What do interlocks *actually* do.

Javier Garcia-Bernardo[1,*]

[1] *Department of Political Science, Amsterdam Institute for Social Science Research (AISSR), CORPNET Lab, The University of Amsterdam, 1018 WV Amsterdam.*

I. BRIEF DESCRIPTION OF THE PROJECT (FIRST DRAFT)

The study of corporate networks dates back to the beginning of the 20*th* century [1, 2], when Jeidels and Lenin noticed an increase in the relationships between banks and industry. These relationships were created through interlocking directorates – the appointment of directors with positions in both industries. Nowadays, 25% of the top one million companies worldwide are connected by interlocks (Fig. 1). However, not much is know about the consequences of shared directors. In the past decades, researchers have investigated the effect of interlocks in firm performance, innovation, acquisitions, mergers, capital growth, firm reputation, and adoption of structures and strategies (for a review see [3]). In spite of the extensive literature, only the spread of structures and strategies has been consistently associated to interlocking directorates [refs], while the other consequences of interlocks are still controversial [ref]. These contradictory results can be attributed to biases in the data used in the studies, namely a few dozen companies in a particular sector and country. Using data from the Orbis database, comprising 200 million companies and 100 million directors, we develop a theoretical and methodological framework to answer the question: "Do (and if so, how) interlocks facilitate structural transformation?", where structural transformation is the reallocation of economic activity across the broad sectors agriculture, manufacturing and services.

Firstly, we will create a network of related economic activities (sectors), where two economic activities are closer in the network if companies from both sectors are often co-located in the same city. Similarly to previous research at the country level [4–7], we expect that structural transformation occurs through the development of new sectors that are close in the network to the existing sectors. Moreover, we anticipate that this diffusive process will be a predictor of economic growth [7]. Secondly, we will quantify to what extent the presence of interlocks facilitates this diffusive process. Finally, we will test if interlocks facilitate diffusion trough an increase in collaboration and innovation between companies.

This research proposal is organized as follows. In section II, we give an overview of complexity methods that have been applied to social sciences. In section III, we focus on the research questions that we will study, and explain the propositions, concepts and hypotheses of the project. Finally, we describe the data in section IV, and detail how we are dealing with its biases.

II. COMPLEXITY THEORY

"One way to explain the complexity and unpredictability of historical systems, despite their ultimate determinacy, is to note that long chains of causation may separate final effects from ultimate causes lying outside the domain of that field of science."

— Jared Diamond.

A. Complex systems

Complex systems are those where many similar parts interact with each other using simple rules to create the whole, which exhibit characteristics different than the parts – "more is different" [8]. Complex systems contrast with complicated systems, where many different parts with defined roles are put together to create the system. Complicated systems, such a watch, can be studied by analyzing their parts. Moreover the failure of a piece produces the failure of the system. Complex systems, such as bird flocks, cannot be studying by analyzing only the characteristics of the parts, but the interactions among them are also needed. Although a consensus definition of complex system does not yet exist, a complex system is characterized by the following attributes. (i) Multi-scale: Many individual parts interact to create the whole. (ii) Networks: The individuals usually interact only with a few other individuals, creating a network of interactions. For social systems, the networks created are "small-world" networks, where the distance between two random people in a network is small. (iii) Emergent properties: The whole has properties that none of the individuals have. (iv) Spontaneous order. There is not a global organizer of the system. For instance, the collective behavior of flocks is an emergent property of the interaction between birds. (v) Memory: The individuals remember previous interactions. (vi) Feedback loops: The interaction between two individuals affect other individuals in the system. For example the decision of a bird to turn affect the probability that other birds turn as well, which affect the probability that the first bird keep turning. (vii) Stochasticity: The system lives in a noisy environment. (viii) Steady-states are far from equi-

* garcia@uva.nl

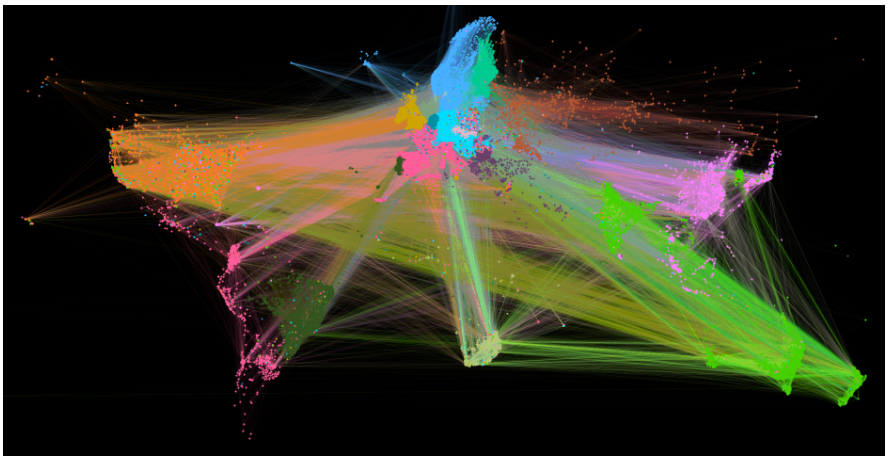

FIG. 1. Global network of interlocking directorates. Color indicates communities – i.e. cities that do business together within each other more often than with others.

librium. Complex systems are usually in a steady-state (except during transition times). However, since they depends on active interactions between people, they stay far from the equilibrium. If no energy is added to the system, the system disappears. (ix) Non-linearity, cascading and hysteresis. The interactions are non additive. In a standing ovation, the addition of a new person standing can produce a cascade of events, resulting on a general standing ovation. (x) Robustness to random failures. The system is highly resistant to the failure of one of the individuals. (xi) Sensitivity to targeted failures if individuals are organized in network: The system is sensitive to the failure/removal of a few specific individuals.

B. The micro-macro problem

Some social systems exhibit different properties than the parts composing them. In those systems, the action of an agent cannot be solely predicted from its characteristics, but the interactions between agents need to be taken into account. For instance, it is more likely that you start using purple hats if your partner thinks that are trendy than if your colleague does, which in turn affect the probability of your friends and colleagues also wearing purple hats. The probability that every person in society start using purple hats cannot be predicted by their perceived trendiness, but networks must be taken into account. Similarly, the performance of a firm depends not only on their product, but also on the interaction between firms, governments and other groups, which in turn are composed and influenced by individuals. Moreover, the success of a company's product – which is partly based on the interaction with other firms and groups – affects the actions of your suppliers and competitors. A classic example are format wars, where inferior products can (and often do) succeed. The problem where the whole depends on both the parts and the interaction between parts is often called 'the micro-macro problem'.

The micro-macro problem is found across disciplines. In physics, snowflakes form by the interaction between low-energy H_20 molecules, where the specific shape of the flake depends on the interactions between the individual molecules. In biology, organs are composed of individual cells that do not have the properties of the organ. A cardiac cell alone lack the capability to beat, however a few hundreds of cells together spontaneously start to beat. In ecology, ant colonies are efficiently organized to collect food, clean and defend the colony, and reproduce. However an individual ant cannot perform all those tasks, but requires chemical stimuli from other ants to coordinate. In social systems, complex behaviours are the results of the interactions between the agents. A mediocre play can receive a standing ovation if a group stands up immediately after the play ends, creating a cascade of people standing up [9].

Importantly, although predicting the whole in individual situations is difficult, this does not imply that we cannot observe correlations at the macro scale. In our physics example, humidity and temperature increase the probability of a specific shape of the snowflake. Cardiac cells will be more likely to beat if specific chemicals are added. A colony of ants will be more likely to leave the colony to scavenge for food if the temperature is moderate. A play will be more likely to receive a standing ovation example if it is good. However, it is important to observe that the correlations are indirect. The humidity does not affect the shape directly, but affect the probability of two molecules of water binding in an specific way. The quality of the play will increase the probability that some people stands up, creating a cascade of people standing. Because macro outcomes are non-linear aggregations of the micro agents, an appropriate model for the micro-macro problem is required.

C. The micro-macro problem in social sciences. A complex systems perspective

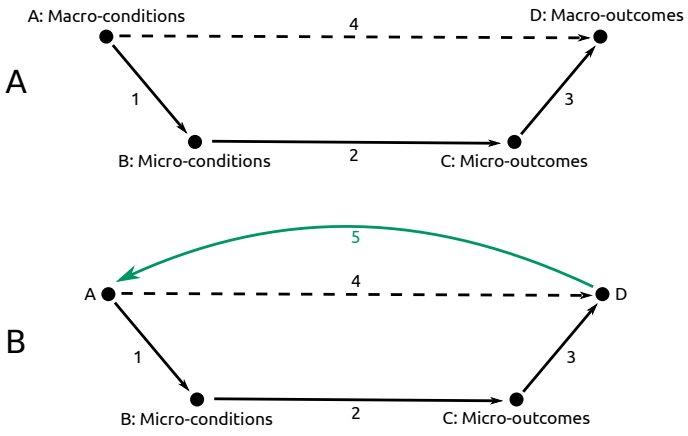

FIG. 2. **Coleman's scheme**. (A) Original scheme, where the arrow 4 is explained by the arrows 1-¿2-¿3. (B) Adapted scheme, where a feedback loop is implemented.

In social sicences, Collemans scheme [10] (Fig. 2A) is the standard framework to represent the micro-macro problem. Nodes A and D in Figure. 2A are the macro-conditions (composed of the environment where the system is situated) and macro-outcomes of the social system. Node B corresponds to the micro-conditions (composed of the perceived environment, as well as genetic and other individual factors). Micro-conditions are affected by the macro-conditions. This arrow is usually labelled as bridge assumptions. Micro-outcomes are the decisions of the individual among the possible options. In the standing ovation example, this corresponds to each individual decision to stand up or not. This arrow is non-trivial, since the decision does not depend only on each agent's conditions, but also on interdependent relationships with other individuals. In Granovetter's model [11] of riots, individuals are characterized by a threshold ϕ that summarizes their micro-conditions. Each person will then riot if there are at least ϕ other people rioting already. Similarly to the standing ovation, the population will end up in a generalized revolution, or it will die off depending on the distribution of thresholds in the population. For instance, consider population A, where 10% of a population starts rioting, but the other 90% will not riot unless 20% of the population is rioting. In this population, the 'average' person will riot if 18% of the population are rioting, but no revolution will occur. Imagine now population B, where 10% of the population starts rioting, another 20% will riot if 10% is already rioting and the final 70% will only riot if 30% of the population is rioting the revolution will spread. In population B, the 'average' person will riot if 23% are already rioting, however due to the non-linearity in the aggregation a revolution will occur. The aggregation of the micro-outcomes (the individual decisions to riot or not riot) corresponds to label 3. This arrow is usually labelled as transformation rules.

Colemans original scheme [10] uses Webbers origin of capitalism [12] as an example, linking it to the micro-macro problem. He explains the rise of capitalism (node D in Fig. 2A) from protestant religious doctrine (A). A protestant religious doctrine creates specific values in the individuals (B) that produce certain economic behaviours (C). The aggregation of these economic behaviours gives rise to the capitalism (D). Capitalism is the result of the economic behaviour of people, which in turn is caused by the individual interpretation of the values of protestantism. Similarly, a revolution is the result of the decision of people to riot given their angriness level caused by the macro-conditions. While apparently coherent, this reasoning has two main flaws, Firstly, it obviates the link from (D) to (A). The origin of capitalism was a process that lasted decades, where the economic behaviours produced some intermediate macro-outcomes that affected the macro-conditions. This in turn affects the micro-conditions (values), which affect the micro-outcomes (economic behaviour) from the previous time point. Capitalism is one of the steady states of the cycle. Our view of the process, using Granovetter model as an example, is summarized in Figure 2B. In Granovetter example there are some specific macro-conditions at time zero, such as a level of hunger, a level of police reprisal, or a sense of collective. Moreover, no people are rioting ($\Phi_0 = 0$, where Φ_0 is the percentage of the population rioting at time 0). Given the micro-conditions, a few people (x) with threshold zero ($\phi_i = 0 \geq Phi_0$) – very prone to riot – make the individual decision to riot (micro-outcome). The macro-outcome is that a group of people are rioting. This affects the macro-conditions ($\Phi_1 = x$). The micro-conditions in Granovetter model are fixed (constant thresholds, although this is a simplification). Now, every individual compares their own threshold with Φ_1, and riot if $\phi_i \geq \Phi_1$. The cycle continues until we reach a stable state (for example revolution or just a few people rioting). This class of systems where the system is continuously evolving correspond to Complex Adaptive Systems.

The second flaw is that it creates an illusion of determinism. In any complex system, random events can create a cascade of events that cannot be explained solely by the macro-conditions. Furthermore, we can always find a logical explanation of the end result. For instance, we can deduct that that few people started rioting because police reprisal, and that produced the cascade. However these kind of 'ad hoc' explanations makes police reprisal a necessary condition, while the same could have happened with no police reprisal, or no rioting could have happened with the same reprisal. While we could probably assume that some degree of police reprisal increased the propensity of rioting, we cannot conclude that reprisal is a necessary cause of rioting, or that police reprisal will cause another revolution in similar conditions. When the end result is the non-linear aggregation of many interconnected actors, the results cannot only be generalized unless

we can compare many independent, similar cases [13].

In the past decades we have witnessed the recollection of large datasets for many complex systems. These large dataset include vast information on biology (e.g. interactions between thousands of proteins in the cell), social interactions (e.g. social networks, movement trackers, etc.), or the economy (e.g. Orbis, Reuters or LexisNexis provide information on firm indicators and directors for millions of companies worldwide). As a consequence, new methods and models have been developed for the study of these rich datasets. These tools can be grouped in three categories. 1. Descriptive tools: To characterize the macro-outcomes and find patterns in the data. 2. Generative modelling: To explain emergence of macro-outcomes from the micro scale, how the macro-conditions affect the micro scale, or both. 3. Predictive tools: A model is useful if it provide insights in the causal mechanisms and can predict future events. This include predicting what has already happened using only a portion of the data. An example that combines the three categories is the work that resulted in the Atlas of Economic Complexity that will be the basis of next section [4–7]. Firstly, they described every country (macro scale) with respect to the type and amount of exports and imports (micro scale). Secondly, they assumed a series of capabilities required to produce a product (for example institutions, materials, human capital). If countries manufacture the products when they acquire the capabilities required to do so, then products that are often exported together will require similar capabilities. Since it is not possible to quantify those capabilities, they created the one-mode projection of the model, namely a network of products (named the 'product space'), where two products were closer in the network if they were usually exported together. Thirdly, they modeled the development of countries, showing that countries develop by acquiring new capabilities and producing products that are close in the product space. For instance, a country may develop an electronic industry if chemicals is already an important export, but not if they rely on exporting cereals [4]. Finally, they used the current products produced in the country to create a better indicator for the economic growth of the country [5]. Linking back to Colemans modified scheme (Fig. 2B) we see how the products that a country currently export (macro-conditions at time zero) affect the products that companies can produce (micro-conditions at time zero). This in turn affect the products that they actually produce (micro-outcomes at time zero), which results in the total exports and imports (macro-outcomes at time zero). In this model, the macro-outcomes at time zero correspond to the macro-conditions at time one. Understanding the cycle has utility to focus investment on certain sectors.

D. Complex system toolbox for network analysis

We next (very) briefly summarize general methods to study data in complex systems. Because most social and economical systems are embedded in networks, we will focus on complex networks. The standard notation treats networks (graphs) as a series of nodes and edges, where nodes are the agents and edges the interactions between agents.

1. Descriptive tools

Descriptive tools are used to find patterns in the data. Summary statistics, correlations and visualizations allow for a rapid exploration of the data, and an assessment of data quality. Moreover, they produce clusters of agents and quantify the importance of agents in the network. Two of the main descriptive tools are community detection and centrality analysis. Community detection finds nodes that interact with each other more than with the rest of the network. For example in a recent paper we studied a network of firm interlocks (ref), where the nodes correspond to firms and the edges to shared directors. We showed that in some regions the business communities are organized along national borders, whereas in other areas the locus of organization is at the city level or international level. Centrality analysis grades the importance of a node in the network. Different centrality measures have been developed. For instance in betweenness centrality a node is important if it connects far regions in the network. The specific measure will depend on the problem. If we are measuring the spread of information, as Granovetter did in the strength of weak ties [14], betweenness centrality would be indicated.

Communities and centralities have been used abundantly to describe data. In Clauset et. al. [15] description of inequality in academia, they mapped the academic trajectory of $19,000$ faculty in three disciplines. From the mapping, they rank each institution by its relative "prestige centrality". Institution A is more prestigious than B if people can do their PhD in A and move to B, but PhD graduates from B cannot find a job in A. They showed that the top 25% of the institutions positioned 71-86% of all tenure-track faculty, revealing steep inequality and clear hierarchical networks.

2. Modeling

Modeling allows us to understand the world and predict future events better than people [17]. For example, modeling the network of bank co-risk (Fig. 3) allowed authorities to assess the risk of not bailing out banks. The end result was the bankruptcy of Lehman Brothers and the bailing out of AIG. The goals of modeling are to make causal inferences about a phenomenon and to predict future events. Three main models are used: First-

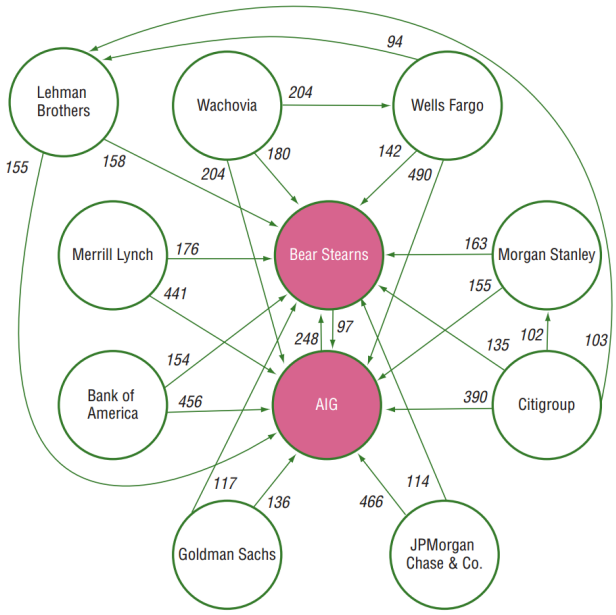

FIG. 3. Sources: Bloomberg, L.P.; Primark Datastream; and IMF staff estimates. Note: This figure presents the conditional co-risk estimates between pairs of selected financial institutions. Only co-risk estimates above or equal to 90 percent are depicted. See Table 2.6 for further information. [16]

ly, traditional statistical models explains an independent variable in terms of dependent variables. They are characterized by a formula whose parameters are estimated. The parameters usually have a direct real-life interpretation. They emphasize inference and work well when the number of dependent variables is small. The most common example of the group is regression. Secondly, machine learning are also used to explain the results of an output (our dependent variable) in terms of an input (independent variables). Opposing statistical models, they are usually black-boxes, meaning that real-life interpretation of the weights do not exist. Consequently, they emphasize prediction and are used when the number of dependent variables is large (big data). Finally, mathematical and computational modeling allows to recreate the physical system and understand the causes that drive the output of the system.

Mathematical and computational modeling

As previously discussed, it is not possible to explain emergent properties just by studying the agents. Mathematical and computational models allow to close the gap between the micro and macro scales. In Granovetter rioting example, a mathematical model explains why a population will end up rioting, while other 'more prone to riot on average' will not. In the segregation example of Schelling [18], a neighborhood of houses is simulated in a square lattice (chess board), where most houses are occupied but some are empty. Having some empty houses allows people to move between houses. Two different types of person live in the houses (lets call them Belgians and Dutchies) in equal proportion. None of them are racists, but they would like to have at least two neighboors that are similar to themselves. If more than 70% of the neighbors are from another country they move at random to one of the empty houses. While the at random assumption will not hold in reality, it is a conservative scenario. If segregation is found with this simple model, selectively moving to neighbourhoods with a large population of your kind will only produce higher segregation rates. The result is the clear segregation showed in Figure 4A , that resembled the segregation in life (Fig. 4).

Mathematical and computational models can help us discover how simple rules can produce complex macro results. Models can be classified in several categories. According to their assortativity, they are classified into perfect mixing models, where all the agents can interact with all other agents, and network models, where the agents can only interact with a subset of other agents. According to the presence of noise they can be classified into deterministic and stochastic, where some amount of noise is included.

According to the method used to solve the model, they can be classified into equation-based models (analytical and numeric), where the population is represented with an equation, and agent-based models, where each individual is modeled individually (see Epstein [19] for an excellent review and implications of agent-based models in social sciences). The type of model will depend on the application. For example if we want to measure the effect of time and distance on terrorist attacks, using an analytical model (Hawkes process), where an attack produces a cascade of events, will be useful [20]. If you are interested on modeling smoking behavior in high schools, an agent-based model that includes selection and influence terms may be more appropriate [21].

In political sciences, such models have been successfully applied to understand the spreading of ideas [refs], to understand the factors affecting polarization vs homogenization [refs], to find the micro-motives in the origin of inequality [ref], show why coordination and collaboration emerge [ref], among others [refs]. In general, modeling allows us to find and quantify which micro and macro-conditions are associated to the macro-outcomes observed.

3. *Prediction (to be written)*

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravi-

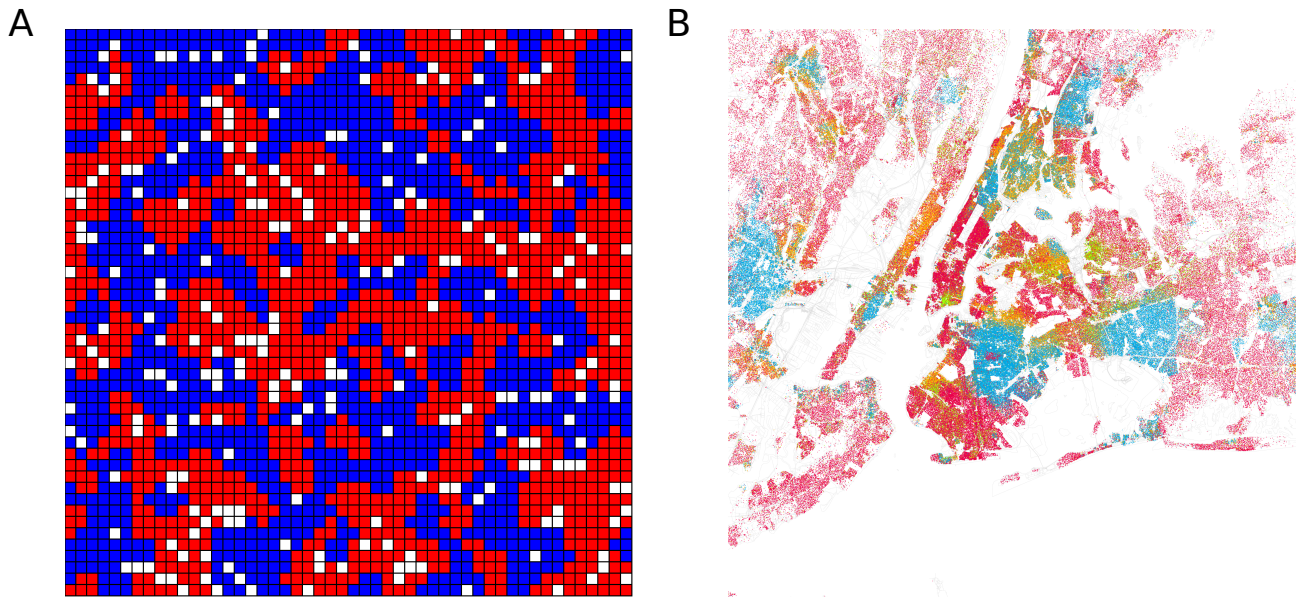

FIG. 4. **Segretation in the simulation and in New York City.** (A) End result of the simulation from `http://nifty.stanford.edu/2014/mccown-schelling-model-segregation/`. (B) Segregation in New York City. Blue colors reflect areas with a majority of black people, orange areas are Hispanic. green areas are Asian and pink areas are white.

da placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

E. Summing up

Social science systems are often embedded in complex networks of interconnected agents, where the action of an agent affect the actions of the others in a non-linear fashion. With the advances in technology, we are for the first time able to gather large datasets of social systems. Data availability has come together intensive research in the area of complex systems, and we can now use formal models to describe, explain and predict these challenging structures.

Let's finish with an extract from Duncan J. Watts's 'The Collective Dynamics of Belief': *"When collective behavior is generated from individual behavior via a non-linear, stochastic aggregation process, it simply is not 'explainable' in the usual way of 'A caused B'. Once they have taken place, however, collective outcomeseven those that are highly counterintuitive and unpredictable– can always be rationalized. Some story can always be told, no matter how complicated, that traces a path from some initial state A to some final state B, and thus sounds like a causal story. [...] But these descriptions are necessarily statistical in nature, comprising, for example, descriptions of probability distributions, rather than descriptions of particular outcomes. Thus while we may hope to understand better the processes by which collective outcomes are generated, even a perfect understanding of those processes will not correspond to a causal explanation of why one particular outcome pertains and not another. Indeed, no such explanation is possible."*

III. RESEARCH QUESTIONS, CONCEPTS AND PROPOSITIONS

A. Interlocking directorates

Companies are embedded in complex networks of 'interlocking directorates' – created when directors from a corporation sit on multiple boards. The relationship between two corporations resulting from a director sitting in both of their boards is often referred to as a direct interlock [3]. Interlocks can also be indirect when two directors from two firms sit together on the board

of directors of a third firm. Several explanations (micro-motives) have been pointed out for the creation of interlocks. These include facilitating cooperation to limit competition (collusion), absorption of potentially disruptive elements (cooptation), creating legitimacy by hiring respectable directors, career advancement, and social cohesion [3].

Although the micro-motives for the existence of interlocks are relatively well understood, their effects are still object to controversy. In the past decades, many papers have investigated the effect of interlocks in firm performance [ref], innovation [ref], acquisitions [ref], mergers [ref], capital growth [ref], firm reputation [ref], and adoption of structures and strategies [ref]. See Mizruchi [3] for a review. However, only the spread of structures and strategies has been consistently associated to interlocks [refs]. We attribute this to limitations in data analysis. Previous papers have focused on a small number of top companies (10–1000), many times restricting the study to only one sector or country. The result of this approach is a misrepresentation of local patterns as global patterns. For example, [this example with this data, ref] found that the presence of interlocks increases firm growth. At the same time, [this example with this other data, ref] found that the presence of interlocks decreases firm performance.

Here, we study a large dataset comprising 200 million companies and 100 million directors (see IV) to bring definitive answers to the field . Our main research question is "How is the network of interlocking directorates structured in time and space, and what is its effect on structural transformation?". For the sake of brevity, we focus here on the second part of the question "what is the effect of interlocks on structural transformation", where structural transformation corresponds to the evolution of the distribution of economic activities within the city. In section III B we explore the concepts of structural transformation and the product-service space. The product-service space determines how closely two economic activities (sectors) are depending on how often companies from both sectors are co-located in the same city. In section III C, we determine if (and how) interlocks affect structural transformation. Section III D shows what factors influence the presence of interlocks. Our unit of analysis can be the company itself, the city, the country or the region. For clarity, because cities have became innovation hubs [22], and because we should not impose a national structure on interlock data [23], we will focus on the city level for the rest of the manuscript.

B. Product-service space

Cities and countries develop economically by moving from producing simple products and services to specializing in more expensive ones – a process referred to as 'structural transformation' [4, 24–26]. This transformation can be explained using differences in productive factors and technology (see [5] for a review). In order to connect these differences to development, current models usually abstract from the products and look at macro-indicators of productive factors and technology. However, development occurs when new products are created or existing ones improved, and it is not clear if these models can explain the variability observed in countries with similar macro-indicators. Moreover, the products that are developed depend on the current products being produced – there is a relatedness between products. Many explanation have been proposed for this, such as similar institutions, infrastructure, physical factors, technology, or some combination of those factors (see [4] for a review). To reconcile previous research without abstracting from the products, Hidalgo, Hausmann and Klinger [4–7] assume that related products require similar underlying factors ('capabilities'). They next developed the 'product space' as a map of the relatedness between products, where products are related if countries have competitive advantages respect to both products. They showed that the product space capture information about the set of capabilities available in a country, is strongly correlated with income per capita, and predictive of future growth. Furthermore, they proved that structural transformation at the country level occurs by moving from existing products to related products, where two products are related if they are close in the 'product space'.

In a world full of multi-nationals, innovation is happening at the city level [22]. To account for an exploration of the process at the city-level, we need adapt the 'product-space' concept. We will create the 'product-space' using relationships between economic activities at the city level, instead of using product categories at the country level as in [4–7]. Using cities allows us to explore not only products, but also services, thus we will define the 'product-service space'. The subquestion here is: "Does structural transformation at the city follows the links of the product-service space". We expect to find comparable trends at the city level to those found by Hidalgo, Hausman and Klinger at the country level. The structural transformation should follow the edges in the product-service space, and the diffusion process should correlate with gdp per capita in the city, and maybe with innovation, for which we can use the city Innovation index [27] or the number of patents developed in companies from the city. The causal argument is symmetrical from Hidalgo, Hausman and Klinger's argument [4–7] but at the city level. Cities require some underlying capabilities (infrastructure, education, institutions, human capital) to maintain specific companies, and those capabilities are similar for economic activities close together in the product-service space. When a city acquire new capabilities, new products can be developed along the product-service space. Since we are more interesting in understanding the role of interlocks in the process, the aim of this section is descriptive, but hinting on causality. Negative results would suggest an extreme

globalization of the cities, where the benefit of having 'in situ' capabilities is minimal.

C. Interlocks and the product-service space

Next, we will analyze if interlocks are a good predictor of diffusion between economic activities in cities. Development itself influences the presence of interlocks. Companies situated close geographically, or in places with similar language or colonial ties have greater chances to interlock [refs]. Since the establishment of companies within a city allows for greater possibilities of interlocks, we expect a relationship between the product-service space and the number of interlocks between economic activities. However it is not clear if interlocks are only a cause of development but also an effect. Interlocks provide a communication channel between companies, and serve as a link for the spread of strategies and structures [refs]. For instance, [examples of previous research]. We hypothesize that interlocks serve as a communication channel for opportunities, thus increasing investment and R&D to sectors close in the product-service space. The increased investment and innovation has been linked to economic development [4, 25, 26]. In a first step, we will test if interlocks affect the diffusion process in the product-service space. In a second step, we will analyze collaboration between companies using patent data to show if this diffusion process is mediated (at least partially) by innovation.

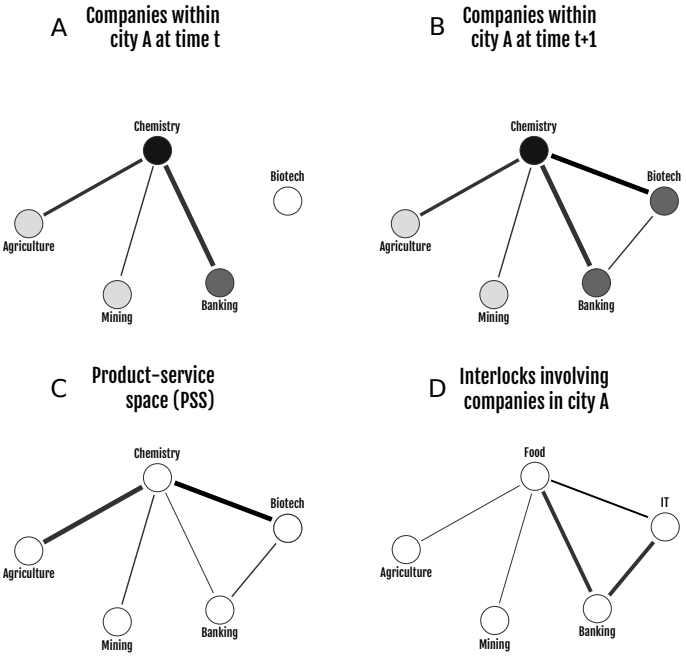

FIG. 5. **Relation between interlocks, the product-service space, and the distribution of economic activities** (A–B) Distribution of economic activities within a city for times (A) t and (B) $t+1$. The weight of the edges is the product of the number of companies in the nodes (the number of possible opportunities to interlock). (C) The product-space service. The weight of the edges indicate the relatedness of products. (D) Number of interlocks from directors sitting on boards in city A.

1. *Interlocks affect the product-service space*

Figure 5 shows our approach. Given the number of companies in city A at time t (Fig. 5A), we want to explain the evolution to time $t+1$ (Fig. 5B). This evolution occur in the edges of the 'product space' (Fig. 5C). We can create a network of interlocks (Fig. 5D), where two economic activities are connected if a director sits in companies from both sectors. Importantly, the network of interlocks have relationships between sectors that are not present in Figure 5A. This is because interlocks are not restricted to the city itself. A director can sit in the banking sector in city A and in the IT sector in city B, even if there are no IT companies in city A. Our research question here is "To what extent do interlocks increase the diffusion rates in the product-service space?".

One method to study if interlocks are predictors of the diffusion process is conditional entropy $H(X|Y)$. The conditional entropy $H(Companies_{t+1}|Companies_t)$ quantifies how much extra information we need to define the structure of the network at time $t+1$ knowing the structure at time t. If the network of interlocks between economic activities affect structural transformation we will find that

$$H(Companies_{t+1}|Companies_t, Interlocks_t) < H(Companies_{t+1}|Companies_t). \quad (1)$$

Other methods include using generative models such as ERGMs [ref] or SIENA [ref] to assess if the presence of interlocks at time t is predicted by $Companies_{t+1}$, better than for $Companies_t$. Such a models allow to control for node attributes (see below) and for recursive network effects – e.g. the probability that there is an interlock A–C increases if there are interlocks A–B and B–C (transitivity), and should be controlled.

There are several problems with this approach. Firstly, we have explained that development creates interlocks (endogeneity problem). The use of longitudinal data allows us to quantify the effect of interlocks on development, independently of the effect of development on interlocks. Secondly, there is a chance of self-selection bias. Companies that want to develop a product in another sector may create interlocks with a company from economic activity beforehand. However, if this is true, interlocks would also facilitate diffusion. Finally, a most important bias is omitted variable bias. If there is an underlying mechanism that produces the interlock at time $t+1$ and the diffusion at time $t+2$, we would find a false effect of interlocks on the diffusion process. For example, cities that are developing fast (for whatever reason) may attract more interlocks than those who are not developing. In order to investigate this possibility,

we need to control for city economic indicators, such as infrastructure, resources, education, city size, population density and growth. Other variables to control are sector size, country indicators and type of interlocks (within city versus between cities). We can create random models using the product-service space and these variables to investigate their effect. Importantly, this approach allows us not only to discover if interlocks play a role in the diffusive process, but also to unravel what other factors also play a role on it. Moreover, positive effects of interlocks in the diffusive process would imply that companies should seek interlocks in companies that are adjacent in the product-service space.

2. Interlocks affect the product-service space (at least) through innovation

In section III C, we investigate if interlocks affect the diffusive process in the product-service space. In this section, we hypothesize that the effect of interlocks in the diffusive process is caused at least partially by an improvement in collaboration and innovation. Innovation has been linked to gaining competitive advantage [28], expanding market share [29] and increasing firm performance [30]. Thus, a correlation in section III B between the product-service space and innovation would not be surprising. The subquestion corresponding to this subsection is "Do interlocks foster collaboration between companies and innovation?". In order to test this hypothesis we will use patent data. Similarly to the interlock case, two companies are connected if they share a patent. We can use generative models such as ERGMs or SIENA to assess if the presence of interlocks at time t facilitates the presence of a shared patent at time $t + 1$. Since we expect other factors to affect the probability of collaboration (such as the presence of a university in the city), we would need to control for those factors (see III C).

D. Which factors affect interlocks

The literature for the factors influencing interlock creation is more consistent. Geographical distance, colonial history, language, education and social networks have been pointed as factors influencing the presence of interlocks [refs]. Time allowing, we will explore this idea at the micro-scale using generative models such as ERGMs [ref] or SIENA [ref].

E. Other projects

The data allow for the exploration of other projects, such as: (i) Describe of the inequality among directors, studying if the inequality has its origin in education. (ii) Analyze the homogenization of coordinated and liberal market economies. (iii) Quantify the independence of a given sector (e.g. food or media). (iv) Measure the transference of power from domestic corporation to transnational corporations. (v) Network motifs, which combination of interlocks between sectors are more likely to occur than random. (vii) Importance of the nation-state in economic networks.

IV. DATA

"The most satisfactory sampling design for structural analysis is a saturation sample of the entire universe or population; however, this alternative is clearly not feasible for large social structures"

— M.P. Allen, 1974

A. Data description

The data from this project was extracted from Orbis [31]. Orbis contains standardized information from 200 million firms and 100 million people. The firm data includes economic indicators (such as turnover, employee number, profit ratios or economic activity using the standard NACE rev. 2), as well as 90 million ownership ties. The directors data includes biographic information (such as name, education, nationality or gender), as well as 151 million position information. Importantly, a person sitting on the board of two companies creates a connection between the firms (interlocks). Since directors can sit on more than one boards, the database has more than 1,000 million interlocks.

The concept of interlocking directorates is related to the corporate elite, part of the power elite. Mills [32] defined the power elite as *"those political, economic, and military circles, which as an intricate set of overlapping small but dominant groups share decisions having at least national consequences. Insofar as national events are decided, the power elite are those who decide them"* [32]. Moreover, Mills determined that there is an 'inner core' of the power elite involving individuals that are able to move from one seat of institutional power to another. In the context of corporations, this inner core corresponds to the corporate elite. A strength of using the concept of interlocking directorates as opposed to corporate elites is that we do not assume a priori characteristics of the elite. Corporate elites on the interlock network are the actors with high centralities values, as measured by network algorithms.

The first step of the project consists on downloading, structuring and storing the data (months 1-4), quantifying the quality of the data (months 4-7) and assessing the effect on the results (months 8-10). For brevity we will skip most the details of the quality assessment. Figure 6A compares the number of companies in Orbis with the OECD estimates. We can see that the quality is extremely good for large companies, but relatively bad for small

companies. However, the companies in Figure 6A are those with available revenue and employment information (60 million instead of 200). Thus, we have most companies in the database, but without financial information.

We developed a two-step approach to assess the quality of the data. In the first step, we developed interactive visualizations to rapidly explore the data (see `https://github.com/uvacorpnet`). The results showed that richer countries (measured by GDP per capita) have larger companies and better data completeness. The visualizations also showed that the observed average revenue for the companies in a country depends on the completeness. Those with higher completeness include also small companies, decreasing the revenue of the 'average' company. In the second step we characterized the data and extrapolate the quality to other countries. The distributions of revenue for a country follow a lognormal distribution, thus can be defined using two parameters (loc and scale). Moreover, the scale is fixed for all countries and the loc can be estimated using macro-economic indicators. This allows us to quantify the type of missing companies (Fig. 6B). In the next months we will assess the effect of completeness in network measures. This assessment will determine to what extent our results are generalizable.

B. Data bias

We position ourselves very far from the data. Given the scope of the project (study the role of interlocks globally) this is the only viable solution. However, the validity of the measurements is still high because we focus on physical phenomena – the distribution of companies in a city and the presence of interlocks between companies. In order to operationalize the concept of city, we understand city as a geographic region with high density of settlements. To find regions with high density of settlements we use the *MeanShift* algorithm [33]. This allow us to automatically merge Boston and Cambridge, New York City and Brooklyn, or Amsterdam and Amstelveen.

Next, we talk about completeness (how much of the data we have), bias (is our sample a random sample), and accuracy (is the data we have true).

Company data: 1. Bias and completeness: We have assessed the bias and completeness of the data (Fig. 6B). In general, Scandinavian countries have the highest quality (very complete and fairly unbiased), while poor countries only have information about their biggest companies. We can now use our method to reconstruct missing data and assess the effect of including it. We expect the effect to be small since small companies are not usually connected in the network – for example the owner of a small shop is not likely to sit in the board of directors of any company.

2. Accuracy: We have checked the accuracy of a small sample and the information in the database is almost always correct. However, there are some problems in this area. Some information providers send their data faster than others, which makes the information of some countries be outdated up to one year. Moreover, large company generally are required to fill consolidated accounts, including in their reports the profits, employees and other economics of their subsidiaries. Because we also have information about the subsidiaries in our database, the profits of the subsidiaries are registered twice. However, also because we have information about the subsidiaries, we can unconsolidate the accounts. Figure 6 is based on

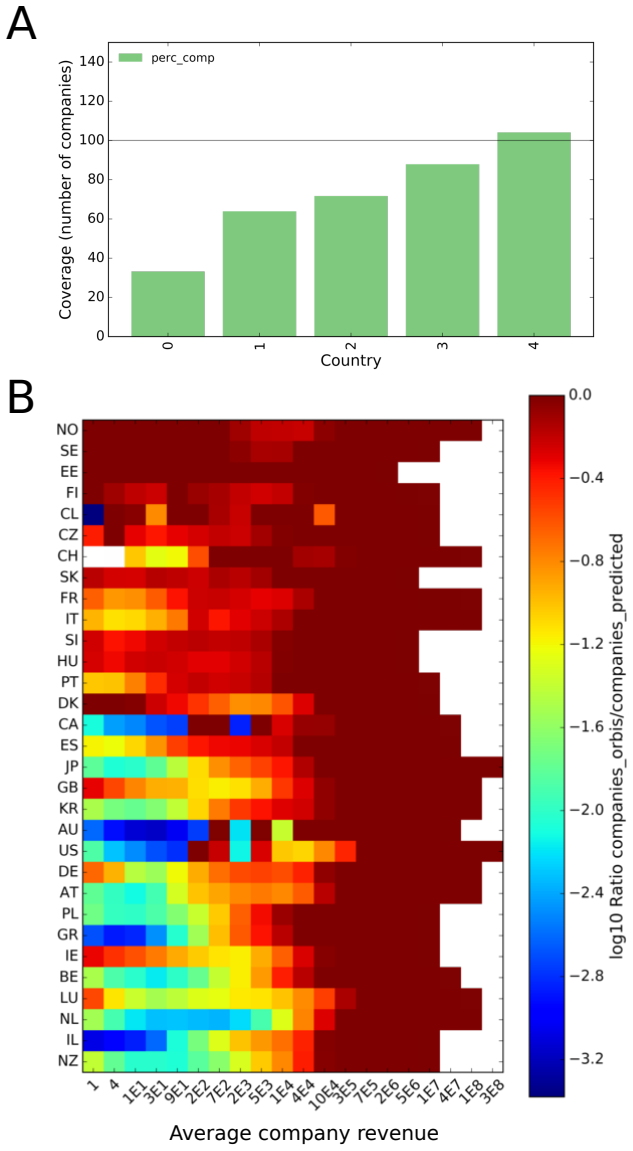

FIG. 6. **Data quality** (A) The coverage of companies with financial information decreases with the company size. (B) Data quality in the OECD. Red colors indicate good coverage, while blue colors indicate bad coverage.

unconsolidated accounts.

Directors data: Missing actors in social networks produce less bias in the network than missing companies [34]. This is important since data quality about directors is more difficult to assess.

1. Bias and completeness: Based on manual inspection of a small sample of the data we found that some directors from small companies are missing. However, we can create confidence intervals by imputing missing directors (rewiring the network) at random multiple times.

2. Accuracy: We found that large companies have extra directors that have already left the company. However, we can create confidence intervals by deleting directors (rewiring the network) at random multiple times. Moreover, the type of position of directors in some countries is unspecified, meaning that we do not know if the person is an administrative (e.g. secretary, lawyer, etc) or a director. However, we can remove administrative ties by using the ownership database. If a director has positions in a shareholder and a wholly owned subsidiary we can filter that tie. A concern in this approach is that two subsidiaries that are wholly owned by company A may not have any ownership relationship between them, but they are still the same corporation. We can correct this by deleting all ties within the same corporate structure.

A final consideration is in the concept of interlocks. We want to measure relationships between companies. However it is not clear that formal interlocks – those occurring by people sitting together on boards – are more important than interlocks created by directors being part of the same social clubs or the same families. By restricting ourselves to formal interlocks we may be missing an important part of the corporate network. However, we expect to capture a significant part of the relations between companies.

V. SCHEDULE

- Year 1:
  - Months 1-4: Get data from Orbis. Set up server.
  - Months 4-7: Analyze data quality and bias.
  - Months 8-12: (i) Study the effects of data quality. (ii) (side project) Analyze inequality in last names in the USA. (iii) (side project) Examine the role of offshore centers in economy.
- Year 2:
  - Create the product-service space.
  - Analyze the relationship between the product-service space and structural transformation.
  - Describe the relationship between interlocks and product-service space.
- Year 3:
  - Model the relationship between interlocks and product-service space and what other factos affect it (causal inference).
  - Analyze if interlocks affect structural transformation by fostering innovation.
- Year 4:
  - Either analyze the factors that determine the presence of interlocks or some other project.
  - Write thesis and get a PhD!

[1] O. Jeidels. *Das Verhältnis der deutschen Grossbanken zur Industrie: mit besonderer Berücksichtigung der Eisenindustrie*, volume 24. Duncker & Humblot, 1905.

[2] V. I. Lenin. volume 1. Zhizn' i znanie, 1917.

[3] M. S. Mizruchi. What Do Interlocks Do? An Analysis, Critique, and Assessment of Research on Interlocking Directorates. *Annual Review of Sociology*, 22(1):271–298, 1996.

[4] C. A. Hidalgo, B. Klinger, A.-L. Barabasi, and R. Hausmann. The Product Space Conditions the Development of Nations. *Science*, 317(5837):482–487, 2007.

[5] R. Hausmann and C. A. Hidalgo. The network structure of economic output. *Journal of Economic Growth*, 16(4):309–342, 2011.

[6] R. Hausmann and B. Klinger. Structural Transformation and Patterns of Comparative Advantage in the Product Space Ricardo. *Faculty Research Working Papers Series Structural*, 2006.

[7] C. A. Hidalgo and R. Hausmann. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, jun 2009.

[8] P. W. Anderson et al. More is different. *Science*, 177(4047):393–396, 1972.

[9] J. H. Miller and S. E. Page. The standing ovation problem. *Complexity*, 9(5):8–16, 2004.

[10] J. S. Coleman and J. S. Coleman. *Foundations of social theory*. Harvard university press, 1990.

[11] M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

[12] M. Weber. Die protestantische ethik und der geist des kapitalismus. 1904.

[13] D. J. Watts. *The collective dynamics of belief*. Stanford University Press, 2007.

[14] M. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[15] A. Clauset, S. Arbesman, and D. B. Larremore. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, 1(1):e1400005–e1400005, 2015.

[16] I. Währungsfonds. Global financial stability report–responding to the financial crisis and measuring systemic risk. *April, Washington, DC*, 2009.

[17] P. Tetlock. *Expert political judgment: How good is it?*

How can we know? Princeton University Press, 2005.

[18] T. C. Schelling. *Micromotives and macrobehavior*. WW Norton & Company, 2006.

[19] J. M. Epstein. Agent-Based Computational Models and Generative Social Science. *Generative Social Science*, 4(5):4–27, 2006.

[20] J. Garcia-Bernardo, H. Qi, J. M. Shultz, A. M. Cohen, N. F. Johnson, and P. S. Dodds. Social media affects the timing, location, and severity of school shootings. *arXiv preprint arXiv:1506.06305*, 2015.

[21] L. Mercken, T. A. B. Snijders, C. Steglich, E. Vartiainen, and H. de Vries. Dynamics of adolescent friendship networks and smoking behavior. *Social Networks*, 32(1):72–81, 2010.

[22] R. Belderbos, S. Du, and D. Somers. Global Cities as Innovation Hubs: The Location of R&D investments by Multinational Firms. page 30, 2014.

[23] E. M. Heemskerk, F. W. Takes, J. Garcia-Bernardo, and M. J. Huijzer. Where is the global corporate elite? A large-scale network study of local and nonlocal interlocking directorates. *Acta Sociologica*, Forthcomin, 2016.

[24] A. Smith and M. Garnier. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London, 1776.

[25] P. Romer. Endogenous technological change. *J Pol Econ*, 98:S71–S10, 1990.

[26] G. M. Grossman and E. Helpman. Quality ladders in the theory of growth. *The Review of Economic Studies*, 58(1):43–61, 1991.

[27] `http://www.innovation-cities.com/innovation-cities-index-2015-global/9609`.

[28] M. a. Hitt, R. E. Hoskisson, R. a. Johnson, and D. D. Moesel. The Market for Corporate Control and Firm Innovation. *The Academy of Management Journal*, 39(5):1084–1119, 1996.

[29] L. G. Franko. Global Corporate Competition: Who's Winning, Who's Losing, and the R&D Factor as One Reason Why. *Strategic Management Journal*, 10(5):449–474, 1989.

[30] G. K. Morbey. R&D: Its relationship to company performance. *The Journal of Product Innovation Management*, 5(3):191–200, 1988.

[31] `orbis.bvdinfo.com/`.

[32] C. W. Mills. JSTOR, 1957.

[33] K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40, 1975.

[34] G. Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3):247–268, 2006.