Sensitizing social data science

Title of the Paper:

Sensitizing social data science: combining empirical social research with computational approaches to the analysis of career data

Authors:
Gian Marco Campagnolo[1¶], Robin Williams[1], Beatrice Alex[2], Alberto Acerbi[3] & Duncan Chapple[4]

Corresponding Author:
Gian Marco Campagnolo
Lecturer & Alan Turing Inaugural Fellow
Science, Technology and Innovation Studies
School of Social & Political Science
The University of Edinburgh
High School Yards
Edinburgh EH1 1LZ
email: g.campagnolo@ed.ac.uk

Running title:
Sensitizing social data science

[1] Science Technology & Innovation Studies, University of Edinburgh
[¶] Alan Turing Institute, London
[2] School of Informatics, University of Edinburgh
[3] Eindhoven University of Technology
[4] Edinburgh University Business School

Abstract

How can social science contribute to data science? The paper responds by presenting a new model for disciplinary integration based on conceiving subjects and objects of the research as well as methods as co-participants in shaping research design and questions. Identified with the notion of participative epistemology, this model is exemplified through the presentation of an on-going social data science project combining ethnographic, computational and inferential approaches to analyse career data of thousands IT industry analysts. By focusing on the project's research design and pilot application of a descriptive approach to test the framework in preparation to the full-scale analysis, we contribute to digital sociology debate by showing how social science - and the sociology of scientific knowledge in particular - can sensitise data science from within, thus surpassing divisive attempts as well as scholarship that conceives social science as offering an 'external' contribution to data science.

Keywords: social data science, participative epistemology, sensitizing concepts, sequential analysis, career studies

1. Introduction

In this paper we discuss how the notion of participative epistemology can contribute to sensitizing social data science. We will do so by offering the case of a project where ethnographic, computational and statistical methods have been used in combination. By exemplifying it through the analysis of the project's research design and early pilot results, we suggest a new view on the collaboration between data science and social science.

Several scholars have indicated that the next frontier of empirical sociology inquiry is about unpacking latent regularities in observed patterns of association of social elements across contexts (Latour, 2010; Leonardi and Barley, 2008; Sandber and Tsoukas, 2009). However, the interest so far has concentrated on how computational approaches can complement dominant ideographic modes of enquiry (Gaskin, Berente, Lyytinen & Yoo, 2014). Received wisdom is that qualitative approaches can help identify case-specific idiosyncrasies (Leonardi, 2011; Pentland and Feldman, 2008), while computational methods offer strengths when used to detect larger scale patterns. In this paper, we discuss a new model for collaboration across the social and computational sciences. We argue that locating social research at the 'micro'/'local' end of the science spectrum - while leaving the task to detect larger scale patterns to computational approaches - is a misleading characterization (Bittner, 1965; Coulter, 1996). It is misleading not only because social research has always been about social order (Garfinkel, 1988). This characterization also contributes to create unhelpful and artificial disciplinary boundaries.

Divisive attempts, unfortunately, are not limited to computational scientists claiming expertise over social science. Accounts that present ethnographic and computational approaches as mutually exclusive are also found in the work of ethnographically-trained scholars developing criticisms of data driven research (Boyd & Crawford, 2012; Stoller, 2013) or claiming uniqueness in inductions derived from ethnographic methods (Wang, 2013).

Rather that dividing disciplines based on the scale of their object of study, our approach will be to ask how methodological tenets deriving from empirical social research can contribute 'sensitizing concepts' (Blumer, 1954) to computational and inferential approaches. Sensitizing concepts are constructs that are derived from the research participants' perspective, using their language or expressions, and that sensitize the researcher as to 'where' to look as opposed to 'what' to look for. In doing so, we share Blok & Pedersen (2014) argument and ask: how do the *kinds* of knowledge that can be obtained from computational and ethnographic methods, respectively, relate?

Based on lessons learnt from a project that combines insights from the sociology of profession and the sociology of expertise involving the application of text mining techniques and sequential analysis, we respond by exemplifying ways in which empirical social research can help sensitize social data science. By referring to how it became apparent in our project, we will propose a participative epistemology for social data science. The notion will be articulated in three points: (1) a view on how research subjectivities can shape research design; (2) an attention for the research object as co-participant in the research and (3) a sensitivity for the object-relation regimes as having an agency in shaping research outcomes.

The paper is structured as follows. In the following section, we discuss existing attempts to articulate the relationship between social sciences and data science, finding that the debate is dominated by a preoccupation for disruptions brought to empirical sociology by computational social science. Next we will discuss theoretical paradigms which focus on models of interdisciplinary integration and how these can be interpreted to guide social science intervention in data science. We will do so by presenting the research design and early pilot results of a project driven by this epistemological paradigm. The project's selection of data sources and methodology will be presented in detail, together with aspects of sampling and data cleaning as well as how data visualisation has been used to inspire the

inferential approach. In conclusion, we discuss the project's research design and early development as a case study of participative epistemology in social data science and call for more empirical examples that substantiate new models of disciplinary collaboration.

2. Literature review

In this section, we will take into account some of the most advanced accomplishments in the debate concerning the relationship between social science and data science, highlighting possible areas for further development. We divide the review into two streams. One addresses digital sociology scholarship aiming at developing a sociological understanding of social data science. The other comprises work asking what social data science (and social media analysis in particular) can add to traditional social science research methods and strategies.

In her recent book, Noortjie Marres (Marres, 2017) helpfully summarises the contribution of digital sociology scholarship aiming at developing a sociological understanding of social data science. Marres distinguishes three ways to discuss the social aspects of digital research: the techno-centric, the data-centric and the practice-centric (Marres, 2017: 58). In the first group are scholars claiming that social science should approach computational research by revealing the non-neutral aspect of digital platforms such as social media (Coleman, 2012; Gillespie, 2010; van Dijck, 2013). In the second group, we have those who challenge the 'naturalness' of the data generated by digital platforms. Internet-generated data might give near real-time access to new social groupings (Housley et al., 2014). However, platform data are low-fidelity in that they are said to be strongly marked by platform effects (Shaw, 2015) and exhibit non-representative samples (Duggan et al., 2015). Finally, we have social scientists whose preoccupations address the way digital technologies are used (Suchman, 1997; Slater, 2002). Advocates of the practice-centred approach claim that features of digital technologies cannot be identified without considering how people use them.

The second stream comprises work asking what social data science (and social media analysis in particular) can add to traditional social science. This line of inquiry is exemplified by research by Edwards and colleagues (Edwards et al., 2013). Here the authors plot the distinctiveness of social media analysis in relation to more traditional research according to two dimensions: research design and research strategy. Social media analysis is said to offer an extensive research strategy, as opposed to qualitative methods (such as ethnography) that represent a more intensive strategy. The possibility to capture data in real time at the level of population and on an ongoing basis makes social media more processual than survey-based or experimental approaches (Sayer, 1992). The authors then move on to ask what social media analysis can do for social research in terms of either replacing, complementing or re-orienting what Mills defined as 'sociological imagination' (Mills, 1959).

Capturing extant literature at the forefront of digital research, these accomplishments provide context to frame the contribution of this paper. Work in digital sociology "stream one" shows that sociology scholars have achieved what Collins and Evans (2002) in their influential contribution to the sociology of scientific knowledge would characterise as 'interactional expertise' in data science. Work within this stream demonstrates social researchers' expertise in interacting interestingly with computational social science participants and carrying out a sociological analysis of data science. Moving from methods of observation to means of direct participation, with our paper we want to bring the debate forward and ask: can social scientists gain a cultural foothold in the area of data science they want to analyse? Continuing with our parallel with studies of expertise and referring to the notion of 'contributory expertise' (Collins and Evans, 2002: 254), we can rephrase our question as follows: can social scientists directly contribute as well as analyse data science by proposing alternative solutions as well as identifying deficiencies? While addressing this question, we will also provide a new empirical case to advance scholarship on what social data science (and social media analysis in particular) can add to traditional social science

research methods and strategies. However, rather than asking what social media analysis can add to social research, our contribution to digital sociology "stream two" will be to address the symmetrically opposite question and ask: how traditional social research can sensitize computational social science?

3. Participative Epistemology

To explore these questions, we need concepts to frame the epistemic trouble (Marres and Weltevrede, 2015) generated by the encounter of social and data science. Before moving to discuss the case study, we briefly present here concepts that guided our project's research strategy and method.

The case discussed in detail in the following section comprises a project using text mining of professional networking data to understand professionalisation in the IT sector. This research has been influenced by notions offering a nuanced understanding of the encounter of social and data science such as consilience (Wilson, 1998), abduction (as in Kitchin, 2014) and symphonic social science (Halford & Savage, 2017). Whilst notions in the sociology of scientific knowledge have led us to frame the project as a case study of participative epistemology (Heron & Reason, 1997).

Consilience – i.e. convergence of evidence from multiple, independent, and unrelated sources, leading to strong conclusions - is gradually gaining currency in the field of social data science after being proposed by George, Haas & Pentland (2014) in their editorial article of Academy of Management Journal on Big Data and Management. Quoting from Wilson they say: "The Consilience of Inductions takes place when an Induction, obtained from one class of facts, coincides with an Induction, obtained from another different class. This Consilience is a test of the truth of the Theory in which it occurs." (Wilson, 1998).

This call for multiple inductions is complemented by Kitchin's claim for abductive reasoning in big data analytics (Kitchin, 2014). Beyond inductive or deductive approaches, abductive reasoning (Bateson, 2002 [1979]) focuses on the unfolding interplay between data, method and theory. In this approach, if theory and methods are initially framing analysis, data are used to feed-back into the initial framework, making the interpretation process iterative.

The symphonic social science approach put forward by Halford and Savage (Halford & Savage, 2017) summarises these ideas in a claim for overcoming skepticism about big data analytics, thus liberating sociology from its defensive position. The approach suggests an end-to-end model of disciplinary integration, with a particular focus on data visualisation as a means for data to resurface at different points in the research process (Healey & Moody, 2014).

In this paper, we expand the repertoire of concepts in digital sociology to see how notions deriving from the sociology of scientific knowledge can positively guide engagement of social scientists in data science. Through her work on epistemic cultures (Knorr Cetina, 1999), Karin Knorr Cetina suggests that topical topics to investigate scientific knowledge production are (1) the epistemic subjects i.e. the authors of scientific research: the human collaboration as well as the digital platform producing data and how agency can rotate between them; (2) the objects of scientific research, and the reconfigurations through which the object gradually becomes amenable to the analysis; (3) the object-relation regime, which in our case means coming to terms with data science's cultural habit to suspends the quest for timeless qualities (such as causation) in favour of the quick identification of time-bound occurrences, as famously captured by Anderson's claim for the 'End of Theory' (Anderson, 2008).

In order to orient these epistemological categories towards informing a sociological intervention in data science, we introduce the term participative epistemology. Born within

an empiricist conception of knowledge production, participative epistemology (Heron and Reason, 1997) is a notion that inspires one particular articulation of action research, whereby all those involved in the research endeavor (i.e. the epistemic subjects, the objects of scientific research and the object relation-regime in our case) are both co-researchers, whose agency contributes to generating ideas, designing and managing the project, and drawing conclusions; and also co-subjects, participating in the activity that is being researched.

4. From qualitative fieldwork data to digital datasets

As a way to illustrate our claim for a participative epistemology in inter-disciplinary interventions, we present here in greater detail research design and early results of an ongoing social data science project using professional networking data to understand the careers of a particular class of IT influencers: IT industry analysts. IT industry analysts are professionals who provide strategic research for enterprise technology buyers. We discuss this project to show how participative epistemology has informed the research design and the iterations between data and theory as well as collaboration within the research team. Following the framework introduced in the previous section, we start by presenting the 'epistemic subjects': the authors of scientific research and their co-participation in framing the research.

Contributing to develop the first extended academic study of the industry analyst profession (anonymized reference), two of the authors came to the project after having completed over thirty interviews, undertaken over 100 hours of observation at conferences, listened and participated in more than 20 webinars and after having engaged in dozens of informal discussions with IT industry analysts. The aim was to see how the collective study of

analysts' work experience profiles crawled from the web could generate additional insights to one of the fundamental questions deriving from ethnographic study: how do people become industry analysts? Given the absence of institutionalised professionalization systems - i.e. regulated career progression - could large volumes of internet-generated data concerning work experience help understand how analysts gain the credibility that makes their expertise so highly reputed? We knew from Abbott's study of professions (1998) that reputation is linked to occupationally rigid professions, with early career entry and longer time spent in education. However, our extended empirical study (anonymized reference) pointed out that industry analysts in their career go through 'distinctive transitions'. Hypotheses of distinctive transitions the team came up with are: *(i)* people come to be industry analyst after heavy industrial experience in IT field; *(ii)* people become industry analysts after a career in market research and technical publishing (anonymized reference). How could we test this conjecture emerging from qualitative fieldwork by using a larger internet-generated digital dataset?

What we knew from fieldwork was also that, together with few incumbent firms, this is an industry with a high turnover of new entrants, including many smaller firms such as single-person analyst houses. While field access e.g. through participation in large IT Conferences (anonymised reference) allowed us to gather qualitative data mainly about the role of large companies (which were often the organisers or the main speakers at these events), less developed was our knowledge of the role of smaller analyst houses. One conjecture emerging from conversations with our informants was that practitioners consider establishing a one-person 'star-analyst' company as the pinnacle of analysts' career. It offers comparable earnings and more autonomy than a job in the biggest companies. Our study of internet-generated data about analyst careers would also help complement our ethnographic findings by extending our enquiry to these less accessible informants. To test the conjecture about the role of smaller houses we would look for analyst career pathways.

Once having joined top companies, do analysts tend to remain and exploit the reach offered by their position or do they prefer to trade their expertise independently, as our conjecture suggests?

To answer these questions, our research team entered a collaboration with members of the Language Technology group from the School of Informatics of the University of Edinburgh, who had previously developed a software (TeXt mining Verticals) capable of processing large amount of textual CV data crawled off the web (Alex et al., 2008). At the time, the idea was indeed to parse IT analysts CVs that could be found on the web to yield information on job sequences and throw light upon the development of industry analysts careers.

However, through contacts garnered from fieldwork, the lead author started a collaboration which granted the research team access to the largest known industrial database of global industry analysts (IAs), maintained by the US based Analyst Relations firm ARInsight. Analyst Relations such as ARInsight are professionals who advise IT vendors and clients on how to interact with Industry Analysts. This particular firm does this by offering their clients a database (called ARchitect) that records the expertise and 'power' of all known industry analysts, including analysts in transit (e.g. retired analysts).

Although realising the unique opportunity of using data from an industrial database, we were confronted with the need to understand how this was structured. It was at that point that we asked [name of co-author, anonymized], a PhD student at Edinburgh University as well as an active IT influencer, to join the team. From the mass of information available in this sort of customer relationship management (CRM) database including all IT analysts in business, we wanted to understand what data could be used to tell where analysts came from and how their provenance correlated with career outcome. ARchitect database helped in this respect as it sorts analysts based on their 'power' - i.e. how many times clients interact with them - and coverage, i.e. what industry they analyse. We thought we could use

information from the database's ranking of 'Power 100' analysts and tell, based on a study of their career data, where the most successful analysts came from as well as how their careers pan out within the industry. However, our co-author and domain expert revealed that ARchitect defines "power analysts" based on the number of times a user of the database queries an analyst profile. By defining "power" on this basis, the database cuts off all analysts that, although successful, do not interact with clients. For example, Peter Sondergaard, the most prestigious analyst of the largest analyst firm in the industry[5], is not in the ARchitect list of 'power 100' analysts for the simple reason that he does operate client-business. The early realisation of this bias was a fortunate circumstance, as it allowed us to unpack the logics black-boxed in the algorithm. From a study of CVs crawled from the web, our project would become a study of IT analyst careers as made apparent on a professional CRM database and associated external links (e.g. analysts' LinkedIn and Twitter profiles). While our ethnographic fieldwork and expertise in the study of professions as well as the participation of a domain expert in the research team would make us aware of how the community of IT analysts uses the database.

5. Social media data vs professional networking data

In this second empirical section, our narrative regarding the project's research design continues by focusing on how our research object (i.e. 'careers') and its agency (i.e. the tendency of careers to respond to mimetic pressures) influenced selection of our data source and, in turn, of our methodology.

In research addressing detection of IT influencers (of which IT analysts are an important sub-category), the preferred source of Internet generated data is the social media platform

[5] Gartner is the largest Industry Analyst firm. Founded in 1979 by Gideon Gartner, employs more than 4,000 offices in over 80 countries and has a market value of $ 2.5 bn.

Twitter. Relying on Twitter data and computational network analysis techniques, social media analytics scholars endeavour to tell who the most influential actors are, based on their position in a network. Some argue that influencers are the most highly connected or the most central people in a network (Pastor Satorras and Vespignani, 2001; Friedkin, 1991). Others find that the most efficient are those located at the core of the network, with distance between influencers being another factor (Kitsak et al., 2010).

However, in the case of industry analysts, a number of factors made us skeptical about the possibility to use Twitter data. First, there were factors considering the nature of how cognitive authority is established in the field of IT industry analysis. Not all types of experts have the same legitimators. There are types of experts – hidden experts (Habermars, [1985], 1987), who exercise power in terms of a distinctive culture that is neither understood nor accountable by the public. At first glance, as sellers of advice, industry analysts (our target group) might appear to belong to the type of experts whose personal expertise is tested by individuals (i.e. the buyers of their expertise). In such a case, more (Twitter) followers might legitimately mean more cognitive authority. However, industry analysts as well as the more general type of IT influencer discussed in the literature are different from the prototype of private expertise (Turner, 2001: 140). IT Influencers' expertise is legitimated by particular types of buyers: government or corporate bureaucrats with discretionary power (Chief Information Officer – CIO, Chief Marketing Officers - CMO), who are experts in their own right and whose views are accepted as authoritative by Executives and their Boards. Turner defines as sectarian (Turner, 2001: 138) those experts whose validating audience does not correspond with the general public. As such, their influence is hardly measurable from simply surveying any overt measure of association (e.g. a vote, a like, number of followers, a re-tweet).

Furthermore, Twitter data are affected by platform dependent factors as well as factors

related to community dynamics that are difficult to control. Digital sociologists remind us that digital data stand out for its artificiality (Marres, 2017: 52; Halavais, 2013; Passman and Gerlitz, 2014). As Shaw (2015) points out, social media data such as Twitter data are strongly marked by platform effects, such as search terms suggested by auto-complete functions or users taking up hashtags that are trending on the platform in question. Alongside platform effects, are factors related to community structure. Despite attempts to isolate a generic topology (Kitsak et al., 2010) each network has its own confounding factors such as homophily (McPherson, Smith-Lovin and Cook, 2001), geographical proximity (Hedstrom, 1994) and interpersonal affect (McAdam and Paulsen, 1993). These factors contribute to make influencer detection based on network effects more problematic. One important approach in our research has been to use sociological insights to identify deficiencies in data science approaches as well as suggesting alternative ways forward. Therefore, together with being able to identify problems in available data sources and methods, we wanted to investigate different solutions that would be compatible with our social sensitivity. We thus considered work experience data from analysts' LinkedIn profiles as apparent in the ARchitect database as a candidate data source for our research.

By making this choice we were not denying that LinkedIn data might have its own "artificialities". For example, the presence of cosmetics effects in how people present their experience through a publicly available and profit-oriented platform (Va Dijck, 2013). However, we know much more about conformity in how people present their career than what we know about how influence spreads in a community. According to a dictum regularly quoted by sociologists of professions (Gunz, Mayrhofer & Tolbert, 2011), careers are 'the moving perspective in which persons orient themselves with reference to the social order' (Hughes, 1937: 413). As social order depends upon confirmation of anticipated behaviors considered as appropriate, people will link their jobs in a sequence to conform with socially acceptable representations of "career". To explain this phenomenon, the

sociologist Neil Fligstein uses the concept of 'organizational field'. According to Fligstein, a group of social actors within which there is a high degree of interaction or comparison constitutes an organisational field. In his study of corporate control (Fligstein, 1990), Fligstein argues that by being part of the same 'organisational field', top managers copy each others' strategy, making these conceptions become increasingly prevalent and institutionalised. Professionalisation, i.e. the collective struggle of members of an occupation to control the "production of producers" (Larson, 1977: 49-52) is indeed referred to by Di Maggio and Powell (1983) as one of the central sources of what they call isomorphism, i.e. the result of processes that makes organisations more similar. "Professions" they say "are subject to the same coercive and mimetic pressures as are organisations" (Di Maggio and Powell, 1983: 152). Given our interest for capturing patterns in careers, tendencies to conformity embedded in the way the platform is designed and used would help prepare the data to respond our questions.

6. Social network analysis versus sequential analysis

These reflections on careers as our object of research had implications for our methodology selection (i.e. what Knorr Cetina refers to as object-relation regime). Aspects of agency of the research object will resurface at multiple points in our project, especially when experience profile will turn into job sequences. In this section, we discuss early evidence of the interaction between our object of research and methodology selection.

Social network analysis is the dominant method in the vast literature in social media analytics (Jamali & Rangwala, 2009; Szabo & Huberman, 2010; Suh et al., 2010) and marketing (Aral & Walker, 2011) on influencer identification. This to an extent it becomes a default choice. Indeed, as remarked by Marres (2017: 187) in what she calls a *laissez-faire*

attitude, work in this field displays a tendency to go with whatever ontology or methodology happens to be hard-wired in the apparatus that generates the data (e.g. co-follower networks for Twitter generated data). The problem with network analysis was, for our research, that these approaches do not tell how actors come about to occupy that particular position in a network. In other words, these approaches fail to respond to the question: what do influencers' careers look like? We were aware of other approaches combining social network analysis with ways to capture the evolution of networks over time (Vendres & Stark, 2010; Gaskin & al., 2014). However, adopting these approaches would mean combining Twitter and LinkedIn data[6], an option that we discarded for reasons explained in above section when we discussed community effects. Another popular methodology in career studies is sequential analysis (Abbott, 1995). Presented by the author as a solution to reconciling the theoretical sequentiality of careers with the 'unrelentingly non-sequential character of sociologists' preferred methodologies' (Abbott, 1995: 95), sequential analysis has a long tradition in career studies (for an historical review see Rosenfeld, 1992). The method would allow us to work on just a sub-set of data from professional networking experience profiles (i.e. job titles, company names and data ranges). However, textual data from experience profiles had to be turned into job sequences. In the following section, we discuss how we addressed the text-mining component of our project by following the transformation of data from text to horizontal bars representing sequences of color-coded states.

7. Starting small

As suggested by Bialski (2016), seemingly arcane methodological activities like data cleaning and data coding inevitably have an epistemic and methodological dimension. Therefore, to

[6] Same issues would apply to the option of using networking information available in LinkedIn only.

allow all members of the teams to theoretically inform sampling as well as critically evaluate and intervene on the minutiae of data cleaning and normalization, we decided we would start with a pilot before scaling up to the full dataset. This would also allow us to use visual comparison of descriptive results to inform further development of the inferential model.

Using a sub-sample of random analysts as our reference group, we wanted to see if there is a pattern in top industry analysts' careers that is different in terms of provenance or career pathway from other typologies of analysts. In order to do so, we created 4 distinct sets of industry analyst samples, all with a publicly accessible profile containing an experience section:

- 100 random industry analysts
- 99 top industry analysts
- 100 industry analysts with experience of ERP
- 33 industry analysts whose current job title is CEO

We collected a set of random analysts to compare and contrast with top analysts (i.e. 'power' analysts) and a set of analysts with heavy industrial expertise e.g. analysts with experience in ERP. To mitigate the client-business bias identified in the ARchitect database by the domain expert, we added to our sample all analysts whose current job title is CEO, assuming job title as another indicator of career outcome. We therefore included a list of people with Chief Executive Officer (CEO) listed as their current job title. From that sub-set of industry analysts, we exported their job title, the company they work for and their LinkedIn URL if available. The HTML profiles were then further processed to extract the LinkedIn Experience section for each person. Each file has a top-level career element which contains a name element referring to the name of the person and one or more job elements referring to the different jobs the person held or is holding. Each job element contains further information on the job title (title), the company where the job was based (company), the date-range of the job (as listed in LinkedIn and normalised to the number of years and

number of months the position was held). The jobs are ordered as listed in the Experience section. All 332 profiles which we collected contained a total of 1,659 jobs. The company names associated with all of the jobs collapse (after lowercasing and unique-ing them) to 1,038 distinct surface forms. It was reassuring to find confirmation that IT influencers were well represented on professional networking media. Out of the 100 top analysts, only one did not have a publicly accessible LinkedIn profile. By turning experience profile in job sequences, we could test the distinctive transition hypothesis (i.e. that the IT and the market research industries are the main feeders for the analyst industry) and find an answer our question on the role of smaller analyst houses in career formation.

8. Why only industry analysts

As well as starting from a smaller-scale pilot to facilitate inter-disciplinary collaboration, we decided to focus specifically on one typology of IT influencers: the IT Industry Analyst. The decision was informed by a number of factors. One is the opportunity to draw upon extensive empirical evidence accumulated through decades of social research. As demonstrated by social studies of influencers, the majority of them are "monomorphic": their opinion is decisive only on certain matters (Katz and Lazarsfeld, 1955; Merton, 1968; Cialdini, 2007; Cheshire, 2011). In the likely absence of a general-purpose definition of influencer, there is value in focusing on the entire population of a particular typology of influencer. Furthermore, we wanted our previous research on this particular professional community to provide context as well as hypotheses to inform the data-driven approach. Inductions from fieldwork (“ground truth” in machine learning parlance) would complement and control the ontological decisions taken as part of a data driven approach, in ways that will become apparent in the next section. Limiting the study to this typology of influencer could also maximize the contribution of the member of the team with domain expertise in this area.

By admission of the database owners, ARchitect contains industry analyst firms as well as firms from cognate industries such as consultancy and sourcing advisory. In order to keep our pilot as specific as possible to IT industry analysis, we decide to use a second expert list of analyst companies and consider only companies that appear in both lists. The first expert list is Barbara French's list of industry analyst firms[7]. From the site, we automatically collected all 430 firms. The second list is the list of all current analysts companies stored in the ARchitect database (963). 464 company names, the intersection, are in both lists.

9. Ontological decisions

One further element in data-driven social research is that the researcher claims ontological authority over which entities he will be researching on. This is to say that, after normalization and categorization, the raw data input for sequential analysis is actually created by the researcher. As arbitrary as it might sound, this is a necessary step. With 464 company names and 1,659 job titles from only 332 profiles there would be no way to otherwise produce sequences whose similarity could be analysed.

We first devoted our attention to creating company categories. We based categorization of companies on size and seniority. Size resulted from how many times the company was mentioned in our sample. Seniority from how long ago a company was joined in the average (See Appendix 1). We expected that operationalizing size in this way would confirm evidence from previous ethnographic study of a few companies dominating the market. We found that the three companies that were mentioned most frequently in our data where indeed Gartner, Forrester and IDC (the biggest three companies in the industry) while none of the other companies was mentioned more than six times.

If we wanted to look at career pathways to confirm our hypothesis of feeder industries as well as find confirmation of role of one-man companies, we also needed a way to identify

[7] http://analystdirectory.barbarafrench.net

which other companies occur later in a career, other than the big three. Distinguishing by seniority would give us ways to capture which companies analysts form or join when they want to trade their expertise later in a career. We thus created three further categories of companies (oldest, old, new) based on when in the career people joins them. Six years is the median number that emerges from querying our database on how long ago people joined their last company (with 71 companies being those that people worked for in their last job). 'Oldest' would be the companies nobody joined in their last job and arguably not the companies people join when they want to trade their expertise independently. 'Old' are the companies joined earlier than in the last 6 years, while 'new' are companies joined in the average during the last 6 years of a career.

(Table 1 here)

[Anonymised author name], an expert in industry analysis and co-author of this paper, has provided us with a ranked list of job title normalisation for various surface forms (see Appendix 2). One additional text processing step involved identifying the seniority of each job title by matching it against a set of keywords (see Appendix 3). To conclude, from an initial sample of 332 profiles (100 random, 100 profiles with ERP expertise, 99 top analyst profiles, 33 CEO profiles), 259 remained after filtering out profiles without any Industry Analyst company in them as well as eliminating duplicates. Zooming in the attribute structure of an individual state at this point we would obtain what follows:

(Table 2 here)

Which means that 'Analyst X' in 1979 was ceo (without seniority mark) in a small analyst industry company (i.e. YY), one that usually people join in the first six years of their career

(i.e category OLD).

Each job sequence will be a sequence of states e.g.

[manager NA N SMALL OLDEST] [ceo NA YY SMALL OLD] [ceo NA N SMALL NEW]

where each state is usually held for a number of years.

Ontological decisions taken at this stage of the research process raise serious questions from a sociological perspective. By creating our ontology through classifying companies and ranking jobs according to fewer categories based on seniority we were programming in the data the expected career pattern ahead of time. However, there were strong arguments for using this simplified categorization. Our decision were indeed supported by inductions from pre-existing data sources, namely "ground truth" garnered through fieldwork. Our ontological decisions were also systematically returned to our domain expert for feedback. More generally, if we wish to find typologies of successful career patterns and already know that certain typologies of companies are more indicative than others of positive outcome, then surely we wish our approach to take advantage of that information.

10. Evolving research focus towards inter-professional career transitions

We then started addressing sequential analysis using the TramineR package. TraMineR is a R-package for mining and visualizing sequences of categorical data (Gabadinho et al., 2011). The purpose on this phase was to adopt an exploratory research approach to generate descriptive findings. This would help test the categories of companies and the hierarchy of jobs to be used in the full-scale study as well as gather preliminary findings that will contribute to refine research questions. Visualizing data was seen at that stage as part of the analytical process to convey further interrogation of data in the inter-disciplinary research

team: an instrument for reasoning about quantitative information rather than presenting final results (Tufte, 1998:9). After having shown in the previous section how we combined inductions from multiple sources (from fieldwork data as well as from digital data) to guide ontological decisions in data preparation, this section will focus on how abductive reasoning became apparent in our project. In particular, we will make visible how early descriptive findings provided useful information for a retrospective adjustment of our categorization of sequence states. Within this and the following section we will also describe how the sharpening of our data categorization and consequent identification of correlations with career outcomes influenced our research questions and led to inclusion of further statistical tests in our research design.

Approaching our data with sequential analysis, we soon realised how difficult it could be to answer our question on 'distinctive transitions'. Our hypothesis was indeed that the IT and the market research industries are the main feeders for the analyst industry. However, our dataset shows that our 259 analysts have worked for 186 different companies before taking up an analyst job. Taking into account that 62 started their career already in an analyst company, it is nearly one company per worker. The variation concerning company names appearing before transition to analyst industry (nearly one company per person) is poised to increase with the analysis of the full data. Together with lack of authoritative lists of companies in feeder industries, it meant we would not be able to tell where (which industry) people worked before taking an analyst job. We therefore decided to merge all possible states concerning jobs had in a non-analyst company into one single state (irrespective to company size or seniority). While we retained the remaining states, as they would provide information on how analyst careers progress from small houses to big houses and eventually to one-person companies[8]. However, although in its simplified form, information concerning inter-professional career transitions remained crucially important for our research strategy.

[8] For further details on how we dealt with overlaps of jobs - e.g. one person has two jobs at the same time at some point in her career, see Appendix 4.

One of the immediate findings from plotting states in a sequence was indeed that industry analyst jobs appear later in a career. This has been found by simply calculating the average time spent in a career before transitioning to an analyst company job. We found that analysts spend half their career into non-analyst jobs before becoming analysts. Results divided by sample look like as in Figure 1.

(Figure 1 here)

Early results generally point in the direction that the careers of 'Power 100' analysts as well as those of analysts with CEO as a job title (i.e. arguably those representing a positive outcome) can be distinguished from others based on inter-professional career span. Successful analysts tend to have longer inter-professional careers. This is particularly apparent from comparison of time spent in careers in other industries. TOP/CEO spend 9.9 years before turning analysts with random sample only 7.5 years.

When these results emerged, we saw our research taking a different direction. Although not directly responding to our original question on distinctive transition, what we found about inter-professional career moves was highly relevant for how it seemed to correlate to positive outcome. Ways to parse careers based on events happening in the initial segment of a sequence are seen as the most efficient when the purpose is early identification of outcome. Career studies scholars' preferred approach in this respect has been so far to look for correlation between career outcome and education. In their study of careers in the IT sector, Bidwell & Briscoe (2011), correlate education with occupational career, measuring occupational intensity in the industry based on years of education of members in the focal occupation. However, it emerges from a recent LinkedIn study (Berger, 2016) that, in terms of achieving a CEO position, changing four different jobs within the same company has nearly the same impact as getting an MBA from a top-five program. This points to the fact

that education might not have a crucial role in entrepreneurial careers such as those of IT influencers. Through early results we were finding that aspects of inter-professional careers (i.e. that long and rigid[9] careers in another industry are conducive of inter-professional success) could be considered an alternative candidate measure to produce early predictions concerning career outcome. This was especially relevant vis-à-vis to scholarship in career studies given recent criticisms to the role of education data in predicting career outcome.

11. The mutual relation between inductions and inferences

In this final empirical section we show how an exploratory research approach led to the identification of appropriate inferential models to refine the research questions and prepare for the application of these models to the full-scale project. In particular, we describe how the use of visualisation of pilot results led to quickly explore alternative interpretations of the data as well as identify relevant, although of not statistically significant, career patterns. We also make apparent how, reciprocally, the prospect to adopt an inferential approach brought into question the compatibility of sampling strategies based on our descriptive research questions with an interest for making predictions.

The exploratory use of visualisations in the research process made possible to quickly compare and contrast multiple vistas on our data in addition to the one based on dividing the sample base of "success". By using sequence visualization to compare careers started after 2000 with careers started before 1990, for example, we saw a stark difference in the distribution of jobs. While in careers starting before 1990 (see Fig. 2) the prevalent strategy was to access the analyst industry (yellow, green or orange coloured sections of timeline) after having spent considerable time in other industries (purple sections of timeline), the

[9] To seek confirmation of the extent to which the study of inter-professional careers can be a viable identifier of influencers, we also looked for how many jobs did analysts change in the average before taking a job in an analyst company. See Appendix 5 for further details.

trend has changed with careers beginning after 2000, with many more starting directly in the analyst industry (see Fig. 3).

(Fig. 2 Here)

Another particularly interesting aspect in this visualisation is the changing role of big companies (green section in timeline). From the areas marked in green in the graph plotting careers started before 1990, it is apparent how the Big Three dominated the market.

(Fig. 3 Here)

While in more recent years (careers started after 2000) there is more variability, with jobs in smaller analyst companies (yellow or orange) becoming prevalent.

Observations on the evolution of careers over time as well as an emerging interest for making predictions about the future evolution of careers led us to identify the need for further controls to be run in preparation to the analysis on the full population (10k industry analysts). Observation of the evolution of careers over time suggest that it is important that if we want to make inferences about how analysts' career might evolve in the future, more weight should be given to data from careers that have started more recently (i.e. career starting after 2000) as they are arguably more representative of the current evolution of the industry. In order to test our emerging conjecture on the relevance of inter-professional career moves and see how other types of career move would score in terms of correlation with outcome, the team identified the option to run three regression models, one for each different way to parse careers. We could parse careers based on moves *(i)* between industries or (ii) between companies within the same industry or *(iii)* between jobs within

the same company. Based on data on more recent careers, a regression coefficient will tell us whether alternative ways to parse careers (such as those based on inter-organisational or intra-organisational moves) offers better promises than the inter-professional transitions. Results can be then controlled for age and for generation.

Another benefit of an exploratory approach is to find trends that might not emerge when relevance in determined by statistical significance. Our exploratory approach based on running a separate sequential analysis of the successful analyst sample (1% of the total population) and descriptively comparing the emerging clusters with typologies found in the random sample had the benefit of giving a central role to key career moves in top analysts' careers as opposed to diluting them in the full sample[10]. It became apparent that, although the majority of people working in the Big Three still complete their career there, a pattern that seems to occur frequently in TOP/CEO sample is a move out of larger companies to establish one's own business. Notably, nearly all moves of this type are accomplished by analysts in the TOP/CEO sample[11]. Although their number is not statistically representative, it is certainly a common trend within the small number of most successful analysts.

However, our emerging interest for testing the hypothesis of inter-professional transitions brought into question our focus on the role of one-person companies in career formation and more in general our focus on factors emerging in careers of most successful analysts. First, by considering marks of success that are only found in from looking at a limited sample of TOP/CEO analysts we were running the risk of attributing unjustified additional relevance to more senior, and therefore relatively longer, careers.

Also, wanting to understanding what analysts do once having joined top companies (i.e. our second set of descriptive questions introduced in Section 4) would also mean systematically

[10] As noted in careers literature, the ideal-typical career is the rarest of all (Rosenbaum, 1979).

[11] Out of our 259 analysts, 39 have done a move out of the big three and up to a company that more often occurs later in a career. If one focuses on descriptive results from the sequential analysis of this sub-sample of careers, it results that the majority of moves out of big three (19 out of 39) correspond with what we call personal portfolio type of move. Notably, nearly all of the personal portfolio types of moves (16 out of 19) are accomplished by analysts in the TOP/CEO sample.

taking into account the entire job sequence of each analyst.

In other words, the prospect of applying an inferential approach revealed a sampling bias in our descriptive approach towards more senior careers.In order to avoid this we would need to expand our definition of success to from a dichotomous variable - based on whether or not an analyst is in the “Power 100” list - to an ordinal variable. This can be tentatively done for example by using last job titles ranking as an indicator of success. We could then run a cluster analysis of the full population of analysts and assign a success score to each cluster. With this approach, we could identify career clusters which are conducive of success as well as clusters containing careers that have an higher success score although they still have not reached the top jobs.

If cluster analysis applied to the entire population could help address our sampling bias towards more senior careers, the problem deriving with the need to respond to our descriptive question by systematically including the entire job sequences would remain unaddressed.

Informing our inference based on the entire career span might result in overloading the model with information that in most cases has nothing to do with our attempt to understand what causes positive outcome. The latest development in our project’s research strategy suggest that this can be addressed by considering application of event history analysis using sequences as a time varying predictor (Rossignon et al., 2016). Combination of event history with sequential analysis allows to retain the thickness of the descriptive account combining it with the sharpness of a focus on predictions.

12. Discussion and Conclusion

In conclusion we want to summarise how the notion of participative epistemology as it became apparent in the project’s research design and early results can contribute to discussion on sensitizing social data science. Sensitising social data science meant in our

project, at a primary level, to juxtapose ethnographically gained inductions from empirical fieldwork to computer generated data-driven inductions. As discussed in Section 4, qualitative insights provided the research motivation as well as crucial information on how the primary data source of our research (ARchitect database) was built and used by the analyst community.

Our notion of sensitising data science also certainly includes, in an equally direct form, not to ignore generations of social science scholarship. This has been visible in the project phases described in Sections 5 and 6, where notions from the sociology of professions provided support for the decision to use professional networking data as opposed to social networking data. Social science insights also helped decide to scope our project to a specific typology of influencers and re-orient our research question by indicating which findings were relevant to fill gaps in existing research. However important, these are yet examples that limit the contribution of social science to that of representing categories that are external to the core of data science. The social scientist can be seen as the one able to "speak" the language of the users of data science or the expert of a method (ethnography) that can be triangulated with computational and inferential methods typical of data science.

While acknowledging and taking stock of these basic understanding of inter-disciplinarity, our case analysis shows a new and more radical sense of disciplinary integration by which the social scientist can operate within a data-science framework. By suggesting epistemic subjectivities, epistemic objects and methods of the research as having an agency in the shaping of social data science research process and results, we wanted to show how social science - and the sociology of scientific knowledge in particular - is capable of producing categories to be used in the making of social data science. Findings from the cases study are summarised below.

12. 1 Epistemic subjectivities as an analytical category

We first suggested epistemic subjectivities in the inter-disciplinary team as research participants, whose complementary and evolving perspectives shape the research design. One crucial incarnation of this epistemological project is in the fact that part of the team have studied the evolution of IT professionals' careers for a decade, with one author (anonymised author) having worked on the topic for nearly 20 years (anonymised reference). This long-term engagement with the research domain, accomplished through ethnography-inspired fieldwork practices, allowed the team to gain in-depth insight on the phenomenon through an extended perspective. In particular, approaching a research problem with different methods over time emerges as an important element in our understanding of how social, computational and statistical approaches can co-exist. Much of the epistemological discussion on the data revolution involving research practice has been on the possibility of concurrent application of complementary paradigms (i.e. inductive and deductive) in a resarch project/teams (Kitchin, 2014; Wagner-Pacifici, Mohr and Breiger, 2015). One more compelling avenue to discuss co-existence of approaches, our case suggests, should be found in how different epistemological postures can be used at different points within an extended research programme.

As well as considering the evolution of epistemic subjectivities, a second element of participative epistemology regards their interaction. This is exemplified in our project by collaboration with the computer scientist. Empirical social scientists base their knowledge claims from expertise deriving from longstanding engagement with a particular field of practice (the study of IT professions in our case) as well as from (or, more appropriately, as part of) mastery of a particular method. Similarly, when building the interdisciplinary team, members have been identified based on their technical expertise as well as previous engagement with the research domain. The computer scientist working in the project

(anonymised name of co-author), for example, has been involved in previous research on parsing CVs of IT professionals (software engineers) and developed matching algorithms for recruiters to identify best candidate for IT jobs. Sharing substantive domain knowledge among an interdisciplinary research group allowed negotiating knowledge claims based on expertise deriving from longstanding engagement with a subject matter, as well as on mastery of a particular method. This contributed to the successful adoption of a reflective perspective in data preparation as described in Section 9 and discussed in section 12.3 below.

A third aspect that derives from considering epistemic subjectivities as co-researchers in social data science is suggested in our project by the collaboration with practitioners as practical theorists (Hoffman, 2004). An active industry analyst practitioner as well as a PhD student, one of our co-authors (anonymised name of co-author) has been able to reveal aspects of database designed, thus contributing to understanding factors affecting commercial data (as discussed in Section 12.2 below). He also provided feedback on crucial decisions such as those related to ranking job titles, while helping seek confirmation from subject matter experts for ontological decisions that were driven by direct interrogation of the data – for example, our ontology of company based on size and seniority.

12.2 The agency of the epistemic object

It is due to the nature of its object of study (i.e. the empirical social world) that empirical sociology concepts are temporary constructs (i.e. sensitizing) and not definitive ones. What social research is referring to by any given concept shapes up in a different way in each empirical instance. And this holds equally - if not more - valid for Social Data Science: there is no “ground truth” that stands the test of time. To achieve the close and self-correcting relation with the social world despite this variability, including the research object within a participative epistemology framework is key. The variability of our research object (i.e.

careers) becomes apparent in our project in multiple forms. First, being accessed through an industrial, internet-generated database, our research object responded to the logics incorporated in a CRM system as much as to the logics of our research. A particularly interesting aspect of participative epistemology is indeed the interaction of the researcher with 'commercial sociology' insights (Burrows & Gane, 2006). Together with granting access, ARInsight (the company curating the analyst profiles database) provided us with a ranking of analysts based on their 'power' i.e. through sharing their transactional data on how many times analyst profiles were visited by users of their database. Rather than conceiving of this index as a biased assessment of influence, we used commercial insights to inform our early research strategy. While commercial insights (as well as insights deriving from platform effects) were integrated at later stages with inductions from qualitative fieldwork and statistical tests.

Further mediations showing agency included the transformation of our research object from textual data into a format that could be read by a sequential analysis algorithm. Some of these transformations (such as those due to the mimetic pressures of publishing work experience onto public platforms or those related to the creation of synthetic ontological categories in data preparation) made data more docile to our questions. Other transformations backfired, such as when we linked our data points together to form sequences as temporal contructs. We realised that not all individual sequences in our data set and not all points in a sequence representing individual analysts' career were equally relevant for our research question. As discussed in Section 11, without controlling our career data for age, or generation, sequence data could give unjustified additional relevance to longer careers. Also, considering information concerning career events occurring after having achieved what is defined as positive outcome could result in inclusion of irrelevant factors, which could confound results.

12.3 Object-relation regime

The agential role the research object is inextricably linked with how the object becomes available to the analysis (i.e. what Knorr Cetina defines object-relation regime). Working in a social data science epistemological environment means that the object of research is approached, within the same project, from a diversity of apparently contrasting epistemological angles: the computational, the ethnographic and the inferential. Working within a computational frame means working in a regime where data is heavily designed. One characteristic of social research is methodological reflexivity (Alvesson and Sköldberg, 2009): a deeper understanding of the subjective, institutional, social, and political processes whereby research is conducted. To the extent a social sensitivity is applied to maintain a methodologically reflexive approach throughout the 'artificial' production of data in a computational regime, data transformations described in Sections 7, 8 and 9 remain a fundamental component for the progress of research and legitimately participate the production of findings. These transformations allowed for example to gradually programme into our data factors that were indicative of the phenomenon under investigation (i.e. career patterns).

The last point of how a participative epistemology should incorporate the object-relation regime as a research actor concerns the interaction between inductive and inferential approaches. As shown in Section 10 and 11, we found the descriptive, ethnographic induction-driven approach to be helpful in identifying patterns that would be otherwise lost in a search for statistically significant results. Quick changes of perspective on the data such as those made possible by the use of visualization also contributed to refine research questions. While by not assuming any pre-existing way of slicing data and by opening up the possibility to ask a different question to our data (i.e. what the careers of tomorrow's analysts will look like) the inferential approach revealed inconsistencies due to the

retrospective character of our descriptive approach.

13. A Stream Three Social Data Science

As a way to conclude we want to go back to the parallel with studies of expertise. In particular, and by drawing on Collins and Evans' understanding of the social basis of expertise (Collins and Evans, 2002), we want to use the parallel to suggest directions for future developments for the integration of social science and data science. By mobilizing notions from the sociology of scientific knowledge to expand the repertoire of concepts in digital sociology and showing their application in the analysis of our case study, we hope to have contributed to dissolve some dichotomies between inductions from computational and ethnographic approaches as well as between the hypothesis-driven and data-driven *modus operandi*, demonstrating that if conceived within a reflective approach, there is no special status in either sides that prevents end-to-end integration. However, with Marres (2017: 61) we attribute to social science a role to intervene in how society is becoming visible and researcheable through digital data sources and data science methods. Therefore, and going back to our original question - can social scientists directly contribute to data science? – we also introduced the notion of participative epistemology as a guide for social scientists' intervention in data science. The notion has been developed based on considering (i) the evolution and interaction of epistemic subjectivities in the inter-disciplinary team, (ii) the research object in its different incarnations and (iii) the shifting object-relations in social data science as co-researchers as having agency in constantly shaping research design and questions.

We recognise that finding a rationale for a special place for social science in data science entails much more work than the analysis of a single case study. We hope our contribution will inspire more scholarship offering reflective empirical accounts of 'before-consensus' social data science practice with a willingness to disclose the iterations in the research

design and the shifts in the research questions which, we argue, are at the core of 'real'

social data science.

References:

Abbott, A. (1995). Sequence Analysis: New Methods for Old Ideas, Annual Review of Sociology, Vol. 21, pp. 93-113.

Abbott, A. 1990. "A Primer on Sequence Methods," *Organization Science* (1:4), pp. 375-392.

Abbott A. (1988). *The System of Professions*. Chicago: Univ. Chicago Press

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Tobin, R., and Wang, X. (2008). Automating curation using a natural language processing pipeline. Genome Biology, 9(Suppl 2):S10.

Alvesson, M., Sköldberg, K. (2009) Reflexive Methodology: New Vistas for Qualitative Research. 2nd Edition. London, Sage.

Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired Magazine*, 06.23.08, https://www.wired.com/2008/06/pb-theory/

Bakshy, E. Hofman, J.M. Mason, W.A. and Watts D.J. (2014) Everyone's an influencer: quantifying influence on twitter, Proceedings of the fourth ACM international conference on Web search and data mining, 65-74.

Bateson, Gregory (2002 [1979]) *Mind and Nature: A Necessary Unity* (Cresskill, NJ: Hampton Press).

Berger, G., Gan, L., Fritzler, A. (2016) How to become and executive, https://www.linkedin.com/pulse/how-become-executive-guy-berger-ph-d-?published=t

Bidwell, M. and Briscoe F. (2010) The Dynamics of Interorganizational Careers Organization Science 21(5), pp. 1034–-1053.

Bittner, E. (1965). The concept of organization. Social Research, 32(3), 239–255.

Blank, G. (2007). *Critics, ratings, and society: The sociology of reviews.* Lanham, MD: Rowman & Littlefield.

Block, A., Pedersen, M.A. (2014) Complementary social science? Quali-quantitative experiments in a Big Data world. 1-6.Boyd D, Crawford K (2012) Critical questions for big data. Information, Communication & Society 15(5): 662–679.

Blumer, H. (1954). What is wrong with social theory? American Sociological Review,18, 3-10.

Burrows, R. and Gane, N. (2006) 'Geodemographics, Software and Class', Sociology 40, 5, 793.

Centola, D. (2010) The spread of behavior in an online social network experiment. *Science*

329, 1194–1197.

Coleman, G. (2012) *Coding Freedom: The ethics and aesthetics of hacking*. Princeton: Princeton University Press.

Collins, H.M., Evans, R. (2002) The Third Wave of Science Studies: Studies of Expertise and Experience, *Social Studies of Science* 32/2(April 2002) 235–296.

Collins, R. (1994) Why the Social Sciences Won't Become High-Consensus, Rapid-Discovery Science, *Sociological Forum*, 9(2), 155-177.

Coulter, J. (1996). Human practices and the observability of the 'macrosocial'. Zeitschrift für Soziologie, 25, 337–345.

De Vaan, M., Stark, D., Vedres, B. (2014) Game Changer: The Topology of Creativity, *American Journal of Sociology*, 120(4), 1-51.

Dijck, J. van (2013) The culture of connectivity: A critical history of Social Media. Oxford: Oxford University Press.

Di Maggio, P.J., and Powell, W.W. (1983) The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields, *American Sociological Review*, 48(2), 147-160.

Duggan, M., Ellison, N.B., Lampe, C., Lenhart, A,. and Madden, M. "Social Media Update 2014," Pew Research Center, January 2015. Available at: http://www.pewinternet.org/2015/01/09/social-media-update-2014/

Edwards, A., Housley, W., Williams, M., Sloan, L., and Williams, M. (2013) Digital social research, social media and the sociological imagination: Surrogacy,augmentation and re-orientation. International Journal of Social Research Methodology, 24: 313-43.

Fine, Gary Alan (2001), *Difficult Reputations: Collective Memories of the Evil, Inept and Controversial*. Chicago: University of Chicago Press,.

Finger, L., Dutta, S. (2014) ASK, MEASURE, LEARN. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Fligstein, N. (1990) The Transformationof CorporateControl. Cambridge, Mass.: Harvard University Press.

Friedkin, N.E., (1991) Theoretical foundations for centrality measures, American Journal of Sociology, 96 (6): 1478-1504.

Gabadinho, A., Ritschard, G., Mueller, N. S. Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1-37.

Gaskin, J., Berente, N., Lyytinen, K., and Yoo, Y. 2014. "Toward Generalizable Sociomaterial Inquiry: A Computational Appoach for Zooming in and out of Sociomaterial Routines," *MIS Quarterly* (38:3), pp. 849-871.

George, G., Haas, M. and Pentland, A.S. (2014), "Big data and management", *Academy of Management Journal*, Vol. 57 No. 2, pp. 321-326.

Gillespie, T. (2010) The politics of 'platforms'. *New Media and Society*, 12(3): 347-64.

Gower. B. Scientific Method: An Historical and Philosophical Introduction. Routledge, Abingdon, UK, 1997.

Gunz, H., Mayrhofer, W., & Tolbert, P.S. (2011): "Career as a social and political phenomenon in the globalized economy." *Organization Studies,* 32 (12), 1613–1620.

Habermars, [1985] 1987 *The Theory of Communicative Action* Vol.2 (Boston, MA: Beacon).

Halavais, A. (2013) Structure of Twitter: Social and technical. Twitter and Society, Weller, K., Bruns, A., Burgess, J., Mahart, M., and Puschmann, C. (eds). New York: Peter Lang: pp. 29-42.

Halford, S., and Savage, M. (2017) Speaking sociologically with big data: symphonic social science and the future for big data research *Sociology*, pp. 1-18.

Heron, J., & Reason, P. (1997). A Participatory Inquiry Paradigm. Qualitative Inquiry, 3(3), 274-294.

Hoffman, A. (2004). Reconsidering the role of the practical theorist: on (re) connecting theory to practice in organization theory. Strategic Organization,2(2), 213-222.

Housley, W., Procter, R., Edwards, A., Burnap, P., Williams, M., Sloan, L., Greenhill, A. (2014) Big and broad social data and the sociological imagination: A collaborative response. Big Data and Society, 1(2).

Hughes, E.C. (1937): "Institutional Office and the Person." American Journal of Sociology, 43, 404–413.Fligstein, 1990

Hyysalo, S (2010). *Health Technology Development and Use: From Practice Bound Imagination to Evolving Impacts*. London: Routledge.

Katz & Lazarsfeld (1955). "Personal Influence". *New York: Free Press*.

Kitchin R (2014) 'Big Data, new epistemologies and paradigm shifts' *Big Data & Society,* April–June 2014: 1–12.

Kitsak, M. et al. (2010) Identification of influential spreaders in complex networks. Nature Physics 6, 888–893.

Knorr Cetina, K. (1999) "Epistemic Cultures: How the Sciences Make Knowledge". Cambridge, Mass.: Harvard University Press.

Larson, M, S. (1977), The Rise of Professionalism: A Sociological Analysis. Berkeley: University of California Press.

Latour, B. 2010. "Tarde's Idea of Quantification," in The Social After Gabriel Tarde: Debates and Assessments, M. Candea (ed.), London: Routledge, pp. 145-162.

Leonardi, P. M. 2011. "When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies," MIS Quarterly (35:1), pp. 147-167.

Leonardi, P., and Barley, S. 2008. “Materiality and Change: Challenges to Building Better Theory About Technology and Organizing,” Information and Organization (18:3), pp. 159-176.

Marres, N. (2017) “Digital Sociology: The reinvention of Social Research”, Cambridge: Polity Press.

Marres, N., Weltevrede, E. (2015) Scraping the social? Issues in real-time social research. In La Médiatisation de l’Évaluation. J. Bouchard, É. Candel, H. Cardy, H. Gomez-Mejia (eds) Berlin: Peter Lang.

Merton, R. K. (1968). Social theory and social structure. Glencoe: Free Press.

Mills, C. W. (1959), “The Sociological Imagination” Oxford, Oxford University Press.

Passman, J., Gerlitz, C. (2014) Good platform-political reasons for bad platform-data. Zur socio-technischen Geschichte der Plattformaktivitäten Fav, Retweet und Like. Media Kontrolle, Working Paper.

Pastor-Satorras, R., Vespignani, A. (2001), Epidemic spreading in scale-free networks. Physical review letters 86 (14), 3200.

Pavlos Basaras, Dimitrios Katsaros, and Leandros Tassiulas (2013) Detecting Influential Spreaders in Complex, Dynamic Networks, Computer, 26-31.

Pentland, B. T., and Feldman, M. S. 2008. “Designing Routines: On the Folly of Designing Artifacts, While Hoping for Patterns of Action,” Information and Organization (18:4), pp. 235-250.

Pentland, B. T., and Feldman, M. S. 2007. "Narrative Networks: Patterns of Technology and Organization," *Organization Science* (18:5), pp. 781-795.

Pollock, N. & Williams, R. (2016) How Industry Analysts Shape the Digital Future, Oxford University Press.

Pollock, N. & Williams, R. (2008) Software and Organisations, Abington: Routledge.

Rogers, Everett M. 1983. *Diffusion of Innovations*. 3rd ed. New York, NY: The Free Press.

Rosenbaum, James E. (1979) "Organizational career mobility: promotion chances in a corporation during periods of growth and contraction." American Journal of Sociology, 85(1): 21-48.

Rosenfeld, RA. (1992) Job mobility and career processes. *Annual Review of Sociology*, 18:39-61.

Rossignon, F., M. Studer, J.-A. Gauthier, & J.-M. Le Goff, (2016) Childhood family structure and home-leaving A combination of survival and sequence analyses. Proceedings of the International Conference on Sequence Analysis and Related Methods (LaCOSA II). Lausanne, Switzerland, June 8-10, 2016

Sandberg, J., and Tsoukas, H. 2009. “Being-in-the-World, Practical Rationality, and Organizational Research: Notes for Theory Development,” paper presented at the First

International Symposium on Process Organization Studies: Sensemaking and Organizing, Pissouri, Cyprus, June 11-13.

Sayer, A. (1992) Method in social science: A realist approach (2nd ed.). London: Routledge.

Slater, D. (2002) Social relationships and identity online and offline. Handbook of new media: Social Shaping and consequences of ICT. Lievrouw, L., Livingstone, S. (eds). London: Sage Publications: pp. 533-46.

Shaw, R. (2015) Big data and reality. *Big Data & Society*, 2(2).

Stoller P (2013) Big data, thick description and political expediency. Huffington Post, posted 16 June 2013.

Schuman, L. (1997) Centers of coordination: A case and some themes. *Discourse, Tools and Reasoning*, Berlin and Heidelberg: Springer: pp. 41-62.

Turner, S (2001) What is the Problem with Experts? Social Studies of Science, 31: 123-149.

Vedres, B., & Stark, D. (2010). Structural Folds: Generative Disruption of in Overlapping Groups. American Journal of Sociology, 1150-1190.

Weng, L. Menczer, F. & Ahn, Y. (2013) Virality Prediction and Community Structure in Social Networks, Nature Scientific Reports, 3 : 2522.

Williams, R. & Procter, R. (1998). Trading Places: A Case Study Of The Formation And Deployment Of Computing Expertise. In Williams, Robin Et Al (Eds) Exploring Expertise. Basingstoke, Macmillan, Chap. 13, P.197-222

Wilson, E. (1998). *Consilience:The Unity of Knowledge*. Knopf.

Tables and Figures

	Number of Companies	Number of Observations
YY Oldest Small	18	37
NN Oldest Small	697	891
YY Old Small	32	98
NN Old Small	9	11
YY Old Large	3	284
YY New Small	14	44
NN New Small	29	44

Table 1. This is a table showing number of companies for each category and number of observations per category. YY stands for Yes Yes and it means that the company is in both expert lists as explained in Section 8.

	(x) 1979				
	Job element		Company element		
	Job title	seniority	industry	size	seniority
Analyst X	**ceo**	**NA**	**YY**	**SMALL**	**OLD**

Table 2. The attribute structure of an individual state in our sequences.

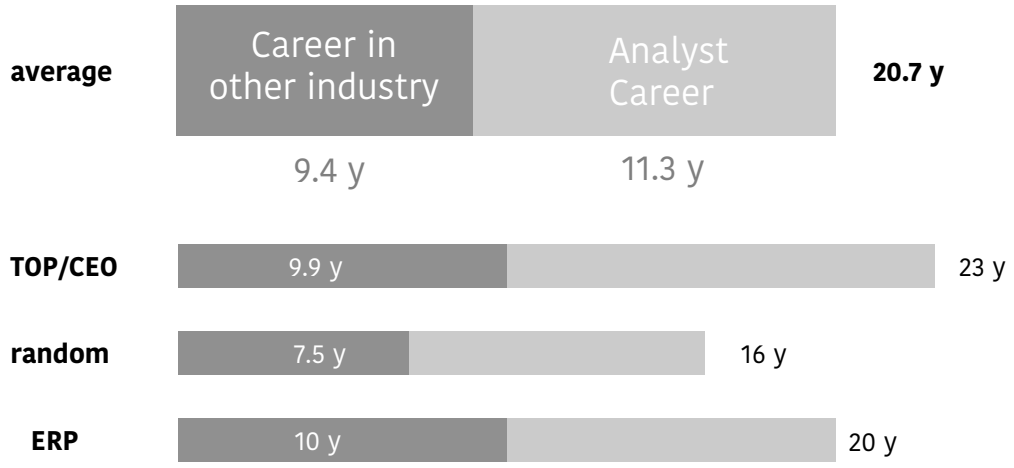

Figure 1. The thicker bar on top of this figure represents analyst average career length (20.7 years) as well as average time of inter-professional transition (i.e. after 9.4 from beginning of career) and average length of career in the analyst industry. The three thinner bars below represent career length by sample (TOP/CEO analysts, random analysts and analysts with expertise in ERP) – please refer to Section 7 in the paper for details about sampling. The darker section of the bar represents average number of years spent in jobs in non-analyst companies (e.g. analysts in the TOP/CEO sample spend 9.9 years in non-analyst job before taking up an analyst job). The lighter section towards the right represents average number of years spent in jobs in analyst companies. The letter "y" stands for "years".

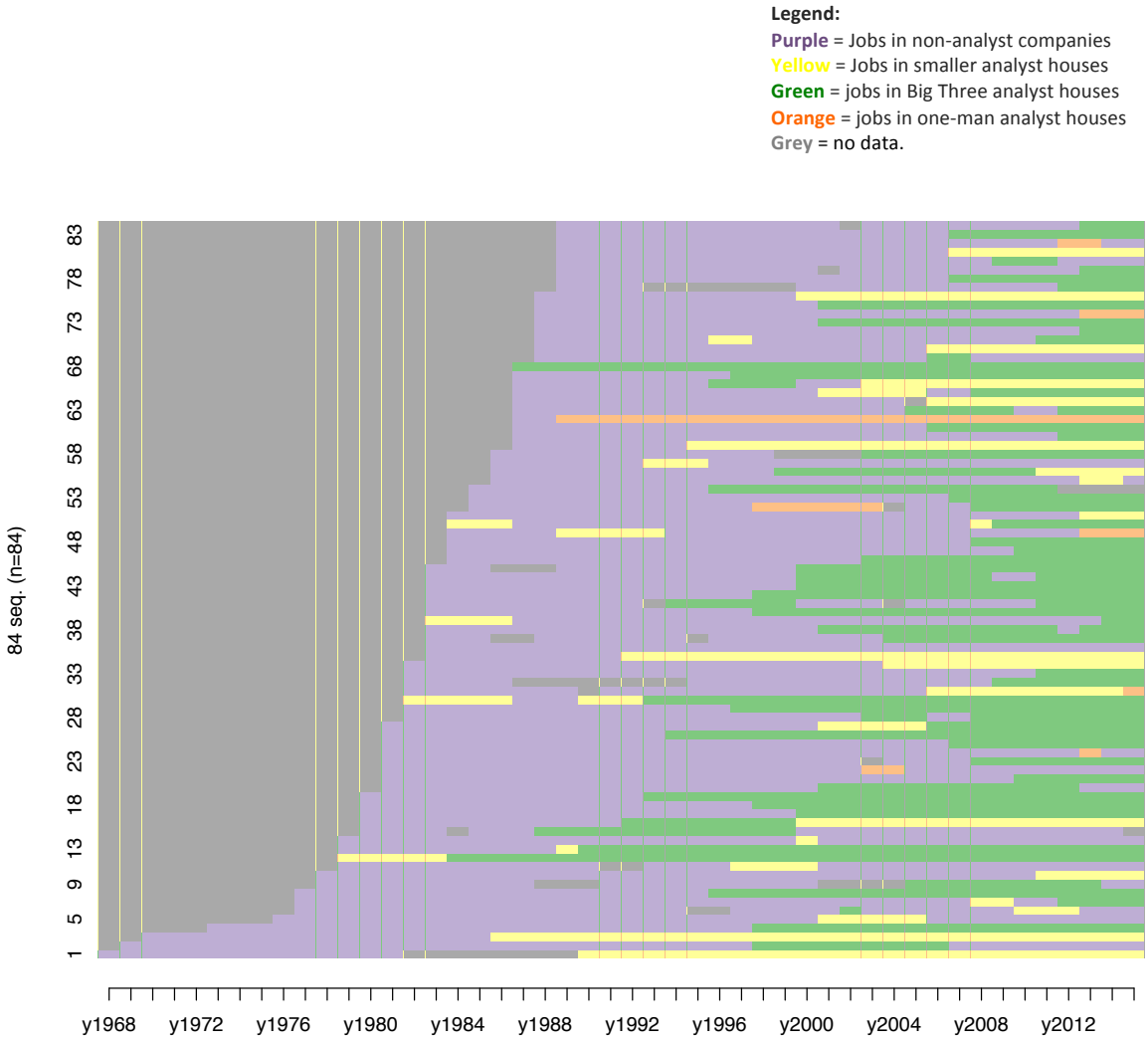

Figure 2: This visualization represents careers started before 1990. Horizontal axis represents time from beginning of career to time of the analysis. The vertical axis represents number of sequences. Careers are sorted from longer (no 1 at the bottom of the data region) to shorter (no 84 at the top).

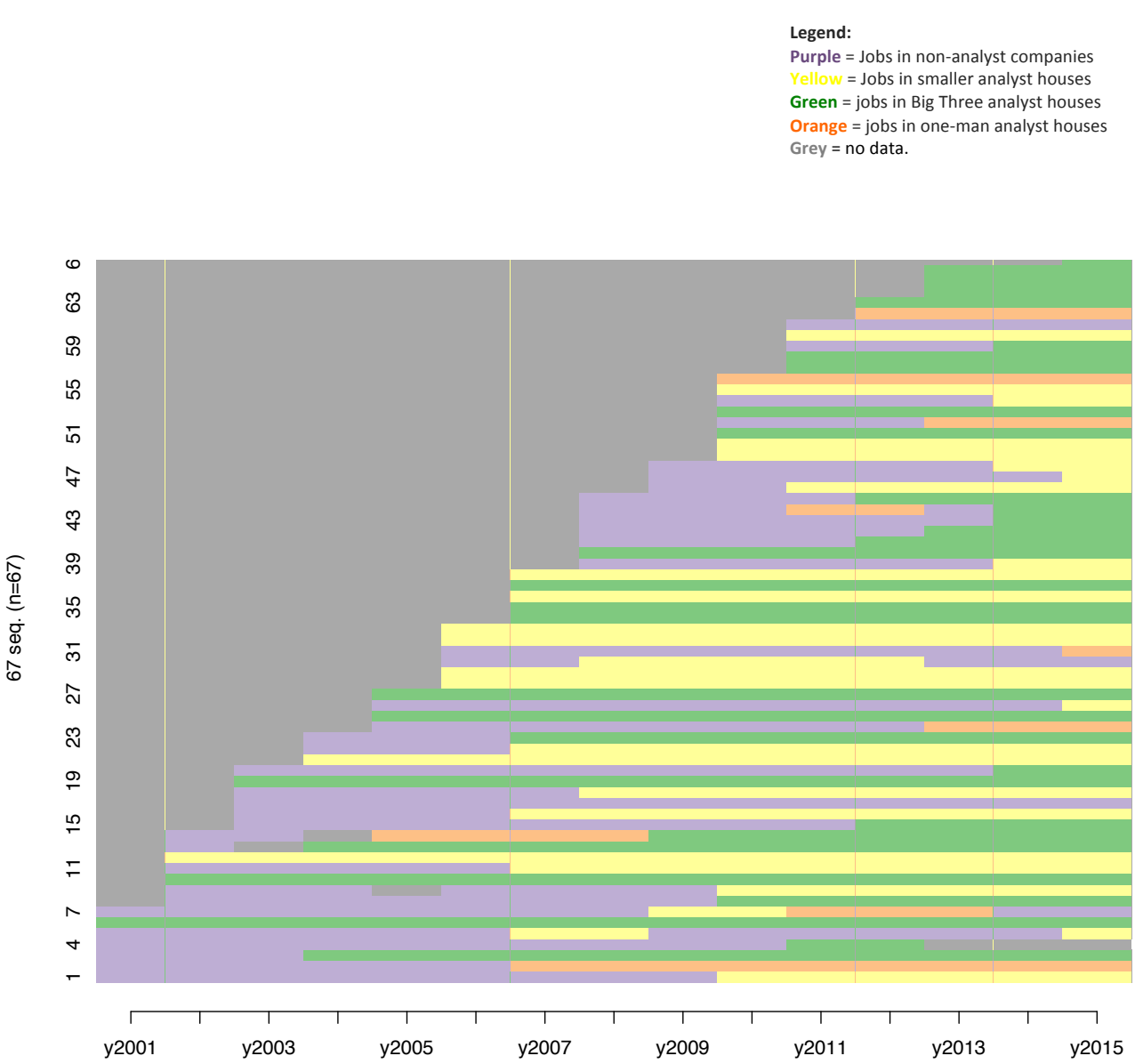

Figure 3: This visualization represents careers started after 2000. Horizontal axis represents time from beginning of career to time of the analysis. The vertical axis represents number of sequences. Careers are sorted from longer (no 1 at the bottom of the data region) to shorter (no 67 at the top) to facilitate pattern identification.

Appendix 1 - Categories of companies

We created two categories of companies based on size.

Table 1: Classification of Companies

Size	How many times a company is mentioned in out dataset
Large	More than 12 times
Small	Less than 12 times

We created three further categories of companies based on when in the career people join them. **6 years** is the median number that emerges from querying our database on how long ago people joined their last company (with 71 companies being those people worked for in their last job).

Table 2: Seniority classification

Seniority	Date the company is joined
New	Company joined during the last 6 years of a career
Old	Company joined earlier than in the last 6 years
Oldest	Company nobody worked for in their last job

We produced a classification of companies according to 7 possible states.

Appendix 2 - Normalising Job Titles

The list of normalisations (before the colon) and surface forms (after the colon) is reported below:

ceo: ceo, chief executive officer
founder: founder
president: president
cxo: chief .* officer, cto, coo, cmo, cio
vice president: vice president, vp, vp.
fellow: fellow
director: director, directeur, directrice
manager: manager, mgr, mgr.
editor: editor
consultant: consultant
analyst: analyst

Out of the 1,659 job titles in the four data sets (TOP, CEO, random & ERP expert analysts), we were able to find a short from for 1,279. Job titles were subsequently ranked in the following way.

1.ceo
2.founder
3.president
4.cxo
5.fellow
6.director
7.manager
8.editor
9.consultant
10.analyst
11.vice president (vp)

Appendix 3 - Seniority Matching

One additional step involved identifying the seniority of each job title by matching against a set of keywords. A job title was marked as senior if it matched case-insensitively at least one of the following regular strings:

senior, snr, snr., sr, sr., sn, sn., chief, distinguished, managing, group, general, head, principal, lead

Currently 381 of our 1,659 job titles are marked as senior. To conclude, from an initial sample of 332 profiles (100 random, 100 profiles with ERP expertise, 99 top analyst profiles, 33 CEO profiles), **259** remained after filtering out profiles without any Industry Analyst company in them as well as eliminating duplicates.

Appendix 4 – Overlaps

To manage overlaps of jobs - e.g. one person has two jobs at the same time at some point in her career - we applied the following rules. Jobs related to analyst types of companies always overwrite non-analyst company related jobs. If the overlap is complete, the non-analyst company related job will disappear from our dataset. If the overlap is partial, the non-analyst company related job will be included in the analysis only for the time period it does not overlap with analyst company related job. Out of a total number of 1371 jobs, jobs that are lost from dataset due to our overwriting decision are 64, less than 5%. With other types of overlaps e.g. overlaps of jobs in the same company categories that still partially overlap in time, we decided to keep the job starting earlier. While, with the last job, we decided to keep the job starting later in accordance to our interest for companies that analysts join later in a career.

Appendix 5 – Further explorations of how inter-professional career correlate with career outcome

If we look at number of jobs changed before the inter-professional transition (see Figure 1 below), it is striking that random analysts change more jobs than TOP/CEO and a similar number to ERP experts in a considerably shorter career.

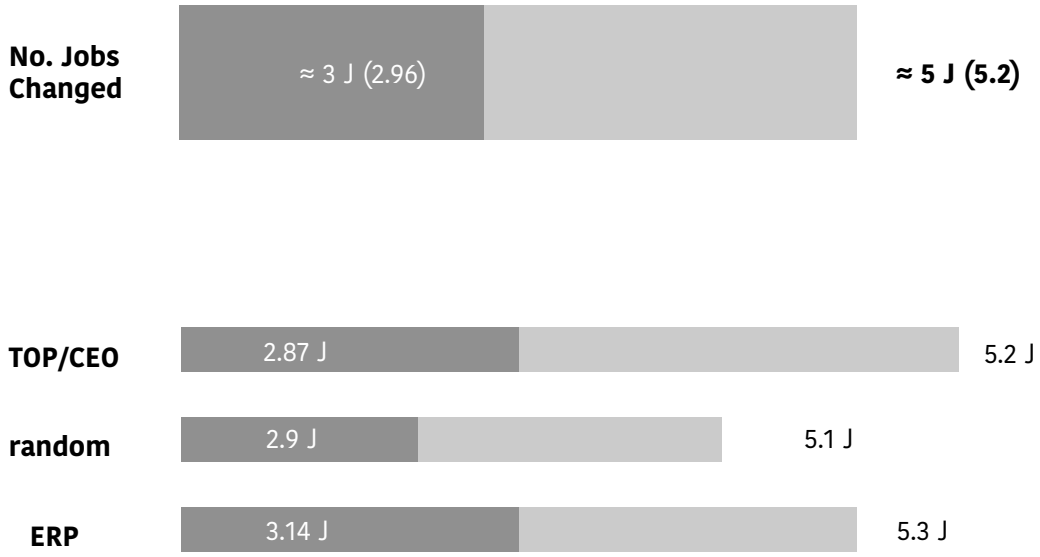

Figure 1. The thicker bar on top of this figure represents analyst average number of jobs changes in analyst career (5.2 jobs) as well as average number of job changes before the inter-professional transition (i.e. 2.96 jobs). The three thinner bars below represent average number of jobs changes by sample (TOP/CEO analysts, random analysts and analysts with expertise in ERP) – please refer to Section 7 in the paper for details about sampling. The darker sections of the bar towards the left represents average number of job changes before the inter-professional transition (e.g. analysts in the TOP/CEO sample change 2.87 jobs before taking up an analyst job). The letter "J" stands for "jobs".

According to our ontology, number of jobs changed before joining the analyst industry might derive from number of job title changes within the same company as well as from changes of company (See discussion in Section 10). In other words, when it comes to job changes before joining the analyst industry, our alphabet does not distinguish between inter- and intra- organizational moves. Still, it can be argued that TOP/CEO having comparable number of job changes in a longer career might point to the fact that long and rigid careers in another industry are conducive of inter-professional success. It takes indeed longer to change job through promotions within the same company than changing jobs through changing company. One way to formalise the notion of rigidity using sequential analysis techniques is to use measures of sequence heterogeneity. Measure of turbulence, for example, calculates the diversity between two sequences (e.g. AAAABBBB & ABAABBAB) by taking into account which one is shifting more from a state to another as well as the continuity of time spent in each different state.