Structure and power dynamics in economic networks

A quantitative analysis of labour flow and company control networks in the UK

Aron Pap

A thesis presented for the degree of
Master in Data Science

Barcelona Graduate School of Economics
Universitat Autònoma de Barcelona and Universitat Pompeu Fabra
Spain
25 June 2020

Structure and power dynamics in economic networks

A quantitative analysis of labour flow and company control networks in the UK

Aron Pap

Abstract

In this thesis project I analyse labour flow networks, considering both undirected and directed configurations, and company control networks in the UK. I observe that these networks exhibit characteristics that are typical of empirical networks, such as heavy-tailed degree distribution, strong, naturally emerging communities with geo-industrial clustering and high assortativity. I also document that distinguishing between the type of investors of firms can help to better understand their degree centrality in the company control network and that large institutional entities having significant and exclusive control in a firm seem to be responsible for emerging hubs in this network. I also devise a simple network formation model to study the underlying causal processes in this company control network. I perform numerical simulations, sensitivity analysis and model parameter calibration, obtaining a set of parameters for the model with which it can approximate reasonably well the empirically observed patterns in the data.

Acknowledgements

I want to thank Omar A. Guerrero and Joan de Martí for their help, insights and guidance for my thesis.

Contents

List of Figures

List of Tables

I Introduction

In this project one of the main interests is in exploring the labour market from a network science perspective, since it can provide a more granular understanding of the dynamics of labour flows between specific companies, industries or regions. Also, this approach allows for modelling these dynamical processes with heterogeneous agents, which is of interest in economic modelling. Understanding the dynamics of these labour flows can also help in predicting the propagation of shocks through the labour market, which is especially relevant in today's world, disrupted by the COVID-19 pandemic. It is even more relevant for the country of interest, the UK, since it is also facing the EU-exit at the end of 2020, which most probably will generate an additional shock to the UK's labour market. Labour flows are also useful for understanding how quickly these shocks might spread through the network and thus help to predict unemployment trajectories. However, in more general terms labour flows offer a practical monitoring tool to follow "skill-paths" in an economy. The uncovered "skill-paths" could help to identify potential for productivity gains in certain sectors, occupations or regions. These examples show that analysing labour flows could provide crucial insights for policy-makers when designing interventions in case of labour market shocks or when implementing retraining programmes to increase productivity. Combined with other techniques like agent-based modelling, labour flow models could serve as counterfactuals to evaluate these policy interventions.

I also analyse company control data, which can help to identify the interconnectedness of firms through their owner structure. These ownership structures might also be related to labour flows between the companies. Community structure and node centrality in such company control networks is of primary interest, because they are closely related to the influence that certain agents might have in the related economic interactions. Since the ownership structure and investor composition of a firm could have significant impact on its future growth prospects, the analysis of such company control networks could have several real-world applications in supporting firms to optimise their ownership structure.

To provide a thorough account of the structure and dynamics of these economic networks first of all a detailed analysis of their empirical characteristics is needed with the appropriate scientific tools. Degree distributions, community structures and dynamical processes of these networks need to be understood. Then these

empirical observations should be explained by causal models in order to be able to design better mechanisms for the functioning of these economic networks.

This thesis document evolves as follows: in the later parts of the current section I the relevant literature is reviewed, then I elaborate on the special aspects of analysing data on officers and investors. In the beginning of section II the data collection method is introduced, then some descriptive statistics are provided regarding the labour flow networks, the company control network and other company characteristics and then I proceed to a more detailed network analysis. Finally in section II.4 the results of the empirical analysis are summarised. In section III.1 the main empirical observations and stylized facts of the network analysis are discussed again and highlighted. Then in the subsequent parts of section III a theoretical model is devised with the goal of being able to explain the empirical results based on microeconomic first principles. Finally, I conclude in section IV with the main results of the project.

I.1 Literature review

Labour flow networks

The term *labour flow network* is introduced in (Guerrero & Axtell, 2013) as a network analysis approach to understand the heterogeneous dynamics and frictions of labour markets. It is a well-suited computational and network science methodology to inform labour market policies. The building blocks for such a network are the employee-employer matched records, which represent the employment histories of workers. These type of matched records are usually available at statistical agencies or bureaus in every country and can be used to empirically test the validity of, or estimate certain parameters of the labour flow networks. Such analysis is done in (Axtell et al., 2019) for the universe of workers in Finland, along with comparing it to the predictions of an economic model. That paper shows that frictions in the network create correlation in the hiring behaviour of firms and that the aggregate unemployment depends strongly on the labour flow network topology. Also utilising the empirical employee-employer matched records from Finland, the authors show in (Guerrero & Lopez, 2015) that aggregate matching function models cannot explain the empirically observed labour flows, which further strengthens the view that the alternative, network science methodology might be more suitable to account for the labour market frictions and flows. Furthermore, the *labour flow network* methodology can help policy-makers to use more realistic assumptions and models for their decision-making, since as discussed in (Guerrero & Lopez, 2017), simulations show that classical models tend to underestimate the unemployment effect of certain shocks to the economy, whereas network-based models could more accurately capture the patterns of shock dissipation through labour markets.

Job search on networks

There are of course other approaches to accommodate networks in labour markets and an important such idea is the referral-based job search networks. Testable implications of such models are derived in (Dustmann et al., 2016) and tested on unique empirical employee-employer matched records. The authors find suggestive evidence that workers earn higher wages and are less likely to leave their firms if they obtained the job through referrals. These effects decline with tenure at the firm, which suggests that referral-based job networks help firms to learn about skills

of the workers and contribute to productivity gains in the labour market. A similar approach is taken in (Glitz, 2017), where coworker networks in labour markets are analysed with empirical data for Germany. Using the exogenous variation due to massive lay-offs, the author finds that the employment rate of a worker's former coworkers has significant effect on his/her re-employment likelihood after the lay-off. This result also shows the strong effects of networks for employment/unemployment. In an even more recent study (Glitz & Vejlin, 2019), the authors show again that coworker referrals play a substantial role in the job search process and they also perform counterfactual simulations which provide some evidence that wages and productivity would decrease without these referral markets due to information deficiencies and less efficient learning about the characteristics of agents.

Community detection in networks

One of the aims of the current study is to explore the community structure of the labour flow and company control network in the UK, therefore the literature of *community detection in networks* is studied extensively here. A comprehensive survey on community detection algorithms can be found in (Fortunato, 2010). For theoretical purposes, several algorithms were considered, however due to practical matters and computational constraints, finally the emphasis is given to efficient methods that also work on large networks like the labour flow and company control networks. Probably the most well-known and commonly used such method is the Louvain-method, introduced in (Blondel et al., 2008). This heuristic method is based on modularity maximization. This is an agglomerative method which starts from isolated nodes and then merges them into communities if that increases the modularity[1]. Then in the next iteration of the algorithm the newly-formed communities are considered as nodes in a new, induced network and then the same steps can be applied at that "higher-level" as well. One of the main reason for the algorithm's computational efficiency is that the change in the modularity due to a "merging step" can be easily computed as:

$$\Delta Q = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot+k_i}}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

where $\sum_{in}$ is the sum of the weights of the links inside the current community, $\sum_{tot}$ is the sum of the weights of all links of nodes in the current community, k_i is the sum of the weights of the links of node i (its degree), $k_{i,in}$ is the sum of the weights of links between node i and the nodes in the current community and m is the sum of all the link weights in the network. A similar expression can be used when considering node removal from a certain community as well. The authors then apply their algorithm for a Belgian mobile phone network with French- and Dutch-speaking nodes. The quality of the community detection (measured by modularity) is high compared to other methods on different benchmark datasets and the computational time of the Louvain-method clearly outperforms the other considered algorithms. The Louvain-method works for undirected and potentially weighted networks, but it is not directly applicable to directed networks. Some of the configurations that I am going to study are directed, therefore I am interested in community detection

[1]Modularity measures the difference of the link density of the community from what this link density would be in case of a randomly rewired network. If modularity is high, that means there is a dense subgraph, which is unlikely to emerge if there would be no community structure in the network.

algorithms for directed networks as well. Several modifications are proposed for the Louvain-method, some of which also make it applicable to directed networks. One such example is the Leiden-method (Traag et al., 2019), which I utilise for the directed network analysis. An insightful empirical example for community detection in labour flow networks is in (Park et al., 2019), which is closely related to what I aim to do in this project as well. The authors in that paper use LinkedIn data with the employment history of more than 500 million users over 25 years, together with 4 million firms globally. They use the Louvain-method for community detection in the network, which reveals hierarchical structure and strong geo-industrial clustering. They also find suggestive evidence that labour inflows to these geo-industrial clusters are linked to the growth (measured by market capitalization trend) of these clusters and the relationship is stronger than if one uses traditional, administrative aggregation units. These results provide useful benchmarks for my current work on labour flow networks as well.

Scale-free networks

Empirically observed networks seem to have some common characteristics, which might be due to some universal underlying organising principle, which was first documented in (Albert & Barabási, 2002), a highly influential paper in network science. One of the main observations is that empirical networks tend to have power-law degree distribution (that is: $\mathbb{P}(d) \propto d^{-\gamma}$, where d is the degree of a node and γ is the exponent of the degree distribution). The labour flow and company control networks also have heavy-tailed degree distributions, therefore the results from (Albert & Barabási, 2002) seem important, especially their preferential attachment model, since some variation of that could be present in both labour flow networks, but even more in company control networks. A detailed overview of power-laws in networks is given in (Clauset et al., 2009) as well and the authors also provide derivations for a maximum-likelihood estimation method for the exponent of the degree distribution, which I use in order to fit power-law functional forms to the empirical degree distributions of labour flow and company control networks.

Network formation models

Network science is an empirical discipline of science by definition (as it was its main distinguishing factor from graph theory in the early days). However after observing some interesting and useful empirical results, it is a natural next step in understanding, to try to come up with models which can reproduce the results documented in practice. Network formation models are part of this scheme, however as network science is a highly multidisciplinary field, there are several approaches for devising generative models, mainly stemming from physics and economics. In this literature review section I focus on the economics-related network formation models (as it is more relevant for the current work), nevertheless keeping in mind that these have some overlap with the physics generative models as well.

A comprehensive survey paper on stochastic network formation models is by (Pin & Rogers, 2016). In this paper the authors survey one-shot models with a fixed population, growing random network models, dynamical models with a fixed/steady-state population and they also extensively investigate the economic literature on homophily. For my purposes, the surveyed growing random network models and the dynamical models are most useful. But for a particular application one might need a combination of the characteristics of the introduced models, there is no universal recipe. However, the authors do provide a general framework to keep in mind when

constructing network formation models, which also turns out to be useful for this project:

- An understanding of the properties of random formation processes
- A set of theoretical frameworks with which to model the incentives of agents and understand their optimal behavior
- Empirical work that identifies the relevant characteristics of agents and their environments, to better understand their decisions
- Structural work to estimate the resulting models

Another highly influential paper is (Bala & Goyal, 2000), in which the authors introduce noncooperative models of social/economic network formation, where agents have some specific costs and some potential benefits from establishing links. They derive the architecture of equilibrium networks in these models and they also analyse the social efficiency of these equilibria. Certain aspects of these models inspired my work as well, since in my network formation models there are some steps which are noncooperative (e.g. when an investor sells his/her stake in a given company), but this model also has cooperative elements, when decisions are based on mutual agreement (e.g. when an investor acquires significant control in a given company).

A recent survey paper on the econometric models of network formation is (de Paula, 2019). The author also starts off with discussing random graph models, detailing the general class of exponential random graph models. Then dyadic models are introduced, where the formation of links is usually a Bernoulli trial, whose mean is dependent on node characteristics. For instance, in the formulation of (Dzemski, 2018) from the mentioned survey paper:

$$G_{i,j} = 1_{\left(X_{ij}^T \beta + \alpha_i^{out} + \alpha_j^{in} + \epsilon_{ij} > 0\right)}$$

where G is the adjacency matrix, X is the matrix of dyadic covariates, 1 is the indicator function and $\epsilon \sim \mathcal{N}(0,1)$. It turns out that parameter estimations in these settings can be carried out using tools from **panel econometrics**. The performance of estimations is usually tested with simulations and empirical applications. One of the main improvements was the introduction of "tetrad logit", which offers a conditional maximum likelihood estimator. Influential models building on the assumption of pairwise stability are also introduced, as well as incompatible/incomplete models and subgraph generation models. Strategic network formation models are also prominent in the literature and well-established through game theory principles. For iterative strategic network formation models, Bayesian estimation techniques are used with a Markov Chain Monte Carlo procedure and empirically tested on the AddHealth benchmark dataset. All in all, this survey paper is an important reference point for my project and for my final network formation model, since that involves both random and strategic elements from the more simple models discussed here as well, nevertheless building on robust microeconomic principles.

Company control networks

Since I have data on the control structure of firms in the UK, I am also interested in this literature. In (Battiston & Catanzaro, 2004) the authors analyse the statistical properties of corporate board and director networks with empirical data from the

US and Italy. They find several similarities across the different countries and also persistent characteristics over time. The main observations are that these networks are "small-world" (small average shortest path, surprisingly small diameter), assortative, highly clustered and have giant maximal connected components. Moreover, the degree distribution seems to follow a power-law here as well (considering the one-mode projection of the original bipartite network). These results are aligned with my findings for the UK, except for the giant component, but this is probably due to the fact that here the authors only analyse sufficiently large companies, whereas my analysis also includes small enterprises.

In (Battiston, 2004) the authors analyse further company control networks, but now focusing on the topologies of shareholding networks, which introduce "interpretable" weights to the networks. Two new metrics are introduced to compactly represent some characteristics of these networks, these are the **number of effective shareholders of a stock** and the **number of companies effectively controlled by a single holder**. These quantities are insightful for me, since I derive similar quantities for the company control network. The empirical analysis with American and Italian data reveals interesting inner structure of these capital control networks, but also substantial difference between the US and Italy. The Italian market can be partitioned into several separated groups of interest, whereas the US markets is characterised by very large holders sharing control on overlapping subsets of stocks.

Then in (Vitali et al., 2011) the authors extend the analysis to the global level and they investigate the architecture of the international ownership network, along with computing the control held by each global player. The paper finds that a significant share of control flows to a small, strongly-connected core of financial institutions. These results are interesting for my work as well, since some of the financial institutions highlighted in the paper (e.g. Lloyds Plc.) are also found to be important for the UK control network and these strong players can also help their holdings to become more "central" in the company projection network.

Finally in (Battiston et al., 2003) the authors go one step further and they analyse the decision-making dynamics of these board of directors, focusing on how the topology of the projections of the original bipartite control graph affects the decision making dynamics. They derive some indirect network characteristics which turn out to be good predictors for the outcomes of dynamic decision-making processes (applying it to empirical data from the US). This is an illuminating result from my point of view, since it provides another real-world example when some non-trivial quantities of networks can substantially enhance the understanding of a social phenomenon.

I.2 Special considerations for officers and investors

As it is highlighted throughout the literature review, there is significant work done for labour flow networks in general and also for networks of board of directors. There are important conclusions to be drawn from these results, which can inform my current work as well. However it needs to be taken into consideration that my empirical data is unique. It is unique in the sense that I only have employment history and thus labour flow data for officers of companies and not workers in general. One could argue that officers have more expertise in one particular industrial area and therefore less likely to change industries than the general population of workers. On the other hand, one could also argue that officers have such general management skills,

which can be useful across all industries, therefore they should be more likely to change industries. A similar dilemma arises when thinking about regional mobility, since officers might have really strong social ties in a particular geographical area, while they might also have more financial resources in order to take the decision of relocation if a better job offer is presented to them. These examples indicate the strong limitations to how the empirical results of this analysis might reflect the general population of workers.

Regarding the data that I collect on the company control network, there are also significant limitations there, since I get the list of all legal persons with significant control in a given company, however I do not have access to their quantified share or stake in the company, therefore all the related empirical analysis has a "discrete" nature instead of the continuous shareholding observations presented in (Battiston, 2004), for example.

With these considerations in mind, the exploration of structure and power dynamics in economic networks might begin.

II Empirical Analysis

First of all a clear definition is needed for some of the "agents" whose actions and activities correspond to the observations in the data. **Officer** of a firm here means any employee or stakeholder, who has significant role in the management of the firm, such as directors or secretaries. **Persons with significant control** are natural persons or legal entities that have significant financial or decision-making influence in the company (for instance owning 50 percent of the company's shares). Using the previous definitions I can now define the objects of interest, the **labour flow network** and the **company control network**. For the labour flow network (denoted by LFN from now onwards), I consider and analyse 2 versions, an undirected version and a directed version.

Undirected LFN: A weighted network where each node is a company, which appeared in the employment histories of the officers at least once and an edge is created between two companies A and B if:

- An officer started working at company B after leaving company A, with no other employment inbetween (and vice versa)
- An officer started working at company B while working for company A (and vice versa)

If there is already an edge between A and B and there is a new "link forming" observation, then the weight of the edge between A and B is increased by 1.

Directed LFN: A weighted network where each node is a company, which appeared in the employment histories of the officers at least once and an edge is created between 2 companies A and B if an officer started working at company B at most M months ($M = 2$ in the current implementation) after leaving company A, with no other employment inbetween.

If there is already an edge between A and B and there is a new "link forming" observation, then the weight of the edge between A and B is increased by 1.

A small example to illustrate how these networks would be constructed from empirical data is shown on figure 1. Both the undirected and the directed LFN-s are shown for three cases, with one, two or three officer observations subsequently.

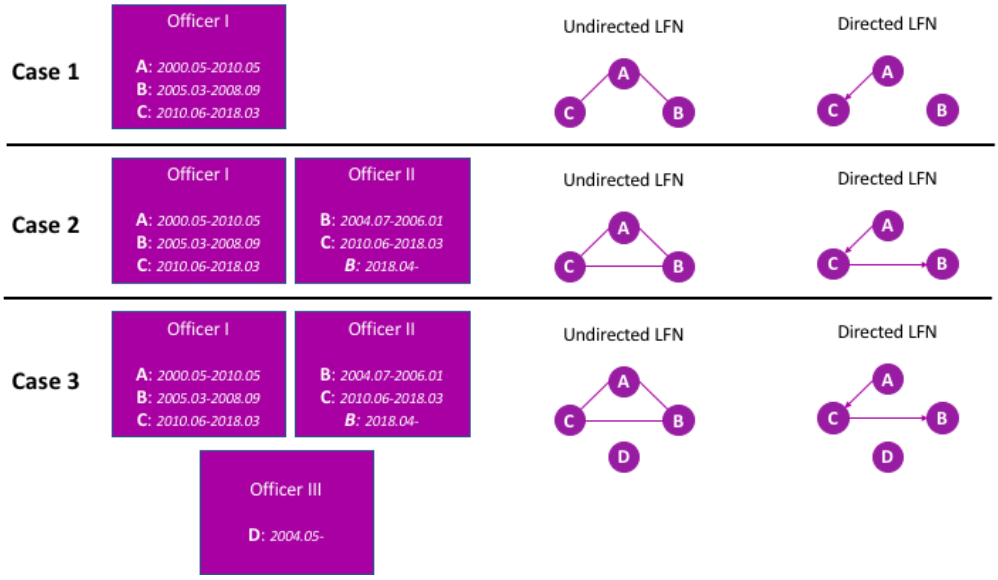

Figure 1: Example labour flow networks based on employment history observations

Company control network: A weighted, undirected network which is constructed from the bipartite graph of company control observations, using its one-mode projection onto the companies. There is a link between company A and B in this projection network, if there is at least one legal person with significant control in both companies and the weight of the link is the number of such legal persons with significant control in both companies. The legal persons can either be natural legal persons (any people who invest in the company), or institutional legal persons such as other firms or other organizations (e.g. charities, associations). Significant control means that the legal person has significant financial or decision-making influence in the company (for instance owning 50 percent of the company's shares).

An example to illustrate how the one-mode projection is done is shown on figure 2, where the **Company projection network** is the final object of interest.

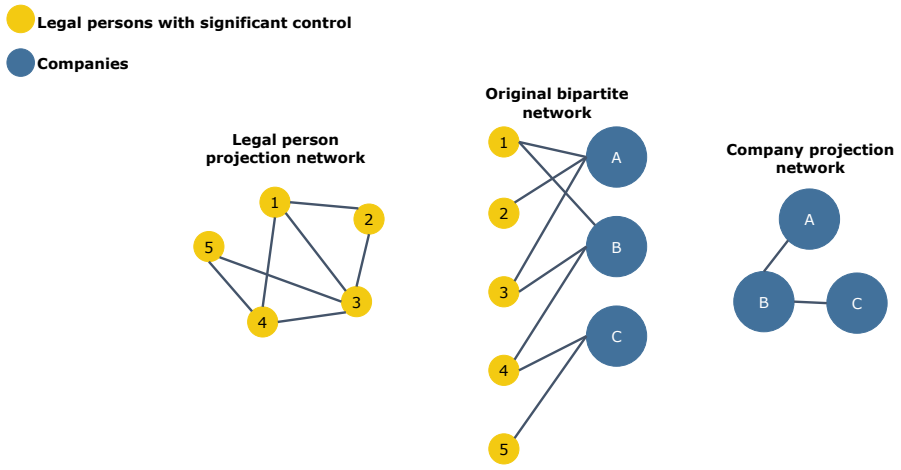

Figure 2: Example of the company control network, a projection of the bipartite control network

There could be other ways to define these networks, but for the questions that this project investigates, the definitions presented above seemed to be most practical and insightful. Also, it might not be immediately clear why it is useful to analyse both the directed and the undirected network as different objects. However, the subtle differences between these approaches might bring interesting conclusions, since the directed network is expected to be closer to the "general" labour flow networks in terms of characteristics, whereas the undirected version is expected to be more specific to officers. The reason for this is that most of the workers usually only have one job at a time and they only move to another company after (or while) leaving their previous employer. The definition for a link in the directed labour flow network captures exactly this type of patterns. On the other hand, officers are more likely to have several positions at the same time, serving as directors or board members of different companies. Then it is reasonable to argue that there should be connection between the companies co-directed by the same person. Since for example this person will know about vacancies at both firms and will also know about employees looking for a career shift in both firms, therefore this person could act as a "facilitator" for a potential labour flow between the two companies. These kind of connections are only captured by the undirected labour flow network. Thus the undirected labour flow network might explain the special characteristics of the labour market from the officers' perspective, while the directed labour flow network might provide more insight into how average workers move through the job market.

II.1 Data collection

For this project I utilise the CompaniesHouse database[2] in novel ways. First of all, I use their API service to get data on general company information (e.g. unique company number, date of creation, industry by Standard Industrial Classification code, region of headquarters, insolvency indicator, indicator whether the company is active or not and company legal type variables) and also to get data about the legal persons with significant control in each company. To extend these data sources, I also implement a web scraper in Python to be able to collect the employment history of each listed officer of the firms on the CompaniesHouse database. The web scraper and the API collection script is constructed in such a way to respect the rate limiting rules of the CompaniesHouse. All the data collection procedures were conducted through Amazon Web Services, using EC2 instances as virtual machines.

Since the rate limiting of the service provider and the large amount of available data on CompaniesHouse prevent users from being able to collect all the data, I also have to choose a sampling procedure. Due to practical considerations, a **snowballing**-type of sampling procedure is chosen, since I start with collecting data on some randomly chosen companies, but in the next round, I consider companies which had direct connections to the firms of the previous round, in terms of labour flows of officers between those companies. This also means that the procedure would stop once it hits the "boundaries" of a connected component in the officers' labour flow network. In such case, the data sampling algorithm would randomly choose a new company to explore from the list of currently still unexplored firms and then the snowballing-procedure would continue.

[2] An executive governmental agency in the UK, registering company information and making it available to the public.

Note that this sampling design might introduce some bias to the results[3], but since there is still a significant portion of UK companies not listed on the CompaniesHouse database, the results would not be representative of the whole UK economy, even if all the data from the CompaniesHouse could be collected. The random restarts in the sampling design can help to explore a huge variety of the "space of companies" in the UK, but still being able to get a detailed view of interconnections between and within neighbourhood of companies due to the snowballing procedure.

II.2 Descriptive statistics

In table 1 some descriptive statistics about the analysed networks are provided. Note that in case of the directed LFN, X/Y refers to in- and out- quantities respectively and the number of components measures the number of weakly connected components. Also note that the reported degree assortativity for the directed LFN is the average of the values for all in-out combination. For fitting degree exponents to the degree sequences a power-law functional form is assumed[4] and the maximum likelihood estimates are reported here (the technical derivation for this estimation is provided in the Appendix A), following the equations from (Clauset et al., 2009). Also note that the reported summary statistics refer to the analysed networks after removing isolates (nodes that are not connected to anyone[5]). For example, initially the number of nodes in the undirected and directed labour flow networks is the same, but due to the different definition of link creation in the two networks, more nodes will be isolates in the directed labour flow network and that is the reason why these networks differ in these fundamental characteristics as well.

Col.: Networks *Row: Characteristics*	**Company control network (undirected, weighted)**	**Labour flow network (undirected, weighted)**	**Labour flow network (directed, weighted)**
Number of nodes	78,844	175,501	57,027
Number of edges	463,536	2,375,811	48,666
Number of components	15,280	68	11,088
Size of giant component	1285	174,060	26,649
Max edge weight	24	20	7
Average degree	11.76	27.07	0.85/0.85
Fitted degree exponent (γ)	1.60	1.36	7.46/5.39
Degree assortativity	0.93	0.35	0.09
Regional assortativity	0.64	0.39	0.2
Industrial assortativity	0.33	-	0.12

Table 1: Summary statistics and general characteristics of the studied networks

[3]For example the average degree might be overestimated this way, since when the next company to consider is chosen during the snowballing procedure and not the random restart, then high degree firms are more likely to be chosen than firms with a low degree.

[4]The minimum degree is set equal to 1, instead of optimizing that parameter of the distribution as well, as some authors do.

[5]Neither in- nor outgoing links in case of the directed network.

Also note that the - symbol in table 1 represents that the corresponding value could not be calculated due to its computational cost (industrial assortativity for the undirected labour flow network). Based on the values in table 1 and also based on the goodness-of-fit tests not shown here, a pure power-law would not be an optimal fit for any of the degree distributions. But it is also documented in (Barabási et al., 2016) that heavy-tailed distributions are rarely approximated well with pure power-law distributions, but rather they need some modifications, for example exponential cut-offs. Since this issue is not the main focus of this work, these modifications and alternatives are not explored here. But more generally, heavy-tailed degree distributions already imply that there are some hubs in the networks. These hubs in economic networks are important since they are orders of magnitude larger than other actors in the network, therefore they have a crucial role in any kind of activity that takes place on this network. For example in financial networks hubs are usually classified as *systemically important financial institutions*, since their operations are essential for the functioning (in terms of liquidity for instance) of the financial network and the failure of these hubs might result in a cascade of failures throughout this system. This example illustrates that the existence of hubs in economic networks implies substantially different system characteristics in terms of robustness, spreading processes or average distance between nodes and therefore economic networks with hubs require special considerations and analysis. The labour flow and the company control networks both have hubs, thus they are also expected to exhibit the characteristics mentioned previously.

It is also interesting to see that the assortativity coefficients are quite large in most cases that are considered here. Assortativity measures whether similar nodes are more likely to link to each other or not. For example, the high degree assortativity in the company control network suggests that companies with a large number of investors usually share at least some investors, whereas small companies might be connected by one or a few small, but locally influential investors. The positive and quite high regional and industrial assortativity shows the first signs that there might be substantial "geo-industrial clustering" in the company control and the labour flow networks. Assortativity might also be useful to derive relevant features for link prediction in networks.

The age distribution of firms in the sample is shown on figure 3:

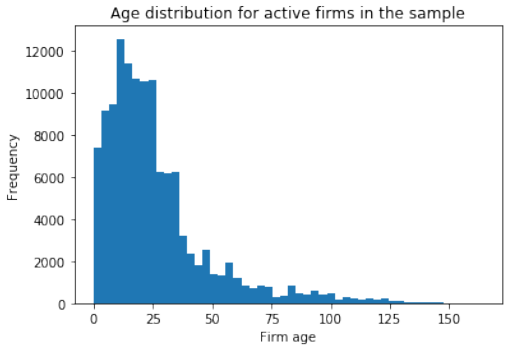

Figure 3: Age distribution of active companies in the sample

The mode of the distribution for the age of the firms in the sample seems to be around 15-20 years, then it decays more or less exponentially. The best continuous

parametric fit to the distribution of firms' age is a **generalized inverse Weibull** probability density function, shown in the Appendix (figure A.2).

There are some hubs both in terms of regions (clearly London is the largest one in the case of the UK) and industries in the sample, which contain large amount of companies, whereas most of the regions/industries only contain a moderate amount of firms. I visualise this observation through rank plots[6] on figure 4, first going until rank 200 then going only until rank 10. It is interesting to see that London (as the region with rank 1) is still an outlier on this rank plot. But apart from the rank 1 observations, the "shrinking factor" of both the regional and the industrial rank plot might be approximated reasonably well through a power-law. The same analysis with the actual names of the 10 largest industries and regions can be found in the Appendix, on figure A.3.

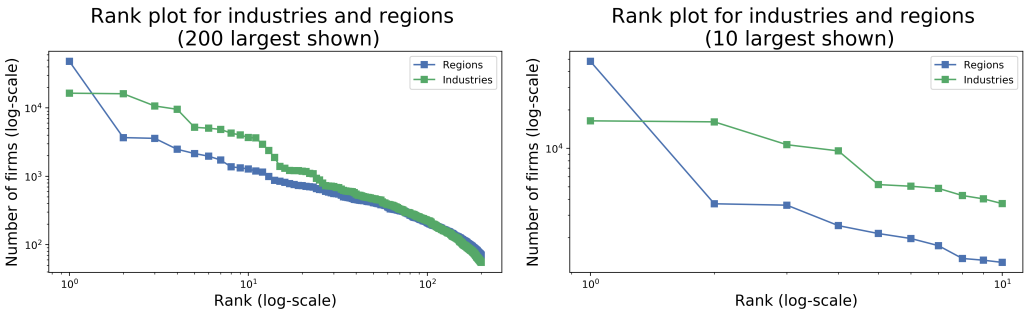

Figure 4: Firm distribution across regions and industries in the sample

II.3 Network analysis

In this section the labour flow and company control networks are analysed from different aspects. Several empirical characteristics of these networks are reported, visualised and discussed.

II.3.1 Undirected labour flow network

Degree distribution

The undirected labour flow network has several "hub" companies as it is exhibited on figure 5. This means that there are certain companies which attract and then "distribute" officers who are active in several companies and who traverse a significant portion of the network of companies throughout their careers. This also suggests that there are large groups/families of companies, among which officers might flow. But it might also be noted that the majority of the companies are only connected to a few others, which means that most of them only have a few officers, who might spend their entire career at those particular firms. This result might also be explained by administrative reasons, since numerous small enterprises are present in the data, with their founders being their only officer (self-employed individuals).

[6] A popular way to present power-law distributions.

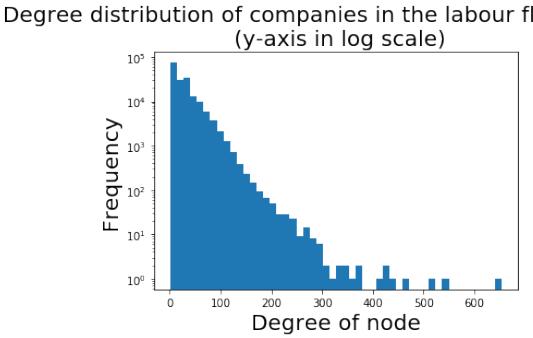

Figure 5: Degree distribution of the undirected labour flow network

Edge weights

The weight of the edges between companies can also be approximated with power-law distributions as it is depicted on figure 6 (log-log plot with logarithmic binning, since that is the correct approach to plot power-law distribution based on (Barabási et al., 2016))[7], but the decay is quite fast in this case. Most of the edges have a small weight (1 or 2) but there are a few which have 20+ weight. Based on manual inspection, these extreme values might be due to reorganisation efforts within a group of companies, controlled by the same holding company for instance.

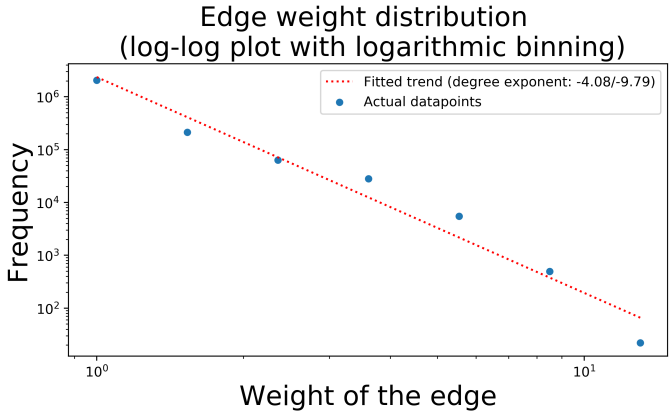

Figure 6: Power-law fitted to the edge weight distribution in the undirected LFN

Connected components

In case of the undirected LFN, there is one giant component, which is extremely large compared to the other connected components. This might also be due to the data collection procedure, nevertheless it shows how interconnected the world of company officers and directors is.

Community detection

For community detection, I used the **Louvain-method** and the **Infomap method**, which resulted in quite similar communities, therefore in the subsequent analysis what is shown is the result of the Louvain-method.

On figure 7 the sizes of the detected communities are shown. These community sizes exhibit interesting patterns, since the majority of them are single-firm commu-

[7]The fitted trend means the degree exponent of the assumed power-law distribution, which is estimated in 2 different ways: fitting a linear trendline to the log-log plot with logarithmic binning/using the MLE estimate that is described in the Appendix.

nities, but then there is a long "plateau" of communities with sizes $\in [0, 2000]$ and then there are a few large communities with more than 3000 firms.

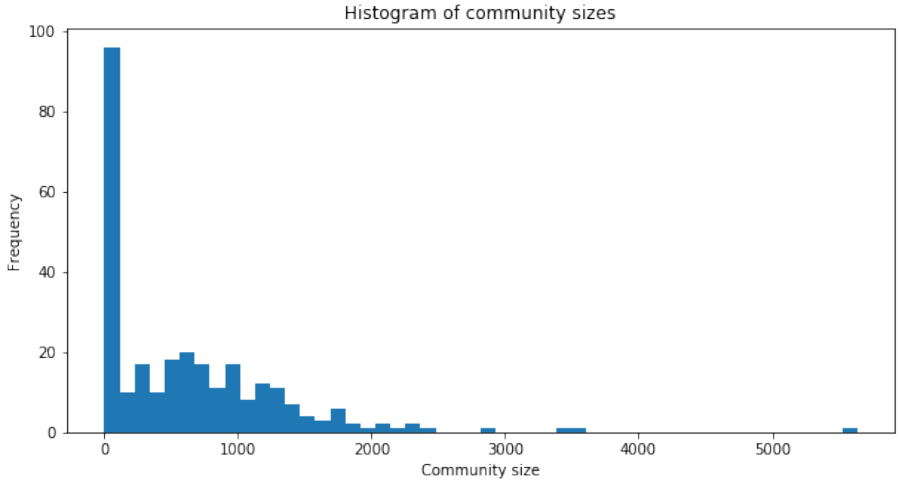

Figure 7: Community size distribution in the undirected LFN

Micro-level analysis of these communities might also provide some hints for the general local characteristics of the network, as shown on figure 8 for an arbitrarily chosen community for illustration purposes:

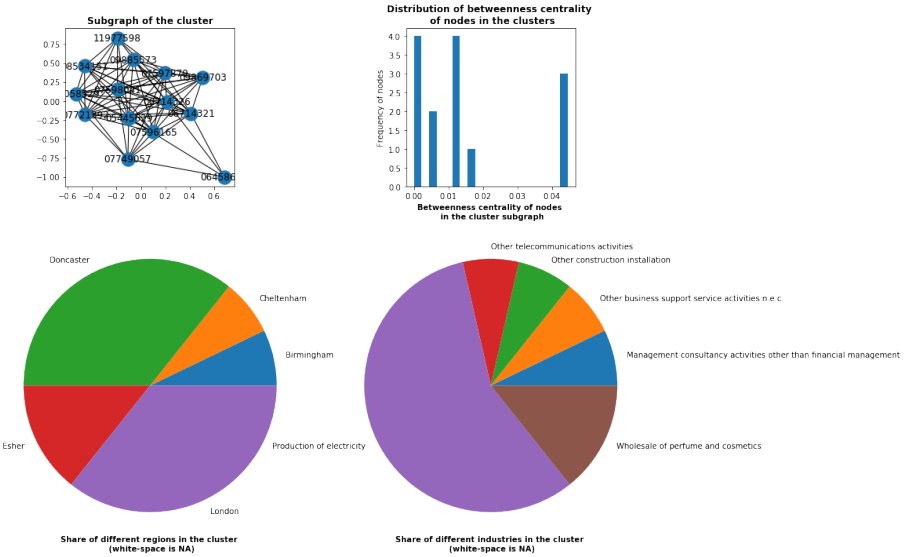

Figure 8: Community micro-analysis example in the undirected LFN

It is depicted well on the subgraph of the cluster that the nodes are quite interlinked and that there is no central company which connects otherwise disparate parts of the network. This is also visible from the histogram of the betweenness centralities of the nodes in the community, where a significant portion of nodes have high betweenness. However, strong **geo-industrial clustering** is present, since 1-2 regions and industries account for a majority of the nodes (*London, Doncaster* and the *production of electricity* respectively). This provides an indication of the "geo-industrial clustering" that is aligned with the global assortativity values, shown in 1.

The induced network of community detection is shown on the left plot of figure 9, where each node is a community and the weight of the links is equal to the sum of the links between the firms of the different communities. The size of the nodes represents the community sizes and the color scale represents the degrees of nodes in the induced network. It is evident that there are hubs among the communities as well and that large communities are also tend to be more interconnected with other communities (the degree of nodes in the induced network exceeds 100 in several cases), rather than being isolated "universe" in the undirected labour flow network. The right plot on figure 9 highlights the 10 largest communities from the induced network. The width of the depicted edges is proportional to their weight and it is interesting to see that the strongest "ties" are between the third and fourth largest communities.

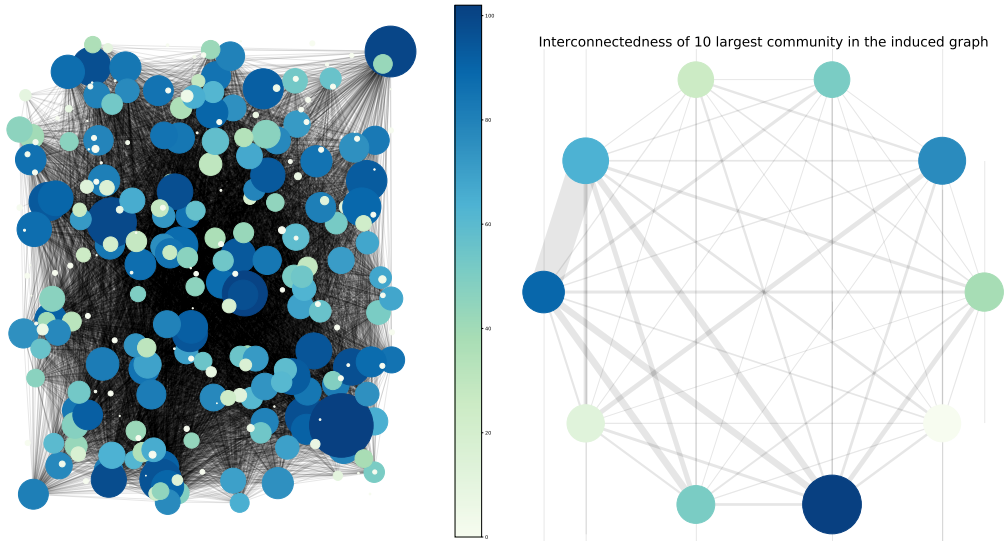

Figure 9: Global community structure in the undirected LFN

II.3.2 Directed labour flow network

Degree distribution

It is also of interest to investigate whether the distribution of in- and outdegrees differ in case of the directed labour flow network. As it is depiced on figure 10, these degree distributions are quite similar, but the out-degrees have more extreme values.

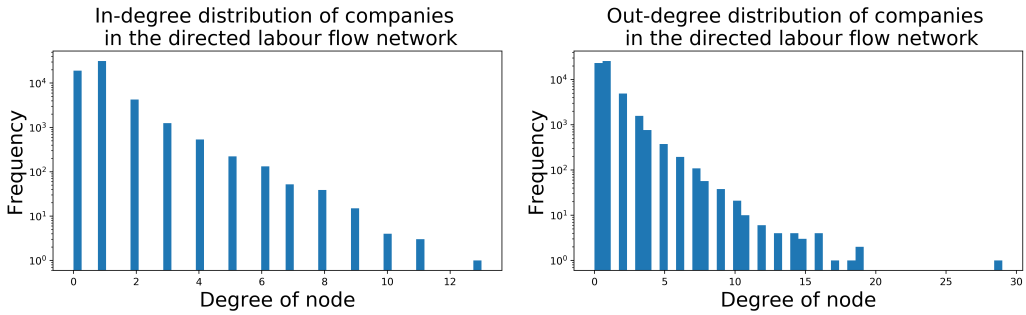

Figure 10: Degree distribution in the directed labour flow network

Edge weights

Regarding the edge weights among firms one important hypothesis is that the strong connections (larger edge weights) should be partitioned into the same communities by any community detection algorithm. Figure 11 reassures this hypothesis, since it is straightforward to see from this figure that the edge weights between communities are small (compared to edge weights within communities):

$$\forall i, j, \text{ such that } C_i \neq C_j \quad w_{g_{i,j}} \leq 3$$

where $w_{g_{i,j}}$ is the weight of the edge between i and j and C_i is the community of node i.

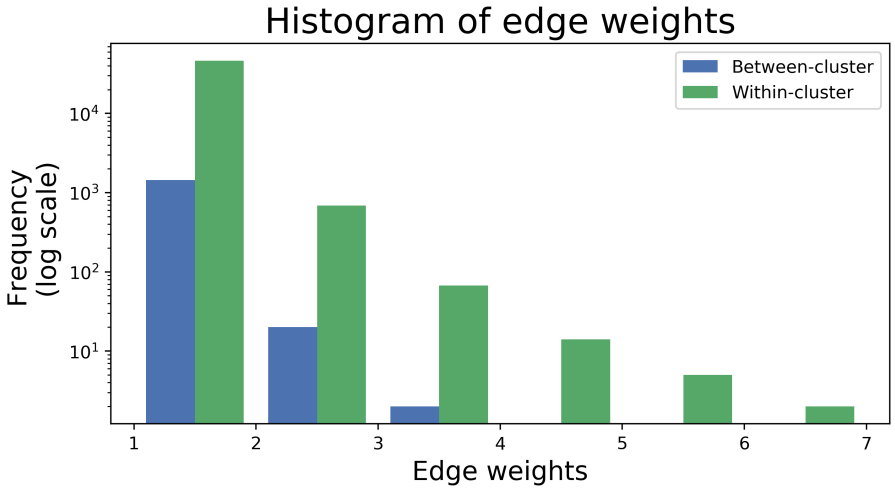

Figure 11: Edge weights in the directed LFN and its variation with communities

Connected components

For the connected components, a similar pattern is observed for the directed LFN as what has been documented for the undirected LFN. Here I consider **weakly connected components**. There is one giant component, along with thousands of small ones.

Community detection

In case of the directed LFN, the **Leiden-method** is used for community detection. The sizes of these detected communities follow interesting shapes, just as in the case of the undirected LFN. These community sizes, along with a micro-community analysis are shown on figure 12.

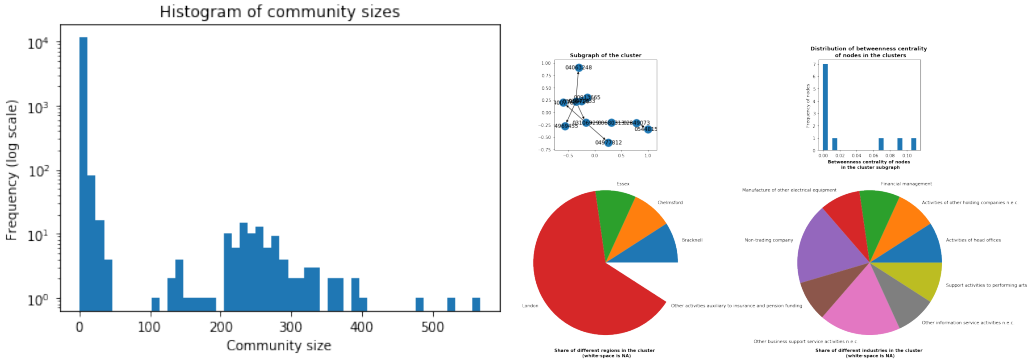

Figure 12: Community detection in the directed LFN

Dynamical analysis

I also carried out some dynamical analysis for the directed labour flow network with the aim to provide answers to the following questions [8]:

1. How the community of a given node/company is evolving over time?

2. How stable/persistent communities are over time?

On figure 13 the characteristics of the community of the same company (which is chosen arbitrarily again for illustration purposes) are shown with all data up to 2005 and then with all data up to mid-2020. It is straightforward to see how the community is growing, but still a particular region/industry tends to dominate its composition.

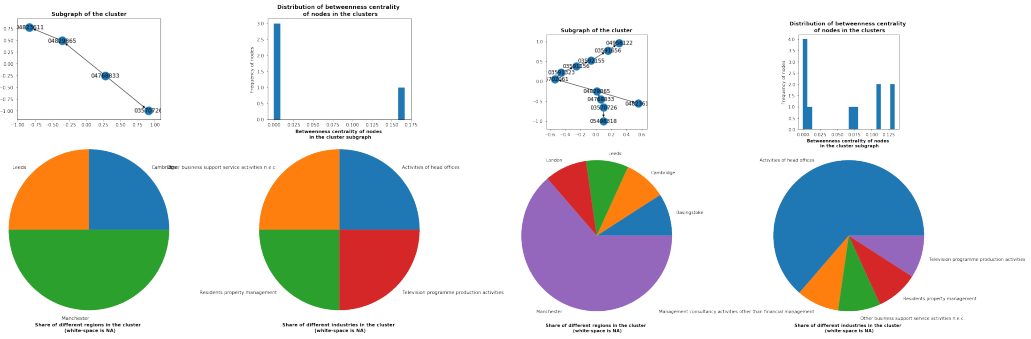

Figure 13: Community evolution of a single node in the directed LFN

But of course the approach discussed above is only applicable at the local, neighbourhood-level, therefore I also use some global quantities, which can concisely indicate the stability of communities over time. Therefore I am running the Leiden-method for community detection at two distinct points in time (2015 and 2020 specifically) and I am analysing the communities at those times. On figure 14 on the left it is shown how the size of the community at the earlier point in time (in 2015) correlates with the number of unique communities the companies "cover" by the later point in time (in 2020), that is for all the companies in the original community, in how many unique communities are they "partitioned into" at the later point in time. For example if we have a community in 2015 with four companies, but then these four companies are actually in two different communities in 2020, then that would result in a dot at the (4,2)-coordinate on the left plot.

A simple global quantity that I am using to measure the stability of the communities will be called "dynamical community stability indicator". The dynamical community stability indicator value is defined as follows for a particular initial community (building on the ideas from the previous paragraph):

$$\mathrm{DCSI}_i^{(t_0,t)} = \frac{S_i^{(t_0)}}{UC_i^{(t)}} \quad \forall i = 1, ..n_c$$

where $DCSI_i^{(t_0,t)}$ is the dynamical community stability indicator value for the i-th initial community between the dates of t_0 and t, $S_i^{(t_0)}$ is the size of the initial

[8] I also analysed whether inflows create more outflows later for a particular company, that is trying to predict out-degree growth from the growth of the in-degree, but this analysis has not revealed robust connections, therefore it is not included in this document.

community at t_0, $UC_i^{(t)}$ is the number of unique communities that the companies from the initial cluster are "partitioned into" at t and n_c is the number of detected communities at the earlier point in time.

Figure 14 shows the results from computing these statistics for a sample of 1000 clusters and not the whole "population" due to computational constraints. It is evident from the figure that the the small communities, which represent the vast majority of all detected communities "stay together", whereas for the larger communities, an approximately linear trend can be seen with a few outliers. This positive relation is expected, but it is unclear in advance whether an exponential, linear or logarithmic trendline might be the best fit. Also, the slope of this linear trend is of great interest, since it might help to forecast the number of new communities emerging from one initial large community. The result that communities with up to 500 companies do not get "partitioned into" more than 20 clusters over a 5 years horizon suggests quite strong stability in this network. The distribution of the dynamical community stability indicator values is highly correlated with the cluster sizes, however a consistent and unbiased estimate of the mean of this dynamical community stability indicator value can be useful in practice. Since this estimate provides an approximation of the "equilibrium size of communities" in the directed labour flow network. It also suggests that each community with a smaller size than this mean value will not fall apart until the next observation, whereas for larger communities, it offers an estimate of how many new communities might be "reachable" from this original large community. One practical application of this analysis might be related to the prediction and optimization of spreading ideas and innovations through networks, when not individual nodes, but communities are the focus of interest. In this case the dynamical community stability indicator values might help to predict the spread of the idea/innovation from some initial nodes[9].

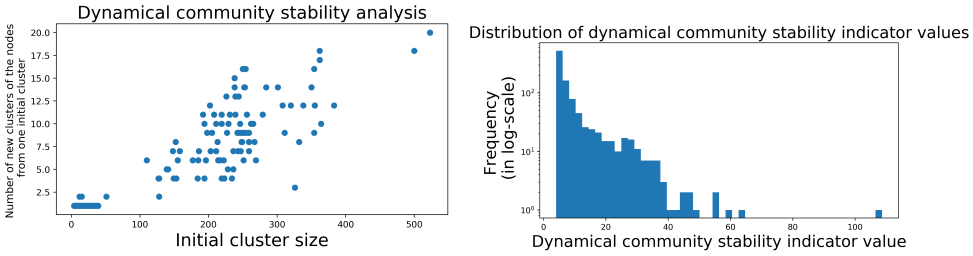

Figure 14: Dynamical stability analysis of communities for the directed LFN

II.3.3 Company control network

Degree distribution

The degree distribution of nodes in the company control network follows a heavy-tailed distribution. This means that most of the companies have small degree in this network, but there are a few hubs, with extremely large degrees. One potential reason for the existence of these hubs is shown on the plot on the right of Figure 15, since some legal persons have significant control in 100+ companies, therefore having such type of investors can contribute to emerging hubs in this network.

[9]Usually called seeds in the related literature.

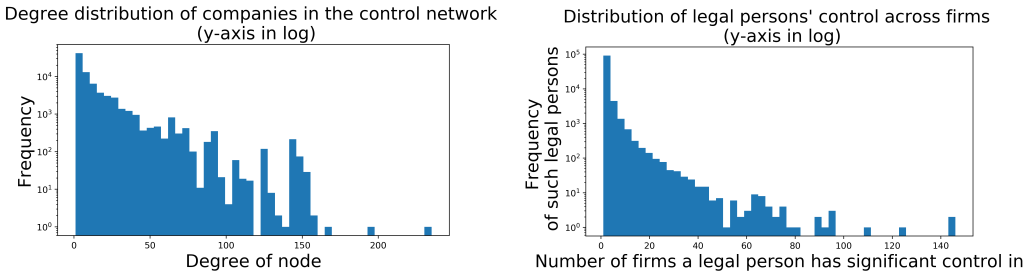

Figure 15: Degree distribution in the company control network

Edge weights

The weight of the edges in the company control network also seems to follow a heavy-tailed distribution as it is shown on figure 16. However, there is surprisingly many edges with 20+ weight. This might be due to the fact that some legal persons might have significant control in all affiliates of a holding company, therefore these edges might connect affiliates of the same holding.

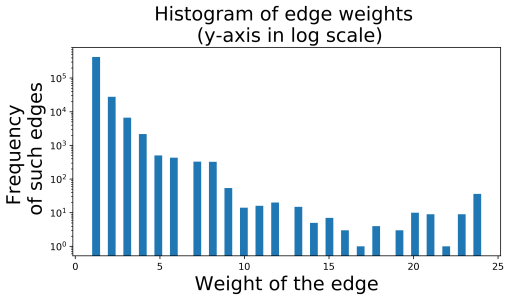

Figure 16: Edge weights in the company control network

Connected components

In this network there are also a few connected components, which are significantly larger than the others, however there is no single giant component, since the largest one only contains approximately 2 percent of all nodes in the network. The histogram for the sizes of these components is shown on Figure 17.

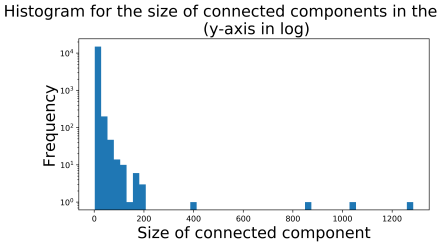

Figure 17: Connected components in the company control network

Community detection

For community detection in the company control network I use the **Louvain-method** and the **Infomap** equation again. However, since both result in communities with similar macroscopic characteristics (e.g. size distribution, number of communities) I focus the discussion on the result of the Louvain-method. The size of these communities seems to decay like a power-law as it is illustrated on Figure

18. Also shown on the figure a micro-analysis for an arbitrarily chosen community, which reveals geo-industrial clustering again, but also shows an example where one company/node acts as a "bridge" between parts of the network and therefore has a significantly higher betweenness centrality value than others.

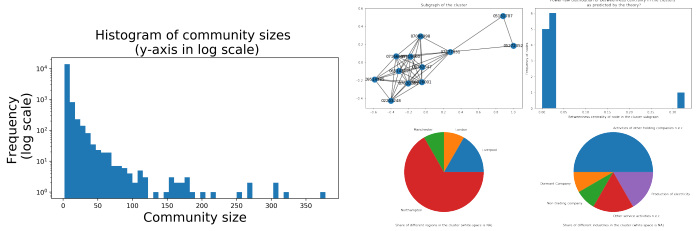

Figure 18: Community detection in the company control network

It is also of particular interest to quantify the "geo-industrial clustering" that has been qualitatively identified through the community detection procedure. The metric that I use for this purpose is the **Herfindahl's concentration index** (also described here (Rhoades, 1993)). This index usually measures how market shares are distributed among companies and provides an estimate for the concentration of the market this way. For example, if there is a single company on a specific market, which is responsible for all the sales on this market, then this market would have a Herfindahl's concentration index of 1. On the other hand, if all firms have equal share on a given market and the number of firms grows to infinity, then the Herfindahl's concentration index goes to 0.

For the analysis here, the concentration is understood in terms of regions and industries, meaning that a set of firms is considered and their respective industries and regions. Then it can be computed how much share a given industry or region has in the set of firms (normalised frequency of each industry or region basically). These shares (normalised frequencies) then can be used to compute the Herfindahl's concentration index, obtaining a regional/industrial concentration index for the set of firms in question. I apply the outlined procedure for all communities which are identified by the Louvain-method for the company control network. On figure 19 on the left plots the size of a community is plotted against the regional/industrial Herfindahl's concentration index for the community[10]. The concentration is decreasing as the communities grow, however even for larger communities, the concentration values are significantly higher than the corresponding global values (that is the regional/industrial Herfindahl's concentration index when using the whole sample of firms as a set).

However, a different experiment and analysis is needed in order to provide suggestive evidence that the regional/industrial concentration values of the communities of the company control network indeed exhibit "geo-industrial clustering". I carry out a randomization experiment to do this, which works as follows:

1. Take the original communities in the network
2. For $t = 1, 2, \ldots T$
    - Choose 2 communities randomly

[10]Communities of size 1 are excluded from this analysis.

- Choose 1-1 firm randomly from the communities chosen in the previous step
- Switch the chosen firms between the communities

3. Output the new, randomized communities and compute the regional/industrial Herfindahl's concentration index for them

This procedure preserves the characteristics of the communities (e.g. the distribution of community sizes) and also the regional/industrial composition of the whole sample. Therefore, if the regional/industrial Herfindahl's concentration index is substantially smaller for the randomized communities than for the original ones, then the data provides suggestive evidence that there is significant "geo-industrial clustering" among these firms. Thus the analysed networks can also help to detect "geo-industrial clustering" as an interesting alternative to standard industrial classification for example.

On the right plots of figure 19 the comparison between the regional/industrial concentration of the original and randomized communities is shown. The randomized communities are obtained via running the outlined procedure for one million iterations ($T = 1,000,000$). It is clear to see from the figure that the randomized communities have lower Herfindahl's concentration index both for regions and industries. Thus this experiment provides suggestive evidence that there is substantial "geo-industrial clustering" in the company control network.

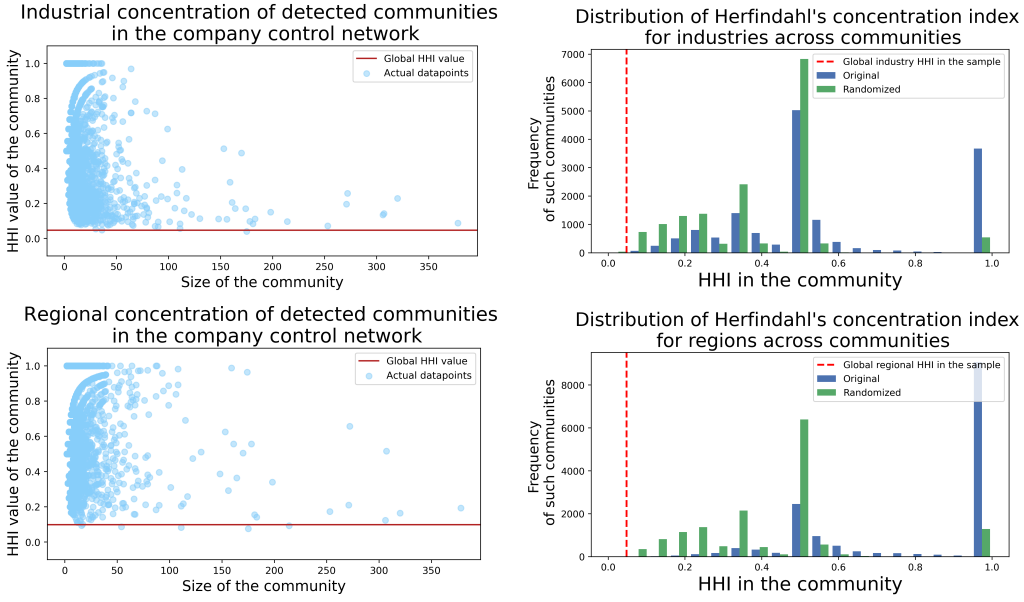

Figure 19: Geo-industrial clustering in the company control network

Investor type analysis

The CompaniesHouse data source does not directly classify the legal persons with significant control as "natural" or "institutional" entities, however I realized during manual inspection of the data that this could be possible with the text of the name of the legal person. Natural legal persons are any people who invest in the company, whereas institutional legal persons are other firms or other organizations (e.g. charities, associations). Having institutional legal persons with significant

control in a firm might indicate that the firm became part of a group of companies, controlled by the same holding companies. This could mean better access to capital, information or human resources, therefore it might be beneficial for the firms to have institutional investors. It might also help the firm to attract more investors, since the presence of an institutional investor might signal to the rest of the market that this firm has solid growth prospects. However, institutional investors might be more reluctant to share control with others than natural legal person investors, which would not let the firm to increase its number of investors significantly (after giving significant control to an institutional legal entity). Given these tensions and trade-offs in the potential impact of having institutional legal persons with significant control in the company, the analysis of the company control network with respect to the type of investors seem interesting. Also, it has direct real-world applications, since a firm might use the insights from such an analysis to support its decision on whether to try to attract capital from institutional or natural legal person investors. In the company projection network, the degree of the node (firm) represents how interconnected the node (firm) is through its investors. This is a reasonable estimate for the "combined influence" that its investors have. Therefore this is a quantity that is of primary interest throughout this analysis, since exploring the relationship between the degree of a node in the company projection network and its investor composition help to understand what "mix" of investor control might be the most beneficial in the current phase of a firm's lifecycle.

A simple natural language processing approach is used to get the distinction, basically looking for predetermined patterns in the names of the legal persons, to classify them either as *natural legal person* or *institutional legal person*. There is an approximately equal share of the two groups in the data. Then the main question is whether there is any systematic difference in network quantities that might be associated with the "investor composition" of a firm. As it can be seen on Figure 20, a really interesting pattern is emerging. At first glance, the number of investors (legal persons with significant control in the firm) of a firm does not seem to be correlated with the degree of the firm in the network. This is counterintuitive, since each of those investors might have significant control in other firms as well, therefore each of them represent potential additional links in the company control network (since it is a one-mode projection of the original bipartite network of companies and investors). Thus the expectation is that the degree of the companies would grow with the number of investors they have. But on the aggregated level, this correlation is not present. However, with a closer inspection and via distinguishing between natural and institutional legal persons, a new pattern is emerging. Now the number of institutional legal persons of a firm seem to be strongly associated with its degree in the company control network, whereas the number of natural legal persons does not seem to have a significant impact. Since this is one of the main empirical observations of this thesis, it is discussed further in Section III.1.

II.4 Summary of empirical findings

In the previous section I carried out a detailed analysis of the undirected labour flow network, the directed labour flow network and the company control network. I introduced the data collection procedure and also discussed some important summary statistics of the networks. I have identified that their degree distribution,

edge weights, connected component sizes and community sizes all seem to follow heavy-tailed distributions, which is frequently observed in other empirical networks as well. I also argued about some potential underlying mechanisms which might be responsible for the emerging hubs in these networks. Strong geo-industrial clustering is observed in the communities of the networks as well. Nevertheless, it is also evident that the network topology and structure provide superior information about labour flows, company connections and power dynamics, compared to a traditional standard industrial/regional classification schemes. Therefore this analysis might provide valuable decision-support for evaluating policy interventions on the labour market.

I also identified an informative distinction between natural and institutional legal persons for the company control network, which helps to find suggestive evidence about the potential relationship between the type and ratio of the investors of a firm and its degree in the network. Then this evidence also relate closely to how "influential" the investors of a particular firm are, which might be of practical interest to any company that is planning to attract capital. But it is also important from the investors' perspective in the real-world too, because they want to understand the power dynamics that they can expect in a particular firm, before they invest in it. Since these relationships are observed even when the degree of a firm is normalized with the number of its investors and even when different subnetworks are considered according to region or industry, this new stylized fact is further discussed in section III.1.

III Company control network formation model

First of all, the main empirical observations and a new stylized fact is discussed in this section. Then I formalise a causal model of power dynamics for the company control network. Next I perform some numerical simulations and compare their results to the corresponding empirical values to see how closely they match. Uncertainty analysis is also conducted to see how robust the results are as certain model parameters are varied. Then finally a Bayesian optimization approach is taken for the model calibration in order to find those combinations of the parameters, which produce results closest to the empirical observations.

III.1 Main empirical observations and a new stylized fact

Throughout the analysis of the networks several important empirical observations emerged, but these are well-documented in the literature (e.g. heavy-tailed degree distribution, edge weights significantly higher within communities than between communities, strong geo-industrial clustering) and there are already several models trying to explain them. However, there is one particular pattern which is related to the relationship between the institutional investors of a firm and its degree centrality in the company control network. Since to the best of my knowledge this has not been documented before, it is an interesting contribution of my current work.

Figure 20 and figure 21 highlight these most important empirical observations which I identified during the analysis of the company control network. First I look at the median degree of companies with a certain number of institutional/natural or total (institutional + natural) legal persons having significant control in the

company. As it is depicted on figure 20 the median does not change significantly as the number of total or natural legal persons with significant control in the company are varied. However as the number of institutional investors increases, the median of the degrees grow significantly, then it peaks around 4 (the 8+ category has only a few observations). This suggests that the most powerful institutional investors, who have significant control in numerous companies and who are partially responsible for creating the hubs in the network, do not like to share their control with too many other parties, especially not other institutional investors. Having multiple powerful institutional investors only happen in special cases, probably for "star companies". The plots therefore also suggest that there seems to be a "saturation point" for the benefit that institutional entities gain from investing in firms, therefore they try to pick investment opportunities where they can be in the "driving seat" and they do not have fear from other players having significant control in the firm as well. On figure 20 on the right plot a similar analysis is shown, but this time for the *degree per investor* quantity, to control for the number of investors of the firms. The patterns are different, but it is still evident that having the same number of institutional investors yields substantially higher median values for the *degree per investor* quantity than for the other cases.

On figure 21 this idea is extended and on the x-axis I show the ratio of institutional legal persons among all the legal persons with significant control in the firm. On the y-axis I show the *degree per investor* quantity that I introduced in the previous section. Note the discrete nature of these quantities are due to the fact that counts of the investors of certain types are used (since the actual ownership share are not available publicly). Interesting patterns are exhibited here as well. Extreme cases (e.g. only natural or institutional legal persons with significant control) have "majority" of observations and the *degree per investor* value is highest for the firms with only institutional investors. However, inbetween the extremes, we see a bell-shaped "upper-frontier" for the *degree per investor* quantities, which is highest when there is a power-balance between the type of investors. This intriguing observation is highlighted on the right plot of figure 21. I first group the observations into 10 bins of equal length according to the share of institutional investors[11] and then plot the 95-th percentile of the *degree per investor* quantity for the firms in each specific bin. This "upper-frontier" is depicted with a solid line and markers, and the actual datapoints are also superimposed on the plot. Again, these results might be distorted due to the discrete nature of the data, but still raise the question of what kind of network formation processes or mechanism could explain this pattern?

But the implications that this observation might have, are also intriguing from a practical perspective. It seems that firms with influential investors (and thus high degree in the company control network) tend to have a "balance" between natural legal persons and institutional legal entities among their investors with significant control. It is unclear in advance however, whether this is something driven by the firms, or the investors or by some institutional frameworks (e.g. laws, regulations). I also searched for such institutional frameworks in the related UK regulation, but I could not find any rules directly addressing this issue. Therefore I continue with focusing on potential causal models that could account for these new stylized facts.

[11] The extreme cases of only natural or institutional legal persons with significant control are excluded from this analysis.

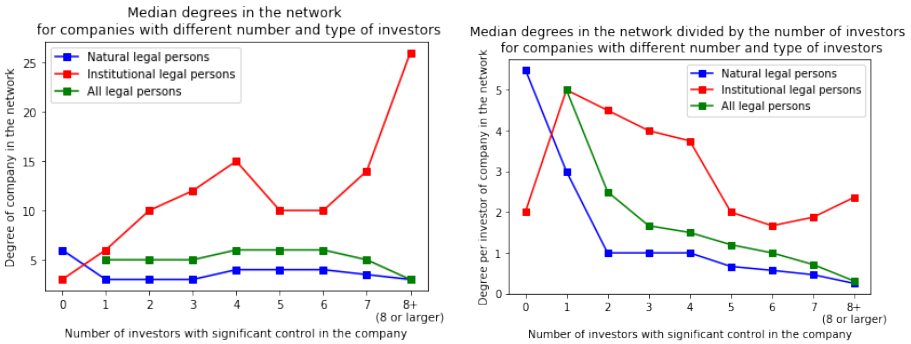

Figure 20: Stylized facts about the company control network (median degrees)

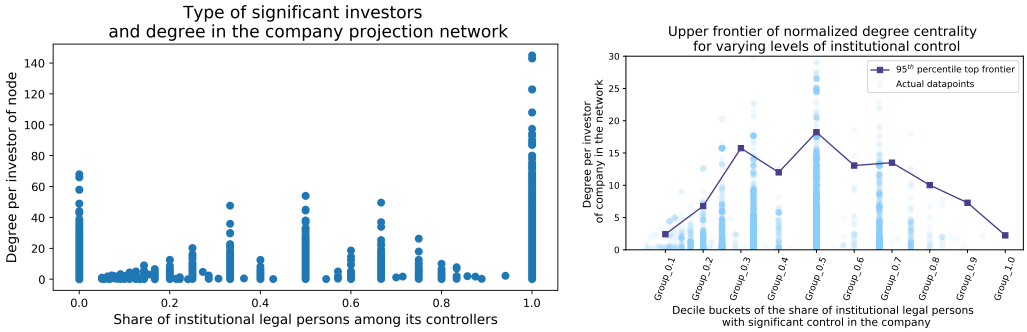

Figure 21: Stylized facts about the company control network (institution shares)

In the next section I devise a theoretical network formation model that helps to understand the potential processes or mechanisms which might be the "underlying drivers" for these empirical observations.

III.2 Formal description of the model

In this model, there are 2 types of agents: **firms** and **people**. Both firms and people have capital, which they can invest in firms, but firms can also use that capital for production in order to generate profit. The profit of firms is a function of the capital they have, but generated stochastically from the following Gaussian distribution:

$$\pi_i \sim \mathcal{N}\left(K_i^{\delta_i} - c_i \cdot K_i,\ \frac{1}{(K_i+1)^{\gamma}}\right)$$

where K_i is the current capital level of firm i, δ_i is a firm-specific productivity parameter, c_i is a firm-specific cost parameter. Then this model also implicitly assumes that larger firms have less variance in their profit generating process (which is a well-established stylized fact). Also each firm can compute its own optimal capital level (with respect to maximizing the expected profit):

$$K_i^* = \left(\frac{c_i}{\delta_i}\right)^{\frac{1}{\delta_i - 1}}$$

and it tries to attract this amount of capital (if it reaches this level, it will not accept more investments and/or move some of its own capital elsewhere). Each firm also

knows the marginal profit that could be achieved by 1 additional unit of capital, given its current capitalization. So companies have perfect information about themselves, but they do not have perfect information about the other companies. People also lack perfect information about the characteristics of the companies. Therefore both people and companies, when acting as investors, use a "noisy signal" to estimate the profitability of each investment option. Realized profits and past capital levels are publicly known,therefore the decision-making heuristic that the investors are using to estimate profitability is the following:

$$\widehat{PI}_i = \frac{\sum_{t=1}^{T} \pi_i^{(t)}}{\sum_{t=1}^{T} K_i^{(t)}}$$

where $\pi_i^{(t)}$ is the profit of firm i in timestep t, whereas $K_i^{(t)}$ is the capital of firm i in timestep t. Then the investor agents use these profitability estimates to weight the investment and divestment choices. In particular, the probability of choosing a particular company i for investment is given by the following expression:

$$\frac{max\left(\widehat{PI}_i + 1; 0\right)}{\sum_{j=1}^{n_c}\left(max\left(\widehat{PI}_j + 1; 0\right)\right)}$$

and then the divestment probabilities are inversely proportional to the above quantity. Using these probabilities, at each timestep, investor agents (people and companies), divest some of their holdings according to the outcome of a Bernoulli trial (with mean equal to the divestment probability). After that they have some available capital to invest (the stakes from the divestment and also the share of profits that they might have received from their holdings). They choose a predetermined number of companies to invest in, and share their available capital among them (again re-weighting between the selected ones only, using the formula described above) to make investment offers.
Then it is the turn of the companies, who first evaluate the amount of capital needed at timestep t. For an arbitrary firm i, it has the following equation:

$$K_i^{(t)(needed)} = K_i^* - K_i^{(t-1)}$$

Then the company considers the investment offers that it received (which are reshuffled randomly in each round) and chooses the first offer, but then the final investment is constrained by the actual capital needs of the company, in particular:

$$I_i^{(t)} = min\left(K_i^{(t)(needed)}; O_i^{(t)}\right)$$

where $I_i^{(t)}$ is the investment received by company i at timestep t and $O_i^{(t)}$ is the first entry in the "offer book" of company i at timestep t. To rephrase this investment process in terms of the strategic network formation literature, mutual consent is needed for forming a link, but investors can unilaterally destroy the link later.
Some special considerations which are also incorporated in the model and can be tuned via parameters:

- Companies are initialized with larger levels of capital than people
- Companies can make more divestment and investment offers in a given round

Below the high-level pseudo-code is shown for the simulation of the model.

Algorithm 1 Network formation model

Initialize network
Create n_c companies with initial capital level, productivity and cost parameters all drawn from a Uniform distribution
Create n_i investors (people) with initial capital level drawn from a Uniform distribution
for $t = 1, 2, \ldots T$ **do**
    for $company = 1, 2, \ldots, n_c$ **do**
        Profits/losses generated stochastically according to the current capital levels
        Profits/losses are distributed among the shareholders according to their respective stakes in the company
    end for
    for $company = 1, 2, \ldots, n_c$ **do**
        Divestment choices and investment offers are made stochastically based on the behaviour rules described above
    end for
    for $investor = 1, 2, \ldots, n_i$ **do**
        Divestment choices and investment offers are made stochastically based on the behaviour rules described above
    end for
    for $company = 1, 2, \ldots, n_c$ **do**
        The first offer from the "offer book" is chosen
        Capital levels of companies are updated
    end for
end for

III.3 Simulations and sensitivity analysis

Simulation results

First I run some simulations to see how the results of the numerical experiments compare with the empirical outcomes. On figure 22 one such simulation result is shown, with a manually chosen set of model parameters. The results capture some characteristics of the empirical quantities, however visual inspection is not satisfactory for the comparison, therefore a more disciplined approach to compare the empirical result with the simulations is introduced in the next paragraph.

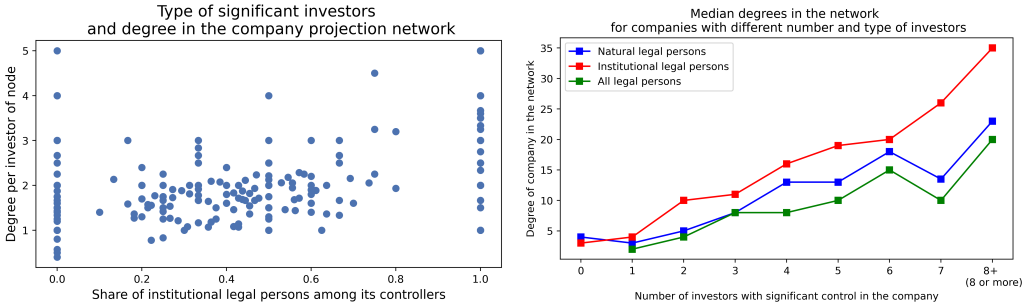

Figure 22: Simulation results for the network formation model

First of all, the "stylized facts" that the network formation model should be able to reproduce are the following:

- Degree distribution of the company control network
- Median degree of companies with a given number of institutional/natural/total investors
- Degree per investor dynamics as a function of the share of institutions among a company's controllers

Therefore a new quantity is devised in order to summarize how closely the simulations match the stylized facts simultaneously. This new quantity is called **"error function"** throughout this document.

$$\textbf{Error function} = \alpha_1 \cdot \Psi_1 + \alpha_2 \cdot \Psi_2 + \alpha_3 \cdot \Psi_2$$

where α_i are just parameters such that $\sum_{i=1}^{3} \alpha_i = 1$ and Ψ_1 is the value of the Kolmogorov-Smirnov test statistic for the empirical and simulated degree distribution in the company control network. Formally:

$$\Psi_1 = sup_d |\mathcal{F}_{empirical}(d) - \mathcal{F}_{simulated}(d)|$$

where $\mathcal{F}_{empirical}(d) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{d_i < d\}}$ and n is the number of companies in the network ($\mathcal{F}_{simulated}(d)$ is defined similarly).

Ψ_2 is a quantity derived from the negative Kendall-Tau rank correlation between the median degree dynamics of the empirical and the simulated results. More formally:

$$\Psi_2 = -\tau_{(I,A)} - \tau_{(I,N)}$$

where $\tau_{(I,A)}$ is the Kendall-Tau rank correlation coefficient between the following empirical and simulated quantities:

$$m^{(a)} \in \mathbb{R}^8 \text{ and } m_i^{(a)} = \frac{Me^{(inst=i)}(d)}{Me^{(all=i)}(d)}, i = 1, 2..8$$

So the i-th entry of this vector is the ratio of the median degrees of companies with i institutional investors and i total investors respectively. For $\tau_{(I,N)}$ the computations are similar, but this time the vector contains the ratios between the median degrees of companies with i institutional investors and i natural legal person investors respectively.

Similarly, Ψ_3 is a quantity derived from the negative Kendall-Tau rank correlation between the degree per investor (as a function of the share of institutional investors, using median values from the discrete bins introduced in section III.1) series of the empirical and the simulated results.

Then as it is stated in the beginning of this section, the final **"error function"** is a linear combination of these 3 quantities and it is used in later analysis as an objective function to optimize (minimize specifically).

Based on the definition of the error function, it is straightforward to see that the values it takes are in the following interval: $[-0.66, 1]$, with smaller values corresponding to a better fit between the empirical and simulated results.

Sensitivity analysis

Here I analyse how sensitive the simulation results are to changing the degree exponent of the variance of the profit generating process. On figure 23 I show the degree distributions for varying γ values, whereas the plot on the right shows the confidence interval for the error function value between the simulation results and the empirical results, for varying variance parameter again, keeping everything else constant. For each γ parameter value, 10 independent simulations are performed and the plot shows the average +/- 1.96*(standard deviation) of the error function value. The sensitivity of the results to changing this variance parameter is lower than expected.

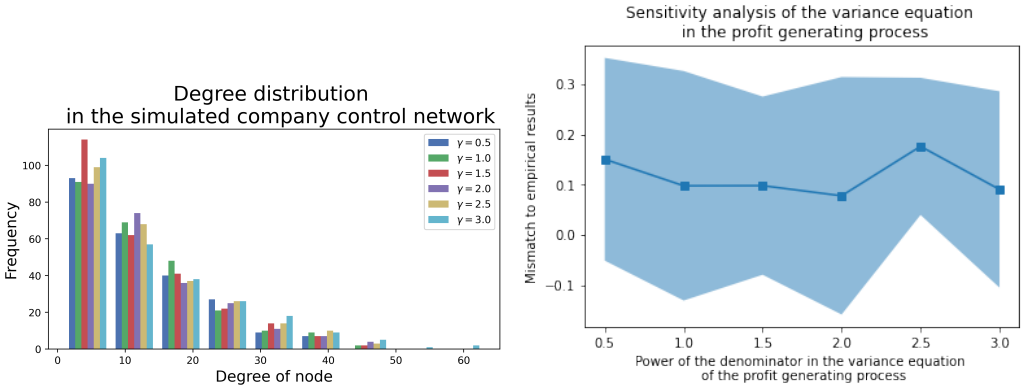

Figure 23: Sensitivity analysis of the variance parameter of the profit generating process

Another free parameter of the model which is expected to make the error function value sensitive is the share of "natural legal person" investors, compared to the

number of companies in the simulation. Numerical experiments match the expectations closer in this case, since as it is evident on figure 24, the confidence interval for the error function is wider than in case of the variance parameter before, whereas the average is also fluctuating around 0.1 in this case (10 independent simulations are done for this analysis as well).

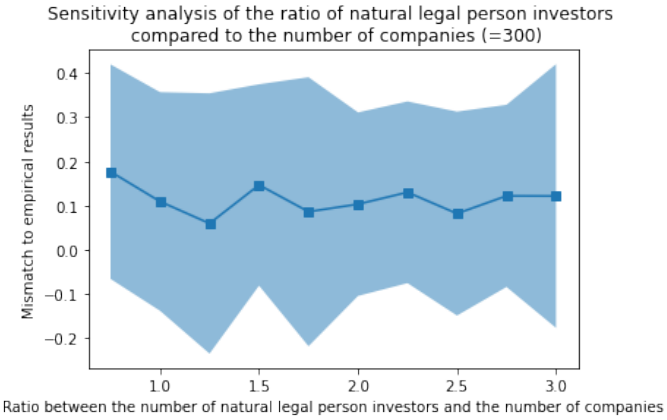

Figure 24: Sensitivity analysis for the number of "natural legal person" investors

Since the error function values have consistently high variance across different parameter values, it seems that the simulation model in general has high uncertainty. However, there is variation in this "instability" as well, since the numerical simulations show that results are more variable when the ratio of investor types is changing compared to changing the variance equation in the profit generating process, which is not straightforward to predict from the analytical form of the model beforehand. On the other hand, the mean of these simulations is not changing rapidly, which suggests low sensitivity. The performance metric (error function) is probably more sensitive to varying a combination of model parameters rather than only one of them. Therefore the partial and "complete" optimization procedures of the next section can indirectly provide insights for sensitivity analysis as well.

III.4 Calibration method for model parameters

I also carried out some experiments to optimize the parameters of the network formation model in order to match several empirical observations simultaneously. For this exercise the previously introduced **error function** is used as an objective function. As an optimization procedure I utilised the Hyperopt Python package, specifically the graphical model-based Tree-based Parzen Estimator, as it is proposed in (Bergstra et al., 2011) for Sequential Model-based Global Optimization. One of the main differences between the Tree-based Parzen Estimator (denoted also as TPE) and the more classical Gaussian processes (denoted also as GP) is that the TPE uses an inverse factorization (with a model for $\mathbb{P}(x|y)$) and the TPE is also cheaper in terms of computational cost.

On figure 25 the partial optimization for several model parameters are shown. Parameters for the firm cost function, productivity, number of different type of agents and number of investments per round are all considered here and optimized (pairwise) partially, keeping the other model parameters at some default value. The

algorithm sequentially evaluates the objective function at those parameter combinations, which give the highest *expected improvement* compared to the current best value achieved based on the samples collected so far. The plots show these sample evaluations and their corresponding error function value. However, since the results of a particular simulation round have a substantial amount of variance, several simulations for each parameter combinations would be needed to get robust estimates. Also, the search space could be extended further as well. It is important to keep these limitations in mind, nevertheless computational resource considerations also matter here, therefore the current approach seems to be a reasonable and practical way to get approximate estimates for the "promising" regions of the parameter space. For example it is straightforward to see that the error function value is lower when firms can make substantially more investment offers in each round than people.

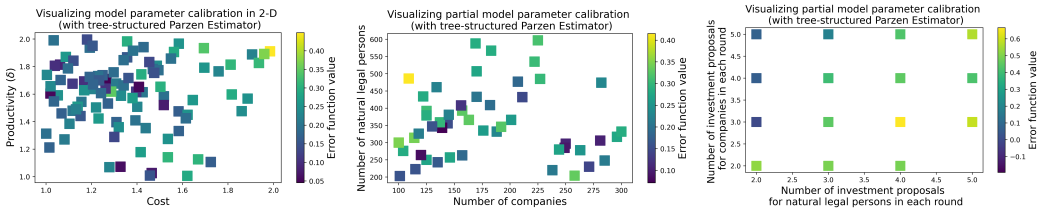

Figure 25: Parameter optimization for the network formation model

After these partial optimization approaches, I also conduct a more complete model parameter calibration method, where I vary all free parameters of interest. However, since the model have nine parameters, this raises the issue of *overfitting*. Therefore as the final model calibration approach, I fixed some of the model parameters (either to the their partial optimum value or to manually chosen values) and I only optimized the model parameters which seem to make the error function sensitive. The *number of natural legal person investors* and the *number of investments per round* that certain investor agents are making seem the most important parameters, therefore these are the free parameters in the final model calibration. Clearly, for the number of natural legal person investors, what truly matters is their ratio compared to the number of firms, however since the number of firms is fixed, the model calibration optimizes the ratio implicitly as well. Then for each considered model parameter combinations I run 5 simulations and average those to get a more robust estimate for the error value. With this approach the minimum "error function" value is **-0.17** which is smaller than the averages from the partial optimization. This is reassuring, since this model calibration is searching a larger parameter space now. The predetermined model parameter values are shown in table 2, whereas the optimal model parameter values found by the calibration method are shown in table 3:

These results also seem sensible from a practical point of view, since in the empirical data, there are also significantly more natural legal person investors than firms to invest in and these natural legal persons on average make fewer investments than the institutional legal persons. This is not surprising, since the model calibration is aimed at minimizing the mismatch between the empirical and the simulation results. Still, the fact that the calibrated model parameters are consistent with the real-world experience offers greater support for the validity of the network formation model.

Model parameter	Fixed parameter value
Number of firms	300
Maximum productivity (δ)	1
Maximum cost	1
Maximum initial capital for firms	1.5
Guaranteed initial capital for firms	0.5
Maximum initial capital for people	1

Table 2: Fixed parameters for the model calibration

Model parameter	Calibrated parameter value
Number of people	587
Number of investments per round for companies	6
Number of investments per round for people	2

Table 3: Optimized parameters from model calibration

The comparison between the empirical results and the simulation results with the calibrated model is shown on figure 26. This exhibits that the model is capable of reproducing the main patterns of the empirical characteristics. Via calibrating more model parameters it would be possible to fit the empirical observations even more closely, however that would raise the issue of *overfitting* again. The current calibrated model seems to capture the most important empirical patterns, however it is also expected to be robust, to provide similar performance results on different empirical data.

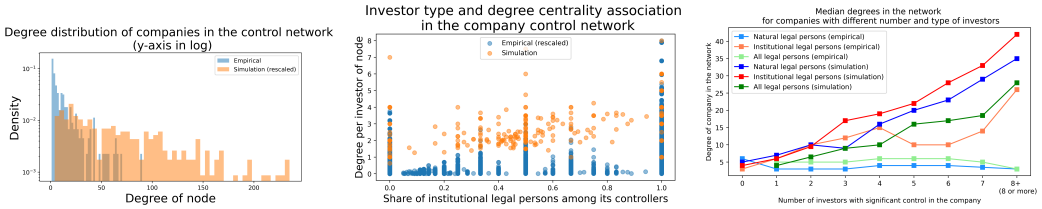

Figure 26: Comparison between empirical and simulation results with the calibrated model

As it is evident from the middle plot of figure 26, the calibrated model can accurately capture the "bell-shape" and the "extremes" for the relationship between the share of institutional legal persons and degree per investor quantities (there is some offset, but the shape is the same). However, the calibrated model performs weaker for the degree distribution, since more mass on small degree nodes and less extremely large hubs would be needed from the simulation results to have a better fit to the empirical data. The median dynamics are reproduced to some extent, as shown on the right plot of figure 26. There the main improvement could be to reproduce the approximately flat curve when considering median degree dynamics as a function of the number of natural and total investors (the dynamics of the institutional legal persons are fitted well).

These results show how methods and techniques from data science, combined with causal network formation models can provide insights for understanding certain

economic activities. Based on the calibrated parameters it seems that the composition of different agents in the company control network, together with their different decision-making rules for investments are the main drivers for the empirical observations documented in this thesis. These insights could not be discovered without using tools from data science in order to perform empirical analysis on large, real-world networks. But causal inference and theoretical models are also needed to understand and explain the underlying economic processes and mechanisms. Then these insights are useful from the practical perspective as well, since both firms and investors can individually benefit from optimizing their behaviour according a principled model of others' actions. Moreover, these insights might also be valuable for policy-makers who design the institutional framework for this economic activity directly. Since if these policy-makers have some objective which might be measured/approximated with network quantities (e.g. average degree/interconnectedness of firms in the company control network), then policy interventions and programmes could be designed while using the presented causal model of network formation as a tool for counterfactual analysis.

IV Conclusions

In this project I collect novel data on the employment history of company officers in the UK and also on the ownership and control structure of UK companies. Then I construct the labour flow and company control networks based on the collected data and I perform a quantitative analysis of these networks. I first observe that the labour flow and company control networks match several characteristics that empirical networks tend to have in general, such as a heavy-tailed degree distribution, small-world effects, clustering and high assortativity. Substantial geo-industrial clustering is also documented and subsequently analysed both at the local and the "global" level, which is important from practical considerations. Dynamical analysis also shows the stability of these clusters and their implications for the firm dynamics on the labour market.

Intriguing empirical patterns and a new stylized fact are documented during the study of the company control network, since there is suggestive evidence that the types and number of investors are strongly associated with how "interconnected" a firm is in the company control network. Based on the empirical data it also seems that the largest institutional investors mainly seek opportunities where they can have significant control without sharing it with other dominant players. Thus the most "interconnected"/central firms in the company control network are the ones who can maintain this power balance in their owner structure.

Then I devise a network formation model with microeconomic foundations to better understand the potential underlying mechanisms for the empirically observed stylized facts about the company control network. I carry out numerical simulations and sensitivity analysis and also calibrate parameters of the model using Bayesian optimization techniques to match the empirical results. With these procedures, the obtained estimates capture the patterns observed in the empirical data. However, these results could be "fine-tuned" at different stages further, in order to have a better empirical fit. First, the network formation model could be enhanced to represent more complex agent interactions and decisions. But also, the model calibration method could be extended to include more parameters and a larger valid

search space for each of those parameters.

This project could also benefit from improvements to the utilised data. For example more granular data on the geographical regions could help to understand the different parts of London more and to have a more detailed view of economic hubs in the UK. Moreover, the current data source provides a static snapshot of the ownership and control structure of firms. Panel data on this front could enhance the analysis of the company control network, numerous experiments related to temporal dynamics could be carried out, for example link prediction or testing whether investors follow some kind of "preferential attachment" rules when acquiring significant control in firms.

Throughout this project I study and analyse several economic networks with empirical data. There are numerous economic interactions taking place through these networks, therefore it is highly relevant for policy-making to be able to use the data on these economic networks for decision-support. Thus the natural next step for this project would be to consider its model as a tool for counterfactual analysis when designing policy interventions that affect the labour flow and company control networks.

Bibliography

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*(1), 47.

Axtell, R. L., Guerrero, O. A., & López, E. (2019). Frictional unemployment on labor flow networks. *Journal of Economic Behavior & Organization*, *160*, 184–201.

Bala, V., & Goyal, S. (2000). A noncooperative model of network formation. *Econometrica*, *68*(5), 1181–1229.

Barabási, A.-L. Et al. (2016). *Network science*. Cambridge university press.

Battiston, S. (2004). Inner structure of capital control networks. *Physica A: Statistical Mechanics and its Applications*, *338*(1-2), 107–112.

Battiston, S., Bonabeau, E., & Weisbuch, G. (2003). Decision making dynamics in corporate boards. *Physica A: Statistical Mechanics and its Applications*, *322*, 567–582.

Battiston, S., & Catanzaro, M. (2004). Statistical properties of corporate board and director networks. *The European Physical Journal B*, *38*(2), 345–352.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization, In *Advances in neural information processing systems*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Chen, J., Zhang, J., Xu, X., Fu, C., Zhang, D., Zhang, Q., & Xuan, Q. (2019). E-lstm-d: A deep learning framework for dynamic network link prediction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, *453*(7191), 98–101.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661–703.

de Paula, A. (2019). Econometric models of network formation.

del Rio-Chanona, R. M., Mealy, P., Beguerisse-Diaz, M., Lafond, F., & Farmer, J. D. (2019). Automation and occupational mobility: A data-driven network model. *arXiv preprint arXiv:1906.04086*.

Dustmann, C., Glitz, A., Schönberg, U., & Brücker, H. (2016). Referral-based job search networks. *The Review of Economic Studies*, *83*(2), 514–546.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486*(3-5), 75–174.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Et al. (2019). Toward

understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, *116*(14), 6531–6539.

Glitz, A. (2017). Coworker networks in the labour market. *Labour Economics*, *44*, 218–230.

Glitz, A., & Vejlin, R. (2019). Learning through coworker referrals.

Guerrero, O. A., & Axtell, R. L. (2013). Employment growth through labor flow networks. *PloS one*, *8*(5).

Guerrero, O. A., & Lopez, E. (2015). Labor flows and the aggregate matching function: A network-based test using employer-employee matched records. *Available at SSRN 2631045*.

Guerrero, O. A., & Lopez, E. (2017). Understanding unemployment in the era of big data: Policy informed by data-driven theory. *Policy & Internet*, *9*(1), 28–54.

Mealy, P., del Rio-Chanona, R. M., & Farmer, J. D. (2018). What you do at work matters: New lenses on labour. *What You Do at Work Matters: New Lenses on Labour (March 18, 2018)*.

Park, J., Wood, I. B., Jing, E., Nematzadeh, A., Ghosh, S., Conover, M. D., & Ahn, Y.-Y. (2019). Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. *Nature communications*, *10*(1), 1–10.

Pin, P., & Rogers, B. W. (2016). Stochastic network formation and homophily.

Rhoades, S. A. (1993). The herfindahl-hirschman index. *Fed. Res. Bull.*, *79*, 188.

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From louvain to leiden: Guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1–12.

Vitali, S., Glattfelder, J. B., & Battiston, S. (2011). The network of global corporate control. *PloS one*, *6*(10).

Youn, H., Strumsky, D., Bettencourt, L. M., & Lobo, J. (2015). Invention as a combinatorial process: Evidence from us patents. *Journal of The Royal Society Interface*, *12*(106), 20150272.

Appendix A

Appendix

Other related literature

Another prominent algorithm for community detection is the map equation framework combined with the Infomap search process, introduced in (Rosvall et al., 2009). This method has completely different foundations than the previous algorithms, since it is based on **information theory** principles and it is basically aiming to minimize the code length that is needed to describe a random walker's movement across the network. To give an intuition for this, this description could be done with assigning unique codes for each node in the network. However, it turns out that utilising the structure in the network can help to get more efficient descriptions. In particular, it is possible to partition the nodes into modules/communities and then have unique codes for these communities, but also have unique codes for each node inside a given community. Since codes can be reused inside different modules several times this way, it might help to reduce the total description length of a random walk on the network. It is straightforward to see that the length of codes used to identify the modules increases as a function of the number of modules, but the length of codes used to identify the nodes inside the communities decreases as a function of the number of modules. Therefore, there is a specific number of modules at which the sum of these two quantities is minimized and that is the minimum description length for the flows/random walks on the network. The map equation is precisely the following:

$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathcal{P}^i)$$

where $L(M)$ is the average code length for a step of the random walk, m is the number of modules or communities,, H(Q) is the frequency-weighted average length of codewords in the index codebook (the dictionary for the unique codes of the modules), $H(\mathcal{P}^i)$ is frequency-weighted average length of codewords in module codebook (the dictionary for the unique codes of the nodes in a given module) i. Furthermore, $q_{\curvearrowright} = \sum_{i=1}^{m} q_{i\curvearrowright}$, where $q_{i\curvearrowright}$ is probability to exit module i. Moreover, $p^i_{\circlearrowright} = \left(\sum_{\alpha \in i} p_\alpha\right) + q_{i\curvearrowright}$, where p_α is the probability to visit node α. Since the derivations are quite involved, the interested readers can consult the original paper for further detail. But what is important for the purposes of this current work, is that this map equation is perfectly compatible with any heuristic/greedy method that finds a network partition which optimizes an objective function. In the authors' main implementation, the neighboring nodes are joined into modules, which subsequently are joined into supermodules, until the map equation is minimized (the

average code length cannot be decreased with merging modules anymore).

Several other methodologies are related to our subsequent network analysis, however since they are not crucial for our main goals and final results, we will briefly discuss such literature in this section. One of the related topics was **link prediction in networks**, which is an active area of research and several approaches are applied for this problem. Just to highlight a specifically interesting one, (Clauset et al., 2008) uses the hierarchical structure of the network to predict missing links in partially observed networks. The authors show that this approach can explain and quantitatively reproduce many commonly observed topological properties of networks and it can also outperform competing approaches, e.g. Common neighbours, Jaccard coefficient or degree product. Machine learning approaches are also applied for link prediction, a successful example is presented in (Chen et al., 2019), where the authors used an Encoder-(Long Short-Term Memory)-Decoder deep learning model to predict dynamic link formation in networks and they achieve state-of-the-art results with their architecture.

Another important aspect to consider in case of labour flows is the **occupational similarity** between the activities that someone will undertake and that he/she has been doing before. The authors show in (Mealy et al., 2018) that people are more likely to transition into jobs which share similar activities to their previous one and that this occupational similarity has better predictive accuracy for job-to-job flows than existing benchmark methods. A similar path is taken in (del Rio-Chanona et al., 2019), but in that paper the authors focus on analysing how employment patterns would change through the occupational mobility network as a result of automation scenarios/shocks. But other authors in (Frank et al., 2019) highlight the challenges and barriers in modelling such automation scenarios, which shows that there is still lot to be done in order to understand and predict the complex dynamics of labour markets.

A loosely related, but insightful reference for our research was the work from (Youn et al., 2015), where they analyse empirical data on innovation, as a process of searching through combinatorial possibilities, highlighting the "exploration-exploitation trade-off", which is one of the main challenges in reinforcement learning. But the reason why this paper is informative for our current work is that investors, and legal persons with significant control in some companies face similar decisions when "optimizing their portfolio" since they might want to invest in some emerging "star-companies", but they also want to extract as much value from their current holdings as possible. This insight was useful when devising our generative model for the company control network.

Maximum Likelihood estimate for the degree exponent in power-law dsitributions

In this section I derive the Maximum Likelihood estimate of the degree exponent, when fitting power-law distributions (to the degree sequence data specifically here). Here we assume real-valued, independently and identically distributed data, which is satisfied by the degree sequence data on the networks. Then, the power-law distribution has the following form:

$$\mathbb{P}(d) = \frac{\gamma - 1}{d_{min}} \left(\frac{d}{d_{min}} \right)^{-\gamma}$$

where d_{min} is the minimum degree and d ($d \geq d_{min}$), the degree is our random variable and $\frac{\gamma-1}{d_{min}}$ is the normalising constant. Then the log-likelihood of our sample data becomes:

$$\mathcal{L}(\gamma) = log \prod_{i=1}^{n} \frac{\gamma - 1}{d_{min}} \left(\frac{d}{d_{min}} \right)^{-\gamma} =$$

$$\sum_{i=1}^{n} \left(-\gamma \left(log \left(\frac{d_i}{d_{min}} \right) \right) + log \left(\frac{\gamma - 1}{d_{min}} \right) \right) =$$

$$-\gamma \sum_{i=1}^{n} \left(log \left(\frac{d_i}{d_{min}} \right) \right) + n \cdot log \frac{\gamma - 1}{d_{min}}$$

then if we differentiate this expression with respect to γ and set the resulting expression equal to 0, then we get the following first order condition:

$$-\sum_{i=1}^{n} log \left(\frac{d_i}{d_{min}} \right) + n \cdot \frac{d_{min}}{\gamma - 1} \cdot \frac{1}{d_{min}} = 0 \Rightarrow \hat{\gamma} = 1 + n \cdot \left(\sum_{i=1}^{n} log \left(\frac{d_i}{d_{min}} \right) \right)^{-1}$$

Network analysis - general company information

Clearly **ltd** is the main legal company type, but there are several other categories with small occurrences which are not shown on figure A.1:

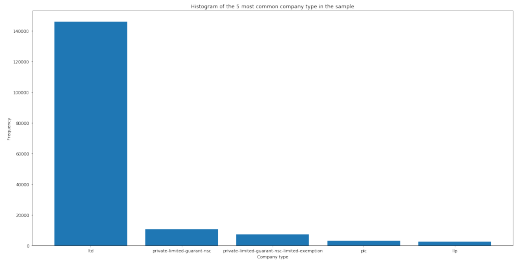

Figure A.1: Legal type of companies in the sample

The best parametric fit with a continuous probability distribution for the age of the firms in the sample is the **generalized inverse Weibull** distribution. The sample age distribution and this parametric fit are shown on figure A.2.

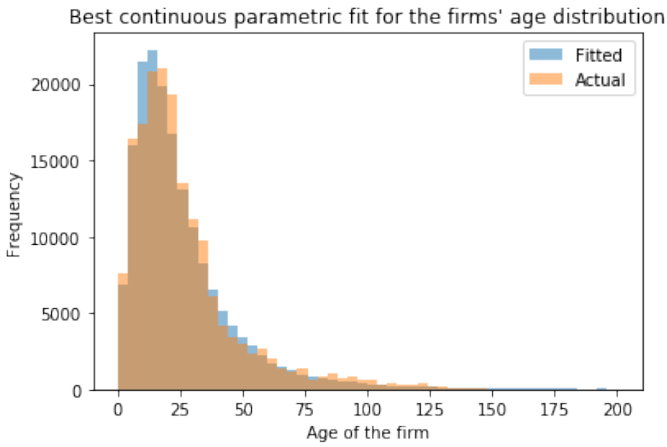

Figure A.2: Company age distribution - best parametric fit

Figure A.3 shows the same rank plot as it is in the main text, but now it also includes the names of those ten largest regions and industries.

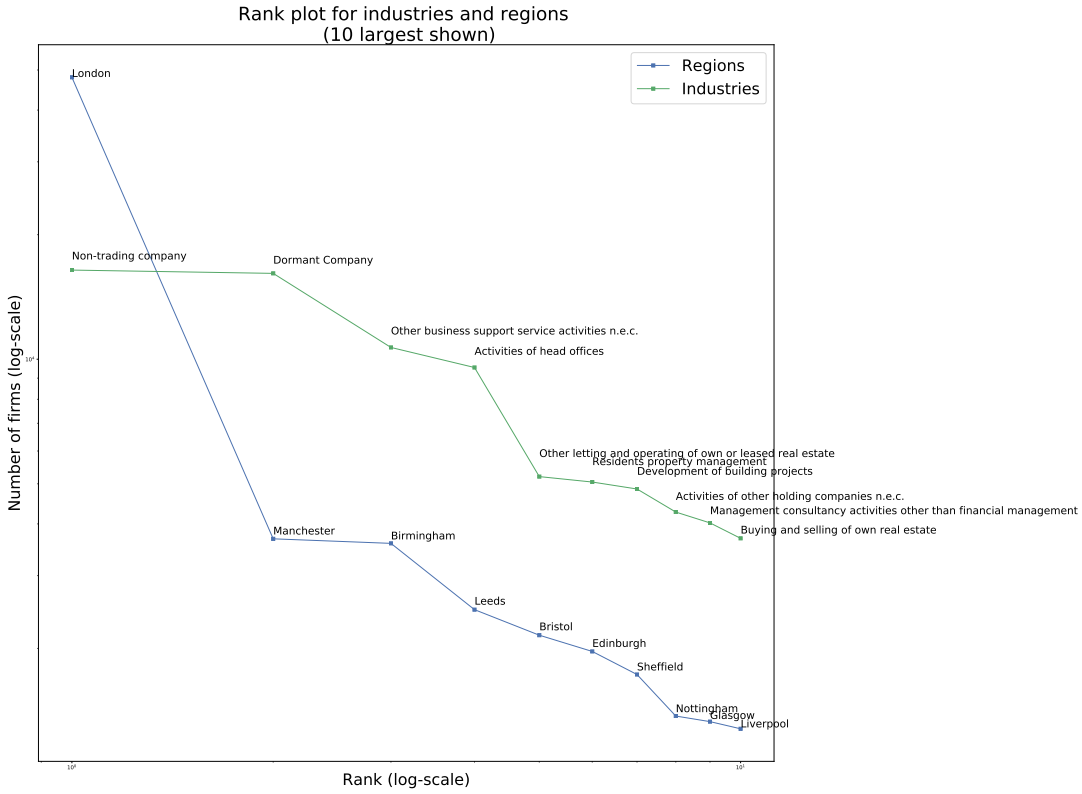

Figure A.3: Firm distribution across regions and industries in the sample

Network analysis - Undirected LFN

It is also interesting to see on figure A.4 how officers are distributed among the companies in the sample. Most companies only have a few officers, whereas a small portion of companies have a different magnitude of officers. Therefore, a power-law might be a reasonable approximation to the original histogram.

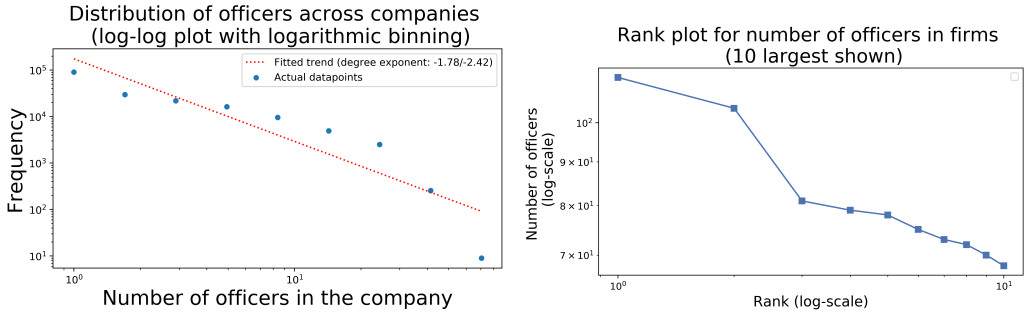

Figure A.4: Distribution of the number of officers in a company (undirected LFN)

Figure A.5 shows the size distribution of connected components. There is one giant component along with several small ones.

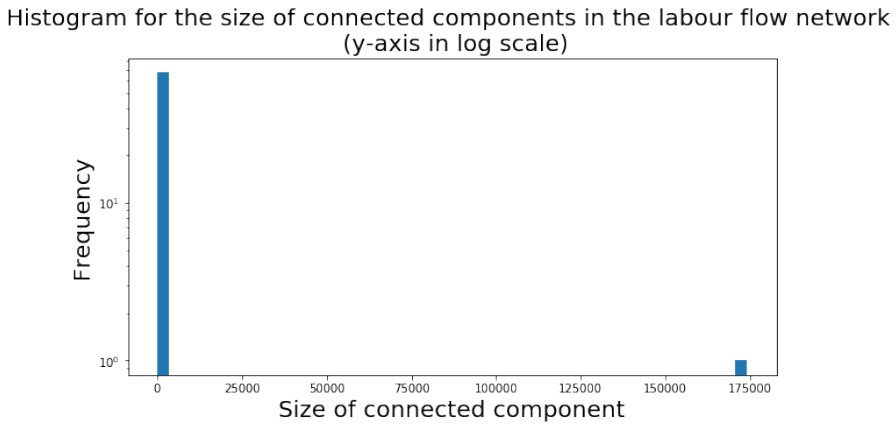

Figure A.5: Size of connected components (undirected LFN)

It is also possible to individually track each company throughout the sample and observe how its links are growing over time as it is depicted on figure A.6.

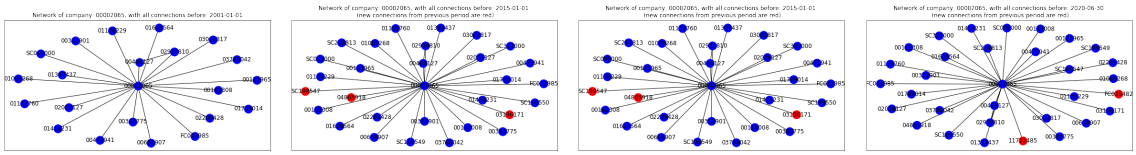

Figure A.6: Following a single node's network evolution over time (undirected LFN)

Analysing the empirical data also reveals that despite the strong geo-industrial clustering, the community detection methods still provide additional value for understanding the interconnectedness of companies and labour flows and they can provide a more accurate picture of labour flows than standard industrial classification. Figure A.7 shows such an example, with several small industries represented in a rather small community detected by the Louvain-method.

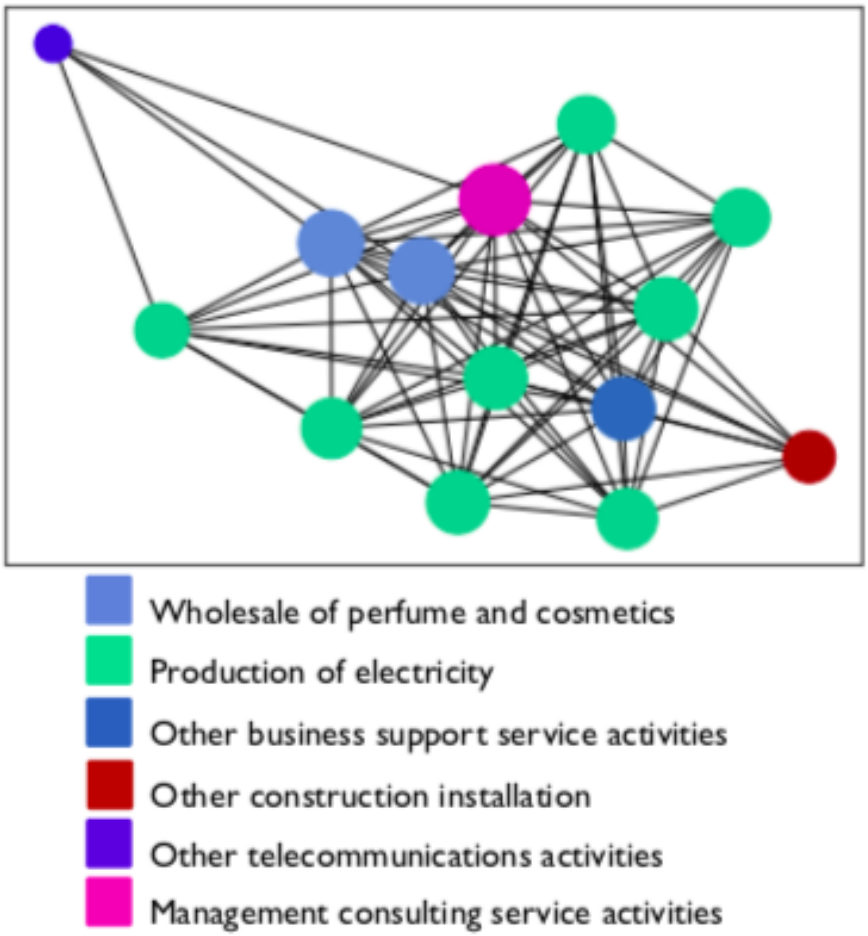

Figure A.7: Industrial decomposition in a detected community (undirected LFN)

Network analysis - Directed LFN

I also implemented a simple **random walker on the network**[1] to see how much time it spends in each detected community, since this is one of the measures that the map equation is using to find the best community structure. Figure A.8 shows running the random walk from a randomly chosen start node for 10,000 steps, with transition probabilities proportional to the weight of the outgoing edges and with a 20 percent chance of "teleportation" to any part of the network. It is evident that the walker usually finds a dense community/cluster and spends most of his/her time there (the share refers to the number of steps inside the cluster divided by the total number of steps).

[1] Also called random surfer due to the "teleportation" probability.

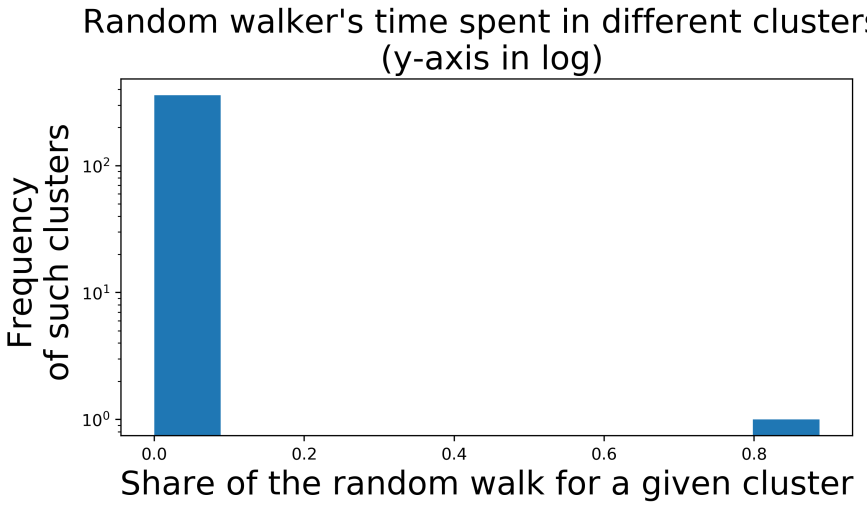

Figure A.8: Random walks among communities (directed LFN)