relationships within the dataset.

In general, networks have to be established using PII data. It therefore comes under the GDPR regulations on data privacy. The data that is used for the network analysis must be held lawfully and either consent obtained or exception permitted. In general law enforcement can use the exception criteria but need to exclude any individuals whose data should not be used within the investigation without consent.

In network analytics based on interviews and/or intelligence, for example, individuals who did not consent to participate can find themselves named by others and thus included. Therefore the criteria for inclusion needs to be very clear and well-defined.

Status of Network Connection

A social network shows that person A is connected to person B through some link attributes. The connection can be direct, level one, indirect via one other entity, or two or three other entities and so on. The possibility of a connection is related to the information that is used. If the connection link is a pub, let's say, the Fox and Hounds" then this will create a large pool of possible contacts without any real evidence whether one person is linked to another. On the other hand, for the law enforcement professional, some of these social contacts may be very significant.

Questions then arise, how far is it ethical to extend a network in the sense of the number of links and what attributes is it ethical to use? Are any attributes or links acceptable?

The question of network level or degree of separation is highly controversial. It also has a practical element. A connection that is at three degrees of separation is less likely to be significant than a direct connection but should this be proscribed or is this a matter for the investigating team to determine? They need to operate efficiently and not waste time on false leads. In practice the very large networks tend to be full of noise.

The question about the nature of the links is different. The risk here is that link attributes are used that are not relevant and will always pull into the network many persons who are not connected to the individual or group under investigation. This leads to the question of data:

Data

Data is at the heart of the analysis. What is the data universe for the network analysis? Typically, this will be a set of databases held by the organisation. So the universe might be all applicants or customers of a bank to take a commercial usage example. It might be all persons rightfully stored on a police force database. It will always be a restricted set and only other persons represented in the data universe are potential network contacts.

One of the problems here is then that the data universe is restricted and connections can appear emphasized as a result. Nonetheless, if a connection can be established within this dataset this seems to be valid intelligence.

New Data

Is it 'new data' to be identified as part of a network? The data exists within the allowable data constraints and so the potential to be thus identified exists already. The network analysis tools realise this possibility.

However, if the network is already identified as a particular organized group then, potentially, being identified by the model as linked to the network makes this person a potential group member and possibly tagged as such. This is then 'new information' in relation to this nominal.

Arguably, this is only making effective usage of data already known and that this process is a necessary part of an investigation. We can make these connections that previously were missed due to data access and analysis limitations.

What requirements are needed to validate this type of 'new' information gained from network analysis? Should only direct connections can be treated as potential group associates and additional evidence required otherwise.

Data Issues

With regard to the data 'universe' this is always restricted in some way, usually by access and date, often by other constraints. In the case of law enforcement usage, the data universe needs to be very well-defined and the following considerations apply:

- Regionality – data can be regional to police force or national.

- Acceptability – data can be derived from suspects, charged nominals, convicted nominals, witnesses, victims. The selection here needs to follow agree protocols in relation to the project and ethical guidelines.
- Time period – when data is available.
- Project constraints – violent offences, etc.

Once all data constraints are applied the data universe may be quite small. Anyone who has escaped the attention of the law will not appear. Anyone excluded by the time constraints or project constraints will not appear. The network will only find connections between, for example, convicted offenders of violent crime to take a very restricted domain. This clearly limits the tool from the crime prevention perspective and mean that existing offenders are trapped within a cycle of constant suspicion.

The results will then reflect the remaining data candidates. This process of winnowing the data will likely result in a more concentrated data pool where networks are more readily identified. In turn this could reflect more seriously against anyone innocently connected (as well as making it more likely the system shows them as connected).

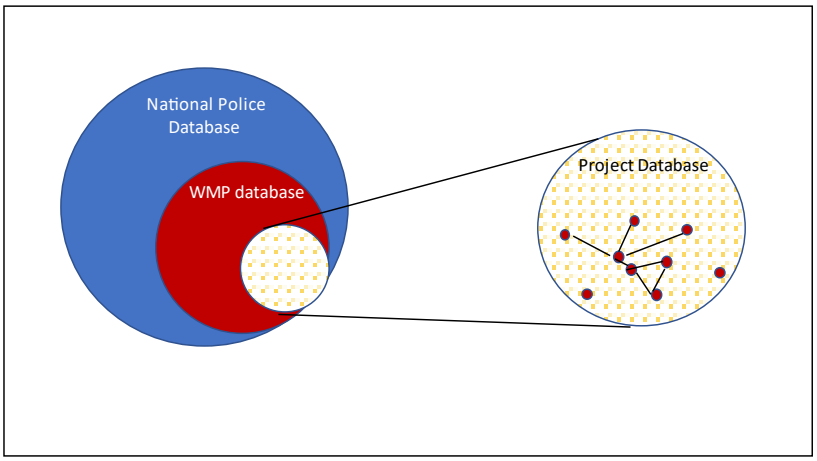


Figure 4: Data Universes

Real World Examples

In their 2018 paper, "Social Network Analysis for Law Enforcement" from the IACA, the authors reference a number of studies looking at the use of social network analysis for law enforcement, including drug gangs and violent crime networks. One of the findings they reported was that: "certain network measures, network capital (connectivity and severity) and structural equivalence

(holds a position in both networks) might be useful in identifying which associates of prioritized targets should also be in the police radar"[3]

Another case study highlights how network analysis, in this case using up to 4 levels of connection, was able to help to explain a situation where two gangs, hitherto peaceful, had become embroiled in a gang war.

Real World Issues

These concentrate on issues encountered when working with financial crime teams in industry who are using network analysis as an investigative tool. This is a very popular tool with financial crime teams and there are many instances where it has enabled successful investigations to be conducted. With so much financial fraud being organised activity a tool that provides the ability to make connections is hugely valuable. However errors do occur and operational policies and training needs to be in place to mitigate against this in the law enforcement context. Usually these take the form of generating false positives and while this is not the norm, in the law enforcement context, the consequences of false positives are much more serious.

In one example a large insurer in the US found their network analysis system was throwing up military bases as suspicious. It turned out that IP address was being used as a data element and military personnel filing claims at the military base were showing up with the same IP address and being linked into what looked like suspicious organised activity. In this case this could be addressed by creating whitelists for these IP addresses so this data would not be used by the system. The alternative is to remove or disable this data element across the entire population if there are too many anomalies of this type.

In another example, a system was delivered to a client in Turkey and huge networks were being generated connecting people called 'Mohammed' or variations thereof. 'Name' had to be removed as a potential connector from the data elements. This illustrates the importance of cultural understanding in the development of these projects. Differences in name usage and naming conventions globally cause many challenges to business analytics.

Employer is sometimes used as a data element in these networks but in some small towns there is an employer that engages large numbers of the people in the town. Employment can sometimes be an important connection element but, for significant employers, it can completely swamp the data by linking large numbers of people and effectively rendering the tool ineffective within this sub-population. If, moreover, we only have a small sample of this population in our dataset, as is usually the case, then this can easily create the impression[4] that we have a significant connection here and they all work at the same employer.

These are all manageable problems once identified but such problems do arise in most systems. In the law enforcement context it is critical that such issues do not arise or are, at least, made transparent and mitigated. Transparency is therefore critical to the effective use of these systems.

[3]International Association of Crime Analysts. (2018). Social Network Analysis for Law Enforcement, Overland Park, KS: Author