Social Networks 61 (2020) 99–106

Contents lists available at ScienceDirect

Social Networks

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

Assessing the missing data problem in criminal network analysis using forensic DNA data

Sabine De Moor[a,*], Christophe Vandeviver[a,b], Tom Vander Beken[a]

[a] Ghent University, Institute for International Research on Criminal Policy (IRCP) Universiteitstraat, 49000, Ghent, Belgium
[b] Research Foundation – Flanders (FWO), Egmontstraat, 51000, Brussel, Belgium

ARTICLE INFO

Keywords:
Missing data
Unknown offenders
Real-world network
Forensic DNA
Police recorded crime data

ABSTRACT

Missing data is pertinent to criminal networks due to the hidden nature of crime. Generally, researchers evaluate the impact of incomplete network data by extracting or adding nodes and/or edges from a known network. Statistics on this reduced or completed network are then compared with statistics from the known network. In this study, we integrate police data on known offenders with DNA data on unknown offenders. Statistics from the integrated dataset ('known network') are compared with statistics from the police data ('reduced network'). Networks with both known and unknown offenders are bigger but also have a different structure to networks with only known offenders.

1. Introduction

It is not easy to map criminal networks. They have fuzzy boundaries and dynamic relationships (Sparrow, 1991). The partnerships between different offenders can be temporary (only for one crime, for example) or more permanent in nature (Weerman and Kleemans, 2002), although offenders do not usually commit multiple crimes with the same co-offender, except in more specialized groups (Reiss and Farrington, 1991). As a result, co-offenders often belong to multiple offending groups at the same time (Warr, 1996). But perhaps the main obstacle to the study of criminal networks is the incompleteness of the available network data. Unlike social networks such as friendships or working relationships, criminal ties are less visible, as offenders try to conceal their crimes and ties with criminal friends. Consequently, criminal networks are incomplete and both nodes and edges are missing (Sparrow, 1991; Xu and Chen, 2005).

Depending on the data collection method, missing data in networks can have multiple causes: the boundary specification problem (BSP), respondent inaccuracy and non-response in network surveys or interviews and the study design. For example, the study design can create a fixed choice effect, where bias is caused by limiting respondents to naming, say, three friends when in reality they have at least ten friends (Kossinets, 2006). The BSP is the most important factor in this research. The BSP refers to the question how accurately network boundaries are defined. Changes in the location of the boundary of the network can have a significant impact on both centrality and density measures (Doreian and Woodard, 1994). The BSP is of particular interest in the study of criminal networks, as only detected offenders and their crimes can be integrated in the network. In other words: the external boundaries of the network lie where the police and court files end (Berlusconi, 2013, p. 63; Campana and Varese, 2011, p. 20). However, these boundaries can be very restrictive, as in many Western countries the clearance rate of registered crimes is very low (De Wree et al., 2006; Lammers and Bernasco, 2013). Unknown offenders and their offences, and the unknown crimes of known offenders, remain out of reach to those studying offending behaviour, and as a result criminal network data is incomplete as part of the existing nodes and/or links are not visible (Coles, 2001).

Current research on imperfect or missing data in networks has an important limitation. Generally, random errors are applied to networks. However, as in many other network contexts, missing data are non-random in criminal networks (Sparrow, 1991). Some offenders may be more likely than others to be absent in police-recorded crime data. The non-random removal of central nodes, as performed in the study by Smith et al. (2017) also doesn't seem an adequate answer, as the missing of nodes and edges may be related to features other than the position one takes in a network. Although not in a network context, Lammers et al. (2012), for example, found that unknown (i.e., not arrested) offenders may differ from arrested offenders as the latter have a longer criminal career (i.e., commit multiple crimes) and have a more versatile offending pattern (i.e., are not limited to only one crime type). Moreover, in light of the low clearance rates, the most central nodes in

* Corresponding author.
E-mail addresses: Sabine.DeMoor@UGent.be (S. De Moor), Christophe.Vandeviver@UGent.be (C. Vandeviver), Tom.VanderBeken@UGent.be (T. Vander Beken).

https://doi.org/10.1016/j.socnet.2019.09.003

Available online 24 October 2019

police recorded crime data may not be the main offenders in the network. Indeed, Sparrow (1991, p. 256) stated that "the determination of centrality will depend upon *who you know most about,* rather than who *is central or pivotal in any structural sense"*.

The central question in this research is whether unknown offenders have an impact on the offender network containing both known and unknown offenders. This research question is further operationalized in two sub questions. First, does the offender image changes when unknown offenders are included in a network analysis (RQ1)? Second, do known and unknown offenders differ in network centrality (RQ2)? In other words, are unknown offenders random missing nodes or not? To assess these questions, police-recorded crime data are used to construct the 'reduced network' and the integrated dataset of police-recorded crime data and forensic DNA data is used to construct the 'real-world network', combining DNA data on unknown offenders and police-recorded crime data on known offenders. By comparing network measures of both networks, the effect of missing data (i.e., unknown offenders) in police-recorded crime data can be evaluated.

The remainder of this article is structured as follows. The next section gives a brief overview of the literature on missing data in criminal network analysis. We then present our data, followed by a method section. The main finding are presented in the results section. A final section concludes this paper with a discussion and some limitations before highlighting possibilities for future research.

2. Literature review on missing data in criminal network analysis

Criminological researchers have used several sources to extract network data: surveillance data from communications using telephone, e-mail or personal contact (e.g., Campana and Varese, 2011), interview data with offenders (e.g., Vlaemynck, 2016), experiential knowledge of police officers and other criminal justice agencies (e.g., McGloin, 2005), police crime reports (e.g., Papachristos and Wildeman, 2014), transcripts of court proceedings (e.g., Reid et al., 2013) and open source media reports (e.g., Burcher and Whelan, 2015; Morselli et al., 2007). In recent years, a number of researchers have evaluated (the completeness of) these different data sources for use in network analysis. Bright et al. (2012) recommended the use of judges' sentencing comments. The researchers admitted that these data might not provide as valuable information as that gained from other sources, but highlighted that, in contrast to offender databases or wire taps for example, judges' sentencing comments are free of charge, publicly available and concise.

Other researchers have focused on methodological issues to evaluate the impact of incomplete data on criminal networks. Malm and Bichler (2011) concluded that the use of multiple data sources can provide a more comprehensive picture of drug market networks. Berlusconi (2013) used wire taps, arrest warrants and judgment data on groups operating in an Italian province. She noted that the number of nodes and ties present in the data decreases from wire taps to arrest warrants to judgment data, although the statistical measures that describe the position of an actor within a network (degree and betweenness centrality) remain quite robust.

A more systematic way of evaluating the impact of incomplete network data is by simulating network errors in an observed network: a certain percentage of nodes and/or edges are extracted from or added to a known network (i.e., the error % mentioned in Table 1). Statistics on this reduced or completed network are then compared with the statistics from the known or real-world network (Borgatti et al., 2006; Kossinets, 2006; Smith and Moody, 2013; Smith et al., 2017; Xu and Chen, 2008). This allows the different error types that are possible in network data to be assessed. For example, the impact of false negative nodes and false negative edges can be studied by deleting nodes or edges in criminal offending networks. A false negative node refers to the absence of a person in the network who should be present as he is an offender. A false negative edge means that the relation between two offenders is not observed in the network. The two offenders are not registered as co-offenders even though they actually are. The impact of false positive nodes and false positive edges can be assessed by adding nodes or edges in criminal offending networks. A false positive node refers to a person registered as an offender who is not the offender of the crime and should therefore not be part of the network. False positive edges appear when relationships between offenders are incorrectly present in a network: two offenders are registered as co-offenders, but have not committed any crimes together (Frantz et al., 2009; Wang et al., 2012). A less commonly studied measurement error is false aggregation and disaggregation. Two nodes are falsely aggregated when they are wrongly regarded as one node. The opposite applies to false disaggregation: one node is wrongly regarded as two separate nodes in the network. The impact of these errors is assessed by aggregating or disaggregating nodes. In the former, edges of node A are connected to node B and node A is removed afterwards. In the latter, node A is split into two nodes A and B. Some of the edges of node A are randomly removed and added to the new node B (Wang et al., 2012).

Most researchers using simulated network errors apply random errors to real-world (i.e., observed) networks (e.g., Huisman, 2009; Smith and Moody, 2013; Wang et al., 2012) or to simulated, random networks (e.g., Borgatti et al., 2006; Frantz et al., 2009) to assess the impact on centrality measures. In a recent study, Smith et al. (2017) applied non-random errors to 12 real-world networks by removing nodes proportional to their centrality. Most of these studies simulating network errors conclude that the effect of missing data depends on a number of factors (Smith et al., 2017). First, the lower the sample coverage, the more the network estimates are corrupted (Galaskiewicz, 1991). The nature of the missing data is a second factor. For example, bias is worse when more central nodes are missing. In other words, non-random missing data cause a higher bias. Third, missing data do not have the same effect on every network statistic (Costenbader and Valente, 2003). Centrality measures seem to be more robust against missing network data than other network statistics like topology and homophily for example (For an elucidation of the robustness of these other network measures, see Smith and Moody, 2013; Smith et al., 2017). Finally, the characteristics of the network will also influence the effect of missing data. Smith and Moody (2013), for example, found in their research on random missing nodes in different empirical networks that larger, more centralized networks are generally more robust to missing data. Borgatti et al. (2006) concluded that, except for edge deletion, centrality measures of dense networks are the most robust against random errors in network data. Frantz et al. (2009) also found that network errors may have a different impact on centrality measures, depending on the network topology (uniform random, small-world, core-periphery, scale-free or cellular networks). Table 1 gives an overview of the main characteristics and results of the studies discussed above.

3. Data

The study makes use of two databases: the Belgian General Police Database and the Belgian National Genetic Database (NGDB). The Belgian General Police Database is the main source for nationwide crime statistics in Belgium. All detected and reported crimes are registered in this database. The database contains, inter alia, information on crimes, offenders, modus operandi, and victims. The NGDB was set up in 2002 and is managed by the National Forensic and Criminology Institute NICC/INCC, a federal scientific institution within the Ministry of Justice. On 31 December 2015 the NGDB contained 34,784 reference profiles and 43,224 forensic profiles or crime scene profiles (NICC/INCC, 2016).[1] The former are 'known' profiles obtained from stains taken directly from known individuals (for example, a buccal swab

[1] The NGDB also contained 1,207 reference profiles of suspects. As from 1 January 2014, suspect profiles can be stored in the NGDB under specific conditions.

Table 1
Overview of main characteristics and results of studies on missing data in networks illustrated in the text.

Study	Type of network	Network measure	Type of error		Error %	Replications	General conclusions
Borgatti et al. (2006)	Random networks	Centrality: - degree - betweenness - closeness - eigenvector	Random	Node removal/ addition Edge removal/ addition	1 5 10 25 50	10,000	- Accuracy of centrality measures declines smoothly and predictably with the amount of error - Different types of error had relatively similar effects on centrality robustness
Smith and Moody (2013)	Empirical networks (directed and non-directed)	Centrality: - degree - closeness - betweenness - Bonacich power score Centralization Topology Homophily	Random	Node removal	1 2 5 10 15 25 30 40 50 60 70	1,000	- Measurement bias generally increases with more missing data - Exact rate and nature of increase varies systematically across network measures - Bias dependent on the features of the network
Smith et al. (2017)	Empirical networks (directed and non-directed)	Centrality: - degree - closeness - betweenness - Bonacich power score Centralization Topology Homophily	Non-random	Node removal	1 2 5 10 15 25 30 40 50 60 70	1,000	- Bias is worse when more central nodes are missing - Bias dependent on the features of the network
Wang et al. (2012)	Empirical networks	Centrality: - degree - eigenvector Clustering coefficient Network constraint	Random	Node removal/ addition Edge removal/ addition Aggregation/ disaggregation of nodes	From 5 to additionally removing up to 95	10	- Networks with low average clustering and less positively skewed degree distributions are most resistant to measurement error - Bias dependent on the features of the network
Kossinets (2006)	Random networks Empirical networks	Degree Clustering Assortativity Fractional size largest component Average path length	Random and non-random	Node removal Edge removal			- Boundary specification (non-inclusion of nodes or edges) can dramatically alter estimates of network-level statistics
Frantz et al. (2009)	Random networks	Centrality: - degree - betweenness - closeness - eigenvector Local clustering	Random	Node removal Edge removal	1 5 10 25 50	10–250	- The topological form of known network (uniform random, small-world, core-periphery, scale-free or cellular) has a measurable effect on robustness - Results are consistent with Borgatti et al. (2006) for the uniform random topology
Costenbader and Valente (2003)	Empirical networks	11 centrality measures	Random	Node removal	From 20 to 80, in steps of 10	25	- Some measures are more stable than others - Bias dependent on the features of the network
Huisman (2009)	Empirical networks (directed and non-directed)	Degree Reciprocity Clustering Assortative mixing (on degree) Distance	Random and non-random	Node removal Edge removal	From 10 to 90, in steps of 10	100	- Missing data can have large negative effects on structural properties of the network
Galaskiewicz (1991)	Empirical networks	Point centrality	Random	Node removal	25 50 75	10	- Bias increased considerably as sampling percentage decreased

from an offender) whereas the latter refers to an 'unknown' DNA profile of a (hitherto) unidentified offender gathered at a crime scene (i.e., crime scene profile). DNA traces involving the same unknown offender, found at different crime scenes at different time points (i.e., a serial offender) can be linked. Moreover, the involvement of other unknown co-offenders can be revealed through the presence of their DNA traces at shared crime scenes. Consequently, even though the crime may not be solved and the associated offenders may not be identified, information on the offenders' network may still be available for researchers.

The dataset contains six years of recorded crime data (2010 through 2015) relating to the four most frequently recorded crime types in the NGDB: violent theft, aggravated burglary, lethal violence, and sexual offences.[2] All known offenders involved in crimes that matched these criteria were selected from the Belgian General Police Database, resulting in a police dataset of 73,837 known offenders. The police dataset was enriched with offender data from the DNA dataset to

[2] Violent theft refers to crimes like robbery, carjacking and home jacking. Aggravated burglary refers to crimes like raid, burglary in a dwelling and theft using false keys. Lethal violence refers to crimes like manslaughter, murder and poisoning. Rape is an example of sexual offences.

Table 2
Data selection of unknown offenders from the NGDB.

Crime	Offender	
C1	Crime scene profile	O1
C1	Reference profile	~~O2~~
C2	Crime scene profile	O3
C3	Crime scene profile	O4
C4	Reference profile	~~O5~~
C5	Crime scene profile	~~O5~~
C6	Crime scene profile	O6
C6	Reference profile	~~O7~~
C7	Crime scene profile	O6

construct a dataset with both known and unknown offenders. Some precaution is needed when combining data from the Belgian General Police Database and the NGDB. The same person may be registered in both as a known offender. Profiles of known offenders (i.e., suspects or convicted offenders) are stored in the NGDB using a DNA code number to protect the privacy of the offender. However, possible matches between known DNA profiles and known offenders stored in the police dataset could not be checked because access to the corresponding identity of the offender is prohibited for scientific research. This could lead to what Wang et al. (2012) define as false disaggregation in network data. To avoid this type of false disaggregation, only unknown offenders were selected from the NGDB.

Table 2 illustrates the process used to select unknown offenders from the NGDB. The same DNA profile can be linked to different crimes, whether it is a reference profile and/or a crime scene profile. For example, the DNA profile of offender O5 is found at a crime scene, and a reference profile is obtained from the same offender O5 in the context of crime C4. Offender O5 is therefore a known offender and would not be included in the analysis. In another example, although two offenders are related to crime C6, only O6 would be retained in the analysis as the other offender, O7, is a known offender. This selection procedure resulted in a DNA dataset of 16,092 different unknown offenders.

4. Methods

4.1. Network analysis

A social network analysis was performed making use of the R packages 'igraph' (Csárdi and Nepusz, 2006) and 'Matrix' (Bates and Maechler, 2017) to identify different components of offenders in the two datasets. Component analysis allows sub-networks within larger networks to be identified. Only offenders and their other co-offenders who commit a crime together are part of the same component or sub-network. That way, each sub-network consists of at least two offenders connected to each other, directly or indirectly, but the offenders within a sub-network have no links with other offenders outside the sub-network (Wasserman and Faust, 1994).

In this study, some basic network measures were computed. The size of a network is equal to the number of nodes or links. The geodesics is the shortest path between a pair of offenders. The *geodesic distance* is equal to the length of the shortest path. The *average path length* is the average length of the shortest paths for all possible pairs of nodes. *Density* describes the network cohesion. It is the proportion of the actual present edges to all the possible edges. A network containing all possible edges is a *clique* (Rossy and Morselli, 2018; Wasserman and Faust, 1994).

This study was limited to two centrality measures (i.e., degree and betweenness) for substantive reasons. Not all centrality measures are meaningful for disconnected networks composed of several distinct components like the network data in this study (e.g., closeness, see Haythornthwaite, 1996; Prell, 2013; Wasserman and Faust, 1994).[3] The *degree* refers to the number of direct links an offender has with other offenders by committing a crime together (Freeman, 1979; Wasserman and Faust, 1994). The minimum degree is 0 (or 0%), which means that the offender committed all his crimes without any co-offender. An offender with degree 0 is called an isolate. The maximum degree is equal to the number of nodes in the network minus 1 (or 100%). Offenders with a maximum degree committed at least one crime with every single other offender in the network. The degree only takes the local positon of the actor into account, as it is not concerned by how the other offenders are connected in the network (Morselli, 2009, p. 39; Wasserman and Faust, 1994). *Betweenness* centrality is the proportion of times an offender is located along the geodesics between any two other offenders. In other words: to what extent is an offender the direct link between two other offenders? Unlike degree, the quantity of direct contacts is not important, but the quality of the (direct and indirect) connections is. An offender with a relatively low degree may play an important 'intermediary' role and so be very central to the network (Scott, 2013, p. 87). As such, a network can easily be disrupted when an offender with a high betweenness centrality is arrested and thus removed from the network (For an extensive elucidation of network disruption, see Bichler and Malm, 2015; Duijn, 2016; Duijn et al., 2014). Prell (2013, p. 107) describes the differences between these two centrality measures as degree centrality, emphasizing activity, and betweenness centrality, emphasizing potential control over information flow. An offender with many co-offenders will be central according to the degree centrality measures. However, offenders with fewer contacts may become more central when the betweenness centrality is measured.

4.2. Monte Carlo simulation

Previous research on the effect of missing data in networks, or network errors in general, usually performed Monte Carlo simulations to evaluate bias. This procedure entails six steps: (1) Identify or simulate a real-world network G(V, E). This network is assumed to be complete. (2) Calculate the network measures of interest for this real-world network. (3) Apply (random or non-random) network errors to the real-world network by adding or deleting a certain fraction of nodes and/or edges. The result is the distorted or reduced network G'(V', E'). (4) Calculate the network measures of G'(V', E'). (5) Repeat step 3 and 4 to obtain distributions and confidence intervals of the network measures. (6) Compare network measures of G(V, E) with those of G'(V', E') to assess the impact of the different error levels (Kossinets, 2006; Smith and Moody, 2013; Smith et al., 2017; Wang et al., 2012).

In this study, three (instead of two) network types were identified: a real-world network (G(V,E)), a real-reduced network (G'(V',E')) and different simulated-reduced networks.[4] The real-world network is composed of known offenders from the police database and unknown offenders from the DNA database. Based on the police database, the real-reduced network is created. It is a real network as it can be observed, but it is reduced because a lot of unknown offenders are missing. Finally, the simulated-reduced networks are obtained by randomly removing a certain percentage of nodes from the real-world network. Eleven different levels of missingness were applied to the real-world network by randomly removing 2, 5, 10, 15, 20, 25, 30, 40, 50,

[3] In addition, an analysis on a network composed of 73,837 known offenders and 16,092 unknown offenders demands significant computational power. To illustrate, an undirected network with n nodes can contain a maximum of n(n-1)/2 distinct edges (Scott, 2013). In theory, in the integrated network with 89,929 known and unknown offenders, in total 4,043,567,556 distinct edges are possible between two offenders.

[4] The VSC (Flemish Supercomputer Center) provided the computational resources (Stevin Supercomputer Infrastructure) and services to compute the simulated-reduced networks and the corresponding network measures. The VSC is funded by Ghent University, the Hercules Foundation and the Flemish Government – department EWI.

60, 70% of the nodes of the real-world network. Each error level was repeated 999 times.

Generally, the impact is assessed by calculating the correlation between the network measure in G(V,E) and G'(V',E') at the network level (e.g., Costenbader and Valente, 2003) or at the individual node level. In the latter case, only nodes present in both the real-world network and the reduced network can be taken into account (e.g., Wang et al., 2012). In this study, the networks with the random generated errors (i.e., simulated-reduced networks) are compared with the real-reduced network G'(V',E'). This allowed the impact of random missing nodes versus non-random missing nodes to be assessed by measuring the degree and betweenness for each of the different error levels. Furthermore, the network measures were evaluated at the network level in this study. Correlations at the individual node level would only take the known offenders into account, as only these nodes are present in both the reduced and the real world network. As the goal of this study is to assess whether and how the network measures change when integrating unknown offenders, the analysis was done at the global (i.e., network) level. The two centrality measures are averaged across the 999 replications for each error level, generating one value for each of the two measures for each of the eleven error levels.

5. Results

5.1. Descriptives

The characteristics of the real-world network and real-reduced network are summarized in Table 3. The real-reduced network is composed of known offenders derived from the police-recorded crime data. The real-world network contains 16,092 more nodes (21.79%) than the real-reduced network. All these additional nodes are unknown offenders from the DNA data. An extra 21,329 edges (43.83%) between offenders are created by integrating unknown offenders in the network. These edges can be between an unknown offender and a known offender who have committed a crime together, but also between unknown co-offenders. Almost half of all the components have a size of only one node (i.e., isolates), both in the real-world network and the real-reduced network (43.36% versus 45.41%). These offenders did not commit any crime with another offender.

The existing components can be supplemented with additional offenders, or several components could be merged, when DNA data is integrated with police data. This means that the composition of the 44,743 components in the real-reduced network may have changed after the integration of DNA data. Moreover, the real-world network contains 8305 more components (18.56%) than the real-reduced network. These components are composed of only unknown offenders.

Table 3
Characteristics of the real-world network and the real-reduced network.

	Real-world network G(V,E) Police data and DNA data	Real-reduced network G'(V',E') Police data
Nodes	89,929	73,837
Edges[a]	69,995	48,666
Average degree	1.56	1.32
Number of components	53,048	44,743
Size of largest component	5838	5282
Size of smallest component	1	1
Number of isolates	38,989 (43.36%)	33,531 (45.41%)
Density	1.731021e-05	1.785309e-05
Average path length	17.44058	17.50912

[a] The number of edges corresponds to the number of edges present in the simplified networks. Simplified networks do not contain multiple edges between two nodes. In a simplified network, only one edge between two offenders is possible, even if these offenders may have committed multiple crimes together.

Table 4
Degree of real-world network and reduced networks.

Network type	Error percentage	Mean degree	Nodes	Mean edge count
Real-world network G(V,E)	0	1.56	89,929	69,995
Simulated-reduced networks	2	1.53	88,130	67,223
	5	1.48	85,433	63,181
	10	1.40	80,936	56,697
	15	1.32	76,440	50,584
	20	1.25	71,943	44,793
	25	1.17	67,447	39,369
	30	1.09	62,950	34,304
	40	0.93	53,957	25,206
	50	0.78	44,965	17,492
	60	0.62	35,972	11,200
	70	0.47	26,979	6,305
Real-reduced network G'(V',E')	17.89	1.32	73,837	48,666

The *average path length* is around 17 for both network. This is quite large, given the average degree is below 2. However, this mean value can be distorted as both the real-world network and the real-reduced network contain one huge component, respectively 5838 and 5282 nodes. The second largest component contains only 144 and 136 offenders respectively. The density or network cohesion, by contrast, is low, caused by the high number of components in both the real-world and real-reduced network.

5.2. The effect of missing data on degree and betweenness

Table 4 gives an overview of the impact of missing nodes on the degree in network analysis. Results on the degree for the real-world network, the different simulated reduced networks and the real-reduced network are summarized in the table. Obviously, the real-world network has an error level of 0%. This network contains almost 90,000 nodes or unknown and unknown offenders and almost 70,000 links between these offenders. On average, every offender in the real-world network has 1.56 co-offenders. At the bottom of the table, the values for the real-reduced network, which only includes known offenders from the police-recorded crime data, are presented. The average degree (1.32) is lower than in the real-world network (1.56), which includes both known and unknown offenders. In other words, integrating unknown offenders also revealed more co-offending relationships in the real-world network.

When comparing the number of nodes, the real-reduced network has an error level of about 18% compared to the number of nodes in the real-world network. Table 4 also presents the results of the eleven different error levels applied on the real-world network, ranging from 2% to 70% of the nodes being randomly removed. Removing nodes clearly has an impact on the average degree: the degree decreases as the error percentage increases. For example, an error percentage of 40 or higher corresponds with a mean degree below one, illustrating that this reduced network contains many isolates (i.e., offenders who did not commit any crime with another offender). Logically, the number of edges also decreases with an increasing error level. [5]

A simulated error level of 15% results in the same value for degree as the 18% error level in the real-reduced network.[6] Randomly

[5] As different Monte Carlo simulations were performed for degree and betweenness centrality, the number of edges for the different error levels in Table 4 is different from the number of edges for the different error levels in Table 5.

[6] There is a significant difference between the degree of the real-reduced network and the other ten simulated error levels.

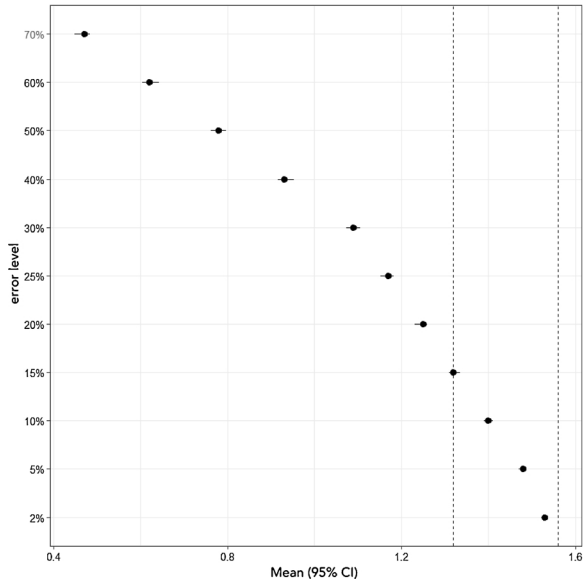

Fig. 1. Forest plot 95% confidence interval for degree of simulated-reduced networks.

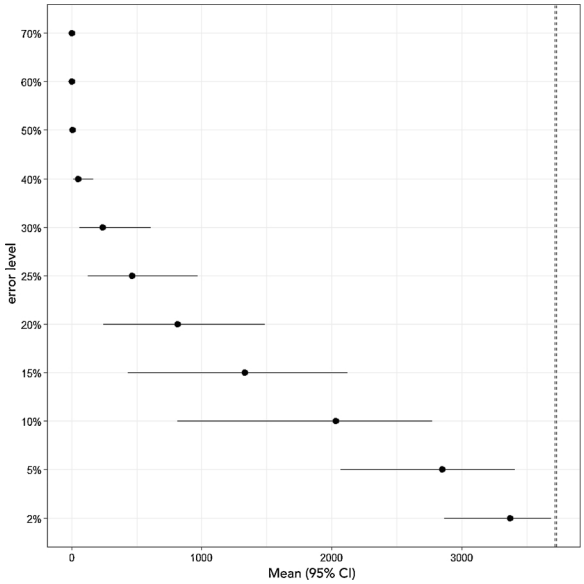

Fig. 2. Forest plot 95% confidence interval for betweenness of simulated-reduced networks.

removing 15% of the known and unknown offenders from the real-world network results in the same degree as non-randomly removing only the unknown offenders from the real-world network. This can also be deduced from Fig. 1, representing the 95% confidence intervals for degree of the 11 simulated reduced networks. The vertical dotted line on the left represents the mean degree of the real-reduced network (1.32) and the vertical dotted line on the right represents the mean degree of the real-world network (1.56). The mean degree of the real-reduced network falls within the 95% confidence interval of the 15% error level network.

Taking into account that known and unknown offenders have a similar impact on the mean degree centrality and that the average degree increased in the real-world network, this means that integrating unknown offenders also changed (i.e., increased) the degree of the known offenders. In other words, co-offending relationships between known and unknown offenders become visible when DNA data and police data are integrated.

The mean betweenness scores across the real-world network, the different simulated-reduced networks and the real-reduced network are presented in Table 5. It is remarkable that the betweenness centrality of the real-reduced network with a non-random error level of about 18% is about the same as the betweenness in the real-world network (3723.58 and 3715.44 respectively). The integration of unknown offenders seems not to have affected the betweenness in the real-world network.

Table 5
Betweenness of real-world network and reduced networks.

Network type	Error percentage	Mean betweenness	Nodes	Mean edge count
Real-world network G(V,E)	0	3715.44	89,929	69,995
Simulated-reduced networks	2	3368.27	88,130	67,238
	5	2845.52	85,433	63,194
	10	2029.64	80,936	56,702
	15	1328.99	76,440	50,606
	20	815.97	71,943	44,811
	25	464.79	67,447	39,403
	30	241.46	62,950	34,336
	40	51.83	53,957	25,208
	50	10.30	44,965	17,462
	60	2.00	35,972	11,219
	70	0.42	26,979	6,317
Real-reduced network G'(V',E')	17.89	3723.58	73,837	48,666

Randomly removing a number of nodes, both known and unknown offenders, has a clear impact on the average betweenness centrality of the offenders. Fig. 2 illustrates that the mean betweenness decreases as the error level increases. Furthermore, the dotted line representing the mean betweenness of the real-world network (3715.44) is not within the 95% confidence interval of any simulated-reduced network.[7] Therefore, the results for betweenness centrality are totally different to those for degree centrality. Removing about 15% of the offenders randomly would result in a much smaller average betweenness (1328.99) than removing the same percentage non-randomly (i.e., 3723.58 in the real-reduced network). Even randomly removing only 2% of the known and unknown offenders from the real-world network results in a lower betweenness centrality than the 18% error level in the real-reduced network with only known offenders ($p = 0.01$). Known and unknown offenders have a different impact on betweenness centrality.

6. Conclusion and discussion

Criminal networks are a textbook example of hidden networks, as many registered crimes are unsolved and the offenders remain unknown. In order to assess the validity of research on criminal networks it is therefore important to assess the robustness of basic network measures under the condition of missing data (Borgatti et al., 2006). As far as is known, this is the first study to integrate forensic DNA data on unknown offenders with police data on known offenders in order to study the missing data problem in criminal networks. The DNA data provides a unique opportunity to integrate missing data into police networks and is an important advancement over prior research.

This study confirms the findings from previous research. First, the higher the error level in the simulated networks, the more the network estimates are affected. This applies both to the degree and to the betweenness centrality. Second, the impact of the error level is not equal for both centrality measures. Betweenness centrality seems to be more

[7] As the mean betweenness of both the real-reduced network (3723.58) and the real-world network (3715.44) are similar, the dotted lines representing these values are very close to each other in Fig. 2.

affected by an increasing error level than does degree centrality. Third, the nature of the missing data is an important factor to consider. Although there does not seem to be much of a difference for degree centrality, there is a difference in randomly and non-randomly removing offenders from the network for betweenness centrality.

The central question in this study is whether and how the image of offender networks is different in a dataset that integrates police data and DNA data, compared to the police data only. Offender networks with both known and unknown offenders may not only be bigger but also have a different structure to networks with only known offenders (RQ1). The results of this study show that integrating unknown offenders has an impact on the degree, but not on the betweenness centrality. The degree is higher in the real-world network, which means that many offenders stored in the DNA database could be linked to the known offenders in the police data or to other unknown offenders. As such, the degree of the known offenders also increased by integrating the data. Removing only unknown offenders from the real-world network (i.e., real-reduced network) had no impact on betweenness. On the contrary, when known offenders are also removed from the network (i.e., simulated-reduced networks), betweenness decreases. In other words, known offenders may be more central nodes than unknown offenders in relation to betweenness (RQ2).

These research findings have implications for both theory and practice. Including the unknown offenders stored in the NGDB in the database resulted in an offender image with not only about 22% more offenders (i.e., nodes) but also about 44% more co-offending relations (i.e., edges), in comparison with a database solely based on police-recorded crime data. Therefore, the generally accepted assertion in criminology that at least half of all crime involves more than one offender (Andresen and Felson, 2010; Felson, 2003; Lantz and Ruback, 2016; Warr, 2002) and that about two-thirds of all offenders commit their crimes with others (Reiss, 1988) is probably an understatement. Furthermore, this research provides a unique view of the position these offenders may take in the whole offending network. Unknown offenders may be more peripheral nodes in the network. The question therefore arises as to whether the unknown offenders remain unidentified by the police because of their peripheral position, or is it, as Sparrow (1991, p. 256) states, just because they stay unidentified by the police that they have a more peripheral position in the network, although they may be more central in reality? This is an important nuance, as in the first case the integrated dataset would give an accurate image of the centrality of the unknown offenders, whereas in the latter case the image would be distorted. It is important to be aware of this uncertainty, because it could mean that arresting unknown offenders has a bigger impact on crime prevention than would be assumed based on their peripheral position in the network.

Some potential limitations of this study need to be acknowledged. First, it is important to bear in mind that the real-world network is unlikely to include all unknown offender from police-recorded crime data. It is, however, an approach that cannot be achieved with any other data. Second, both the real-world and real-reduced networks contain many isolates and small components, which limits the network research possibilities but also has an impact on the mean degree and betweenness at network level. Third, it is not possible to be certain that all profiles stored in the database belong to offenders and not to victims, for example. The impact of this possible error differs according to the research point of view: operational or criminological research. For example, in operational research, an unknown victim connecting two known offenders can provide new investigative leads. For criminological research, this only distorts the results. Finally, false disaggregation could also apply to a known offender from the police dataset and an unknown offender from the DNA dataset. However, according to the Belgian DNA law of 2011, offenders of a crime or attempted crime mentioned in a restrictive list of crimes have to provide their DNA profile upon conviction (Art. 14 DNA law 2011, Belgisch Staatsblad, 2011). As the four crime types selected for this study are part of this list, an unknown DNA profile will become 'known' when there is a match with the DNA profile of a convicted offender. Consequently, except for administrative delays in taking reference samples of convicted offenders or not being able to find the convicted offender to take a reference sample, for example, this problem does not apply to the current study.

Future research might progress the work developed here. In this study only two centrality measures were taken into account. Future research could explore the possibilities to measure the impact of missing data on other network measures. The present study could also be replicated using the traditional approach in missing data research by studying the correlations of the centrality measures at individual node level in networks with different error levels. Finally, to address the limitation of the high degree of isolates mentioned above, future research could focus on only the biggest component(s) present in the real-world network. All these suggestions for future research could foster the theoretical insights on known and unknown offenders.

Funding

This work was supported in part by the BRAIN-be Programme 'Belgian Research Action through Interdisciplinary Networks' (Belgian Science Policy Office) [BR/132/A4/Be-Gen to T.V.B.]. Vandeviver's contribution to this work was supported by the Research Foundation - Flanders (FWO) Postdoctoral Fellowship funding scheme [FWO15/PDO/242 to C.V.].

Declaration of Competing Interest

None.

Acknowledgment

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI. In compliance with the Belgian Privacy Law of 8 December 1992, the Belgian Privacy Commission was notified of our processing of encrypted personal data for scientific purposes. We thank Patrick P.J.M.H. Jeuniaux from the Belgian National DNA Database and Dirk Geurts from the Belgian Federal Police for their efforts in providing and contextualizing the data.

References

Andresen, M.A., Felson, M., 2010. The impact of co-offending. Br. J. Criminol. 50, 66–81.

Bates, D., Maechler, M., 2017. Matrix: Sparse and Dense Matrix Classes and Methods. Retrieved from. R package version 1.2-12. https://CRAN.R-project.org/package=Matrix.

Belgisch Staatsblad, 2011. Wet van 7 november 2011 houdende wijziging van het Wetboek van strafvordering en van de wet van 22 maart 1999 betreffende de identificatieprocedure via DNA onderzoek in strafzaken. Retrieved from. http://www.ejustice.just.fgov.be/eli/wet/2011/11/07/2011009773/staatsblad.

Berlusconi, G., 2013. Do all the pieces matter? Assessing the reliability of law enforcement data sources for the network analysis of wire taps. Glob. Crime 14 (1), 61–81. https://doi.org/10.1080/17440572.2012.746940.

Bichler, G., Malm, A.E. (Eds.), 2015. Disrupting Criminal Networks: Network Analysis in Crime Prevention Vol. 28 First Forum Press, Boulder.

Borgatti, S.P., Carley, K.M., Krackhardt, D., 2006. On the robustness of centrality measures under conditions of imperfect data. Soc. Networks 28 (2), 124–136. https://doi.org/10.1016/j.socnet.2005.05.001.

Bright, D.A., Hughes, C.E., Chalmers, J., 2012. Illuminating dark networks: a social network analysis of an Australian drug trafficking syndicate. Crime Law Soc. Change 57 (2), 151–176. https://doi.org/10.1007/s10611-011-9336-z.

Burcher, M., Whelan, C., 2015. Social network analysis and small group 'dark' networks: an analysis of the London bombers and the problem of 'fuzzy' boundaries. Glob. Crime 16 (2), 104–122. https://doi.org/10.1080/17440572.2015.1005363.

Campana, P., Varese, F., 2011. Listening to the wire: criteria and techniques for the quantitative analysis of phone intercepts. Trends Organ. Crime 15 (1), 13–30. https://doi.org/10.1007/s12117-011-9131-3.

Coles, N., 2001. It's Not What You Know—It's Who You Know That Counts. Analysing Serious Crime Groups as Social Networks. Br. J. Criminol. 41 (4), 580–594.

Costenbader, E., Valente, T.W., 2003. The stability of centrality measures when networks are sampled. Soc. Networks 25 (4), 283–307. https://doi.org/10.1016/s0378-8733(03)00012-1.
Csárdi, G., Nepusz, T., 2006. The igraph software package for complex network research. Inter J. Complex Syst. 1695 (5) Retrieved from. http://www.necsi.edu/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.pdf.
De Wree, E., Vermeulen, G., Christiaens, J., 2006. (Strafbare) overlast door jongeren-groepen in het kader van openbaar vervoer. Antwerpen, Maklu.
Doreian, P., Woodard, K.L., 1994. Defining and locating cores and boundaries of social networks. Soc. Networks 16, 267–293.
Duijn, P.A.C., 2016. Detecting and disrupting criminal networks. University of Amsterdam.
Duijn, P.A.C., Kashirin, V., Sloot, P.M., 2014. The relative ineffectiveness of criminal network disruption. Sci. Rep. 4, 15. https://doi.org/10.1038/srep04238. Retrieved from. https://www.ncbi.nlm.nih.gov/pubmed/24577374.
Felson, M., 2003. The process of co-offending. Crime Prevention Studies 16, 149–167.
Frantz, T.L., Cataldo, M., Carley, K.M., 2009. Robustness of centrality measures under uncertainty: examining the role of network topology. Comput. Math. Organ. Theory 15 (4), 303–328. https://doi.org/10.1007/s10588-009-9063-5.
Freeman, L.F., 1979. Centrality in social networks: conceptual clarification. Soc. Networks 1, 215–239.
Galaskiewicz, J., 1991. Estimating point centrality using different network sampling techniques. Soc. Networks 13 (4), 347–386. https://doi.org/10.1016/0378-8733(91)90002-B. Retrieved from.
Haythornthwaite, C., 1996. Social network analysis: an approach and technique for the study of information exchange. Libr. Inf. Sci. Res. 18 (4), 323–342. https://doi.org/10.1016/S0740-8188(96)90003-1. Retrieved from. http://www.sciencedirect.com/science/article/pii/S0740818896900031.
Huisman, M., 2009. Imputation of missing network data. J. Soc. Struct. 10 (1), 1–28.
Kossinets, G., 2006. Effects of missing data in social networks. Soc. Networks 28 (3), 247–268. https://doi.org/10.1016/j.socnet.2005.07.002.
Lammers, M., Bernasco, W., 2013. Are mobile offenders less likely to be caught? The influence of the geotable dispersion of serial offenders' crime locations on their probability of arrest. Eur. J. Criminol. 10 (2), 168–186.
Lammers, M., Bernasco, W., Elffers, H., 2012. How long do offenders escape arrest? Using DNA traces to analyse when serial offenders are caught. J. Investig. Psychol. Offender Profiling 9, 13–29.
Lantz, B., Ruback, R.B., 2016. The relationship between Co-offending, age, and experience using a sample of adult burglary offenders. J. Dev. Life. Criminol. 3 (1), 76–97. https://doi.org/10.1007/s40865-016-0047-0.
Malm, A., Bichler, G., 2011. Networks of collaborating criminals: assessing the structural vulnerability of drug markets. J. Res. Crime Delinq. 48 (2), 271–297.
McGloin, J.M., 2005. Policy and intervention considerations of a network analysis of street gangs. Criminol. Public Policy 4 (3), 607–636.
Morselli, C., 2009. Inside Criminal Networks Vol. 8 Springer, New York.
Morselli, C., Giguère, C., Petit, K., 2007. The efficiency/security trade-off in criminal networks. Soc. Networks 29 (1), 143–153. https://doi.org/10.1016/j.socnet.2006.05.001.
Papachristos, A.V., Wildeman, C., 2014. Network exposure and homicide victimization in an african american community. Am. J. Public Health 104 (1), 143–150.
Prell, C., 2013. Social Network Analysis: History, Theory & Methodology. Sage, Los Angeles.
Reid, A.A., Tayebi, M., Frank, R., 2013. Exploring the structural characteristics of social networks in a large criminal Court database. Paper Presented at the IEEE International Conference on Intelligence and Security Informatics.
Reiss, A.J., 1988. Co-offending and criminal careers. Crime Justice 10, 117–170.
Reiss, A.J., Farrington, D.P., 1991. Advancing knowledge about co-offending: results from a prospective longitudinal survey of London males. J. Crim. Law Criminol. 82 (2), 360–395.
Rossy, Q., Morselli, C., 2018. The contribution of forensic science to the analysis of crime networks. In: Rossy, Q., Décary-Hétu, D., Delémont, O., Mulone, M. (Eds.), The Routledge International Handbook of Forensic Intelligence and Criminology. Routledge, Oxford, pp. 191–204.
Scott, J., 2013. Social Network Analysis, 3rd ed. Sage, Los Angeles.
Smith, J.A., Moody, J., 2013. Structural effects of network sampling coverage I: nodes missing at random. Soc. Networks 35 (4), 652–668. https://doi.org/10.1016/j.socnet.2013.09.003. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/24311893.
Smith, J.A., Moody, J., Morgan, J., 2017. Network sampling coverage II: the effect of non-random missing data on network measurement. Soc. Networks 48, 78–99. https://doi.org/10.1016/j.socnet.2016.04.005. Retrieved from. http://www.ncbi.nlm.nih.gov/pubmed/27867254.
Sparrow, M.K., 1991. The application of network analysis to criminal intelligence: an assessment of the prospects. Soc. Networks 13, 251–274.
Vlaemynck, M., 2016. Disentangling 'social Supply': A Personal Network Study into the Social World of Recreational Cannabis Use and Its Supply. Ghent University, Ghent.
Wang, D.J., Shi, X., McFarland, D.A., Leskovec, J., 2012. Measurement error in network data: a re-classification. Soc. Networks 34 (4), 396–409. https://doi.org/10.1016/j.socnet.2012.01.003.
Warr, M., 1996. Organization and instigation in delinquent groups. Criminology 34 (1), 11–38.
Warr, M., 2002. Companions in Crime. The Social Aspects of Criminal Conduct. Cambridge University Press, Cambridge.
Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, New York.
Weerman, F.M., Kleemans, E., 2002. Criminele groepen en samenwerkingsverbanden. Tijdschr. Voor Criminol. 44 (2), 114–127.
Xu, J., Chen, H., 2005. Criminal network analysis and visualization. Commun. ACM 48 (6), 101–107.
Xu, J., Chen, H., 2008. The topology of dark networks. Commun. ACM 51 (10), 58. https://doi.org/10.1145/1400181.1400198.