**Giving artefacts a voice: Collecting network data on non-participatory populations.**

Daniel Tischer, University of Bristol

Adam Leaver, University of Sheffield

**Abstract:**

**Introduction:**

Data is at the heart of social network research. As such, it should be of little surprise that literature on data collection and coding is generally well developed within the field. Social network researchers can access discussions of traditional data collection approaches – interviews, questionnaires and observations – in many books dedicated to ‘doing’ social network research (Bellotti 2015; Borgatti et al. 2013; Crossley et al. 2016; Robins 2016). More recently and driven by the ascent of ‘Big Data’, data collection has also become a more technical exercise. Data is collected by ‘scraping’ websites, either manually or through sophisticated machine learning, email correspondence and collecting data from social networking sites (Hogan 2008; Marres and Weltverde 2013[[1]](#footnote-1)). Of course, these developments are welcome and in line with the rise of the digital everything, yet they also change our relationship with the data – a push of the button replaces the more intense relationship with subjects in the field.

Getting *access* to data is a key obstacle for social network scholars. This is particular burdensome for network scholars seeking to engage with organisations because access needs to be negotiated at various levels – individual, team and organisation – and often requires significant commitment and support from managers (Ref Weissinger, this issue?). Capturing ‘all’ relationships within an organisation becomes increasingly difficult the larger the organisation is as they may have thousands of employees in remote locations rendering the endeavour practically impossible not least because resources are restricted. Hence much of organisational social network research is limited to one (or few) organisation(s), or, more likely, a specific activity / team within (Ref). In the event where network scholars manage to negotiate access to multiple organisations, coordination and resource limitations become an even bigger obstacle (Ref Swinburne people, this issue?). But the default response by organisations may be to refuse access to researchers, in particular where the proposed research is seen as counterproductive to the organisation’s own interests. But critical enquiries into organisation are of interest to many academics, and as such not studying a phenomenon because gaining access is too difficult or problematic is not an option.

This paper seeks to remedy the problem of access by providing an alternative approach: using documents as a source of data. Topically we focus on a set of firms that were involved in the structuration of *Collateralised Debt Obligations* (CDO) – the exotic financial derivative at the heart of the financial crisis of 2008. Access to these organisations is generally problematic for a number of reasons: 1) the phenomenon is historic (data up to 2008) and many firms are defunct; 2) there are 120 firms in this study which focuses on a particular activity; 3) some of the firms have been subject to regulatory and legal scrutiny; and 4) the object of the study revolves around what went wrong in collateral management. Hence, we assume a general unwillingness to “let the researcher in”. We focus on a particular set of actors, “collateral managers”, who selected and managed underlying assets for CDOs, a key risk-mitigating function in the structuration process to assuage investor concerns’ of moral hazard (Leaver and Tischer, forthcoming). Each product requires documentation which enters the public realm via a process of listing the securities on stock exchanges, and it is these documents that give us access to our data: career histories of employees involved in the structuration.

This discussion will focus on collecting this data via a series of reflections on discovery, collection, coding and triangulation. First, we will review existing literature on document-based research and relate this to existing discussion on data collection within SNA literature. Second, the paper sets out the research background and topic. In section three we will take stock – what data is available and how do we turn this into network data. Section four discusses how we can triangulate and add to this data using complementary sources from third parties. Section 5 discusses ethics before a concluding section will discuss benefits and downsides of this approach and outline potential areas for research.

**Section 1: Documents as prime data source for organisational network data**

Social network researchers have devoted considerable attention to discussions of data collection. Introductory textbooks on how to analyse social networks in particular feature a number of methods with some detail (see for example Scott 1991; Wasserman & Faust 1994; Borgatti, Everett & Johnson 1993: Crossley et al. 2015). Collecting data is largely discussed in two contexts: 1) going into the field – interviews, observations or questionnaires – and 2) extraction of data from existing sources – surveys, databases & archives. Data collection is usually bound by a particular setting and population of interest (a class or school, MSM, sports teams) or they may extract information from existing surveys, specific archives or internet databases (for example IMDB[[2]](#footnote-2) or Oxford Music Online[[3]](#footnote-3)).

Deciding on the “right” method is usually informed by the particular phenomenon – friendship relations, sexual health or team performance – that is to be investigated, its complexity, access and the size of the population. Ideally, data collection is designed to elicit a particular relational artefact, yet in many cases the method of choice may be prohibitively resource intensive. Interviews and questionnaires become increasingly *burdensome* with network size, particularly in situations in which respondents are asked to report on connections of other actors in the network (Borgatti et al. 2013: pg 60). Yet any research effort may be brought to a halt when access to particular populations or organisations is refused. This problem is of course not new and common to researchers of covert, health or organisational networks. There are obvious reasons for refusal: covert networks, for example, have a natural interest in maintaining secrecy, whereas organisations may refuse access on the grounds of commercial sensitivity or because the research may not yield any benefits for the firms. An even more obvious motive to refuse access may be found in the critical nature of enquiries into how businesses, criminal gangs or hidden population organise their day-to-day activities via networks given the political pressures and potential turbulence arising from such an enquiry (van der Hulst 2009[[4]](#footnote-4)).

Refusal of access to a population thus puts a researcher in a difficult position: should he or she 1) abandon the project or 2) find an alternative way in? In some cases a point blank refusal may make it impossible to continue research; yet, other situations in which direct access is refused may still allow for data to be collected, albeit be it indirectly. In these cases, researchers may revert to documents to access data.

***Beyond the archive***

It is somewhat surprising that documents feature relatively little in discussions of data collection for social network research given that documents are important artefacts of social and organisational life (Glaser & Strauss 1967, Preda 2002, Tischer et al. 2018). Where documents are explicitly discussed in social network research, this generally occurs within the context of archives (see Marsden 2005, Borgatti et al. 2013). Archives present a valuable source of historical relational data (Edwards & Crossley 2009; Edwards 2010[[5]](#footnote-5)) using multiple types of documents – speeches, biographies, letters, newspaper articles, etc. – that enable researchers to construct networks and narratives. The most prominent example of archival research informing SNA is probably Padgett and Ansell’s (1993) work on Florentine families based on a combination of various sources: historical books, tax assessments and neighbourhood co-residence (Crossley et al. 2016). Researchers of organisations find archives appealing because they are rich in content and allow them to explore organisational practices retrospectively (Ventresca and Mohr 2001[[6]](#footnote-6)).

But archives are also exclusive: archival documents are “accumulated by a person or organization in the course of the conduct of affairs and preserved because of their continuing value” (Ellis, 1993; 2). Archives exist because organisations or individuals see value in archiving documents for specific purposes (Welch 2000[[7]](#footnote-7)). This may include board minutes and annual reports are archived for governance reasons, financial records are kept for tax and regulatory purposes. Still, the act of archiving is not an exhaustive exercise but selective. Documents that are not considered of value by the organisation or the individual tasked with archiving data may not be considered or even be excluded (Rojas 2010[[8]](#footnote-8)). And even where organisational archives exist, they may not be (easily) accessible as they are private (Welch 2000).

Documents also exist outside of neatly-bounded archives; they are everyday devices, formal or informal, produced routinely by administrators, civil servants, managers etc. (Coffey 2014[[9]](#footnote-9)). Storing these documents therefore occurs with everyday use in mind and at the heart of evolving social practices (ibid). The US *Securities and Exchange Commission* has established “centralized recordkeeping facilities” to keep comprehensive up-to-date records of financial instruments and transactions to intervene in the day-to-day trading activities (SEC 2017[[10]](#footnote-10)). Here documents are not curated by individuals, but the process is automated and produces vast, ever expanding databases that provides researchers with alternative and equally valuable sources of data. Appealingly, in many cases access is less restricted in the open-access depositories.

***Accessing organisations’ secrets through the back door: documents and organisational fact making***

Documents are peculiar artefacts – they relay events and are snapshots of organisational life – but they are also more than mere descriptions of the past: “Documents provide a useful point of entry into contemporary problems” (Riles 2006: 2). Preda (2002) shows how documents are organizational devices both document and structure actions and relations within and across organisations. Documents may also structure and even “bullet-proof” the future by setting out strategic plans and targets to be achieved (Strathern 2006). Contracts are a prime example of ‘creating’: they contain detail on a particular transaction between two (or more) parties and structure the relationship between those parties as well as associated third parties - for example dispute resolution or trusteeship. In other words, contracts present both technical systems and communities of discourse (Suchman 2003).

But not all documents are of *equal* use to academic research. Documents take on many different forms and may differ in quality: they may contain revised, standardised or less structured observations of events. However, ‘quality’ cannot be generalised: it is situational and context-driven (Rowlinson et al. 17; Scott 1990[[11]](#footnote-11)). A standardised, glossy corporate document may provide useful information for the purpose of comparison (Tischer et al. 2018), but they may illicit fewer ‘interesting’ narratives than hastily written meeting notes or an unconsidered email response[[12]](#footnote-12). Documents receive their use-value from content *and* how this content relates to the problem that is to be researched. In other words, the research object frames the suitability of a document. This is because documents are artefacts that upon their publication become fixed entities; or better said, their content becomes a fixed entity. This durability, or immutability, of the document is what makes it a valuable source of data (Shankar, Hakken & Osterlund[[13]](#footnote-13)). As Rowlinson et al (5) argue, (organisational) documents are not only drivers of (organisational) narratives, but they are important, albeit “under-utilized” artefacts for constructing data. That is, a researcher’s ability to reinterpret and re-associate contents from documents enables them to produce new knowledge (Shankar et al 2017).

What does change, however, is how researchers approach these fixed contents over time as local and global events reshape society and how it contextualises social, economic and political action. Tischer et al. (2018) demonstrate how events change our engagement with documents can be explored in times of crises, such as the recent Great Financial Crisis. Without the collapse of the CDO market it is highly unlikely that we, the authors, would have explored CDO offering circulars. They are mundane, highly structured and technical-legal artefacts (Riles 2006). They are not presumed to be read by external parties, never mind academic researchers (Tischer et al. 2018). Even if we did read these documents in the absence of the GFC, we would have read them differently, in a different context (Scott 1990). Without the exhaustive analysis by the FCIC (2010), the numerous court cases heard and documented by the SEC, and Haldane’s (2009) work on financial networks, we may have never shown an interest in collateral managers and explored how these actors are connected, either directly or via latent ties, to other Wall Street institutions.

Of course, using documents has its downsides. Presuming an exclusive focus on documents, our analysis is firmly bound by what documents exist and the information they contain. In many cases it may be possible to search for additional data elsewhere – in other documents, through interviews or other means – but researchers cannot rely on this being the case. Also chances of locating additional primary data may decline over time as the context-specific knowledge leaves the organisation and employees move on (Welch 2000). Or there may be a general unwillingness by organisations to respond to requests, for example, after paradigm-shifting events such as the financial crisis. Yet on balance documents are valuable and trusted sources for the organisational, social science research. And even though they are not responsive and do not actively answer our questions, researchers can illicit answers from these documents.

**Section 2: Contextualising the document: Crisis, serendipity and careers**

The data gathering approach discussed in this paper is part of a larger research project which examines actors involved in the creation of *Collateralised Debt Obligations* (CDOs), the exotic financial credit derivative that caused the 2008 Great Financial Crisis (GFC). Whilst much has been written about the GFC over the past ten years, Leaver and Tischer (forthcoming) are particularly interested in observing the complex social network structures entangling different financial and legal actors that participated in the structuring of CDOs. The study is based on the analysis of a particular document: the CDO *offering circular* (OC). The OC is a legal document that is issued for a new security that is for sale as required by the US Securities Act of 1933. These documents set out the terms of trade and offers prospective buyers a detailed description of the product, payments schedules, conflict resolution, etc. The content and format of OCs is standardised (Tischer et al. 2018), and hence all OCs contain a key feature that is invaluable for network research: a list of actors that participate in the structuring of a specific CDO (Leaver & Tischer forthcoming).

Yet, despite their prominence, there is no publicly available database containing all CDOs documents. This is because changes to American law (CFMA Act 2000) allowed securities to be traded over-the-counter, hence removing requirements to list trades on exchanges (Stout 2010[[14]](#footnote-14)). Neither is there a conclusive figure for the number of CDOs issued leading up to the GFC, nor do we have access to a list naming all CDO deals; all we know is the value of CDOs issued (SIFMA 2016). This poses problems for researchers seeking to capture the complete population – i.e. all CDOs issued. Still, while recognising this issue, we identified the Irish Stock Exchange (ISE) as a suitable source of offering circulars. The ISE is a leading provider of listing services for debt securities, such as CDOs, and by listing products, offering circulars enter the public domain. To build our CDO database we gathered all CDO offering circulars issued between 2000 and 2008 available from the Irish Stock Exchange. Our final dataset is comprised of 450 CDO offering circulars.

***Serendipitous discoveries: On finding peculiarity in collateral managers***

When we started exploring these offering circulars, we were mainly interested in interorganisational networks leading up to the financial crisis. It was only later that we recognised the relative importance of collateral managers. But this ‘eureka’ moment is not born in the fact that financial literature and SEC announcements inform us about CMs’ relevance; it is the very documents which advertise the product to investors, the offering circulars (and as we later discovered the pitchbooks; see Tischer et al 2018), that alert us to CMs relevance. This actor commands considerable space in OCs and pitchbooks alike (Tischer et al. 2018) and they are the only biographical information contained in these documents. If it was a common feature to display expertise, then why are there no biographies for the investment bankers who directly interact with investors or the trustees? There is a certain peculiarity about the singling out of collateral managers, given all the possibilities.

Serendipitous discoveries, whilst seldom referred to as such, have long had their role to play in exploratory research: from surprise chemical reactions that ‘went wrong’ and created new materials, the discovery of Penicillin, to the apple dropping on Newton’s head sparking his imagination of gravity, serendipity is more widespread than we may acknowledge (Roberts 1989). If it was not for stumbling upon curious data, a certain observation or accidents, academic knowledge arguably would have not progressed in the way it did. Some might consider serendipity to “lack” scientific credentials as it is not preconceived, not defined *ex ante*, not articulated via hypothesis (Copeland 2015 thesis). Or researchers may decide to frame these discoveries through exploratory processes. After all, there was a research interest in OCs’ as sources of social network data, hence why we downloaded and read these documents in the first place. The discovery of data on CMs could therefore be framed as a *research finding*. But in reality, we did not look for the data outright. There was little way of knowing that these documents would provide us with what is a relatively rich source of data and descriptions of CM careers (see Fig.1). This discovery by chance is a fluid expressions of openness to new ideas, without preconception. Thus it was by reading through these documents, we discovered the peculiarity of one actor, the *collateral manager* (CM).

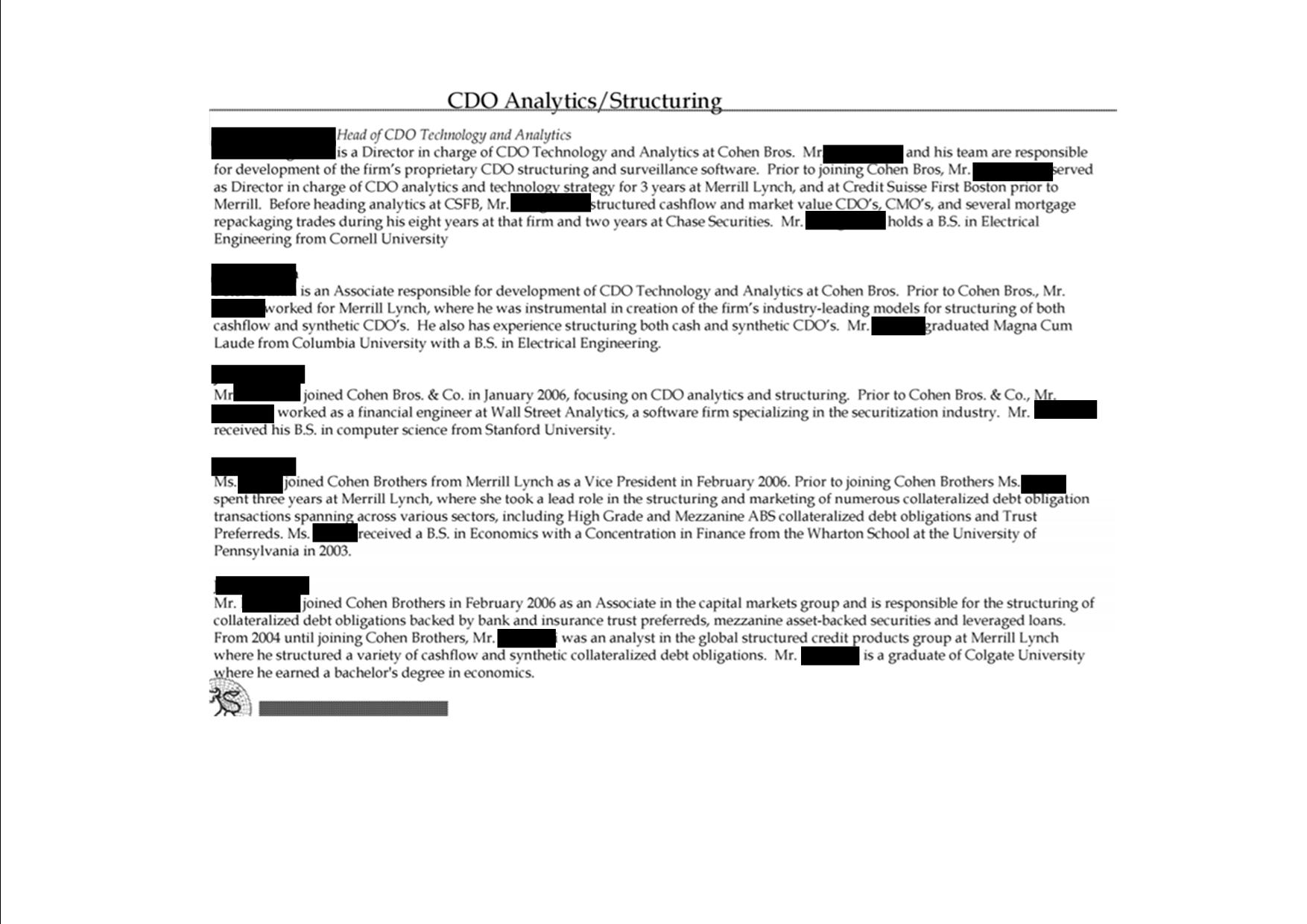
**

Figure 1: Anonymised example of collateral managers’ careers

***Collateral manager careers in context***

CMs are interesting actors because they sit between buyers and sellers and present a key risk-mitigating function (Leaver & Tischer forthcoming). In the context of CDOs, they are hired to select and manage assets for the benefit of the investor. But in the aftermath of the financial crisis it transpired that this was not always the case. Numerous legal actions pursued by the SEC reveal that a number of key collateral managers have failed their fiduciary duties (see SEC 2016[[15]](#footnote-15) for an overview). Instead, CMs acted in the interest of investment banks and hedge funds, e.g. Paulson and Magnetar, who pre-selected asset pools, without disclosing doing so to investors. This was in part facilitated by investment bankers hiring the collateral managers, rather than the investors. Leaver and Tischer (forthcoming) and Chernenko (2017) show how this introduced perverse incentives: investment banks pressured CMs into neglecting their fiduciary duty to investors and accept new terms, including the involvement of third parties in the portfolio selection, unbeknownst to investors (for specific cases, see for example SEC 2013; FCIC 2011[[16]](#footnote-16)).

This puts the CMs’ claim of being independent from investment banks in doubt. Collateral managers and investment banks alike stressed that CMs acted on behalf on their clients without instructions from investment banks. This claim of independence is, however, obscure as it does not detail what kind of independence is referred to. One CM, Maxim Advisory, highlights that they have “no *traditional* ties to investment banking” (Jupiter Highgrade II CDO, 2004a: 25) [emphasis added by the authors]), although employees have previously worked for investments banks such as Investec, Nomura and JP Morgan Chase. Goldman Sachs hired CMs that were founded by and staffed with former Goldman Sachs employees – *Greywolf* and *GSC Partners* (FCIC 2011: 392f.). In both cases Goldman Sachs shorted over a third of assets, that is, Goldman bet on the default of these assets to profit.

Collateral managers are therefore an interesting object of analysis. Legal discussions of collateral managers reflect the relative importance of their role as active managers (see Bavoso 2017; Bethel, Ferrell & Hu 2008[[17]](#footnote-17)). However, academics have shown relatively little interest in CMs bar a few exceptions (Chernenko 2017, Maehlmann 2013). In fact, we know relatively little about collateral managers beyond their technical function. Riles (2011) provides an in-depth account of collateral management as a mundane legal technique but her account does not provide details of who these actors are, or, how they connect to the finance services sector more generally. In part this is because analysing these obscure actors is difficult: 1) they are relatively small and interested in retaining their privacy; 2) they have been subject to intense legal and regulatory scrutiny which may naturally limit their interest in collaborating with external researchers. Nonetheless, we can gather information from secondary sources: offering circulars. OCs give a relatively detailed description of CMs history and ownership. But more importantly, they contain detailed career biographies of employees involved in the deal. These career biographies include information on educational attainment, previous employments and specific expertise gained.

The career biographies are interesting for social scientists concerned with understanding CMs beyond the technical managing of collateral. Examining CM’s careers will not only tell us who these people are, but also how they are connected across Wall Street. CMs may have a common education provided by business schools or Ivy-league universities: a BS in Finance from Fordham University, or an MA in Mathematics of Finance from Columbia Business School. Or they may have received specialist professional training: ICMA more recently introduced two-day training courses on collateral management[[18]](#footnote-18). And whilst this affirms collateral management as integral part of today’s financial activity, this may have not been the case prior to the GFC. Given the relative importance of collateral management as a legal technique that is so fundamental to trillions of dollars-worth of securities suggests that we need to think carefully about how collateral management organises within the financial services industry.

Studying careers in the context of social network research in itself is of course not ground-breaking given Granovetter’s (1974) early work *Getting a Job* and the many subsequent papers examining “careers” (of which there are 146 in *Social Networks* alone). But in this study out treatment of careers transcends the individual. We are keen to understand the entanglement of collateral managers with the wider world of financial services and how this may influence our understanding of financial networks. Thus, rather than focusing on board interconnections, alliances or resource flows, we use the individual careers to gain a broader understanding on how knowledge, experience and personal relationships *connect* collateral management as activity and financial firms. Naturally, strong and weak ties (Granovetter 1973) are of interest to this study because they may connect different actors, for different purposes, in different ways. Even short stints at a firm could create relationships that after years of inactivity/latency may become productive. Consequently, it is not only immediate career steps that are of interest, but we may be able to identify specific sequences or stepping stones within our research that provide individuals with opportunities to (re-)connect with firms or be subject of firms seeking to connect with them.

There is a set of preliminary questions that are of interest. Where did CMs begin their career? Where they recruited by collateral management firms straight out of university or did they gain work experience in other financial intermediaries (investment banks, real estate or even credit rating agencies)? What mix of careers make up a successful collateral management firm? These questions will not be answered in this paper, but they guide our way of thinking about the data we are presented with.

**Section 3: Taking stock: collecting data from documents**

We now discuss the data contained in CDO OCs in more detail. Collecting this data can be thought of in four steps: first, the researcher needs to *gain an overview* of the document’s contents and the specific data that is to be extracted; a second step requires making decisions about the initial *formatting* of the data collection and some, at least preliminary, ideas of how to code this data; before, third, employing the coding of data in a unified format. A final, yet crucial step revolves around issues of data *anonymisation*.

***Step 1: Understanding what data is available***

At first glance the data on CM careers is relatively self-explanatory and features five basic data categories: name, university level education, previous jobs held, roles within those previous jobs, and the CM they are working for. Knowing these categories allows us to formulate a strategy to extract this data in a uniform way. But it also tells us that other facets of the individual’s life – kinship relationships, hobbies and membership in clubs – are not within the scope of this research.

In addition to the individual career data, we can extract a variety of attributes for each individual in relation to the CDOs and collateral management firm they were involved in: name and value of CDO, or the performance of the CDO during the GFC. Furthermore, we can collect on firms that were involved in the structuration of each CDO (see Leaver and Tischer forthcoming) – investment banks, law firms, trustees, etc. – which could provide suitable corroborative evidence for the formulation of a latent tie argument. For example, if Person A worked for Firm X prior to working at CM1, and, now Firm X is structuring a CDO with CM1 for which Person A now works, then we could use this data to support hypothesis around revolving doors or the reactivation of latent ties. This would be particularly intriguing in Person A used to work for an investment bank that now hires Person A’s CM.

Thinking about the available data in this way is convenient; however, it fails to capture the subtle differences in the biographic accounts disclosed for each CM and in each OC. When first examined, OCs appear relatively standardised. Indeed, as Tischer et al. (2018) argue, they are in terms of the fundamental document architecture, the technical-legal jargon they are written in and the appearance of a cut and paste job. However, the detail provided on CM careers contains a set of subtle and not so subtle differences (Table 1).



Table 1: Range of data points across CM career biographies

First, there is *variation in the number* of biographies included. Some OCs make do with less 10 biographies whilst othersprovide between 30-40 individual bios. This raises questions as to whether a) data is complete for those that provide fewer biographies, or b) if certain actors provide overly elaborate accounts and numbers of people involved in the CDO process. Given that these biographies are included to establish the credentials of a collateral manager, it may be a case of the more, the better. Hence, we need to think carefully about how we deal with different types of involvement. Luckily, the documents themselves give us some clues as to the role played by individuals: Declaration[[19]](#footnote-19), for example, offers what appear to be executive profiles, followed by *CDO investment and trading*, *fundamental research* and *quantitative research*; Aladdin and BlackRock on the other hands only list *key personnel*.

Second, data may be *complete* and detail all positions or only reference *career highlights* evidenced through gaps in the timeline. The latter posits the need for locating additional data to ensure the completeness of data. There are various options available to do this as many individuals can be located on other websites including LinkedIn, Bloomberg or company websites. A detailed discussion of triangulation and potential ethics issues are discussed in Sections 4 and 5.

Third, the *detail of information* for each individual may differ across OC documents. Detail on education may only be provided at the level of degree (BA, MSc, PhD) and institution in some cases. Others may also include the course and year of attaining the degree (MA in Sociology 1987 or PhD in Nuclear Physics 2004). Similarly, some biographies offer more or less detail on job descriptions. Some may include a sentence on responsibilities or achievements, but more formal accounts may only disclose “worked at x”. The display of longitudinal data may include length of tenure or years of tenure. Thus, at the very least, some basic mathematical aptitude is required. Moreover, the biographies provided within each CDO OCs description of CMs may not be uniform either, and they may include aggregates of experience such as “has build a long-standing career in the securities industry which spans over 24 years” (see e.g. Maxim Advisory Jupiter Highgrade).

The data baseline includes name of the employee, university name and key positions held by the individual prior to taking on work at the CM but this may not apply to junior members of staff hired straight out of university. Fundamentally, understanding the data suggests that the data collection process may not benefit from automation. The format in which data is displayed is not uniform and can be at times cryptic and require some form of translation by the researcher. The problem is summarised in Fig, 2a and 2b. Fig. 2a illustrates a relatively straight-forward one-option example of capturing the data as sufficient information is available:



Figure 2a: A simple, one-option example of capturing data

Fig. 2b provides a more complex example as information on job tenure is summarised and presents the researcher with three cases and a certain level of ambiguity in the absence of additional information. The lack of specificity on how the “six years” are spent across these two careers can ultimately alter the composition of the dataset considerably. Thus, we need to think carefully about how we can translate this raw data displayed in the OC into an analysable data format that captures maximum detail across all observations.



Figure 2b: A multi-option example of capturing career data

***Step 2: Capturing network data from secondary data***

One of the key considerations when collecting data from documents is that the researcher is not participating in the creation of the data but is presented with a ‘finite’ set of relations and attributes. Whilst we may be able to collect additional data from other sources, an initial step should focus on the original document, the offering circular, and the data it contains.

Unlike primary data collected with a specific analysis in mind and therefore a preconceived measurements of tie strength and number, financial documents, and the specific biographic data we are interested in, are not written with the academic researcher in mind. Moreover, the secondary data used is, in essence, qualitative and therefore requires the researcher to translate data into a quantitative format (Williams & Shephard 2017[[20]](#footnote-20)). Network ties are therefore (re-)constructed by the researchers and these *post-hoc* decisions about what constitutes a tie and the coding thereof are crucial (Ter Wal & Boschma 2009[[21]](#footnote-21)). Selecting a coding scheme *post-hoc* requires thoughtful engagement with the data to align the available data with the research question and the network relations (Borgatti et al 2013, Williams & Shephard 2017). This alignment is relatively straightforward in our case as career data is (easily) quantifiable using an inductive, data-driven approach (Williams & Shephard 2017). The identification of variables that relate to both network content and structure is aided by the relatively standardised format in which careers are presented and the fact that there is limited “noise”.

The data describes individuals’ careers captured by a two-mode or affiliation network where each individual has a set of relationships with particular organisations they previously worked for. The two-mode network is information-rich and can be transformed into one-mode matrices: 1) relationships between individuals that have worked for the same firm or 2) relationships between firms connected by individuals. Of course, one person on its own working for two organisations with 100s if not 1000s of employees may legitimately raise questions about the extent to which this tie matters. And even though (or because) this research is scaled up to capture knowledge flows between organisations based on 1000s of career biographies, capturing the right data on individuals’ careers is hugely important for the results of the study.

To guide not just the ‘what’ but also ‘how’ to collect, we can draw on insides from network studies, not least Granovetter’s (1973) seminal thesis on the *strength of weak ties*. This theory is of relevance to coding this data because the researcher must consider how this data can answer the research question. Given that we focus on careers and ties between organisations and those who were employed there, we may want to consider tenure, that is the length of employment as a useful indicator of the strength of tie, rather than just highlighting that a tie exists. Using valued data is in principle straightforward because tenure (in years) gives us a simple, common-sense measure. But, doing so raises the issues of strong versus weak ties (Granovetter 1973; 1984), and how these should be defined. At the level of the ego, a 15 year tenure at Firm X may be considered stronger than a 1 year term, but of course there is a danger in enforcing an ordinal variable on the relationship, but other indicators may well (co-)determine the strength of the relationship (Table 2):



Table 2: Potential indicators of tie strength

Being an administrator or a senior analyst may well alter the content and strength of a tie. Of course, we could decide on employing a different coding, for example, tenure range by treating all tenures lasting 5 years or more as strong ties with the value ‘3’, and short tenures of less than a year may carry the value ‘1’. But likewise, ties may carry a negative weighting of ‘-1’, for example, if we hypothesise that very short tenures have a negative influence on social capital (Borgatti & Everett, 2014; Labianca & Brass 2006). But making these decisions at the early stages of collecting the data reduces the information we hold for each individual career. Figures 3a and 3b illustrate the difference between networks constructed from basic and comprehensive data collected. Figure 3a only shows basic and shared connections whereas Figure 3b displays a much more complex (if not too complex) story:

* Line thickness to represent tenure
* Positive (triangle up) or negative (triangle down) follow up moves
* Firm specific characteristics (bank = orange; CM = yellow; CRA = blue; insurance = black)
* Seniority (node size) and role (node colour)

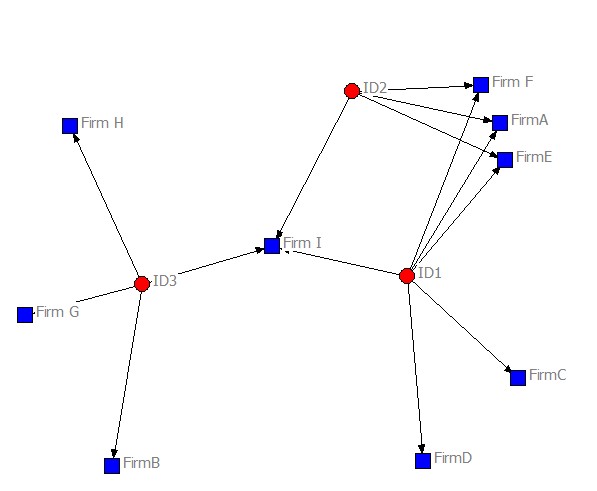
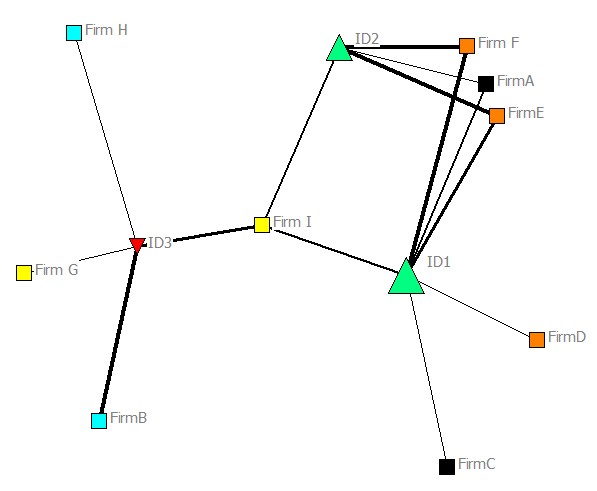
 

Figure 3a and 3b: illustrated effect on basic and comprehensive data gathering

An additional complication arises because the research is not interested in the ego-net *per se.* That is, unlike Granovetter (1974) we are not trying to understand individual’s careers, but we are interested in how numerous careers across a specific activity may influence firm and market behaviour through knowledge transfers etc. As individuals move between firms, they create connections, and multiple individuals moving between two firms create multiple ties between those firms. But it is not simply a question of how many ties exist between two firms, but the strength of ties identified at the individual level also needs to be considered. As illustrated in Figure 4, in a scenario where two firms share the same number of interconnecting individuals the properties of the individual’s ties with the firm and connections between these individuals may very well be relevant in determining the strength of ties. Still, this does not automatically preclude the existence of strong ties between organisations which are connected by unconnected individuals without strong ties to either firm. Importantly, the tie contents collected for individuals

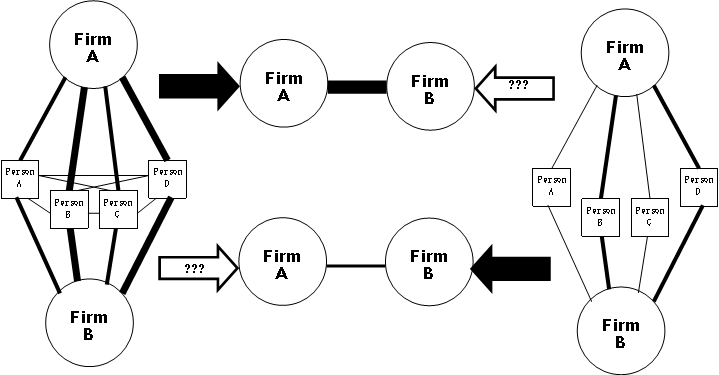


Figure 4: Strong and weak tie scenarios for multiple ties

(line strength illustrates tie strength)

Because of these issues, our attempt is to maximise the data extracted in the richest format.

Given the above discussion it appears insufficient to only collect data on tenure. Instead we should also collect information on the time period an individual was employed at a firm. View cases may give us this information directly, for example Individual A was employed at Firm X from 1994-1997. Yet for those cases where this information is not explicit, we can revert to calculating the time period. A simple example may give us the year (2001) and a qualifier (for 3 years), hence the term would span 2001-2004. A more complex story may require us to research further, because ‘Individual B may have worked for 3 years at Firm X before joining Firm Y for 2 years. In 2003 she joined CM 1’. Calculating the time period Individual B spent at Firm X is not rocket science (1998-2001); however, the unit of analysis, *year*, may also raise concerns about whether people have actually been at the firm at the same time, because one may leave before the other joins. To mitigate this, we could decide to employ a stronger measure, for example, we could only count ties where both employees have spend at least 1 year at the firm at the same time. But of course, we must not discount ties at this stage, because latent ties may very well be used for organising or strategising by both individuals and firms.

***Step 3: Hard versus soft data***

As noted in the previous stage, the data is not uniform and selecting the right level at which to extract the data is difficult. Without gaining an overview of data first, we may simply decide to include *hard facts* into our dataset, that is, name and dates of the university an employee attended and the career trajectory. If we focus exclusively on hard facts, we may miss more nuanced *soft* descriptions of what people actually do in specific jobs. It is one thing to say that a person “worked” at a credit rating agency; it is something entirely different to say that a person “developed and implemented convertible valuation models” (Vacca, Clinton, Abacus AC1). How does one code the latter in a meaningful way? Does “valuation models” provide sufficient detail and does it make sense analytically?

Of course, the former’s coding appears to be more straightforward, yet we still need to think about the pitfalls of what appears to be a relatively simple task. For example, if Individual A has worked in Firm X for 4 years prior to starting a position at Firm Y in 2002, how do we best code this? Should we collect start and end date of the employment, that is 1998 to 2002, or should we collect the tenure, four years, or both? One way of deciding on how to code the data is to collect all hard data available; however, an alternative approach maybe to code data in its information-richest format. For example, collecting the end and start day does allow us to calculate tenure, whereas recording tenure does not allows us to calculate end and start dates (unless our data is complete, which may not always be the case given the selection of key career steps in OC documents).

In addition, researchers need to think about soft descriptions of activity. Some of these descriptions may be softer or more elaborate than others. ‘*Portfolio Manager’*, ’*Research Analyst’* and even *‘ran [Firm X’s] high yield trading arm’* are relatively straightforward descriptions of an individual’s activity, although there may, in fact, be substantial underlying differences between the 120 firms. However, being a ‘*primary decision maker and risk manager in [Firm X] proprietary investing activities’* or an ‘*Assistant VP* *for CDO and Structure Product surveillance/administration’* gives substantially more interpretive scope. In most instances activities can be grouped both hierarchically and according to other principals, for example, ‘*technical’,* ‘*sales-focused’*,or ‘*administrative’*; yet, this may not always the case. The categorisation of softer data is thus a major obstacle for the social scientist.

***Step 4: Anonymisation***

People have a right to privacy and given our current political climate and importance of social networking sites, we, as researchers must ensure that our subjects cannot be identified. And new GDPR requirements have meant that regulations around storing data have been tightened and often we are required to produce a data management plan before being given the green light for our research project. Of course, we could argue that because the data is published and available to the public, that anonymity is waived and that we need not worry about it. But the data provided was provided for a different purpose; its intended use was not for social network research. And whilst we may use that data for our purposes we must make sure that people’s privacy is kept and that we do not, directly or indirectly, damage the individuals.

However, our research is not interested in the individual career, but in larger patterns of organising, i.e. what can careers tell us about a particular, obscure economic/financial agent? Yet this does not mean that we can take the individuals’ privacy lightly. Deciding on a coding scheme to anonymise individuals is both simple and difficult, because we may want to retain certain characteristics in the code that allow us to explore particular patterns. We may retain gender, age, education or role within the code, but generalise it to such a level as to make it difficult to identify any specific individual.

The fact that this network is two-mode adds complexity in dealing with the firms these individuals worked for prior to joining the collateral manager. We could choose to use the firms’ names in the dataset, but doing so may inadvertently make individuals discoverable despite being anonymised. We could also decide to code firms according to their functional expertise: Goldman Sachs could be *Investment Bank A*, Maples and Calder may be *Law Firm 1*, etc. Would calling Goldman Sachs Investment Bank 1 have any implications for the research findings? Journalists may say yes, but as academics we are not interested in exposing specific actors; we are interested in answering larger questions of structure and organisation. Indeed, the decision to code in accordance with some common characteristic may simplify the analysis in a positive way. For example, educational attainment may be more relevantly discussed according to university status – Ivy League, public Ivy, community college – because it may tell us something about the socio-economic status of individuals. An additional benefit of anonymising firms may lie in how readers perceive the research, because it would avert claims of “naming and shaming” of particular actors.

**Section 4: Triangulation and the collection of additional data**

Having collected data from offering circulars it is important to check both for completeness of data and the data’s accuracy. As noted earlier on, some OCs only detail career highlights for collateral managers, in particular for more senior employees that may have move jobs many times. There may also be other reasons for omitting data, for example, they may link individuals with companies that may raise questions by investors. And whilst being vetted by lawyers, these career biographies are self-reported by the individuals and collateral management firms, thus may be subject to human error. But whatever the reasons, ensuring data is as complete as can be and checking for accuracy are basic principles any researcher should abide to.

Locating alternative sources of data depends on the topic of the research, but we can imagine many potential sources – databases, archives, (quality, non fake) news outlets, company documents etc. – to be suitable. Given that this research project focuses on careers of individuals, the primary source of additional data is relatively easily located: LinkedIn, a globally active professional network with over 550 million users. Crucially, LinkedIn user profiles contain similar data points – education, employment history, location – and thus are most suitably sources for comparison.

One key reason why LinkedIn works well in this case is that we have a set of pre-identified individuals we are seeking to locate additional data for. In the absence of having already collected data for these individuals, LinkedIn may be less suitable, because, for example, searching LinkedIn for “collateral management” yields over 750,000 results, and many of these are irrelevant to the structuration of CDOs and of more general nature. Searching for “collateral management CDO” will yield close to 3000 results, but we may not easily discern if these individuals have, in fact, worked for CDO collateral managers. Additionally, we are interested in individuals that partook in structuring CDO prior to the GFC in 2008, hence people joining CM post 2008 would have to be excluded. Searching for CM firms directly equally yields too many or irrelevant individuals. Searching for “CVC Credit Partners” for example generates 157 results, and the CVC LinkedIn page refers to 89 current employees.

Triangulating data using LinkedIn is relatively straightforward in procedural terms as we have key information for each individual (name, employer and education). Using these parameters we can locate individuals and collect data for each individual manually[[22]](#footnote-22). Whilst most LinkedIn profiles corroborate the biographies provided in OCs, it is useful to discuss examples where information differs. The degree awarded by universities represents a data point that may provide different and/or more detailed descriptions. For example, a degree may be clarified as to the thematic studied or a Sociology degree may be specified as “Political Science & Government”; yet likewise, the specific nature of a degree may be reduced, for example, a PhD in “Control and Dynamic Systems” may become a PhD in “Mathematics”. Again, when faced with the choice of two slightly different descriptions, we would recommend collecting as much detail as possible, as we can re-categorise them later, if needed.

LinkedIn CM example.emf

But of course, not every individual has a LinkedIn profile. Where individuals cannot be located on LinkedIn, it may make sense to attempt to locate additional data using other services, such as *Bloomberg Executive* or via corporate websites. This works relatively well with more senior employees, but, we should keep in mind that those in junior roles ten years ago, will have risen through the ranks and may very well feature on these services too. The process of collecting this data is similar as for the discussion above; however, it is worth noting that oftentimes additional information is provided for the most senior individuals, in particular positions within philantrophic or charitable organisations, i.e. board memberships, founding or funding thereof. Because this is not relevant in explaining financial careers in collateral management, and there are comparably few individuals for which such data exists, we have excluded these positions within our dataset.

In our case triangulation is more than *cross-examination* of data. In fact, we used triangulation to fill in gaps in the biographies provided in the OCs. Doing so provides a more detailed and therefore accurate view of individual CM careers.

**Section 5: Ethical considerations**

Re-use of public data

“But the data is already public.” The title of Zimmer’s (2010) paper reviewing a scandalous study involving students’ *Facebook* profiles, resonates strongly with concerns about whether social media data should or should not be used by researchers. Since then, the *Cambridge Analytica* scandal has intensified public discussions about what is permissible and what is not.

Debates about the public *vis-a-vis* private nature of this data are complex and all but settled. It is our opinion that these concerns must be considered on a one-by-one basis, rather than finding a general solution. Whilst we may use social media as a broad category, it contains a large variety of services used for different purposes and that contains different data (see Bos et al 2009, Zimmer and Kinder-Kurlanda 2017[[23]](#footnote-23)). Thus while the use of Facebook data causes ample controversy (Zimmer 2016), it seems more permissible to analyse *Twitter* data, even when it looks at grieving individuals (Glasgow et al 2016). And this makes sense given the nature of the data that is provided by these profiles and its use: Twitter amplifies free speech and personal opinion so that it can be heard globally, whilst Facebook profiles document the personal social life of an individual within a self-built social circle, i.e. his or her friends. Facebook is specific in the sense that it allows users to keep their profile private to safeguard their data and interactions, whereas Twitter is build on the premise of sharing opinions.

LinkedIn provides an alternative case yet again. It does not target naive individuals but, we may argue, more sophisticated *professionals[[24]](#footnote-24)* and this is reflected in the service it offers: hire, market, sell and learn[[25]](#footnote-25). Whilst professionals may have ultimately different reasons for joining LinkedIn, the underlying principle is one of *discoverability* and *opportunity*. By discoverability we refer to the intention of professionals to become visible beyond their traditional network in real life. They connect to specific companies, follow influencers and ask other users to promote skills or provide testimony of their achievements. And the contents of their personal pages are carefully crafted and distinctly work-related in character. By opportunity we consider the purpose for why individuals join the page: to promote themselves to others; or as LinkedIn puts it, to “[c]reate economic opportunity for every member of the global workforce”*10*. People can and do advertise themselves to potential future employers or recruitment agents, and contact firms directly. In this view privacy is arguably of lesser concern to the users, not least because the context is professional, rather than social. In fact, professionalism triumphs over the social to the extent that the social is rendered instrumental: exchanging weekend frolics turns into updating each other on the final stage of implementing Project X; the carefully crafted “best night ever” post with friends is replaced by an equally heavily staged photo of some team work, including the flip chart full of colourful arrows, boxes and words; and the beach-body-perfect pose with cocktail and palm trees becomes the generic two-piece suit image of corporate success. Any socialising occurs through direct communication outside of the public’s eye through private messages. Thus any social networking remains private and only public information is used by the research. Of course these considerations do not provide us with a *carte blanche* to use their data for whatever purpose but it highlights that LinkedIn profiles are distinct to Facebook users. They are carefully crafted for a specific purpose and less social in character.

In our particular case, we use LinkedIn profiles alongside already existing biographies that are publically available. Thus the primary use of LinkedIn is not the discovery of data, but it is used to ensure the accuracy and completeness of existing data. We argue that even where additional data is collected, it is ethical to do so as long as they correspond with already existing categories. What we mean by this is that we may collect additional details of education or employment as these are pre-existing categories, but we accept that it would be more problematic to collect additional data – for example, hobbies, relationship status or sexual orientation – that bares no relation to the initial research question.

Consent is an additional issue that requires careful engagement with the purpose of the research and the private/public nature of the data. As argued by Wilson et al. (2012) it is insufficient to claim that all data is public to avoid consent. However, equally we must consider the difficulty of eliciting consent from all participants, a problem recognised given the distance between research subject and researcher (Zimmer & Kinder-Kurlanda 2017; Buchanan & Zimmer 2018[[26]](#footnote-26)). The question here is really whether or not we require consent at all. This may sound counter-intuitive; however, new GDPR guidelines stipulate that it is legitimate to process data without consent if it is in the public interest (see ICO 2018[[27]](#footnote-27)).

**Conclusion**

With much of the debates around data collection focusing on going out there in the field and collecting the data from scratch, this paper provides a different, but no less valuable account of social network data collection. Using obscure financial documents, such as contracts and memoranda as sources of data comes, of course, with its own problems; but it does allow us to explore relational patterns between actors which may otherwise be impossible to be observed for a lack of resources and/or access.

But ‘collecting’ network data from these sources is not as straightforward an exercise as may be assumed. The first question relates to the “quality” of the source. The data we use is limited in breadth, but relatively com

Limited but completeness

Accuracy

Translation

First, it requires translation as data is presented in a non-social network format and it contains cer

Measuring the quality of this data is in itself a relativistic exercise. Some might say that this data is controversial because it was not collected directly by us. B

Reusing existing data limits what we can do with the data and as such we may perceive of it as less appropriate than

Data ethics

Cut offs:

Even though the origin of the word from the Latin *firma* refers to solid and stable structures, in reality, firms merge and acquire; they divest and close down parts. Likewise workers may leave individually or collectively to join a competitor or to set up a new firm. It is important to not only know that Individuals A and B were working for Firm X, it is relevant to know if they did so at the same time. If they worked at Firm X at the same time, and subsequently moved to Firm Y, this could indicate a) a team move or b) a merger.

If they set up their own business, and subsequently get contracted by Firm X, then we may think of this as an extension of the firm boundaries, or *doing a favour* in either direction. However, we must not discount the possibility of a tie being performative if B started working for Firm X after A has left. Being an alumni from the same firm, for example, having worked for Goldman Sachs, may still produce a productive relationship between A and B, although it being a hidden one in our dataset. After all, boys clubs are well and alive and finance in particular has a reputation for networking, not least because, popular books, such as Michael Lewis’s *The Big Short* and *Flash Boys* (2010 & 2014), vividly illustrate how relations matter to get the deal done.

Lastly, we may raise the issue of relationships between firms and (former) employees being regarded as latent ties. Firms have been shown to not only establish new ties to solve problems, but to reactive latent ties (Mariotti & Delbridge, 2012 org science; Grabher 2004 Org Studies). Such a strategic view is also visible at the individual level, as ties formed may be reactivated many years later to access resources (Kwon & Adler, 2014 AMR).

Latent ties’ association with social capital is an intriguing one, because social capital may be used to activate latent ties for strategic purposes. Individuals are not only hired for the technical knowledge they hold, but also for the social networks they have. Likewise, banks and other firms may benefit from identifying previous employees that now work for collateral managers to gain access to these firms. Latent ties may also be considered vehicles for spreading new ideas, or recruit actors to participate in new activities or specific projects (Grabher 2004; Starkey et al. 2000 org science). Starkey et al.’s idea of latent organisations shows how latent ties may be preferred to constantly coupling and decoupling ties to take advantage of different knowledges held by respective parties (2000: 300). The collective expertise of collateral managers reflects such a resource mix – technicians, analysts, lawyers – that allows them to interact with associated parties through latent ties, and *vice versa*.

1. Analyzing Social Networks via the Internet; Scraping the Social? Issues in live social research [↑](#footnote-ref-1)
2. A Core/Periphery Perspective on Individual Creative Performance: Social Networks and Cinematic Achievements in the Hollywood Film Industry [↑](#footnote-ref-2)
3. Music as Collective Invention: A Social Network Analysis of Composers [↑](#footnote-ref-3)
4. Introduction to Social Network Analysis (SNA) as an investigative tool [↑](#footnote-ref-4)
5. http://eprints.ncrm.ac.uk/842/ [↑](#footnote-ref-5)
6. Archival Research Methods [↑](#footnote-ref-6)
7. The use of archival records in case study research [↑](#footnote-ref-7)
8. Power through institutional work: Acquiring academic authority in the 1968 third world strike [↑](#footnote-ref-8)
9. Coffey A () Analysing Documents in Flick Sage Handbook of Qualitative Data Analysis [↑](#footnote-ref-9)
10. <https://www.sec.gov/divisions/marketreg/security-based-swap-data-repositories.htm> in line with Title VII of the Dodd-Frank Act [↑](#footnote-ref-10)
11. A Matter of Record: Documentary Sources in Social Research [↑](#footnote-ref-11)
12. See, for example, accounts of emails written between bankers prior to the [Great Financial Crisis](http://online.wsj.com/public/resources/documents/WSJ_GSQUOTES.pdf) or the [Libor scandal](https://www.bbc.co.uk/news/business-21358362) [↑](#footnote-ref-12)
13. The Handbook of Science and Technology Studies [↑](#footnote-ref-13)
14. <http://www.hblr.org/wp-content/uploads/2014/09/Derivatives-and-Credit-Crisis.pdf> Derivatives And The Legal Origin Of The 2008 Credit Crisis [↑](#footnote-ref-14)
15. https://www.sec.gov/spotlight/enf-actions-fc.shtml#keyStatistics [↑](#footnote-ref-15)
16. <https://www.sec.gov/litigation/admin/2013/33-9438.pdf> <https://www.hsgac.senate.gov/imo/media/doc/PSI%20REPORT%20-%20Wall%20Street%20&%20the%20Financial%20Crisis-Anatomy%20of%20a%20Financial%20Collapse%20(FINAL%205-10-11).pdf> [↑](#footnote-ref-16)
17. legal and economic issues in litigation arising from the 2007-2008 credit crisis [↑](#footnote-ref-17)
18. https://www.icmagroup.org/executive-education/courses/CollateralManagement/ [↑](#footnote-ref-18)
19. Independence CDO VI [↑](#footnote-ref-19)
20. Mixed Method Social Network Analysis: Combining Inductive Concept Development, Content Analysis, and Secondary Data for Quantitative Analysis [↑](#footnote-ref-20)
21. Applying social network analysis in economic geography: framing some key analytic issues [↑](#footnote-ref-21)
22. It is important to note, that LinkedIn does not allow the use of software to scrape or copy data from its website - see <https://www.linkedin.com/legal/user-agreement#dos> for further information [↑](#footnote-ref-22)
23. <https://www.cc.gatech.edu/pixi/pubs/122-bos.pdf> <http://www.christinehowes.com/papers/Zimmer&Kinder_book_CC_license.pdf> [↑](#footnote-ref-23)
24. https://about.linkedin.com/ [↑](#footnote-ref-24)
25. https://business.linkedin.com/# [↑](#footnote-ref-25)
26. Buchanan, Elizabeth A. and Zimmer, Michael, "Internet Research Ethics", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/fall2018/entries/ethics-internet-research/>. [↑](#footnote-ref-26)
27. https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/ [↑](#footnote-ref-27)