Plan Overview

A Data Management Plan created using DMPonline

Title: Using social network analysis to understand crime and victimisation

Creator: Ben Palfreeman-Watt

Principal Investigator: Benjamin Palfreeman-Watt

Data Manager: Benjamin Palfreeman-Watt

Project Administrator: Benjamin Palfreeman-Watt

Contributor: Benjamin Palfreeman-Watt

Affiliation: University of Manchester

Funder: Economic and Social Research Council (ESRC)

Template: ESRC Template Customised By: University of Manchester

Project abstract:

The Metropolitan Police Service (MPS) aims to keep London safe for everyone. One element is to make the most of the insights that can be gained from data and digital technologies to support ethical and effective crime prevention practices.

Existing research highlights the importance of analysing crime and offender networks, for example to better understand vulnerability to violent crime victimization, of child sex trafficking and knife crime. This project will explore connections between offenders and other agents.

The aim of this project is to develop techniques that use social networking analysis to target enforcement and preventative action to reduce violence. The candidate will work with MPS data holdings (e.g. arrest data, police intelligence) and open source data (e.g. social media) to understand how can we connect known offenders in ways that inform how offending or victimisation may trigger or manifest.

Topics might include determining pre-existing relationships between offenders, how offenders are connected to each other, and the strength, frequency and influence of these connections. These measures can be integrated within existing machine learning forecasting models, or utilised in other ways to inform effective and ethical targeting of police resources.

ID: 116052

Start date: 01-10-2022

End date: 01-10-2026

Last modified: 31-01-2024

Grant number / URL: ES/T002085/1

Copyright information:

Using social network analysis to understand crime and victimisation

Manchester Data Management Outline

1. Will this project be reviewed by any of the following bodies (please select all that apply)?

- Funder
- Ethics

2. Is The University of Manchester collaborating with other institutions on this project?

- Yes - Part of a collaboration and owning or handling data

I will be working with MPS data holdings such as arrest records and intelligence reports following a successful (currently pending *edit: now successful) vetting application.

3. What data will you use in this project (please select all that apply)?

- Acquire new data

Use of existing arrest records, intelligence reports etc to generate models of criminal networks.

Simulation of artificial networks will be used, the extent depending upon data provided by the project partner

4. Where will the data be stored and backed-up during the project lifetime?

- P Drive (postgraduate researchers and students only)
- University of Manchester Data SafeHaven

Data, upon receipt, will be stored securely in the university P drive. Depending on the sensitivity of the data, password controls/restricted permissions will also be employed. If data is highly sensitive, the data safe haven will be used. Raw data files will not be shared, instead links will be used. Sign in from a UoM account will be required.

Raw data will not be shared outside of the supervisory team or project partners.

Physical copies of any data are not expected to be necessary, however if required, they will be stored in a locked cabinet in a password secured room (Criminology Postgraduate office).

5. If you will be using Research Data Storage, how much storage will you require?

- < 1 TB

Data used for this research should not comprise more than 1 TB.

6. Are you going to be receiving data from, or sharing data with an external third party?

- Yes

I will be receiving information from the Metropolitan police, specifically working with the Strategic Insight Unit (located at New Scotland Yard). A data sharing agreement may be necessary, if not already in place. This will be discussed with the supervisory team and project partner.

7. How long do you intend to keep your data for after the end of your project (in years)?

- 5 - 10 years

Data may be kept for further research. 0-4 years will be the minimum (as data is expected to be provided during the integrated Masters stage of the PhD, meaning up to 4 years is required for the duration of the programme). The further 5-10 year timeframe is to facilitate future post-doctoral research as necessary. This plan will be updated following discussion with the project partner.

Guidance for questions 8 to 13

Highly restricted information defined in the Information security classification, ownership and secure information handling SOP is information that requires enhanced security as unauthorised disclosure could cause significant harm to individuals or to the University and its ambitions in respect of its purpose, vision and values. This could be: information that is subject to export controls; valuable intellectual property; security sensitive material or research in key industrial fields at particular risk of being targeted by foreign states. See more examples of highly restricted information .

If you are using ‘Very Sensitive’ information as defined by the Information Security Classification, Ownerships and Secure Information Handling SOP, please consult the Information Governance Office for guidance.

Personal information, also known as personal data, relates to identifiable living individuals. Personal data is classed as special category personal data if it includes any of the following types of information about an identifiable living individual: racial or ethnic origin; political opinions; religious or similar philosophical beliefs; trade union membership; genetic data; biometric data; health data; sexual life; sexual orientation.

Please note that in line with data protection law (the UK General Data Protection Regulation and Data Protection Act 2018), personal information should only be stored in an identifiable form for as long as is necessary for the project; it should be pseudonymised (partially de-identified) and/or anonymised (completely de—identified) as soon as practically possible. You must obtain the appropriate ethical approval in order to use identifiable personal data.

8. What type of information will you be processing (please select all that apply)?

- Special category personal data, or criminal offence data
- Anonymised personal data
- Pseudonymised personal data

Depending on data released by the Metropolitan Police, data such as intelligence reports and arrest records may be accessed. These may or may not be anonymised. The lead researcher (Ben Palfreeman-Watt) is expected to undergo a developed vetting procedure to obtain clearance to access the data (currently pending approval *edit: vetting approved).

If data is not anonymised, anonymising the data will be the first priority upon receipt.

Simulated networks will require no anonymisation, being entirely artificial.

9. How do you plan to store, protect and ensure confidentiality of any highly restricted data or personal data (please select all that apply)?

- Store data in buildings, rooms or filing cabinets with controlled access
- Store data on University of Manchester approved and securely backed up servers or computers
- Store data in encrypted files, folders, computers or devices
- Anonymise data
- Impose suitable data sharing and collaboration agreements
- Access data hosted by the University of Manchester via its secure Virtual Private Network (VPN)
- Access data hosted by a third-party data provider via their secure facilities (e.g. the UK Data Service Secure Lab)
- Where needed, follow University of Manchester guidelines for disposing of personal data

1. Data will be anonymised as a first priority, if not already anonymised.

2. Data will be securely stored on the P drive, or data safe haven as necessary

3. Restrictions on viewing access will be imposed, only named individuals will be permitted access data.

4. Any data stored outside the P drive will be stored on an encrypted device

5. Devices will, themselves, be stored in secure rooms, such as the criminology postgraduate office

6. The UoM secure VPN will be used whenever accessing data

7. Data will be accessed from the Metropolitan Police using whichever secure facilities are available

8. Data is expected to be anonymised in house at the Met, either by the lead researcher on site or by dedicated staff members.

10. If you are storing personal information (including contact details) will you need to keep it beyond the end of the project?

- No

11. Will the participants' information (personal and/or sensitive) be shared with or

accessed by anyone outside of the University of Manchester?

- No

Personal information may be provided by the project partner, this being the case the project partner is obviously free to access it. It will not be shared with anyone else by the members of the research or supervision team, however.

12. If you will be sharing personal information outside of the University of Manchester will the individual or organisation you are sharing with be outside the EEA?

- No

13. Are you planning to use the personal information for future purposes such as research?

- No

14. Will this project use innovative technologies to collect or process data?

- No

15. Who will act as the data custodian for this study, and so be responsible for the information involved?

David Buil-Gil / Tomas Diviak / Nicholas Trajtenberg Pareja

16. Please provide the date on which this plan was last reviewed (dd/mm/yyyy).

2024-01-31

Assessment of existing data

Provide an explanation of the existing data sources that will be used by the research project, with references

This research may utilise MPS data holdings such as arrest data and police intelligence reports, pending a successful developed vetting application. Simulated data is also intended.
Initially, the MPS data will likely be qualitative, consisting of arrest data and/or police intelligence

reports. These will be analysed for relevant data, such as information indicating a network tie, or relevant offense information.

MPS data will be processed and stored as a spreadsheet in a network analysis relevant format. An example of this is having a column indicate offender identity and subsequent columns indicate the identities of offenders they are connected to. It is possible that time will be relevant to network data at various stages. If this is the case, network data can be stored in separate files representing apparent structure at each time point. Other data (such as records of criminal activity) can be stored as part of the network data (such as having an attribute for each individual) or separately in a dedicated file.

This data is not expected to comprise a significant amount of storage space, as such backup and storage will not be problematic (the University of Manchester OneDrive will be sufficient).

Access for this research will be dependent on a successful vetting application (approved in April). Future access and re-use, for example in the case of future research, will depend on consent from the Metropolitan Police, as well as anonymisation of any identifying information contained in the initial data.

Provide an analysis of the gaps identified between the currently available and required data for the research

The primary gap is the lack of currently available open source data on criminal networks. Understandably, many organisations with the capabilities to gather such data (i.e. law enforcement or intelligence groups) are often reluctant to release it. This has led to many studies on criminal networks relying on the few public access criminal network data sets that are available (such as the Caviar network), or relying on generated data such as through agent-based models. As such, findings from novel data sets are critically important to ensure that conclusions in research are not simply artefacts of a limited sample size, as research suggests criminal networks operate substantially differently from non-covert networks.

Furthermore, simulation studies within network analysis provide a useful tool to allow confidence intervals and network parameters to be estimated using maximum likelihood estimation and permutation significance. This research aims to contribute substantively to the field by addressing methodological issues such as missing data in network analysis.

Information on new data

Provide information on the data that will be produced or accessed by the research project

Data accessed

Arrest records (.docx), police intelligence reports (.docx), other (project partner has mentioned possibility of access to other sources such as archived network data, likely stored in .xslx or .csv format). The data is expected to comprise at least one criminal network. The initial volume of this data (in pure word count) may be quite high, however not to the extent that storage will pose an issue. The data will be parsed for network and activity relevant information with the rest being discarded.

Data produced

Network Data in graph format (.csv or .xslx), potentially with relevant actor level attributes (such as any offenses committed for a given time frame). This is the standard format for social network analysis research and allows convenient use of a range of dedicated software packages in R. It is also very efficient in terms of storage and easy to re-use, simply requiring the data to be loaded into a statistical environment with the relevant packages. The environment and packages used for this research are all open source meaning that, if permission is given by the project partner to share the data, anyone with the relevant skills would be able to easily use the data.

Data produced by simulation will be in .xslx or .csv format.

Total scale

Less than 1 TB

Quality assurance of data

Describe the procedures for quality assurance that will be carried out on the data collected at the time of data collection, data entry, digitisation and data checking.

The data is expected to be of high quality, comprising records supplied by the Metropolitan Police.

Guidelines for data extraction will mainly revolve around definition of a tie within the network when using arrest data and intelligence reports. Currently, it is planned to have two levels of tie; weak and strong. Weak ties will refer to network ties with only one data source confirming their existence, strong ties will refer to ties which have more than one data source as support. Single instance co-arrest data will also be treated as a weak tie in the absence of corroborating evidence.

If access is given, archived network data will be extracted as is, as these networks were constructed by police intelligence as part of larger scale operations (not simply built with co-arrest data) and verification of this information would be beyond the scope of the research team. These networks will not be assumed to be entirely accurate, the theoretical relevance of them will instead be in the difference in structure compared to more simplistic networks, such as co-arrest networks.

Simulated networks will have no quality concerns.

Backup and security of data

Describe the data security and backup procedures you will adopt to ensure the data and metadata are securely stored during the lifetime of the project.

Following anonymisation processes at the Met, the initial data will be stored on the P drive or data safe haven as necessary, in a password protected folder with restricted permissions (named individuals only). The password will be given only to individuals comprising the research or supervisory teams. Access to the data will be secured using the universities VPN.

Once the data has been processed (and anonymised if necessary), it will be stored securely on the P drive. When accessed, it will be accessed from a password protected computer. It will not be stored on portable storage devices such as USB sticks, nor on any device which is not password protected. Devices themselves will be stored in secure locations, such as staff or PGR offices. The data is only expected to be stored and processed on one computer, which will be encrypted.

No paper or other hard copies of the data are expected to be required, however if this expectation proves false, they will be stored in a locked cabinet in a securable room (the Criminology Post-

Graduate office of the Williamson building).

Access to the data is expected to take place in person, involving an in-person visit to New Scotland Yard where data will be transferred to an encrypted device.

Management and curation of data

Outline your plans for preparing, organising and documenting data.

Arrest records will be parsed from oldest to newest. They will be grouped into categories (such as from the same month) and searched for co-occurring names (or unique identifiers post-anonymisation), such as individuals being arrested together. Any co-occurrences will be interpreted as a network link and logged as such in anonymised graph format. Intelligence reports will be analysed in a similar fashion, with evidence of individuals interacting with each other being interpreted as a network link and stored accordingly.

Once initial data has been anonymised, the networks generated from it will be stored in separate files. The initial data will be stored in the P drive or data safe haven.

Networks data will be stored with file and folder names indicating crime type of network and time point (in the case of longitudinal data), such as "Drug_Trafficking_Jun22_Jun23" being stored in the folder "Drug_Trafficking_1". In the case of multiple networks of this type they will be numbered in the order they were discovered.

The final network data will be documented, describing relevant details about each node (person) as well as explaining the structure of the file (graph format excel file). The initial data will likely not be sharable due to the high chance of containing identifying information, however this is not expected to leave the Met before anonymisation. Final documentation will also include information regarding the research question/objectives, a description of how the data was collected and analysed and an explanation of the validation measures in place. Data architecture and other relevant metadata and content will be documented in a ReadMe where the data is stored, as well as in any repository the project partner grants permission for sections of the data to be shared publicly to.

Difficulties in data sharing and measures to overcome these

Identify any potential obstacles to sharing your data, explain which and the possible measures you can apply to overcome these.

Sharing the final data should not be problematic, depending on the consent of the project partner. The initial data will not be sharable as described above, however the processed data will be anonymised with all identifying information removed. Precedent for sharing anonymised criminal network information exists in the literature, such as the 'Caviar' and 'Siren' networks.

Consent, anonymisation and strategies to enable further re-use of data

Make explicit mention of the planned procedures to handle consent for data sharing for data obtained from human participants, and/or how to anonymise data, to make sure that data can be made available and accessible for future scientific research.

Consent for sharing will be dependent on the Metropolitan Police. If not pre-anonymised, all data will be anonymised in house before sharing such as by giving individuals a numerical identifier. The final data set will be anonymised to the extent of removing any directly identifying variables, as well as collation or collapsing of key variables which may otherwise be combined with public data to achieve identification, and express permission will be sought from the project partner before exporting any data to a university device. No data will be directly gathered from human participants.

Data which could be used to identify an individua directly will not be included in the shared data.

Copyright and intellectual property ownership

State who will own the copyright and IPR of any new data that you will generate.

After discussion with the supervision team, it was decided that copyright and IPR of any new data will be owned by myself (Ben Palfreeman-Watt), insofar as is possible. Data sharing will not be postponed or restricted for patenting. The Metropolitan Police have ownership over the initial data supplied, however the results of subsequent analyses are the property of the student, in accordance with section 3.1.4 of The University of Manchester Intellectual Property Policy, (my participation in this PhD being funded by the Metropolitan police).

Responsibilities

Outline responsibilities for data management within research teams at all partner institutions

Retrieving and sharing the initial data for the research will be the responsibility of the Metropolitan Police.

Data processing, cleaning, analysis, documentation and storage/other will be the responsibility of Ben Palfreeman-Watt. The supervision team is not expected to manage any data shared or simulated during this PhD.

Preparation of data for sharing and archiving

Are the plans for preparing and documenting data for sharing and archiving with the UK Data Service appropriate?

If consent is gained from the project partner to share the final data, the prepared data, simulated data, and code documenting analysis will be shared via an open online repository, such as that provided by the ESRC or another platform such as GitHub. The final dataset is expected to be simple, comprising a spreadsheet of anonymised actors and their connections, and another spreadsheet detailing actor attributes. Data will be made available after publication, or, failing that, at the earliest opportunity. No restriction of access is expected to be required. No restraints will be imposed upon the simulated data.

If permission is given to share the final data, this would be a useful contribution of itself to criminal network literature. As may be expected, data for this field is hard to come by unless working directly with law enforcement. As such, open data in the field is hard to come by (the 'Siren' and 'Caviar' networks being notable examples of widely used, publicly released data sets). The data will be appropriately documented to allow re-use, with instructions on how to pull the data from the repository and load it into a statistical environment included in the ReadMe file, in addition to relevant metadata. It will be saved in a graph format, which is standard in network literature, to enable convenient re-use.

Is there evidence that data will be well documented during research to provide highquality contextual information and/or structured metadata for secondary users?

Structural metadata will be written as the data is collected and included as a ReadMe file in the final repository. Whilst not all data may be cleared for release by the project partner, any data that is cleared will have information regarding structure, contents and relation to other forms of data documented. Metadata will also indicate how data has been processed using data providence diagrams, and indicate any categories of data which have been removed, their quantity, quality and rationale for removal.