

Geography versus topology in the European Ownership Network

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2011 New J. Phys. 13 063021

(<http://iopscience.iop.org/1367-2630/13/6/063021>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 129.132.202.35

The article was downloaded on 22/05/2013 at 14:45

Please note that [terms and conditions apply](#).

Geography versus topology in the European Ownership Network

Stefania Vitali^{1,2} and Stefano Battiston^{1,3}

¹ Chair of Systems Design, ETH Zurich, Kreuzplatz 5, 8032 Zurich, Switzerland

² Department of Economics, Università Politecnica delle Marche, P.le Martelli 8, 60121 Ancona, Italy

E-mail: sbattiston@ethz.ch

New Journal of Physics **13** (2011) 063021 (18pp)

Received 1 February 2011

Published 10 June 2011

Online at <http://www.njp.org/>

doi:10.1088/1367-2630/13/6/063021

Abstract. In this paper, we investigate the network of ownership relationships among European firms and its embedding in the geographical space. We carry out a detailed analysis of geographical distances between pairs of nodes, connected by edges or by shortest paths of varying length. In particular, we study the relation between geographical distance and network distance in comparison with a random spatial network model. While the distribution of geographical distance can be fairly well reproduced, important deviations appear in the network distance and in the size of the largest strongly connected component. Our results show that geographical factors allow us to capture several features of the network, while the deviations quantify the effect of additional economic factors at work in shaping the topology. The analysis is relevant to other types of geographically embedded networks and sheds light on the link formation process in the presence of spatial constraints.

³ Author to whom any correspondence should be addressed.

Contents

| | |
|--|-----------|
| 1. Introduction | 2 |
| 2. Data | 4 |
| 3. Link reshuffling and spatial constraints | 5 |
| 4. Basic network statistics | 6 |
| 5. Geographical distance and network distance | 8 |
| 5.1. Geographical distance along links | 8 |
| 6. Geographical distance across longer paths | 9 |
| 6.1. A small geographical world? | 9 |
| 6.2. Geographical distance at varying network distance | 12 |
| 7. Conclusion | 15 |
| Acknowledgments | 16 |
| References | 16 |

1. Introduction

A large body of work has shown how the theory of complex networks can be fruitfully applied to a wide range of domains [1–3]. The case when complex networks are embedded in a geometric space (e.g. geographical) has been attracting growing attention (for a recent extensive review, see [4]). Indeed, when the spatial location of the nodes plays a role in the formation of links, the geometric space contributes to shaping the topology of the network [5–7] and, as a consequence, affects the dynamics that may take place on the network.

Most of the works on spatially embedded complex networks have focused on the geographical distance between first neighbours. However, in all networks in which information or matter flows along links, it is of interest to understand how the geographical distance of nodes that are higher-order neighbours relates to the length of their shortest path. In this paper, we investigate the relation between geographical distance and network distance. In particular, we study how the geographical distance between higher-order neighbours is predicted by the distance between first neighbours and how this affects the small world (SW) properties of the network. We address these issues in an empirical dataset of a directed and weighted network using high-resolution geographical locations.

Empirical works have investigated several contexts in which geographical space is relevant, including transportation services (e.g. airport network), mobility (e.g. commuting), infrastructures (e.g. power grids, roads and the Internet) and social networks (e.g. friendship and phone calls). For instance, the worldwide airport network has been thoroughly investigated [8, 9], as well as various public transportation services at the national [10] or metropolitan level [11]. In all the above-mentioned contexts, the effect of space is that links are associated with a cost growing with geographical distance. As a result, the probability that two given nodes are connected decreases exponentially with distance. A related fact that emerges in a number of contexts is the so-called ‘gravity law’, originally introduced in the geography literature during the 1960s. According to this law, the intensity of the connection between two nodes is proportional to the attractivity of the nodes and is inversely proportional to the square (or variant exponents) of the distance.

From a theoretical point of view, models of spatial networks can be grouped into several classes [4]. Typically, nodes are assigned a given location either according to a convenient geometry (e.g. a lattice) or a distribution (e.g. uniformly scattered) or according to some empirical pattern (e.g. the airport location). On the other hand, links are usually assigned stochastically according to a probability that depends on the distance between two nodes. Some models assume that the network grows with a preferential attachment so that the linking probability depends also on the current degree of the nodes. In general, as is done in fitness models, the linking probability depends on a state variable associated with the nodes. In turn, this can be quenched or is dynamically evolving with the topology. Overall, many models have focused on reproducing the scale-free nature of the degree distribution together with the spatial distribution of the nodes [6, 12]. In addition, others offer some insights into how spatial location may constrain the network topology [13].

Among the domains in which geographically embedded networks have been investigated, the case of economic networks has so far received very little attention. Yet, geographical space plays an important role in economic activities, as witnessed by a rich stream of works on economic geography [14]. In particular, in this field there is an open debate on whether a firm's spatial location (e.g. proximity between firms) is more important than the firm's non-spatial relationships (e.g. R&D collaborations, ownership and venture capital, board interlock) for its competitiveness [15]. For instance, it has been shown that certain corporate practices (i.e. the 'poison pill' versus the 'golden parachute') spread among firms via interlock links, whereas others do so via proximity (i.e. firms simply adopt the practice of nearby firms) [16]. Moreover, a related issue is the extent to which network relations enable economic actors to access knowledge beyond their organizational and geographical boundaries [14, 17, 18]. Finally, economic networks have been recognized to be one of the main research challenges in the field of complex networks [19].

In this paper, we analyse the ownership network of European firms in relation to their geographical location. An ownership relation refers to the fact that a firm owns fully or in part another firm. Ownership plays a role not only in the control that shareholders can exert over a firm [20], but also in the spread of information and business practices in the corporate community. Several properties of these networks have been investigated so far. First, small world (SW) properties have been observed in national samples [21, 22]. Moreover, the scale-free nature of degree and scaling laws, relating degree and portfolio volumes, have been observed in various stock markets [20, 23], as well as in foreign direct investments [24]. To some extent, the geographical information has been considered in [24], where ownership links are aggregated within regions at the NUTS3 level (i.e. roughly corresponding to provinces). Similarly, in [25] the aggregation is at the level of countries, and the links are used to simulate the propagation of financial crises worldwide. However, to our knowledge no previous study has used the geographical location of firms at the postal code level in a large sample.

In our analysis, we find an exponential decay of the linking probability with the distance, similar to what was found in transportation and infrastructure networks. We observe, instead, deviations from the gravity law in the weights of the links. However, the main contribution of the paper is to go beyond the level of first neighbours and to analyse the relationship between network distance and geographical distance. So far, a study of this kind has been carried out only for the metropolitan transportation network of Berlin [11], in which distances are on a much smaller scale. We compare our empirical results with those obtained with a random spatial model (RSM) that belongs to the class of geometric graph models [4]. We keep the empirical

location of nodes and we apply a link reshuffling procedure that preserves the degree sequence and is also subject to a tunable spatial constraint. We then analyse how SW properties are affected by geographical distance (i.e. we look at the shortest path of nodes located at different geographical distances). We find that the empirical network is a ‘smaller world’ than it would be without spatial constraints.

In addition, we analyse how the distribution of geographical distance, which is exponential among first neighbours, is affected when we look at pairs of nodes at higher network distance. Such a distribution turns out to be fairly well reproduced by the RSM. Overall, our results show that geographical space allows the capture of a number of features of ownership networks. Nevertheless, the important discrepancies observed between the data and the model point to the role played by economic factors. In particular, the organization of economic control in the empirical data is significantly more hierarchical than would emerge solely from geographical factors.

2. Data

The dataset covers about 200 000 firms located in 14 EU countries (see table 2), their geographical location at postal code level and about 188 000 ownership relationships among them. Ownership data are obtained from the Orbis database of the Bureau Van Dijk’s, 2007 release, by selecting those firms that have at least one ownership relationship with another firm in Europe (i.e. firms owning shares or having a shareholder firm in the same country or in another EU country). The geographical coordinates of the postal codes of the firms are obtained from the database TeleAtlas GIS⁴. Our dataset allows us to construct the European ownership network, in which the nodes correspond to firms and the links correspond to ownership relationships. The adjacency matrix of the network is defined as A_{ij} , where $A_{ij} = 1$ if there is an edge (or link) $i \rightarrow j$. The network is directed (i.e. A is not symmetric) and the *direction* is taken with the following convention: A_{ij} implies that i has a share of j . The network is also weighted and thus it is associated with the matrix W , with $W_{ij} \in (0, 1]$ corresponding to the ownership share, i.e. the *fraction* of the value of j owned by i . There is a constraint on the weights of the incoming links $\sum_i W_{ij} = 1$, but not on the outgoing links, because a single firm can have large shares in many other firms.

A peculiar aspect of these data concerns the paths in the network. We recall that a path from two nodes i and j not directly connected is a sequence of adjacent nodes from i to j . In many networks, it is not possible to assign a precise meaning to paths. For instance, in the airport network, while a connection between airports A and B reflects the number of connecting non-stop flights per day from A to B, one cannot derive much information on indirect connections, i.e. on the number of passengers who travel from airports A to B and then to airport C in the same journey. In contrast to ownership relationships, if firm A owns W_{AB} shares of B and B owns W_{BC} shares of C, this implies an indirect relationship in which A owns indirectly $W_{AB}W_{BC}$ shares of C [20].

⁴ In the Orbis database, on average the postal code is missing for 15% of the firms selected based on their ownership relations. This occurs at comparable frequency across countries, with the exception of the Scandinavian countries where it occurs more frequently. Typically smaller firms are affected. We thus assume the sample remains representative.

3. Link reshuffling and spatial constraints

Throughout the paper, we aim at assessing the extent to which the empirical properties of the European ownership network are consistent with those of a network obtained from a random link formation process. In this section, we introduce an algorithm to reshuffle the links in the network, taking into account some structural and geographical constraints.

Let us denote the network by the graph $G = (V, E)$, where V is the set of vertexes and $E = \{s, t\}$ the set of edges connecting the vertexes, with s being the origin (or source) vertexes and t the target vertexes of the link (s, t) . Let $W_{s,t}$ be the share of firm t held by firm s . There are no loops in the network, i.e. $s \neq t$ in the link (s, t) . We want to construct a randomized network that preserves the following properties:

1. Degree sequence, i.e. the number of outgoing and incoming links of each node.
2. Ownership share. As mentioned in section 2, the total weight of incoming links has to sum up to one.
3. Geographical location. If geography matters, then constraining the origin and the destination of a link should allow us to recover some structural properties of the network.

As in the Maslov–Sneppen algorithm [26–28], the basic idea is to choose random pairs of links and swap the sources of the two links. This automatically satisfies constraints 1–2, provided that each new link of the pair is accepted if and only if: (i) it is not a loop—the no-loop condition; and (ii) it does not already exist—the no-multiple-links condition. A possible implementation consists in randomly choosing pairs of links and rejecting those swaps that violate the no-loop and no-multiple-links conditions. A more instructive way is to determine the pairs of links that need to be excluded *a priori* from the set of those eligible to be swapped. Consider the two links (s_1, t_1) and (s_2, t_2) ; the patterns to be excluded are listed below:

- coinciding sources: $s_1 = s_2$ (case 1);
- coinciding targets: $t_1 = t_2$ (case 2);
- chains: $t_1 = s_2$ or $s_1 = t_2$ (case 3);
- Z-pattern: the edge (s_2, t_1) exists in the original network (case 4);
- Z-mirror pattern: the edge $s_1 = t_2$ exists in the original network (case 5).

An example of each pattern is provided in table 1, where it is straightforward to see that swapping the source nodes s_1 and s_2 would create loops or multiple links or both. One can also verify that these are all the possible patterns to be excluded. After excluding the patterns above, we are sure that each link either will be swapped or cannot be swapped in the current configuration of the network. Compared with the Maslov–Sneppen algorithm, this method makes more apparent the topological limits to the reshuffling procedure. Incidentally, it is enough to loop once along the list of links to ensure that all the eligible links have been effectively swapped.

The randomization procedure described above, which we call the random direct model (RDM), does not consider the geographical location. In each swap, the source s_2 is chosen randomly among all the existing sources in the list of eligible links. Clearly, sources with large out-degree have higher probability of being selected, because they occur in several links. But their location does not matter. We thus introduce the RSM, in which the choice of the source

Table 1. Particular cases of edges to be excluded for correct swapping.

| | Case 1 | Case 2 | Case 3 | Case 4 | Case 5 | | | | | | | |
|-----------------|--------|--------|--------|--------|--------|---|----|----|----|----|----|----|
| The source, s | 1 | 1 | 4 | 5 | 7 | 8 | 9 | 10 | 10 | 13 | 14 | 13 |
| The target, t | 2 | 3 | 6 | 6 | 8 | 7 | 11 | 12 | 11 | 15 | 16 | 16 |

s_2 is stochastically dependent on its geographical proximity to source s_1 , according to an exponentially decaying probability,

$$\Pi_{s_2} \propto e^{(-d(s_1, s_2))/d_c} \quad (1)$$

where $d(s_1, s_2)$ denotes the distance between the two sources to be swapped and d_c is a characteristic distance. The value of d_c modulates the role of geographical distance. For $d_c \rightarrow 0$, the source s_2 tends to be the firm closest to s_1 among all the eligible sources. For $d_c \rightarrow \infty$, the source s_2 tends to be chosen randomly and in this limit RSM coincides with RDM.

Finally, it should be noted that the procedure of the RSM described above shares some basic principles with the class of geometric graph models, as classified in [4]. In particular, in [6], nodes are located in a d -dimensional lattice and are assigned the degree from a given distribution. Then, links are formed by connecting each node to nodes within a given radius until the degree is reached. In contrast, in [12] the probability of forming a link is a power law of the distance with exponent δ , which plays the role of control parameter. Geometric graph models have also been combined with the fitness model [29]. In [30] links are formed depending both on the distance and on the fitness values associated with the two nodes of each pair.

In the present work, the location of nodes is taken from the empirical data of the firm's location. Thus, the concentration of nodes in space is quite heterogeneous (e.g. high density in metropolitan areas and low density in rural areas). The degree is also taken from the empirical data and follows approximately a power law. The probability of forming a link decays exponentially with the distance. Our aim is not to reproduce the distribution of degree nor the distribution of the distance between first neighbours. Instead, we use both degree and distance as input to determine the properties they induce in terms of: (1) geographical distance between higher-order neighbours, (2) small world and (3) connected components.

4. Basic network statistics

As a preliminary step, we perform a coarse grain analysis of the location of sources and destinations of links. For each country r , we compute an integration index, defined as the ratio of intra-country links over the total number of links departing from that country, $\sum_{i \in c_r, j \in c_r} A_{ij} / \sum_{i \in c_r, j} A_{ij}$, where c_r is the set of firms located in country r . This is a common indicator of economic integration in the economic geography literature [31]. We extend this index to a weighted integration index by taking into account the weights of the links, $\sum_{i \in c_r, j \in c_r} W_{ij} / \sum_{i \in c_r, j} W_{ij}$. In both cases, we find that about 80% of the ownership relations (both in number and in weight) are intra-country (i.e. the owner and the owned are in the same country); see table 2. This first result implies that, already at the level of countries, a preference for links to be formed at short distance is apparent. Notable exceptions are Switzerland and Luxembourg. A plausible explanation is that many companies set their headquarters in these

Table 2. The number of firms and the integration level for each country analysed.

| Countries | Number of firms | Integration index | Weighted integration index |
|-----------|-----------------|-------------------|----------------------------|
| AT | 5 418 | 0.844 | 0.839 |
| BE | 7 823 | 0.680 | 0.759 |
| CH | 1 071 | 0.228 | 0.286 |
| DE | 62 798 | 0.902 | 0.950 |
| DK | 11 013 | 0.847 | 0.821 |
| FI | 573 | 0.699 | 0.671 |
| FR | 21 514 | 0.753 | 0.842 |
| GB | 54 205 | 0.923 | 0.959 |
| IT | 6 688 | 0.869 | 0.813 |
| LU | 863 | 0.309 | 0.328 |
| NL | 36 237 | 0.672 | 0.824 |
| NO | 5 401 | 0.856 | 0.894 |
| PT | 1 953 | 0.928 | 0.945 |
| SE | 6 290 | 0.771 | 0.791 |

two countries for fiscal reasons, but operate in foreign countries via other firms connected via ownership links.

Moving to the basic measures of network connectivity, since this is a directed network, we have to distinguish between the in-degree of a node, denoted by k_{in} (i.e. the number of incoming links), and the out-degree, denoted by k_{out} (i.e. the number of outgoing links). As shown in figure 1 (left), in our dataset the out-degree has a very broad distribution. Note that the tail of the distribution deviates from the straight line that fits the bulk of the data. This means that the frequency of firms with out-degree in the upper end of the range exceeds what could be expected even based on a power-law behaviour. In the context of ownership, the out-degree corresponds to the number of firms in which a given shareholder owns shares. This number is related to the level of portfolio diversification of a shareholder [20], and it is known to correlate with the volume of the portfolio of the shareholder. Thus, the result is essentially in line with what was previously observed in other ownership datasets [20, 23, 24]. On the other hand, in-degree corresponds to the number of shareholders of a firm and, consequently, to the level of ownership concentration, although here weights have to be taken carefully into account [20]. In our dataset, as in previously studied ones, the in-degree distribution is bounded because of the reporting system: only important shareholders are recorded in the database. Thus, the distribution of in-degree displays a clear cut-off.

We then analyse the weight of the ownership relations. Interestingly, the empirical observations suggest a tri-modal distribution of weights. As shown in figure 1 (right), the majority of the relations imply full ownership (e.g. 100% of the shares) or a very small share, while there are a minor, but considerable, number of links with weight just above 50%. This evidence can be interpreted as follows. The first pattern corresponds to a relationship between a holding company and a subsidiary. The second one corresponds to an investment without the aim of controlling the target firm. The third one corresponds to an obvious intention to control the target firm without owning all the shares, since $50\% + \epsilon$ is the minimum share that allows full control.

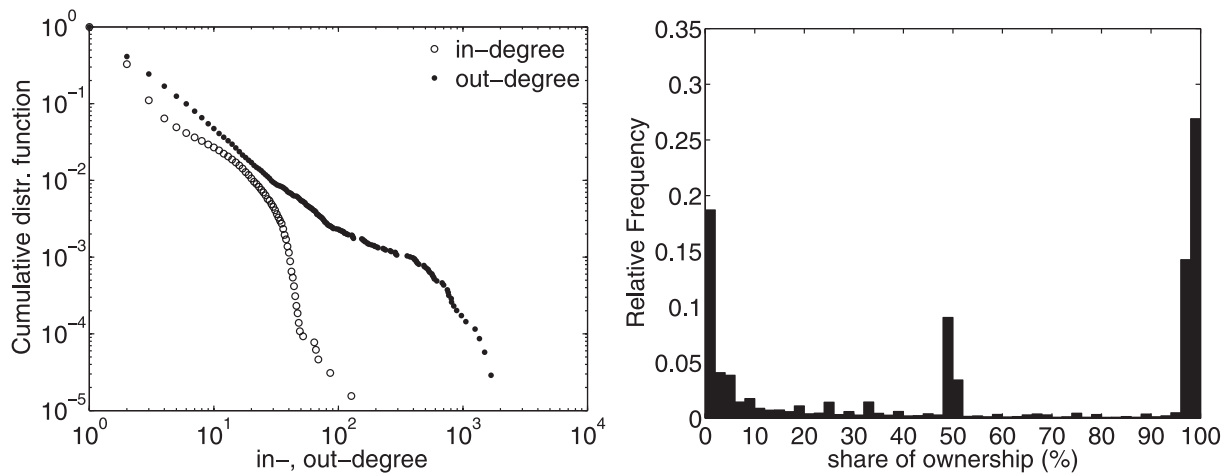


Figure 1. Left: distribution of the nodes' in- and out-degree. Right: distribution of ownership weights.

In terms of components, the largest weakly connected component (LWCC) in the network comprises 32.6% (70 070) of all firms. Since all the other components are significantly smaller (the second largest one contains only 134 nodes), from now onwards we focus only on the LWCC.

5. Geographical distance and network distance

5.1. Geographical distance along links

We first analyse the geographical distance between all pairs of nodes connected by a link. As shown in figure 2 (left), we find that the frequency distribution of links decreases exponentially with the distance: $P_{ij} \propto \exp(-d_{ij}/d_c^*)$, with $d_c^* = 1/0.0027 = 370.37$ km ($R^2 = 0.949$). This result is consistent with previous empirical works on geographically embedded networks. The airport [7] and the Internet router network [32, 33] also display exponential decay in the distance probability. This finding should not be confused with the so-called 'gravity law' from the geography literature [34, 35], which predicts that the strength of a relationship between two entities positioned in the geographical space decreases as a power law of their distance: $W_{ij} \propto (m_i m_j)/d_{ij}^\alpha$, where m_i and m_j represent the attractivity of the entities and α is a parameter. In our context, the role of attractivity could be played by some proxy of the size of the firms, which unfortunately was not available in our dataset. For the sake of completeness, we analysed how the link weight depends on the distance, assuming that $m_i = m_j = 1$. Figure 2 (right) shows the frequencies of edges between two firms at a given distance and with a given weight value. For a given geographical distance interval, the colour code indicates how the edge weights are distributed. We can see that for small distances (i.e. roughly $< d_c^* \approx 370$ km), most ownership relationships are associated with shares close to 100%. This would suggest that many holding companies have subsidiary firms located nearby, e.g. in the same city. In many cases, a subsidiary is located in the same postal code as the owner company; consequently, the two companies share the same headquarter location. In contrast, at larger distance, most relationships involve small shares ($< 5\%$). This finding may be due to the fact that at short

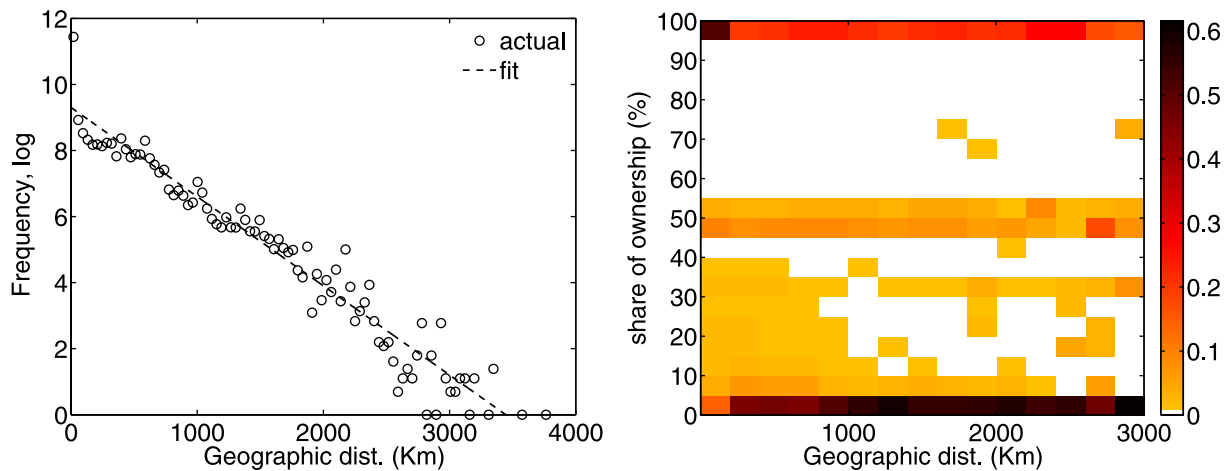


Figure 2. Left: distribution of the geographical distance between a pair of companies tied by a direct ownership link. Right: relative frequency of ownership shares as a function of distances. The range of distance values is divided into bins of 100 km and the range of ownership shares into bins of width 5%. The colour code represents the count of links falling in each two-dimensional (2D) bin. Counts are normalized: each column represents the probability distribution of weight for a given distance.

distances, companies have better possibilities to monitor the management of the companies they own and to influence their decisions. The investment is perceived to be safer and thus a higher share is invested. In contrast, longer distances may imply differences in culture and legal settings other than supervision difficulties, making the investment more risky. Consequently, shareholders are willing to hold smaller shares. However, it is important to observe that about 30% of the relationships are still associated with shares of more than 95%. There is also a smaller but not negligible proportion of shares of around 50% that, similarly, remains constant with distance. Even by averaging weight values in each distance bin, we do not obtain a monotonic decay of the weight with the distance. Thus, our results seem to deviate from the prediction of the gravity law.

6. Geographical distance across longer paths

6.1. A small geographical world?

SW properties have been found in most empirical complex networks, including networks of corporate board membership [36] and ownership [21, 22]. We first verify that the European Ownership Network is also a SW; we then investigate how this property interacts with geographical distance.

6.1.1. Directed small world (SW). A weaker definition of SW is that the average network distance (i.e. the average path length) of the graph is comparable to that of a random graph of the same size. A stronger definition of SW implies that, in addition, the network has a high clustering coefficient when compared with a random graph [37]. We recall that the clustering

Table 3. SW statistics for our European ownership data, the Erdős–Rény model (ERM) and the RDM.

| | Data | ERM | RDM |
|------------------------------------|--------|--------|--------|
| Clustering coefficient (C) | 0.0885 | 0.0001 | 0.0041 |
| Characteristic path length (L) | 7.65 | 8.59 | 7.68 |

coefficient is the average ratio of the number of existing links among the neighbours of a node i to the number of all their possible links (see [38]):

$$C = \frac{1}{N} \sum_i^N \frac{\sum_j \sum_h (A_{ij} + A_{ji})(A_{ih} + A_{hi})(A_{jh} + A_{hj})}{2[k_i^{\text{tot}}(k_i^{\text{tot}} - 1) - 2k_i^{\leftrightarrow}]}, \quad (2)$$

where k_i^{tot} is the total degree of i and k_i^{\leftrightarrow} is the number of bilateral edges between i and its neighbours⁵. On the other hand, the average path length is the average number of edges along all the shortest paths connecting all pairs of nodes,

$$\bar{\ell} = \frac{1}{N(N-1)} \sum_{i \neq j} \ell_{ij}, \quad (3)$$

with ℓ_{ij} being the shortest path connecting i and j [3].

As benchmarks to compare the values of C and $\bar{\ell}$, we consider two synthetic networks. The first one is an undirected Erdős–Rény random graph (Erdős–Rény model (ERM)) [2] with the same number of nodes and edges as that in the empirical network. The second one is generated with the reshuffling method described in section 3 and is referred to as RDM⁶. In the case of ERM, analytical expected values of clustering C and average path length $\bar{\ell}$ are known. For the former, it is simply $C = k/n$, where k is the average degree and n the number of nodes. For the latter, $\bar{\ell} = \ln(n)/\ln(k)$ [3]. In the case of the reshuffled network, we can measure the quantities only empirically. As shown in table 3, we find that $\bar{\ell}^{\text{actual}} \simeq \bar{\ell}^{\text{random}}$ and $C^{\text{actual}} \gg C^{\text{random}}$ for both synthetic benchmarks. $\bar{\ell}$ takes similar values in the three networks. In contrast, C in the empirical networks is almost 10^3 times larger than that in the ERG and about 20 times larger than that in the reshuffled network RDM. The difference in C between ERM and RDM reflects the impact of imposing the constraint on the degree sequence. Our result is in line with those of [21, 22, 39].

6.1.2. SW properties at varying geographical distance. The fact that the European ownership network is an SW (in the stronger sense) suggests that the geographical location of companies does not matter much for the exchange of business practices and information, and for exerting corporate control. Indeed, the existence of shortcuts shrinks the network distance, allowing firms to reach other firms located far away in the network. An important question that, to our knowledge, has not been addressed so far is to what extent these shortcuts also reduce the geographical distance. In particular, when a network is embedded in geographical space, we

⁵ The computation was carried out with the code by David Gleich implementing the algorithm in [38].

⁶ Due to the size of the network, we did not generate a set of reshuffled networks. We simply observed that results were very robust after changing the seed of the random number generator three times.

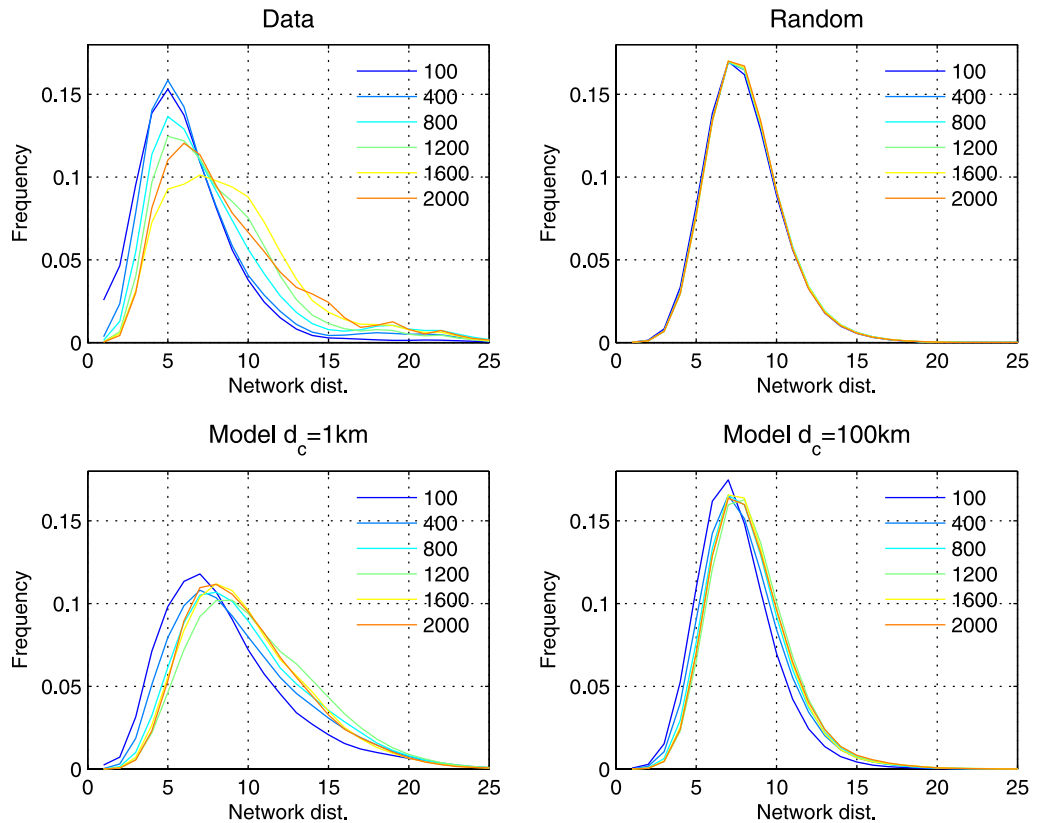


Figure 3. Distribution of network distance for different levels of geographical distance.

can ask how the typical network distance depends on the geographical distance. Since most of the links are between nodes at short distance, one could expect that the further two nodes are geographically apart, the larger is the shortest path that connects them.

Figure 3 (top left) shows the distribution of network distance for several values of geographical distance. We select all pairs of nodes located within a given interval of geographical distance d and we plot the distribution of their shortest path length ℓ . The empirical distribution of ℓ shifts slightly towards the right for increasing values of geographical distance. Moreover, we observe that at large geographical distance, some probability mass builds up on values of ℓ around [15–25]. This means that, proportionally, among pairs that are far away geographically, longer paths are more probable. However, the distribution is essentially stable, which means that the SW property (in the weak sense) is robust with respect to geography. As a comparison, in the random benchmark RDM all distributions coincide. This is not surprising as, by construction, links are assigned without dependence on the geographical distance. On the other hand, the RSM reproduces qualitatively the shifts to the right in the distribution, but does not reproduce the distribution itself. As expected, with increasing characteristic distance d_c , the distribution becomes closer to that obtained with RDM. However, even at very small d_c , the distributions are broader and are centred at higher values of ℓ .

Figure 4 compares directly the empirical distribution with those obtained with the various models. Interestingly, in the empirical network, the network distance tends to be smaller than in its randomized versions, especially when nodes are geographically close. Note that this

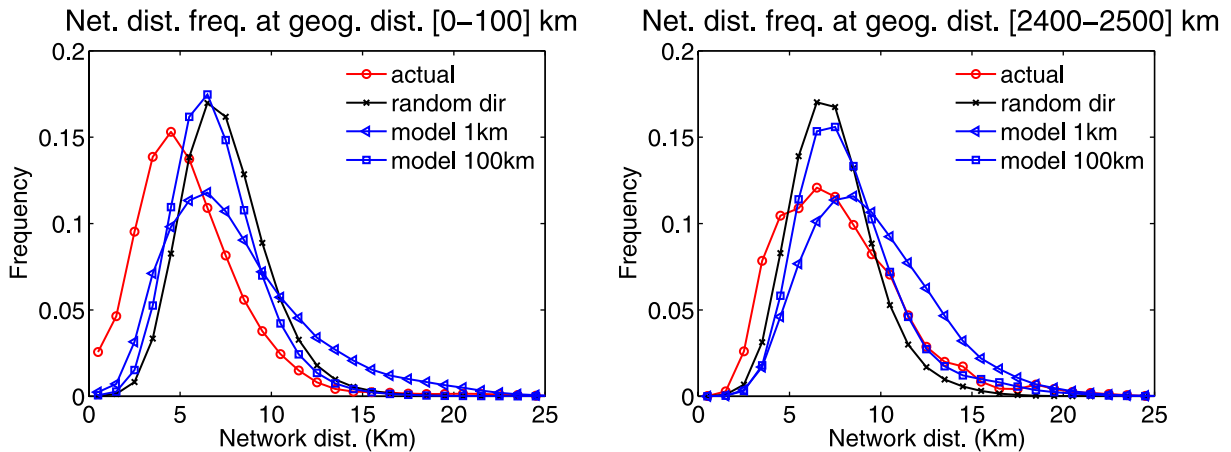


Figure 4. Distribution of network distance for fixed levels of geographical distance and different models.

holds even if the randomization takes geographical distance into account. Indeed, in figure 4 (left), the distribution of the RSM model is to the right of the empirical one at both small and large geographical distance (recall that RDM coincides with RSM at infinite d_c). The values of mean network distance (with their respective error of the mean) are: 7.64736 ± 0.00069 for the empirical network and 9.83233 ± 0.00020 for RSM (1 km). These findings suggest that (i) there is a strong SW property, especially at short geographical distance, and that (ii) SW are quite stable across geographical distance.

An analysis similar to that described above has been carried out only in a few previous works, e.g. in the case of metropolitan public transportation networks. There, the distance between initial and final stations of passengers' journeys scales as a power law of the path length (i.e. the number of stations along the journey) [11]. This behaviour reflects the fact that journeys are typically planned with the purpose of travelling the longest possible distance along the shortest number of stations. In contrast, in our network, the path length is only slowly growing with the geographical distance. This finding means that when firms acquire shares of other firms, they build ownership chains that tend to depart from the origin, but typically are not intended to travel long distances in the shortest number of steps.

From a theoretical point of view, it is interesting to mention here the model of [12] in which, depending on the control parameter δ and the dimension d of the space in which the network is embedded, three different qualitative topological regimes are observed, i.e. random graph for $\delta < d$, lattice-like for $\delta > 2d$ and scale-free in the interval $d < \delta < 2d$. In our case, $d = 2$ and the distribution of distances is exponential, i.e. the closest case is $\delta > 3$. This would imply that our network is located at the border between the scale-free regime and the lattice regime. Indeed in our network a majority of the links are at short distance as in a lattice, but the degree is definitely scale-free. There are, however, important differences between the two models in the way the network is constructed, so an understanding of the relations between their results is deferred to future investigations.

6.2. Geographical distance at varying network distance

Finally, we investigate how geographical distance between two nodes i and j connected by a path varies with their network distance ℓ_{ij} . We already studied in section 5.1 the case $\ell = 1$,

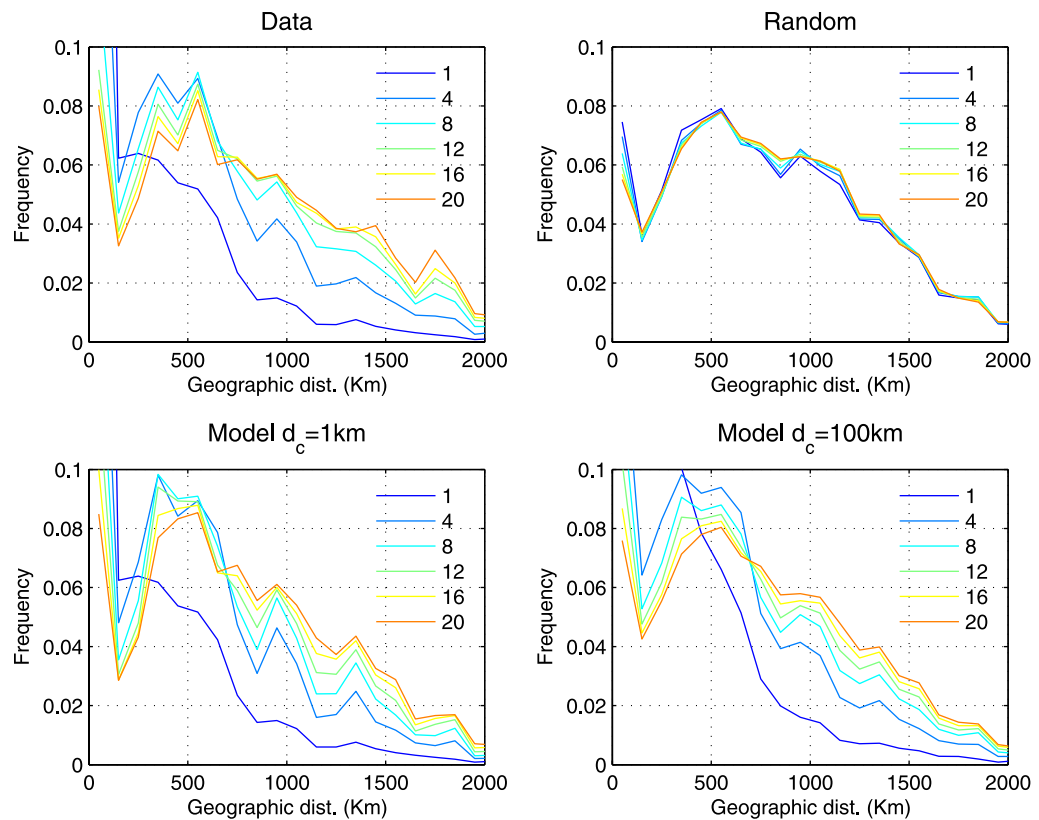


Figure 5. Distribution of geographical distance for different levels of network distance.

i.e. when the two nodes are directly connected, and we found an exponential decay with the geographical distance. We now ask how the geographical distance is distributed when nodes are located at different network distances. One may expect that, for nodes located at large network distance, geography does not matter any more. Then, their geographical distance should be distributed as a random variable subject only to the constraint imposed by the distribution of locations.

Thus, two further questions arise. Firstly, given that for $\ell = 1$ it is an exponential, how does the empirical distribution evolve to the random case? Secondly and more important, is the distribution of distance induced by the RSM consistent with the empirical one? In other words, is the RSM able to reproduce the distribution of geographical distance between nodes that are connected not directly by a link, but via a longer path?

Figure 5 shows that the empirical distribution evolves from the exponential one to a bimodal one. The location of firms in the space reflects the specific geographical characteristics of the countries included in the dataset. For instance, the concentration of firms in urban areas that are disjoint may result in links associated with values of distance that are either small (i.e. two firms in the same urban area) or larger than a minimal distance (i.e. two firms in two different areas). In other words, the bimodality of the distribution may be explained by the fact that firms seek to have relations in the next closest large city, which typically is not closer than 200 km. Hence the gap and the hump in the distribution. The presence of geographical constraints such as mountains, seas, rivers, etc may also contribute to shaping the distribution in a similar way. Moreover, as seen in section 4, firms tend to prefer links within their respective

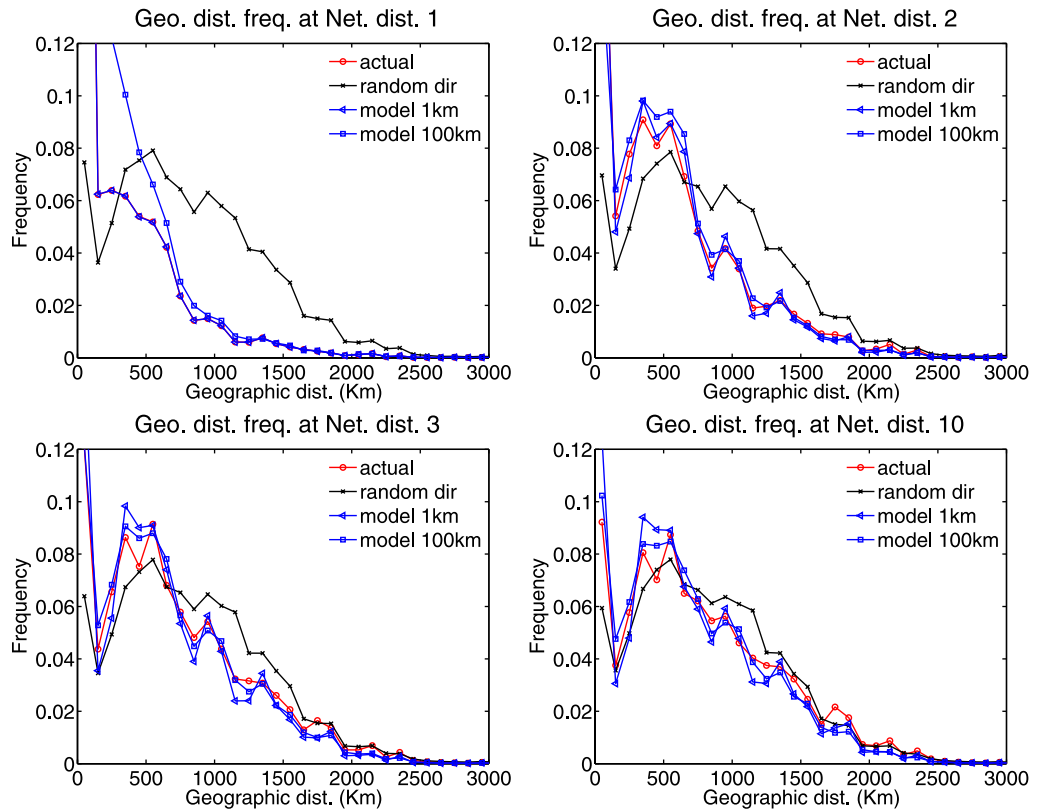


Figure 6. Distribution of geographical distance for different levels of network distance and different models.

national boundaries because of legal settings and cultural similarities. This could create a sort of finite-size effect that could be absent or smaller in a larger country such as the US. However, even a large country is divided by administrative boundaries that may have a similar effect. In any case, we are not able to quantify this effect with our data. From $\ell \approx 8$ onwards, the empirical distribution starts looking like the random one. As shown in figure 5, at large network distance, the distribution in the empirical network and in RSM tends to coincide with that of RDM.

Moreover, as shown in figure 6, the distribution obtained with RSM follows quite closely the empirical one when $d_c < 100$ km. At $\ell = 1$, this is expected because, in the limit of $d_c \rightarrow 0$, RSM replaces the source of a link with the closest eligible source and therefore the new value of distance for a given pair is close to the original value if the area is densely populated. In contrast, the match observed at $\ell > 1$ is not a trivial finding, implying that the network distance induced by the RSM is roughly correct.

As we have seen, RSM is able to fairly reproduce the empirical distribution of geographical distance, for different levels of network distance. In contrast, there are significant deviations in the distribution of network distance. These deviations are also reflected in other topological properties, as reported in table 4. In the randomization process, the size of the largest weakly connected component is preserved within a few per cent. However, the number of strongly connected components (SCCs) and the size of the largest one (LSCC) are radically modified. The empirical network contains 1544 SCC (most with less than 10 nodes) with a LSCC consisting of 502 nodes. In contrast, RSM yields an LSCC that is at least seven times larger. For

Table 4. Network and geographical statistics.

| | Avg net. dist. | Avg geo. dist. | CC | Size largest CC | SCC | Size largest SCC |
|--------------------|----------------|----------------|-----|-----------------|------|------------------|
| Data | 7.65 | 9.56 | 1 | 70 070 | 1544 | 502 |
| RDM | 8.03 | 8.95 | 667 | 68 602 | 2 | 13 836 |
| RUM | 5.31 | 8.75 | 667 | 68 602 | 667 | 68 602 |
| GRM $d_c = 0.1$ | 10.59 | 9.15 | 871 | 67 699 | 570 | 3620 |
| GRM $d_c = 1$ | 9.83 | 8.87 | 802 | 67 989 | 371 | 5763 |
| GRM $d_c = 10$ | 9.05 | 8.77 | 767 | 68 348 | 39 | 9337 |
| GRM $d_c = 100$ | 8.13 | 8.74 | 746 | 68 364 | 2 | 12 980 |
| GRM $d_c = 1000$ | 8.15 | 8.92 | 674 | 68 554 | 1 | 14 091 |
| GRM $d_c = 10000$ | 8.11 | 8.91 | 669 | 68 588 | 1 | 13 791 |
| GRM $d_c = 100000$ | 8.04 | 8.93 | 676 | 68 596 | 1 | 13 900 |

increasing values of d_c , the LSCC further increases (with some fluctuations) up to a size about 28 times larger (14 000 nodes) than the empirical LSCC. Notice that in the undirected case, the LSCC coincides of course with the LCC.

Such topological discrepancy implies that in the empirical network the LSCC is smaller than what would be obtained with a randomization of the links. The LSCC and the set of nodes upstream of it are those where most of the economic control flows [20]. Thus, a smaller LSCC is associated with a more hierarchical organization of economic control. By taking into account the geographical distance, the RSM captures part of this organization. The remaining discrepancy can be explained by the fact that in addition to geographical proximity, economic factors are also at work. Simply put, the likelihood that a firm acquires a share in another firm must depend also on the size of its balance sheet and not only on its geographical proximity to the target, as instead is assumed by RSM. In particular, RSM allows for a shareholder with shares in big firms to become a shareholder with the same degree, but with shares in small firms. Therefore, the economic power of nodes can be modified by the reshuffling. Note that this can destroy the correlation between the degree and the volume of the portfolio [23]. The result is a network with a much larger SCC, which implies many more paths along which control and exchange of business practices can flow.

7. Conclusion

In this paper, we have investigated the network of ownership relationships between European firms and its embedding in the geographical space. Our results, beyond the area of economic networks, are relevant to the field of spatial networks [4], an area of recent development that deserves greater attention. Specific to our dataset is that geographical information is at the postal code level and that even indirect connections have a well-defined meaning.

In terms of basic properties, as already found for other ownership networks, this network displays a broad degree distribution (with quite a fat tail), is a small world and has a trimodal distribution of weights. In terms of geography, we first focused on nodes directly linked, i.e. first neighbours in the network. In this case, the distribution of geographical distance decays exponentially, with a characteristic distance of about $d_c \approx 370$ km. This finding, which is reported here for the first time for ownership networks, is in line with the finding for other

types of networks embedded in geography. In contrast, the dependence of link weights on the geographical distance is not monotonic and seems to be at odds with the gravity law, which is commonly assumed in the geography literature.

An important contribution of this paper is to look also at geographical distance between nodes that are not first neighbours, but are connected via paths of length $\ell > 1$. As ℓ increases, the distribution evolves from an exponential decay to a bimodal distribution that reflects the specific geographical location of firms in the EU. On the one hand, the distribution of geographical distance is qualitatively reproduced, even at $\ell > 1$, by an RSM that reshuffles the links in the network with a likelihood that depends on the geographical distance between the nodes to be swapped. On the other hand, the empirical distribution of the network distance, for a given geographical distance, is not well reproduced by such a model. Empirical distances tend to be shorter than in the model, implying a world ‘smaller’ than it would be by chance, especially at short geographical distance. This is reflected also by deviations in other topological quantities. In particular, the empirical network is characterized by a largest SCC much smaller than would be obtained in the benchmark models. Such a deviation quantifies the extent to which, in addition to geographical ones, economic mechanisms are at work. Their effect seems to lead to a more hierarchical organization of the network.

The relevance of our results for the economic geography literature lies in the investigation of the extent to which the geographical factor is able to explain the organization of links among firms. In particular, we find that the structure of the European ownership network is only partially explained by the geographical location of firms. Although geography matters a great deal, firms establish ownership relations also on the basis of socio-economic factors. For example, additional variables to be taken into account are the size and the sector of the firms or the presence of special structures such as ‘pyramids’ and ‘cross-shareholdings’, which firms build up in order to access corporate control and fiscal advantages [40].

The analysis presented in this paper can be applied to other kinds of geographically embedded complex networks. A comparison of the results across domains could contribute to a better understanding of the mechanisms underlying the formation of spatially embedded networks.

Acknowledgments

We acknowledge F Schweitzer, A Garas and M Gallegati for fruitful discussions and valuable feedback. SV acknowledges financial support from the EC FP7/2007-2013 ‘FINNOV’ Project (no. 217466). SB acknowledges financial support from the ETH Competence Center ‘Coping with Crises in Complex Socio-Economic Systems’ (CCSS) through ETH research grant CH1-01-08-2 and the European Commission FET Open Project ‘FOC’ (no. 255987).

References

- [1] Newman M E J 2003 The structure and function of complex networks *SIAM Rev.* **45** 167–256
- [2] Cohen R and Havlin S 2009 *Complex Networks. Structure, Robustness and Function* (Cambridge: Cambridge University Press)
- [3] Caldarelli G 2007 *Scale-Free Networks: Complex Webs in Nature and Technology* (Oxford: Oxford University Press)
- [4] Barthelemy M 2010 Spatial networks arXiv:1010.0302

- [5] Barthelemy M 2003 Crossover from scale-free to spatial networks *Europhys. Lett.* **63** 915
- [6] Rozenfeld A F, Cohen R, Ben-Avraham D and Havlin S 2002 Scale-free networks on lattices *Phys. Rev. Lett.* **89** 218701
- [7] Barrat A, Barthelemy M and Vespignani A 2005 The effects of spatial constraints on the evolution of weighted complex networks *J. Stat. Mech: Theory Exp.* **2005** P05003
- [8] Colizza V, Barrat A, Barthélemy M and Vespignani A 2006 The role of the airline transportation network in the prediction and predictability of global epidemics *Proc. Natl Acad. Sci. USA* **103** 2015
- [9] Guimera R, Mossa S, Turtschi A and Amaral L A N 2005 The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles *Proc. Natl Acad. Sci. USA* **102** 7794
- [10] Hołyst J A, Sienkiewicz J, Fronczak A, Fronczak P and Suchecki K 2005 Universal scaling of distances in complex networks *Phys. Rev. E* **72** 26108
- [11] Von Ferber C, Holovatch T, Holovatch Y and Palchykov V 2009 Public transport networks: empirical analysis and modeling *Eur. Phys. J. B* **68** 261–75
- [12] Kosmidis K, Havlin S and Bunde A 2008 Structural properties of spatially embedded networks *Europhys. Lett.* **82** 48005
- [13] Serrano M Á, Krioukov D and Boguná M 2008 Self-similarity of complex networks and hidden metric spaces *Phys. Rev. Lett.* **100** 78701
- [14] Ter Wal A L J and Boschma R A 2009 Applying social network analysis in economic geography: framing some key analytic issues *Ann. Reg. Sci.* **43** 739–56
- [15] Castells M 2000 *The Rise of the Network Society* (New York: Wiley)
- [16] Davis G F and Greve H R 1997 Corporate elite networks and governance changes in the 1980s *Am. J. Sociol.* **103** 1–37
- [17] Bell G G and Zaheer A 2007 Geography, networks, and knowledge flow *Org. Sci.* **18** 955
- [18] Almeida P and Kogut B 1999 Localization of knowledge and the mobility of engineers in regional networks *Manag. Sci.* **45** 905–17
- [19] Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, Vespignani A and White D R 2009 Economic networks: the new challenges *Science* **325** 422–5
- [20] Glattfelder J B and Battiston S 2009 Backbone of complex networks of corporations: the flow of control *Phys. Rev. E* **80** 036104
- [21] Kogut B and Walker G 2001 The small world of Germany and the durability of national networks *Am. Sociol. Rev.* **66** 317–35
- [22] Corrado R and Zollo M 2006 Small worlds evolving: governance reforms, privatizations and ownership networks, in Italy *Ind. Corp. Change* **15** 319–52
- [23] Garlaschelli D, Battiston S, Castri M, Servedio V D P and Caldarelli G 2005 The scale-free topology of market investments *Physica A* **350** 491–9
- [24] Battiston S, Rodrigues J F and Zeytinoglu H 2007 The network of inter-regional direct investment stocks across Europe *Adv. Complex Syst.* **10** 29–51
- [25] Garas A, Argyrakis P, Rozenblat C, Tomassini M and Havlin S 2010 Worldwide spreading of economic crisis *New J. Phys.* **12** 113043
- [26] Maslov S and Sneppen K 2002 Specificity and stability in topology of protein networks *Science* **296** 910
- [27] Zlatic V, Bianconi G, Diaz-Guilera A, Garlaschelli D, Rao F and Caldarelli G 2009 On the rich-club effect in dense and weighted networks *Eur. Phys. J. B* **67** 271–5
- [28] Zhou S and Mondragon R J 2007 Structural constraints in complex networks *New J. Phys.* **9** 173
- [29] Caldarelli G, Capocci A, De Los Rios P and Munoz M A 2002 Scale-free networks from varying vertex intrinsic fitness *Phys. Rev. Lett.* **89** 258702
- [30] Masuda N, Miwa H and Konno N 2005 Geographical threshold graphs with small-world and scale-free properties *Phys. Rev. E* **71** 036108
- [31] Manna M 2004 Developing statistical indicators of the integration of the euro area banking system *ECB Working Paper No. 300*

- [32] Yook S H, Jeong H and Barabási A L 2002 Modeling the Internet's large-scale topology *Proc. Natl Acad. Sci. USA* **99** 13382
- [33] Lakhina A, Byers J W, Crovella M and Matta I 2003 On the geographical location of Internet resources *IEEE J. Sel. Areas Commun.* **21** 934–48
- [34] Bergstrand J H 1985 The gravity equation in international trade: some microeconomic foundations and empirical evidence *Rev. Econ. Stat.* **67** 474–81
- [35] Fotheringham A S and O'Kelly M E 1989 *Spatial Interaction Models: Formulations and Applications* (Dordrecht: Kluwer)
- [36] Davis G F, Yoo M and Baker W E 2003 The small world of the American corporate elite, 1982–2001 *Strateg. Organ.* **1** 301
- [37] Watts D J and Strogatz S H 1998 Collective dynamics of 'small-world' networks *Nature* **393** 440–2
- [38] Fagiolo G 2007 Clustering in complex directed networks *Phys. Rev. E* **76** 26107
- [39] Conyon M J, Muldoon M R and Hirsch A B 2006 The small world of corporate boards *J. Bus. Fin. Acc.* **33** 1321–43
- [40] Vitali S, Battiston S and Glattfelder J 2011 The network of global corporate control *PLoS-ONE* at press