Data & Policy (2023), 5: e4
doi:10.1017/dap.2022.41

RESEARCH ARTICLE

From text to ties: Extraction of corruption network data from deferred prosecution agreements

Tomáš Diviák[1,*] and Nicholas Lord[2]

[1]Department of Criminology and Mitchell Centre for Social Network Analysis, University of Manchester, Manchester, United Kingdom
[2]Department of Criminology, University of Manchester, Manchester, United Kingdom
*Corresponding author. E-mail: tomas.diviak@manchester.ac.uk

Received: 03 November 2021; **Revised:** 17 August 2022; **Accepted:** 21 December 2022

Key words: corruption; corruption networks; data extraction; reliability; social network analysis

Abbreviations: DPA, deferred prosecution agreement; NA, not available (missing data); SNA, social network analysis; SoF, statement of facts

Abstract

Deferred prosecution agreements (DPAs) are a legal tool for the nontrial resolution of cases of corruption. Each DPA is accompanied by a Statement of Facts that provides detailed and publicly available textual records of the given cases, including summarized evidence of who was involved, what they committed, and with whom. These statements can be translated into networks amenable to social network analysis allowing an analysis of the structure and dynamics of each case. In this study, we show how to extract information about which actors were involved in a given case, the relations and interactions among these actors (e.g., communication or payments), and their relevant individual attributes (gender, affiliation, and sector) from five Statements of Fact. We code the extracted information manually with two independent coders and subsequently, we assess the inter-coder reliability. For assessing the coding reliability of nodes and attributes, we use a matching coefficient, whereas for assessing the coding reliability of ties, we construct a network from the coding of each coder and subsequently calculate the graph correlations of the two resulting networks. The coding of nodes and ties in the five extracted networks turns out to be highly reliable with only slightly lower coding reliability in the case of the largest network. The coding of attributes is highly reliable as well, although it is prone to missing data on actors' gender. We conclude by discussing the flexibility of our data collection framework and its extension by including network dynamics and nonhuman actors (such as companies) in the network representation.

Policy Significance Statement

Mapping corruption networks is crucial for understanding how corporate corruption unfolds and evolves over time. This makes empirical analysis of corruption networks informative for formulating and testing efficient evidence-based strategies in combatting corruption, which cannot be obtained by more traditional prevalence indices or perception-based measures. In this study, we propose a rigorous and reproducible framework for extracting network data from textual summaries of corporate corruption cases. This framework is flexible and parsimonious, as it can be universally applied beyond our cases and it may incorporate various amounts and types of information. We demonstrate the applicability and reliability of our data collection framework by extracting five network data from Statements of Facts accompanying Deferred Prosecution Agreements.

1. Introduction

Major multinational corporations operating from countries around the world, including the UK, continue to be implicated in bribery, contract manipulation, and other acts of corporate corruption. By diverting financial resources from their intended recipients to corrupt politicians, public servants, or businesspeople, corruption hampers social and economic development (Uslaner, 2008). The severity of corruption and its consequences thus lead many scholars and practitioners to call for evidence-based and data-driven approaches toward its measurement, understanding, and control (Kertész and Wachs, 2020; Luna-Pla and Nicolás-Carlock, 2020).

While conventional measures such as perception indices or public opinion surveys provide useful information about how corruption is perceived, they offer little insight into how exactly corruption unfolds and into its internal structure and dynamics. However, in order to formulate evidence-based anti-corruption measures and strategies, reliable data about the structure and dynamics of corrupt relations are necessary. Social network analysis (SNA; Borgatti et al., 2013; Robins, 2015) offers a way to collect and analyze such data. SNA is an approach concerned with the collection and analysis of data about a set of nodes and ties among them. In terms of corruption, the nodes represent actors involved in corruption (politicians, managers, etc.) and ties represent their communication or exchanges of resources (bribes, favours, etc.; Lauchs et al., 2011; Diviák et al., 2019). Given such data, SNA can describe the structure of the network, identify the key actors within it, and provide many other insights (Diviák et al., 2019; Wachs et al., 2020).

Since 2014, the UK's Serious Fraud Office (SFO) has had the power to enter into so-called Deferred Prosecution Agreements (DPAs) with corporations implicated in foreign bribery. DPAs provide a tool for prosecutors to negotiate a formal, but voluntary, agreement with corporations to bring a conclusion to alleged corrupt behaviors. Each DPA is accompanied by a Statement of Facts (SoF) that provides a detailed account of the corruption that took place, including extensive relational data. These statements are textual records of the given cases, which summarize the evidence of who was involved, in which way, what they committed, and most importantly with whom. This information can be translated into networks amenable to SNA using content analysis.

In this study, we apply a data collection framework developed by Diviák (2019) to extract information from each SoF about actors involved in corruption, their attributes, and ties among them. We do so with two independent coders. We subsequently assess the reliability of their coding, an important step that is frequently omitted in extant research on criminal networks. The goal of this article is to provide a step-by-step guide on how to extract network data from the text while exploring DPAs as a data source.

2. Deferred Prosecution Agreements

The UK Government introduced DPAs as a means of holding corporations to account without having to pursue full prosecution in the criminal courts. DPAs were enacted under the Crime and Courts Act 2013, and came into force in February 2014. A DPA is a discretionary tool available for use by certain prosecutors in England and Wales that enables a formal but voluntary agreement to be negotiated between the prosecutor, and a corporation, with judicial approval, in order to defer a criminal prosecution for alleged criminal conduct in exchange for the fulfillment of certain terms (see King and Lord, 2018). In the case of corporate bribery in international commerce, the relevant prosecutor is the SFO, and at the time of writing (2021), seven DPAs have been negotiated with corporations implicated in foreign bribery.

As above, once a DPA has been negotiated, approved, and made public, the Agreement is accompanied with a Statement of Facts (SoF) that provides a detailed account of the criminal behaviors of the case, including the persons involved, natural and legal, and the nature of their relations, the "offending locations," and agreed information as to when the criminal acts took place and how. Whilst recognizing that these documents are negotiated and agreed statements between the prosecutor and the corporation, and that they do not require an admission of guilt on behalf of the corporation, they do represent

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

reasonably accurate depictions of the bribery that took place. The statements can be accessed via the website of the SFO and are available to the public for download.

The seven DPAs have been negotiated with the following: Standard Bank Plc; Sarclad Limited; Rolls Royce Plc; Güralp Systems Limited; Airbus SE; Airline Services Limited; and, Amec Foster Wheeler Energy Limited. These cases involved varying offenses, including substantive bribery offenses (i.e., offering or making a bribe), conspiracy to corrupt, and failures to prevent bribery by employees, agents, or associated persons. In some cases, the corrupt conduct related to a one-off transaction (e.g., Standard Bank), in others there was corruption over many years and many contracts (e.g., Sarclad). In some cases, company directors and senior employees were implicated (e.g., Güralp), yet in others the criminal conduct related to third parties and agents (e.g., Airbus). Many of the cases involved bribery in multiple jurisdictions (e.g., Rolls Royce). At the center of all these cases, however, was the bribery of foreign public officials in order to win or maintain business interests in particular jurisdictions, often low-income countries where the loss of public funds through corruption creates major social harms. The results in this study pertain to five DPAs only, as at the time of writing, the SoF for the Amec Foster Wheeler DPA was not available and the information contained in the SoF for the Rolls-Royce case was too vague to yield any meaningful network representation. We come back to the Rolls-Royce case in the discussion.

DPAs can only be negotiated with companies, and not individuals, which has led to variation and inconsistency across the statements of facts in terms of disclosed information on the individual actors in the network, as this is supplementary to the agreement with the company. In 2020, the SFO published a new chapter on DPAs in its Operational Handbook, which directly addressed the naming of individuals, stating that " [c]onsideration must be given to the necessity for and impact of the identities of third parties being published."[1] However, prior to this, some statements included the names of individuals involved in the alleged corruption, as well as agents and other third parties who may have enabled it. This was problematic as some individuals named as part of corruption conspiracies were later acquitted at trial. Usually, if criminal proceedings against individuals involved are being pursued or considered, individuals are anonymized to avoid prejudice in court proceedings. In some cases, the name of the negotiating company has been anonymized to ensure fairness in criminal proceedings against individuals. Whilst the SFO guidance does not prohibit the naming of individuals, doing so is increasingly the case. Where anonymization takes place, unique information of individuals is usually included so that distinctions between those involved can be made, although as above, this has not always been done satisfactorily. A consequence of anonymity in the SoF is that triangulation with other sources, such as media reporting, is very limited, or outright impossible, as it cannot be said with certainty who are the involved individuals.

In the following section, we place the DPAs in the context of other data sources used in criminal network analysis and focus on what type of information relevant for network analysis can be extracted from them.

3. Social Network Analysis

SNA is concerned with the study of entities and connections among them. SNA has been increasingly applied in criminology to study criminal networks, in which criminal actors cooperate or communicate in committing crimes and avoiding detection as it is in the cases of organized crime, gangs, or terrorism (Morselli, 2009, 2014; Cunningham et al., 2016). The research of corruption networks draws upon the analysis of criminal networks, as the actors involved in corruption networks are also collaborating or communicating with others when they commit bribery, kickbacks or when they manipulate contracts (Lauchs et al., 2011; Diviák et al., 2019; Luna-Pla and Nicolás-Carlock, 2020). By mapping who bribes whom, who conspires with whom on manipulation of contracts and corrupt acts, we can construct a network representation of a given case.

[1] https://www.sfo.gov.uk/publications/guidance-policy-and-protocols/guidance-for-corporates/deferred-prosecution-agreements-2/.

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

Collecting data on criminal networks poses a problem for researchers, because they need to collect observations on actors who may actively try to conceal themselves and their ties (Morselli, 2009; Diviák, 2019). In practice, this makes primary data collection via direct observation or questionnaire surveys impossible. Thus, researchers in this area are left to rely on secondary sources of data. In their recent review of criminal justice data sources, Bright et al. (2021) classify these sources into five categories: offender databases, investigative records, prosecution files, court files, and reports of department inquiries and commissions. In this perspective, Statements of Facts accompanying each DPA can be subsumed under the court files category, as they are a part of the judicial process of deferring prosecution from a company implicated in bribery in exchange for fulfilling certain negotiated criteria. None of the mentioned data sources is primarily intended for research purposes, which comes along with various issues regarding data validity and accessibility (Bright et al., 2012). For instance, actors in court may be incentivized to withhold incriminating information which may bias the resulting court file or in the case of police investigative records, the information may be classified and thus not accessible for researchers.

As a part of court proceedings, DPAs combine all the available information produced in the preceding stages of the investigation. Thus, each SoF may contain information from multitude of other sources such as transcripts of e-mail or phone conversations, surveillance summaries, summaries from interrogations, or information from public databases deemed relevant for the given case by the Serious Fraud Office. Statements of Facts thus provide a form of triangulation within one source of data similarly to other prosecution or court documents (Bright et al., 2012, 2021). Given that for criminal networks, there is usually no pre-specified "ground truth" network and network representations based on just a single source may sometimes substantially differ between different sources (Rostami and Mondani, 2015), such built-in compilation of multiple data sources in each SoF increases the likelihood of obtaining a more complete representation of the studied case. This also partly compensates for the limited possibility to triangulate or validate the information in DPAs by other sources due to issues with anonymization described above. Additionally, the Statements of Facts are publicly available, which is an advantage over other data sources in criminal justice domain (cf. Bright et al., 2012).

The information that can be extracted and subsequently constructed into a network dataset can be thought of as consisting of up to six different aspects, namely nodes, ties, attributes, modes, dynamics, and context (Diviák, 2019). In this study, we will demonstrate how to extract and code information on the first three network data elements that together allow to construct a network representation of a given case. The first such element is the set of actors involved in the case or the node set. The second element, ties, represents the relations and interactions among the actors. These two elements comprise information that is necessary in order to conduct SNA (cf. Bright et al., 2021). Finally, the third element, attributes, refer to variables that describe relevant individual traits such as affiliations or gender. In the following section, we detail how to extract the nodes, their attributes, and ties among them from textual summaries contained in the Statements of Facts that accompany each DPA.

4. Data Extraction

Since each SoF accompanying a DPA consists of textual summaries of what happened in a given case, they can be used as a data source for content analysis or other text mining techniques (Krippendorff, 2012). In criminal network studies, content analysis is frequently used for extracting network data (van der Hulst, 2009; Bright et al., 2012; Campana and Varese, 2012) and we also use it as a methodological basis for data extraction here. Specifically, we focus on extracting the content that relates to nodes (actors), their attributes (relevant variables), and ties representing their connections. We propose and illustrate a simple three-step approach applicable not only to any SoF, but also to any textual summary in general, which enables to collect data about nodes, their attributes, and ties among them. The goal is to create a transparent approach that may be applied to various sources of textual records while enabling reliability assessment. Note that reliability assessment is not usually carried out or reported in current studies on criminal networks. For instance, Bright et al. (2021) do not report any study doing so in their systematic review of criminal justice records used for SNA.

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

4.1. Step 1—Establish the node set

The first step in constructing a network representation of a given case is to establish the node set. In other words, it is necessary to determine which actors will be included in the network. In SNA, this is referred to as the boundary specification problem (Laumann et al., 1983). In general, there are two approaches to specifying boundaries of a network. The first is a realist approach, wherein the actors themselves define what they consider the network and its boundaries to be (such as in snowball sampling or in qualitative studies). The second approach is called nominalist and here, the boundaries are set by the researcher using some sort of objective criterion (such as a list of nodes or some node-level variable). As we stated above, research on criminal and corrupt activities is almost exclusively based on secondary data sources and this rules out the possibility of adopting the realist approach for specifying the boundaries or the network (Diviák, 2019; Bright et al., 2021). This means that researchers in this area are relying on the nominalist approach and thus they have to adequately define a criterion that allows to clearly include or exclude individual nodes in the network. In our case, we included any actor that was mentioned in each SoF as participating in the case by collaborating or communicating with others. Such a definition of boundaries is the broadest possible as it allows one to extract the most actors and therefore the most data as well (cf. Campana and Varese, 2012). The broad definition of boundaries is suitable for early exploratory stages of research while it is not precluding from constraining the criteria for future stages, where the boundaries may be narrowed in order to answer a specific research question.

4.2. Step 2—Identify node attributes

Attributes are variables that describe relevant individual qualities of the nodes. Researchers of criminal networks typically face a dilemma when they want to extract attribute data. On the one hand, there are reasons for extracting information of specific attributes that stem from theoretical interests or previous research. On the other hand, such information may not be available in the data source, because it is not of interest to the practitioners who produced the document. For instance, there has been some interest in specific skills and expertise and how they shape network positions of the actors who posse them (Bright et al., 2015; Diviák et al., 2020). However, information about whether each actor has or does not have a particular skill or expertise is rarely mentioned in the textual summaries of the cases. This is also the case of each SoF, where the information may be occasionally mentioned, but in order to analyze such information reliably, it is necessary that the information is mentioned systematically. Otherwise, one risks making inferences from highly incomplete data. In our analysis, we extracted information on actors' gender, organization they work for, and whether they are working in public or private sector.

4.3. Step 3—Identify and assign ties

The crucial element that goes beyond actors and their attributes (as in traditional quantitative data analysis) and that makes it possible to construct a network is the extraction of ties or edges, which represent the interactions and relations among the chosen set of nodes. Ties have two basic properties: direction and strength (Borgatti et al., 2013; Robins, 2015). If it is possible to define whether the tie goes from a node another, it is directed (e.g., "A sent an e-mail to B"), whereas when it is not possible to distinguish the direction, the tie is undirected (e.g., "A and B met together"). In the final network, all the ties have to be either directed or undirected. This is problematic when extracting the data from a SoF, because ties that are theoretically directed (such as phone calls or e-mails) may not be reported as such, which is the case in formulations such as "A and B exchanged several e-mails" or "during that month, A and B spoke on the phone multiple times". This implies that in most of the cases in our dataset, we cannot ascertain the direction of each tie and thus we work with all ties as if they were undirected. As for strength, the ties can have different strengths depending on the number of times an interaction occurred between a particular pair of nodes. Thus, if a tie between A and B has a strength of four, it means that there are four recorded interactions between these two nodes in the corresponding SoF. In some DPAs (e.g., the Sarclad case), the granularity of the information contained in the SoF allows to clearly count the number of

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

interactions between any given pair of actors, while in other cases, the information may not be as granular, but the number of recorded interactions still functions as a proxy of tie strength between the two actors. For the following section, we therefore consider the ties to be undirected and weighted for the purposes of reliability assessment and demonstration here. Note that if the granularity of information in any given case does not allow to specify the tie strength, the ties can still be constructed as binary indicating a presence of a relationship between the two actors.

4.4. Demonstrating the data extraction procedure

Let us demonstrate this data extraction procedure on a specific example of an SoF related to the ASL case. Among many other things, this SoF states the following:

> *In acting on behalf of ASL, ASL Agent 1 worked closely with ASL Senior Employee 3 and ASL Senior Employee 4 who were based in Germany. A large part of the business that ASL Agent 1 introduced to ASL was with Lufthansa. At the same time as acting as an agent for ASL, ASL Agent 1 was also retained by Lufthansa as a consultant project manager in a department named Product Competence Centre Cabin Interior & In-flight Entertainment. Former Lufthansa Senior Employee 2 allocated work and gave instructions to ASL Agent 1.*

This quote mentions four actors, namely *ASL Agent 1, ASL Senior Employee 3, ASL Senior Employee 4,* and *Former Lufthansa Senior Employee 2.* Since they all are reported as actively participating in the case of corruption, they fulfil the boundary specification criterion that we defined above and so we include them all in the network. Furthermore, the affiliation of each actor can be determined from this quote as well - *ASL Senior Employee 3* and *ASL Senior Employee 4* are affiliated solely with ASL and *Former Lufthansa Senior Employee 2* is solely affiliated with Lufthansa, whereas *ASL Agent 1* is affiliated with both companies. This is captured in a part of the attribute list in Table 1. Additional attributes, such as gender, cannot be established from the quote above, and so the remainder of the SoF has to be searched in order to extract the information (if it is available). The quote above also contains information which can be translated into three ties: two ties in the part "*ASL Agent 1 worked closely with ASL Senior Employee 3 and ASL Senior Employee 4…"* and one tie in "*Former Lufthansa Senior Employee 2 allocated work and gave instructions to ASL Agent 1."* A part of the edgelist that captures these three ties is displayed in Table 2. Edgelist is a common data format in SNA, where each row refers to an edge and each column then describes the properties of a given edge. Two of ties in Table 2 have no direction, while the third is directed from the *Former Lufthansa Senior Employee 2* to the *ASL Agent 1.* None of the ties contains any weight, but these ties may not be all the ties between the given pairs of nodes. Further interactions or relations between the connected actors may appear in the rest of the SoF. If that happens, they can be recorded in the same way in the edgelist and their frequency summed up in the final network representation, which we also do in this study.

Note that the data extraction procedure described above can be used not only straightforwardly once from step 1 to step 3, but it may also be used iteratively when new information becomes known or

Table 1. *Example attribute list*

Example attribute list		
Actor	ASL	Lufthansa
ASL Agent 1	1	1
ASL Senior Employee 3	1	0
ASL Senior Employee 4	1	0
Former Lufthansa Senior Employee 2	0	1

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

Table 2. *Example edgelist*

Example edgelist		
Source	Target	Direction
ASL Agent 1	ASL Senior Employee 3	0
ASL Agent 1	ASL Senior Employee 4	0
Former Lufthansa Senior Employee 2	ASL Agent 1	1

established by the coders. For instance, if an involvement of a previously unmentioned actor is discovered, such actor with her attributes and ties may simply be included on top of all the information that had been available so far. This data extraction framework thus enables considerable flexibility.

5. Reliability Assessment

As the example in the previous section shows, an unstructured document such as an SoF may be ambiguous in what information it contains and how that information can be extracted from it. This induces an element of subjectivity into the data extraction and coding. In order to assess to what extent is the resulting solution prone to subjectivity, researchers usually use two (or more) independent coders to extract the data and subsequently assess the reliability of their coding using an appropriate measure of agreement (Krippendorff, 2012). We also follow this logic by employing two independent coders, assessing the reliability of their coding, and constructing the resulting network dataset by combining the coders' datasets. For the purposes of demonstration, we simplify some of the decisions here, but in general, the decisions should be guided by the research question at hand.

Regarding the coders, they were two criminology research master students with experience in data analysis and brief exposure to SNA provided by us together with instructions on how to code the material. The coders conducted manual coding of the material for two reasons. Firstly, such coding allows researchers to get an insight into the qualitative dimension of each case and familiarize themselves with its context (van der Hulst, 2009; Campana and Varese, 2012). Secondly, the interactions and relations among actors are latent content, which means that they rarely if ever explicitly mentioned as interactions or relations and are instead described or referred to in various different and context-specific ways (Campana and Varese, 2012). The latency of the content together with unstandardized structure of each SoF makes any automated extraction much more difficult to carry out validly, although it does not entirely rule it out. We come back to the possibility of automated text extraction in the discussion.

5.1. Step 1—Establish a common node set and its reliability

Extracting and coding network data from a textual summary similar to an SoF comes with one key problem—there is no "ground truth" in terms of which units to code, that is there is no prespecified set of actors or ties among them for coders to code. However, in order to assess the coding reliability, the coding has to be comparable between the two coders. To achieve comparability, the very first step in the coding process is therefore to establish a set of units, on which all the subsequent coding will be anchored. In our case, the most suitable such "anchor" is the node set defined by the boundaries. As we explained above, the boundaries for the present study are the broadest possible—all actors actively involved in the case and reported as such in the SoF. The specification of the node set is done in two steps. First, each coder codes the SoF on their own and tries to establish all the actors who satisfy the boundary definition criterion. Second, the solutions of the two independent coders are compared and the nodes that appear only in the solution of one coders are independently verified by a third person (in our case, one of the authors) with regard to the SoF and the final node set is established by taking the union of the verified nodes. This additional verification helps with clearing the data of errors such as inconsistencies (a single actor being

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

Table 3. *Reliability assessment results*

Reliability assessment of data extraction and coding					
Case	Güralp	ASL	Std Bank	Sarclad	Airbus
Nodes	1	1	0.82	1	0.84
Attributes					
Gender	1	NA	1 (7 NA)	1 (5 NA)	NA
Organization	1	1	1	1	1
Sector	1	1	0.89	1	1
Graph correlation	0.99	0.98	0.94	0.97	0.82

Abbreviation: NA, not available, missing information.

referred to with two different names in the SoF), when an organization is mistakenly included as a node, or when typos in the coding produce two different nodes who are in fact the same actors (e.g., "ASL Agent 1" vs "ALS Agent 1"). As a measure of intercoder reliability for extracting the nodes, we calculate the percentage of actors who were identified by both coders in the number of actors identified in total in their initial coding of the data (before combining the solutions of both coders for steps 2 and 3). Table 3 shows the number of nodes that has been established this way for each case as well as the measure of reliability.

5.2. Step 2—Node attributes coding reliability

With an established set of nodes, coders may code their attributes and ties among them. This task closely resembles standard content analysis, in which coders code a unified set of units on predefined variables. Here, we chose to extract information on actors' gender (binary variable), sector they are employed in (binary variable with values "private" and "public"), and the organization they work for (for simplicity, a binary variable indicating affiliation to the company that is the subject of the given DPA). To assess coders' reliability on each of these attributes, we use a simple matching coefficient defined as the number of the actors coded in the same way by both coders divided by the total number of actors. The matching coefficients for each variable in each case are reported in Table 3.

5.3. Step 3—Ties coding reliability

Even with a unified node set, coding the ties presents the greatest challenge in the entire coding process. Similarly, as in the case of establishing the node set, this is due to the absence of a prespecified and/or unified set of ties to code. Instead, the coders have to extract the ties themselves. Moreover, there is a certain degree of ambiguity in the text of each SoF, where the same part of text may be used to derive different ties or where different parts of the text may eventually yield the same tie. Both these reasons make the reliability assessment of individual ties quite impractical or outright impossible. However, the ties each coder extracts can be used to construct a network, which is a representation of the network from a particular coder's point of view. Since the networks contain the same node set, they can subsequently be compared using graph correlation, which is an adaptation of Pearson's correlation coefficient for two adjacency matrices with the same dimensions (Butts, 2008). The graph correlation is in this view a measure of coding reliability appropriate not only for our undirected and weighted ties, but also for other types of network data provided they are based on the same node set. The graph correlation coefficients of each case are reported in Table 3.

The process of assessing reliability of the coding of ties is illustrated in Figures 1 and 2, where we continue with example of the ASL case. The two figures visualize the networks constructed from the coding of each coder. In both visualizations, the positions of nodes are fixed in a circular layout for to enable easy visual comparison. Both coders largely agree on the structure of the network: both capture strong ties representing frequent interactions between *ASL Agent 1* and *ASL Senior Employee 3*, and

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

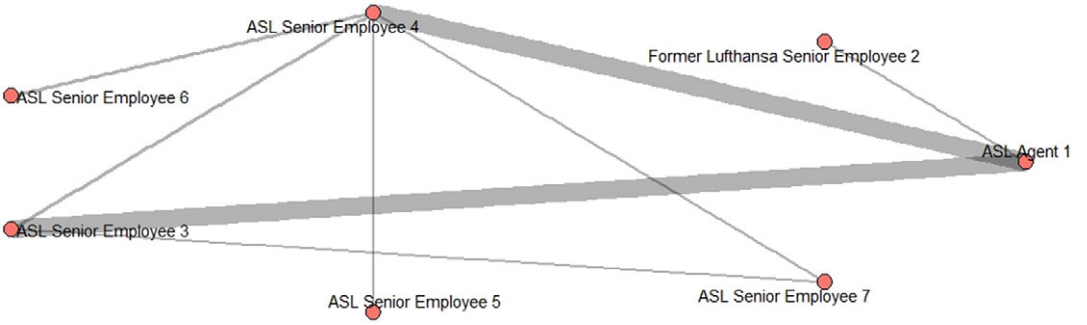

Figure 1. *ASL case network constructed by coder 1, thickness of edges corresponds to their weight.*

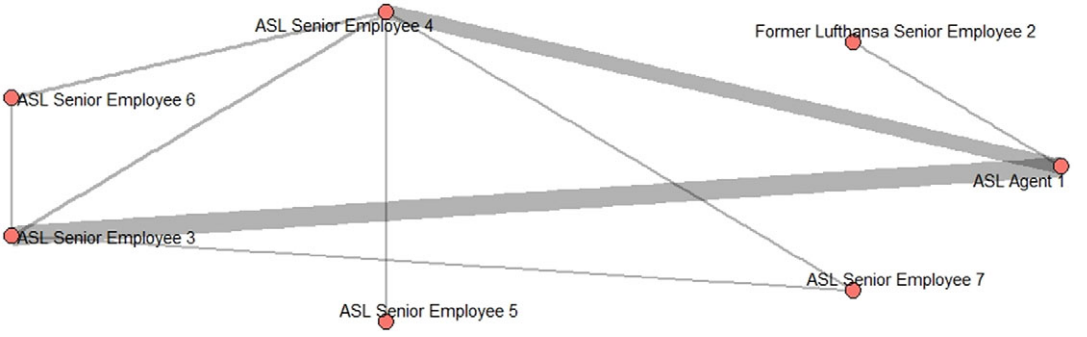

Figure 2. *ASL case network constructed by coder 2, thickness of edges corresponds to their weight.*

between *ASL Agent 1* and *ASL Senior Employee 4*. Furthermore, in both coders' solutions, *ASL Senior Employee 5* and *Former Lufthansa Senior Employee 2* have ties only to one other node (*ASL Senior Employee 4* and *ASL Agent 1* respectively). The only marked difference is that coder 2 finds a connection between *ASL Senior Employee 3* and *ASL Senior Employee 6*, while coder 1 does not. While the omission of this tie changes the structure of the network slightly, it does not alter the overall conclusions in any substantial way. The underlying adjacency matrices of the two networks are subsequently used to calculate the graph correlation. Union, intersection, or any other criterion for aggregating of the two networks may then be used to produce the final network representation. For instance, if we combined the two versions in this case using union, the resulting network would be the same as the network depicted in Figure 2, because it contains all the information captured in Figure 1 plus the tie between *ASL Senior Employee 3* and *ASL Senior Employee 6*. The choice of how to aggregate the two (or more) versions of the network into a single network for final analysis should reflect the research question one is aiming to answer and also reliability of the coding. In general, if the coding reliability is high, the criterion used for deriving the final network representation should not make a substantial difference as the results would not differ due to the underlying similarity of the extracted networks.

Table 3 shows that the size of the networks may vary quite substantially across the case ranging from five participating actors (Güralp) to 36 (Airbus). In general, the cases with small number of actors display very high reliability in terms of extracting the node set. The small cases Güralp and ASL together with the Sarclad case display perfect reliability of 1 indicating that our coders were able to retrieve the same set of actors when initially coding the Statements of Facts. The Standard Bank and Airbus cases display nonperfect reliability in terms of establishing the nodes set with initial nodeset overlap between the coders of 82% and 84% respectively. The Standard Bank case was the first case the coders had to code together with the Güralp case and so the lower value may be partly attributed to the learning effect. Specifically, coders in this case did not mutually reach three actors and one of them also erroneously coded

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

a company ("SB subsidiary") as an actor. The Airbus case is the largest case with a total of 36 actors identified eventually and with 40 pages long Statement of Facts, which makes the initial coding more challenging for the coders. These two cases demonstrate the usefulness of manual coding and the possibility to revise the initial coding in establishing the node set.

Despite this high variability in the network size, the reliability of the coding of ties is consistently very high between the two coders with the graph correlations above .9 in all cases but Airbus, where the graph correlation reaches .82. To put these values in perspective, the coding of the ASL example detailed above shows a graph correlation of .98 suggesting that while the coders are not perfectly aligned in their coding, their coding overall results in very similar networks. The same can be said about the remaining networks as well with the exception of the Airbus network, as the coding reliability there is lower (.82). The Airbus case is the largest one in terms of number of nodes which may partly explain its relatively lower coding reliability for ties, as the number of potential disagreements between the coders increases by (n-1)/2 for coding ties with each additional node (so while there are only nine more nodes in the Airbus case than in the Sarclad case, there are 279 more pairs of nodes to compare). For the purposes of substantive analysis, the results would be expected to be largely the same for the four cases with high graph correlation regardless of which coding or their combination (i.e., union or intersection) we would use, whereas minor differences could be expected in the Airbus case. For the Airbus case, a deeper examination of where the coders differ and why would be advisable before carrying out a substantive analysis of the case, as this case also exhibits lower reliability in terms of establishing the initial node set.

The matching coefficients reported in Table 3 for coding the attributes reveal two main insights. The first insight is that whenever there is enough information about actors' gender, organization, or sector, the coders usually achieve perfect agreement with the matching coefficient value of 1. The sole exception to this is the coding of sectors in the Standard Bank case, as the coders did not match on the fact that two actors are simultaneously involved in both public and private sectors (thus exposing them to the risk of conflict of interests). The second insight from assessing the reliability of coding the attributes is that there was a lot of missing information related to actors' gender (see the number of NAs reported in brackets in Table 3) in all the case but the smallest one (Güralp). For all the actors on whose gender there was any indication in the text of the given SoF, the coders match on the resulting coding. They also match on the actors where there is no such indication. It may seem somewhat paradoxical that seemingly detailed information about the affiliation of actors is always provided in the text of the SoF, while information about gender, an attribute that is one of the most frequent in the social sciences, is frequently missing. However, this makes sense in the light of the fact that DPAs are legal procedures, not scientific procedures, and thus the Statements of Facts contain legally relevant information and not necessarily information, that would be relevant scientifically only. Therefore, any information about which organization a particular actor is affiliated with is always present as it is essential for each case, while the gender of actors is never explicitly mentioned, because it bears no significance on one's culpability. The information on gender can be only deduced from personal pronouns or other implicit mentions (such as "Mr" or "Mrs" in some of the names) used in the text. Especially in the case of peripheral actors who are mentioned in only few instances, this may not be possible.

6. Discussion

Despite the seriousness of corruption as a social phenomenon and its consequences, scientific research of corruption is hindered by paucity of data on how actors involved in corruption operate. In this study, we proposed using Deferred Prosecution Agreements (DPA) as a data source, from which data amenable to social network analysis (SNA) may be extracted using a simple and reproducible coding. We used a framework proposed by Diviák (2019) for extracting network data from textual summaries accompanying each DPA and for subsequent reliability assessment of this coding. Among the five cases we processed and assessed their coding reliability, we see very high coding reliability of coding actors and ties, except for the largest case where the reliability is still adequate, and we also see very high coding reliability in terms of coding individual attributes of the nodes assuming the information is available on the given

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

attribute regarding a particular actor. Our study thus also contributes more broadly to the literature on data collection for covert and criminal networks (van der Hulst, 2009; Bright et al., 2012, 2021; Campana and Varese, 2012) with explicit emphasis on coding reliability which is seldom discussed in other studies.

We showed a three-step procedure for extracting nodes, their attributes, and ties among them, which is the essential information for constructing and analyzing corruption networks. However, SNA can utilize further information in the data and the entire framework for data collection can be extended further to include additional information (Diviák, 2019). Specifically, two immediate extensions regard network dynamics and multilevel/multimodal network data. Network dynamics refer to the change of networks over time. In our context, we could theoretically code specific time-stamp for each extracted tie that would denote when the tie was created and when it was dissolved. Such granularity may not be realistically achievable in each case depending on the information precision in the corresponding Statement of Facts, but in some cases, ties could be dated for when the interaction they represent occurred with precision on days in some cases or months in others. Reliability of such temporal coding could then be assessed by calculating the correlation of the resulting time series. Multilevel or multimodal network data refers to such network representation that captures multiple separate types of nodes and ties both within between these types. In the case of corporate corruption, multilevel networks may help in capturing the role of nonhuman nodes involved in the corruption, such as companies, the affiliation of actors to them, and the ties among the companies (e.g., contracts). Similarly to network dynamics, including this additional information may enrich the picture of the network and our analytical insights about corruption, but it may not be always or fully available in the Statements of Facts.

As the preceding paragraph together with our results suggest, information availability presents the largest limitation of our framework. There are two specific instances where the information unavailability precluded us with going forward in the research. The first such instance was coding of the Rolls-Royce case. For this case, the Statement of Facts was available, but when coders started extracting the data they stumbled upon the fact that essential information for constructing network data was vague or unclear in this particular case. This related to the way actors are referred to in the SoF. In all the other cases, each actor was referred to by a single distinguishable name or nickname. This enables to distinguish who interacts with whom throughout the entire SoF. However, in the Rolls-Royce case, actors were simply labeled as "Rolls-Royce employee" despite referring to different individuals, making it in turn impossible to distinguish who exactly was the actor participating in a given relation or interaction. As this occurred frequently in the SoF for this case, it was not possible to extract any meaningful network representation from it due to the fact that the very first step in the data extraction, establishing the node set, could not be done unambiguously as in the remaining case. The second instance, where unavailable information restricted the data we were able to extract, was while coding the gender of actors involved in our cases. Only in the smallest case (Güralp, $n = 5$), it was clearly possible to code the gender of all involved actors. In the remaining cases, it was not always possible to infer actor's gender from personal pronouns or their (nick)name, especially for actors who appeared only once or a few times in the SoF. This makes any coding and subsequent analysis of the role of gender in our networks severely limited. Similarly, the anonymization of a given SoF makes it practically impossible to use data sources (e.g., media, criminal justice files, online registers) to triangulate or complement the information with.

Unclear labeling of actors and missing information about their gender both stem from the same underlying issue, which has a more general impact on data collection on covert and criminal networks. The issue is that the original data sources are not primarily intended for scientific purposes, but for legal (in the cases of DPAs or court records) or security (such as police monitoring data) purposes. Therefore, information that may be of scientific relevance may simply not be contained in the data source, because it is of little or no importance with regard to its primary purpose (cf. Diviák et al., 2020). Further, as we have explained above, there is no "ground truth" underlying network against which researchers could compare their data or use to triangulate, which makes unavailability of information an obstacle impossible to overcome in this research area and thus to answer certain research questions. Researchers should keep this in mind when collecting data in a similar way as in this study, as the data source itself is a representation of reality constructed with a specific aim such as prosecution or surveillance. This may in turn introduce

https://doi.org/10.1017/dap.2022.41 Published online by Cambridge University Press

biases in the data such as the spotlight effect, which denotes a disproportionate attention to certain individuals in the data source (e.g., public figures in the media coverage or initially monitored actors in the police investigation) making them appear to be more central in the network even though it may only be that they are more frequently and/or thoroughly observed (Smith and Papachristos, 2016; Bright et al., 2018; Diviák, 2019). Awareness of these potential sources of biases is very important as knowing how any given bias might have affected the data can be used to account for it for instance by including appropriate control variables in statistical modeling or by cautious interpretation of results.

Speaking of further research, there are numerous opportunities for the data we collected open for both substantive and methodological research. In terms of substantive research, SNA guided with appropriate testable theories can illuminate the networks we constructed in this study by describing their structure and testing the mechanisms that generate these structures. Situated in the context of previous research, this data can thus contribute to the research on corruption networks and the way they operate (Lauchs et al., 2011; Diviák et al., 2019; Luna-Pla and Nicolás-Carlock, 2020; Wachs et al., 2020). Exploratory probe into three of the network extracted here reveals remarkable structural similarities in the centralization of their structures despite the differences in the sizes of the networks or their time span (Diviák and Lord, 2022). In terms of methodological research, our data collection framework may be further extended to other sources of textual data and it may also be automated using available methods for automated text analysis (natural language processing methods). The issues with irregular structure and latency of content together with our aim to grasp the qualitative context of each case led us to utilize manual coding here, but for large bodies of text, devising an automated mode of extraction may save time and be worth the effort. For our cases with 31.2 pages on average (range 15–55), it was more efficient to conduct manual coding. Content automatically extracted from voluminous bodies of text may be in turn validated by human coders by coding a subset of the data source.

Using a common framework for data extraction that we presented in this article together with the analytical toolbox in SNA for both descriptive and inferential analysis has the potential to enhance comparability across studies and accumulation of knowledge in the research on corruption. Robust evidence gained with rigorous methods can then eventually be used for designing intervention and prevention measures in combatting corruption, provided that researchers critically examine the potential consequences of the measures with a dose of scientific skepticism.

Funding Statement. Data collection for this study was supported by Manchester Statistical Society grant *Conspiracy to Corrupt: Extraction and Analysis of Bribery Network Data from Deferred Prosecution Agreements* awarded to Tomáš Diviák. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests. The authors declare none.

Author Contributions. Conceptualization: T.D., N.L.; Data curation: T.D., N.L.; Data visualization: T.D.; Methodology: T.D.; Writing original draft: T.D., N.L. Both authors approved the final submitted draft.

Data Availability Statement. Data can be made available to interested researchers upon request by email to the corresponding author.

References

Borgatti SP, **Everett MG and Johnson JC** (2013) *Analyzing Social Networks*. Thousand Oaks, CA: SAGE.

Bright D, **Brewer R and Morselli C** (2021) Using social network analysis to study crime: Navigating the challenges of criminal justice records. *Social Networks 66*, 50–64. https://doi.org/10.1016/j.socnet.2021.01.006

Bright DA, **Greenhill C**, **Reynolds M**, **Ritter A and Morselli C** (2015) The use of actor-level attributes and centrality measures to identify key actors: A case study of an Australian drug trafficking network. *Journal of Contemporary Criminal Justice 31*(3), 262–278. https://doi.org/10.1177/1043986214553378

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

Bright D, **Koskinen J and Malm A** (2019) Illicit network dynamics: The formation and evolution of a drug trafficking network. *Journal of Quantitative Criminology 35*, 237–258. https://doi.org/10.1007/s10940-018-9379-8

Butts CT (2008) Social network analysis with sna. *Journal of Statistical Software 24*(6), 1–51. https://doi.org/10.18637/jss.v024.i06

Campana P and Varese F (2012) Listening to the wire: Criteria and techniques for the quantitative analysis of phone intercepts. *Trends in Organized Crime 15*(1), 13–30. https://doi.org/10.1007/s12117-011-9131-3

Cunningham D, **Everton S and Murphy P** (2016) *Understanding Dark Networks: A Strategic Framework for the Use of Social Network Analysis* (Reprint edition). Lanham, MD: Rowman & Littlefield Publishers.

Diviák T (2022) Key aspects of covert networks data collection: Problems, challenges, and opportunities. *Social Networks 69*, 160–169. https://doi.org/10.1016/j.socnet.2019.10.002

Diviák T, **Dijkstra JK and Snijders TAB.** Structure, multiplexity, and centrality in a corruption network: The Czech Rath affair. *Trends in Organized Crime 22*, 274–297 (2019). https://doi.org/10.1007/s12117-018-9334-y

Diviák T, **Dijkstra JK and Snijders TAB** (2020) Poisonous connections: A case study on a Czech counterfeit alcohol distribution network. *Global Crime 21*(1), 51–73. DOI: 10.1080/17440572.2019.1645653

Diviák T and Lord N (2022) Tainted ties: The structure and dynamics of corruption networks extracted from deferred prosecution agreements. *EPJ Data Science 11*(1), 7. https://doi.org/10.1140/epjds/s13688-022-00320-2

King, C., & **Lord, N.** (2018). Negotiated Justice and Corporate Crime: The Legitimacy of Civil Recovery Orders and Deferred Prosecution Agreements.

Kertész J and Wachs J (2021) Complexity science approach to economic crime. *Nature Reviews Physics 3*, 70–71. https://doi.org/10.1038/s42254-020-0238-9

Krippendorff K (2012) *Content Analysis: An Introduction to its Methodology*, 3rd Edn. Thousand Oaks, CA: Sage.

Lauchs M, **Keast R and Yousefpour N** (2011) Corrupt police networks: Uncovering hidden relationship patterns, functions and roles. *Policing and Society 21*(1), 110–127. https://doi.org/10.1080/10439463.2010.540656

Laumann E, **Marsden P and Prensky D** (1983) The boundary specification problem in network analysis. *Applied Network Analysis: A Methodological Introduction 61*, 18–34.

Luna-Pla I and Nicolás-Carlock JR (2020) Corruption and complexity: A scientific framework for the analysis of corruption networks. *Applied Network Science 5*(1), 1–18. https://doi.org/10.1007/s41109-020-00258-2

Morselli C (2009). *Inside Criminal Networks* (Roč. 8). New York: Springer.

Morselli C (2014) *Crime and Networks*. New York: Routledge.

Robins G (2015). *Doing Social Network Research*. London: SAGE.

Rostami A and Mondani H (2015) The complexity of crime network data: A case study of its consequences for crime control and the study of networks. *PLoS One 10*(3), e0119309. https://doi.org/10.1371/journal.pone.0119309

Smith CM and Papachristos AV (2016) Trust thy crooked neighbor multiplexity in Chicago organized crime networks. *American Sociological Review 81*(4), 617–643. https://doi.org/10.1177/0003122416650149

Uslaner EM (2008) *Corruption, Inequality, and the Rule of Law.* Cambridge, United Kingdom: Cambridge University Press.

van der Hulst RC (2009) Introduction to social network analysis (SNA) as an investigative tool. *Trends in Organized Crime 12*(2), 101–121. https://doi.org/10.1007/s12117-008-9057-6

Wachs J, **Fazekas M and Kertész J** (2021) Corruption risk in contracting markets: A network science perspective. *International Journal of Data Science and Analytics 12*, 45–60. https://doi.org/10.1007/s41060-019-00204-1

Cite this article: Diviák T and Lord N (2023). From text to ties: Extraction of corruption network data from deferred prosecution agreements. *Data & Policy*, 5: e4. doi:10.1017/dap.2022.41