The Digital Layer: Alternative Data for Regional and Innovation Studies

Abstract

The lack of large-scale data revealing firms' interactions has constrained empirical studies. Utilizing relational web data has remained unexplored to remedy this data problem. We constructed a Digital Layer by scraping the inter-firm hyperlinks of 600,000 German firms and linked the Digital Layer with several traditional indicators. We showcase the use of this developed dataset by testing whether the Digital Layer data can replicate several theoretically motivated and empirically supported stylized facts. The results show that the intensity and quality of firms' hyperlinks are strongly associated with firms' innovation capabilities and, to a lesser extent, with hyperlink relations to geographically distant and cognitively close firms. Finally, we discuss the implications of the Digital Layer approach an evidence-based assessment of sectoral and place-based innovation policies.

Keywords: Web Mining, Innovation, Distance, Network, Natural Language Processing

JEL: O30, R10, C80, D85

Introduction

Innovation and its impact on economic growth have been of great interest in the past decades (Marshall 1890; Schumpeter 1911; Jacobs 1970; Romer 1990). The pioneering works suggest that the innovation capability of organizations reflects their competence in combining existing knowledge and materials (Schumpeter 1911; Weitzman 1998). This combinatorial process does not occur randomly. Commonly, this process occurs as organizations interact and observe their colocated peers. Borrowing methodological tools of network science, scholars from a wide range of disciplines studied how the colocation of firms and inter-firm relations facilitate learning and trigger innovation (Kogut and Zander 1992; Powell, Koput, and Smith-Doerr 1996; Kogut 2000; Lobo and Strumsky 2008; Ter Wal 2014; Strumsky and Lobo 2015; Vedres 2021).

Although three decades of studies contributed considerably to our understanding of how innovation occurs, the lack of large-scale and representative data revealing firms' interactions has constrained empirical studies (Bailey et al. 2018). A large number of empirical studies utilized secondary data to approximate knowledge exchange between companies ranging from patent documents to data on strategic alliances, scientific co-publications, and R&D projects (Owen-Smith and Powell 2004; Autant-Bernard et al. 2007; Ponds, van Oort, and Frenken 2007; Breschi and Lissoni 2009; Crespo, Suire, and Vicente 2016; Abbasiharofteh and Broekel 2020; Simensen and Abbasiharofteh 2022). These data sources, however, typically represent innovative activities of larger firms and publicly funded organizations and say nothing about innovation capabilities of smaller firms, and organizational and service innovations (Archibugi and Planta 1996). For instance, Fritsch, Titze, and Piontek (2020) show that relying only on patent data to capture innovation underestimates knowledge interactions, especially those of smaller firms.

Alternatively, a relatively smaller number of empirical studies conducted surveys or collected region-specific data on new products and sales (Delgado 2018; Lorenzen 2018). This approach to data collection is costly and cannot be easily scaled up to a larger population of firms. Therefore, scholars argue that unresolved research questions in regional and innovation studies call for utilizing alternative data sources and linking them with traditional ones (Duranton and Kerr 2018; Kedron, Kogler, and Rocchetta 2020; Bottai et al. 2022). In addition, triangulating different data sources allows researchers to depict a complete picture of business ecosystems and their learning dynamics (Basole et al. 2015).

In a business context, websites serve as a showcase for firms' products, services, credibility, achievements, critical decisions, strategies, and relationships with other firms (Gök,

Waterworth, and Shapira 2015). This information is usually encoded using text. In recent years, techniques to retrieve and analyze textual data coupled with high-performance machine learning enabled researchers to harvest and analyze this information by employing web scraping and Natural Language Processing (NLP) techniques (Gök, Waterworth, and Shapira 2015; Kinne and Axenbeck 2020; Stich, Tranos, and Nathan 2022). This 'big data revolution' opens up new opportunities for developing an alternative data source to study inter-firm relations. Interestingly, company websites provide critical information on the early phases of the innovation process that one cannot capture by patent data (Gök, Waterworth, and Shapira 2015). For example, Nathan and Rosso (2022) show that retrieving and analyzing the textual data of firms' websites, and media reveal information about the launch of new products and services that are not fully captured by formal intellectual property and surveys. Stich, Tranos, and Nathan (2022) developed a method of identifying economic clusters using web data. Multiple studies reliably detect companies' innovation capabilities by analyzing their websites' text and HTML structure (Daas and van der Doef 2020; Kinne and Lenz 2021; Bottai et al. 2022).

In this study, we focus on relational web data (also known as hyperlinks) that has attracted far less attention than the analysis of textual web content. Scholars identify hyperlinks as the essential structural element of the Internet, revealing information on the association and disassociation of two websites (Park 2003; deMaeyer 2013). Hyperlink data promises a particularly up-to-date and extensive view of the digital reflection of real-world company networks (Park 2003). Multiple studies point towards the significance of hyperlinks for uncovering firms' network relations (Heimeriks and van den Besselaar 2006; Vaughan, Gao, and Kipp 2006; Kinne and Axenbeck 2020; Axenbeck and Breithaupt 2021). For instance, Tranos, Carrascal-Incera, and Willis (2022) extracted hyperlinks between geolocated archived commercial websites in the UK and predicted inter-regional trade flows. Nevertheless, researchers have not exploited inter-firm hyperlink data sources in combination with novel machine-learning methods in innovation studies. This approach may open up fruitful avenues for empirical innovation studies and enrich our current knowledge of the interplay between inter-firm relations and innovation capabilities, which so far have been empirically investigated based on traditional relational data sources.

In doing so, the aim of this paper is twofold. First, we construct a Digital Layer that captures the relationships between companies based on their hyperlink networks and the textual content of their company websites. Also, we link the Digital Layer with several traditional indicators at the firm level. Second, we showcase the use of this developed dataset in economic geography,

regional, and innovation research by summarizing several stylized facts associated with inter-firm relations and innovation capabilities. Next, we empirically test whether the Digital Layer data can replicate the stylized facts. It is important to note that the empirical setting of the study does not seek to infer causal relationships. The reported correlations can, however, guide future research in using hyperlinks as an alternative source of data that complements traditional secondary data sources.

Our results reveal that the quantity (i.e., frequency) and quality (i.e., hyperlinks to innovative firms) of firms' hyperlinks are strongly associated with firms' innovation capabilities and, to a lesser extent, with hyperlink relations to geographically distant and cognitively close (i.e., a similar knowledge base) firms. We organize the remainder of this article as follows. In the next section, we review the literature on inter-firm relations and innovation capabilities and summarize the main findings as several stylized facts. In data and constructing the Digital Layer, we present the data and methodology used to construct the Digital Layer as well as the estimation of the innovation capabilities of firms. Next, we discuss how we created the variables of interest and the estimation strategy to test whether the Digital Layer can replicate the summarized stylized facts. We then present and discuss our results and conclude by underlining the policy implications of our research and accounting for the limitations of our study and potential avenues for future research.

Inter-firm relations and innovation capabilities

In this section, we develop five theoretically motivated and empirically supported stylized facts on inter-firm relations and innovation capabilities by building on multiple strands of literature ranging from network science to management to innovation and regional studies and economic geography.

Stylized fact 1: Taking a central position in an inter-firm network is positively related to firms' innovation capabilities.

In the early 1990s, network scientists developed methodological tools that enable researchers to measure relational attributes numerically. Numerous studies on this topic reveal that the patterns of relationships among individuals and organizations determine the outcome of socio-economic processes (Borgatti et al. 2009). For instance, firms with similar portfolios may show different innovation capabilities based on their position in an inter-firm network.

In the context of inter-firm relations, scholars acknowledge that a knowledge transfer network is one of the main ways whereby companies access complementary resources and

improve their innovation capabilities (Gulati and Gargiulo 1999). Brusoni, Prencipe, and Pavitt (2001) argue in their seminal article that the boundaries of firms are beyond where the activities are performed. They can specialize and integrate new knowledge pieces from multiple technological domains through inter-firm relations. In another management study, Shan, Walker, and Kogut (1994) investigate a reciprocal association between cooperation and innovation. They provide evidence that cooperation affects innovation, whereas the opposite is not the case. Bell (2005) finds that taking a central position in a manager network positively correlates with the increase in the innovation capabilities of Canadian firms. Owen-Smith and Powell (2004) show that networks function as channels of knowledge spillover in Boston biotech cluster, and the ability of firms to use their relations to absorb information accounts for their innovative outcomes. This seems to be relevant at the local level as well. Giuliani and Bell (2005) and Eriksson and Lindgren (2008) provide evidence on the uneven distribution of knowledge among firms, and those well-positioned in the networks are the most productive ones.

Similar empirical findings are reported in internet studies and hyperlink research. For instance, Vaughan and Wu (2004) find that hyperlinks to commercial websites can serve as a business performance indicator. Another prominent example of a hyperlink-based study is the work of Brin and Page (1998), who developed the 'Page Rank' algorithm based on hyperlinks to calculate a site's relevance on the web. This is somewhat related to social network studies, where taking a central position (i.e., a node's degree centrality) is a common measure of importance.

Studies point towards a number of reasons why taking a central position in inter-firm relations may benefit firms' innovation capabilities. Some scholars argue that the formation of inter-firm relations is highly selective and may follow the 'rich-get-richer' logic (Giuliani 2007). From a relational point of view, this implies that only a few firms take central positions, whereas the rest are poorly positioned in the periphery (Barabási and Albert 1999). Thus, seeing interfirm relations as a vehicle to carry information means that only a small share of firms has access to required inputs for innovation. Gulati's (1999) work is among the first studies empirically investigating the 'rich-get-richer' mechanism in inter-firm relations. His work suggests that firms taking a more central position in their network tend to involve in more alliances in the future.

As an alternative rationale, Chandler et al. (2013) argue that firms that take central positions in an inter-firm network can detect future high-reputation partners (i.e., higher perceived quality) and establish new ties with them thanks to their centrality in the network (i.e., higher

status). This argument resonates with the work of Lazega et al. (2012), claiming that status is the main driver of tie formation in uncertain situations.

Furthermore, several scholars interpret the relevance of taking a central position concerning the structural holes and receiving good ideas (Burt 2004). The seminal work of Burt (2004) conceptualizes the value of bridging ties that link otherwise separated regions in a network. Sociologists and innovation studies scholars provide a large body of evidence suggesting how such bridging ties trigger the generation of new ideas and innovation (Crossan and Apaydin 2010; Anderson, Potočnik, and Zhou 2014; Aral 2016; Vedres 2021). Although firms that bridge structural holes do not necessarily need to take a central position, empirical studies show that central firms are more likely to span structural holes. For instance, Mazzola et al. (2018) find that if biopharmaceutical firms can manage to maintain their central position in the inter-firm network for a certain time period, it is more probable that they create bridging relations. They also show that such firms more often develop new products. Therefore, we expect that taking a central position in an inter-firm network is positively related to firms' innovation capabilities.

Stylized fact 2: relations with innovative firms are positively related to firms' innovation capabilities.

While the studies mentioned above rightly shift the attention to the relevance of the structure of inter-firm networks, sociologists discuss that researchers should not remain agnostic about the content of exchanged knowledge. While Moody (2011) takes into account the relevance of the structure of a knowledge transfer network, he emphasizes the role of the content of exchanged knowledge. Also, he argues that the content of exchanged knowledge may interplay with the structural properties of a knowledge transfer network. This argument found evidence in management and organization studies. Kobarg et al. (2019) analyze a sample of 218 innovation projects conducted in manufacturing companies and show that the attributes of knowledge transfer relations account for the nature of the outcome. More specifically, they find an inverted U-shaped relationship between intense interactions and incremental innovation capabilities, and between diverse interactions and radical innovation capabilities. In economic geography research, Haus-Reve et al. (2019) discuss firms receive different inputs through various inter-firm relations because a supply-chain network can carry market-related information. In contrast, relations with knowledge-broking organizations (e.g., research institutes) provide firms with more novel ideas and new commercial applications. Thus, firms

can excel in different innovation modes based on the type of exchanged information through their relations (Jensen et al. 2007; Fitjar and Rodríguez-Pose 2013).

One can argue that firms benefit more when they establish relations with more innovative partners. One reason for this claim is that innovative firms may have a better access to market-related information (Haus-Reve, Fitjar, and Rodríguez-Pose 2019) or excel at combining existing knowledge pieces and materials (Weitzman 1998). Considering these aspects, it is plausible that interaction with an innovative firm is of higher quality. Demirkan et al. (2013) investigated the impact of tie-specific attributes on the evolution of an innovation network of 367 US biotechnology firms. They show that, among other factors, knowledge quality is one of the main determinants of how an innovative network evolves. Similarly, the works of Lin (2014) and Lee et al. (2015), studying manufacturing firms in Taiwan and SMEs in the Republic of Korea, suggest that partnership quality is positively related to the technological innovation capabilities of interacting firms. The latter also found that collaborators' experience positively correlates with innovation output.

Moreover, innovative firms may benefit from their ability to identify and find needed knowledge and expertise in a risky and uncertain environment. In other words, the capability of creating, managing, and maintaining relationships (also known as collaboration capability) leads to a higher degree of innovativeness. Blomqvist and Levy's (2006) systematics literature review of conceptual and empirical research in management studies suggests that collaboration capability is an enabler factor in knowledge creation in an uncertain environment. Firms' status and innovation capabilities may be positively related. The empirical work of Arya and Lin (2007) suggests that collaboration with high-status organizations is beneficial because their higher status enables them to identify and derive needed recourses more efficiently. All the discussed factors that bring about innovation capabilities also provide interaction premium for related firms. We, therefore, expect that relations with innovative firms are positively related to firms' innovation capabilities.

Stylized fact 3: Having only long-distance inter-firm relations negatively affects firms' innovation capabilities.

Geographic distance refers to the physical distance or travel time between two firms. One can also define geographic distance as perceived distance based on the degree of embeddedness of two firms in a common spatial context (Micek 2018). There is a long-lasting tradition of research showing that the likelihood of forming social and advice tie relations decreases

substantially if the geographic distance exceeds a certain threshold (Zipf 1949; Verbrugge 1983; Marmaros and Bruce 2006; Sonn and Storper 2008; Kabirigi et al. 2022).

Around the turn of the twentieth century, Marshall (1890) argued that the availability of specialized suppliers (sharing), specialized workers (matching), and informal interaction (learning) are the main reasons for the tendency of firms to collocate in a common spatial context. The sharing, matching, and learning mechanisms create a learning hub for informal social interaction, facilitate inter-firm collaborations, and substantially reduce transaction costs (Bathelt, Malmberg, and Maskell 2004). The geographic colocation also enables firms to benefit from non-interactive learning through observing other firms (Glückler 2013). This large body of literature advanced into the study of industrial clusters and the geography of innovation (Jaffe 1993; Audretsch and Feldman 1996; Asheim and Gertler 2006).

In the early 2000s, several scholars developed an alternative 'the death of geography' argument based on narratives of globalization and conjectured that the formation of inter-firm knowledge relations might not be negatively affected thanks to the recent advances in communication and transport technologies (Cairncross 2001; Friedman 2005). At least for innovation capabilities, recent empirical evidence mainly supports the positive association between geographic colocation and innovation. To name a few studies, Graevenitz, Graham, and Myers (2022) show that the diffusion of innovation is still spatially bounded despite recent advances in telecommunication and transport systems. Similarly, Balland et al. (2020) and Balland and Rigby (2016) empirically show that a wide range of activities, such as scientific research, innovation, and industry, are geographically clustered. Studies in the extant literature of related diversification also provide evidence of the comparative advantage for the colocation of workers with similar skills and local inter-industry matching driven by skill-relatedness. Skill-relatedness mimics the rationale behind Marshallian externalities (Boschma, Eriksson, and Lindgren 2014; Andersson and Larsson 2022). Therefore, it is plausible that the concentration of innovation activities is driven by the need for face-to-face interaction for transferring tacit knowledge in larger teams with ever-increasing complex topics (Broekel 2019; Bloom et al. 2020; van der Wouden 2020).

While recent studies suggest the importance of colocation for innovation capabilities, the conceptual framework of the local buzz and global pipeline suggests that local interactions (buzz) lead to innovation capability if combined with global collaborative relations (Bathelt, Malmberg, and Maskell 2004). Empirical results supporting this conjecture are mixed. For instance, while Bathelt and Turi (2011) and Berg (2018) show that innovation capabilities benefit from both short and long-distance relations, the study of Aarstad, Kvitastein, and

Jakobsen (2016) suggests that only local interactions contribute to the innovativeness of small and medium-sized enterprises. Moreover, firms do not necessarily need long-distance relations to access the global source of novel ideas if they interact with other local firms functioning as a knowledge broker (Morrison 2008; Breschi and Lenzi 2015; Ozman, Balland, and Matta 2018). These lines of argument lead to the stylized fact that having only long-distance inter-firm relations is negatively related to the innovation capabilities of firms

Stylized fact 4: cognitively distant inter-firm relations are negatively related to firms' innovation capabilities.

Since the development of the proximity conceptual framework, it has been theoretically argued and empirically shown that the establishment and effectiveness of interactions between economic agents depend on the distance between firms along multiple dimensions[1] (Andre Torre and Rallet 2005; Boschma 2005; Torre 2008; Balland, Boschma, and Frenken 2015; Broekel 2015; Balland, Boschma, and Frenken 2020). The evolutionary economic geography approach suggests that the cognitive dimension of relations plays a critical role in firms' learning and innovation capabilities. The relevance of cognitive proximity seems to become even more relevant, considering innovation increasingly requires larger teams that consist of experts specialized in similar or related fields (van der Wouden 2020). In other words, although the colocation of firms facilitates inter-firm knowledge transfer, but not enough if firms are cognitively distant. For instance, a joint project between two companies that are active in building products and airline industries is unlikely to benefit the innovation capabilities of the two firms.

Many studies investigated the role of cognitive distance in inter-firm knowledge transfer. For instance, Juhász and Lengyel (2018) empirically prove that cognitive distance is negatively related to inter-firm knowledge relation persistence. Similarly, Broekel and Bednarz (2019) confirm a negative association between cognitive distance and the establishment of knowledge ties. Lazzeretti and Capone (2016) take a dynamic approach and provide evidence of the hampering effect of cognitive distance on the formation of knowledge transfer relations. Cantner and Meder (2008) show that technological dissimilarity negatively impacts collaborative innovation. Therefore, we expect that cognitively distant inter-firm relations are negatively related to firms' innovation capabilities.

[1] Boschma (2005) formulates five proximity dimensions (geographical, cognitive, organizational, social, institutional). We however focus on geographical and cognitive dimensions that have attracted most attraction in the literature, whereas other dimensions are less studied or used mainly as control variables in empirical studies.

Stylized fact 5: inter-firm relations that bridge small cognitive gaps are positively related to firms' innovation capabilities.

Although having relations with cognitively distant peers may have a negative impact on innovation capabilities, the proximity approach notes that too much cognitive overlap also hampers mutual learning (Nooteboom 1999; Boschma 2005). The notion of 'optimal' proximity builds on Nooteboom's (2000) argument that firms must interact with peers with an optimal cognitive distance from them because the exchanged information is useless if it is not new (i.e., a complete overlap of cognitive domains) or if it is so new that it cannot be absorbed and interpreted (i.e., completely separate cognitive domains). This argument aligns with the notion of 'proximity paradox', suggesting a large degree of proximities facilitates inter-firm tie formation but do not contribute to firms' innovative performance (Boschma and Frenken 2010). Empirical evidence for this has been presented by Wuyts et al. (2005), Ahuja and Katila (2001), Cloodt, Hagedoorn, and Kranenburg (2006), and Nooteboom et al. (2007), who discovered an inverted U-shaped relation between the cognitive distance of interacting firms and their innovation capabilities. In other words, firms benefit from links across slightly different cognitive domains. We thus conclude that inter-firm relations that bridge small cognitive gaps are positively related to firms' innovation capabilities.

Data and constructing the Digital Layer

In this section, we first present the dataset used in this study. We then outline how we used web scraping to transfer the base dataset into the Digital Layer - a network of hyperlinked firms with associated web texts. Lastly, we present two innovation datasets (the German Community Innovation Survey and a large-scale dataset of web-based innovation indicators) used in this study.

Firm base data

We use the Mannheim Enterprise Panel (MUP) of 2019 as our base dataset. The MUP is a firm panel database that covers the entire population of firms in Germany. It is updated on a semi-annual basis (Bersch et al. 2014). In addition to firm-level characteristics, such as firm size, age, and location, the MUP also includes the web addresses (URL) for 1,155,867 of the 2,497,412 firms in early 2019 (URL coverage of 46%). A prior analysis of this dataset (Kinne and Axenbeck 2020) showed that URL coverage differs systematically by sectors, regions, firm size, and age groups. Very small and young firms (smaller than five employees and younger than two years), especially from sectors such as agriculture, are not covered as comprehensively

as medium-sized and larger firms from manufacturing and ICT (information and communication technology) services. The MUP, nonetheless, constitutes an exhaustive dataset with a very high URL coverage in those firm groups that are most relevant for innovation development (Kinne and Axenbeck 2020; Rammer, Kinne, and Blind 2020). We removed firms without address information from our dataset and geocoded the remaining firms using street-level geocoding (without house numbers; see, e.g., Zandbergen 2008).

The geocoded firms were also used to calculate a firm-level location control variable by counting the number of other firms within one kilometer of each firm. The resulting local firm densities are used as a control for potential local spillovers. The search radius of one kilometer was selected according to Rammer, Kinne, and Blind (2020), who showed that spillovers from local knowledge sources decay within a few hundred meters.

Constructing the Digital Layer

For the web scraping of the firms' websites, we used ARGUS (Kinne 2018), an open-source web scraping tool based on Python's Scrapy scraping framework. ARGUS was used to scrape texts from the websites of all MUP firms as well as the hyperlink connections among the firms. After the web scraping, we excluded erroneous downloads and potentially misleading redirects from the data due to, for example, resold domains or mergers and acquisitions (see Kinne and Axenbeck 2020). After this step, 684,873 firms remained in the dataset.

We then created a network of firms where the edges are constructed from the extracted hyperlinks between firms (see Figure 1 for a schematic representation). Edges are given either weight 1.0 if the hyperlink connection between a pair of firms is unidirectional or weight 2.0 if the firms are mutually linked (i.e., both firms have a hyperlink connection to the other firm on their respective websites). As an example, in Figure 1, *firm 3* appears two times in the hyperlink vector of *firm 1* because the firms are mutually linked. As a result, the geographic distance between *firm 1* and *firm 3* is weighted by 2.0 when calculating the "mean distance" value for *firm 1*. This method is only one of several possible network operationalizations. Another possibility would have been to use only reciprocal (i.e., mutual) hyperlinks for the construction of edges, to construct a directed network, or to construct an undirected network entirely without considering reciprocal hyperlinks. We chose the approach described here because we think it to be a good compromise in which non-reciprocal links remain included in the dataset. Still, at the same time, the particular implication of reciprocal hyperlinks is considered by giving these relations a higher weight in the calculation of firm-level "mean distance."

After constructing the network, we excluded 150,246 (21.9%) firms without any hyperlink connections to other firms. Firms without links have considerably fewer employees (11.9 vs. 27.7) than those with hyperlinks and are younger (23.0 vs. 24.8 years). Both values are different at a highly significant level, according to a t-test. Both firms with and without hyperlinks were used to calculate a local firm density control variable (see below). Overall, there are 7,076,560 hyperlink connections in our dataset.

[Figure 1 about here]

Firm-level innovation data

We use two datasets with firm-level innovation indicators: The Mannheim Innovation Panel (MIP), a traditional questionnaire-based innovation survey of firms sampled from the MUP, and a web-based innovation indicator developed by Kinne and Lenz (2021).

The MIP survey is the German contribution to the Community Innovation Survey (CIS), conducted every two years in the European Union and has been used in an array of innovation studies (Gault 2013). The survey methodology and the definition of innovation follow the Oslo Manual (OECD 2018) and cover firms with five or more employees from manufacturing and business-oriented services. In the survey, firms are asked whether they introduced new or significantly improved products or services (hereafter, product innovations) during the three years before the study and whether they will introduce such products or services in the current year. In this study, we use the latter indicator from the MIP survey of 2018, which relates to the same year and is available for 2,463 firms.

Our second innovation dataset consists of predicted firm-level product innovator probabilities based on a deep learning model and website texts. For this web-based indicator, an artificial neural network (ANN) was trained on the website texts of firms surveyed in the MIP. After training on this dataset of labeled (product innovator/no product innovator) firm website texts, the ANN can be used to predict the product innovator probability of any out-of-sample firm with a website. Specifically, the authors use the ANN as a machine learning prediction model that receives as input the entire text of a single company website. The words used on the website, which describe the company itself, as well as its products, services, and employees, serve as input signals for the ANN, which processes them and makes a prediction about the probability of the company being a product innovator (i.e., a company that launched new products). During the training step, the ANN has learned the non-linear and multi-dimensional interaction of the individual input signals and their complex relationship to the

product innovator status of a company from the training data. Kinne and Lenz (2021) have shown that this approach can generate a reliable firm-level innovation indicator even in industrial sectors and size groups that are not covered in the training data (i.e., in the MIP survey). Among other things, the authors show that the novel web-based indicator highly correlates with traditional innovation indicators from patents and regional innovation indicators from official statistics. At the same time, the web-based indicator has several advantages, such as significantly greater coverage than survey data, which can only be applied to large company populations via extrapolations, but also patents, which are not relevant and widespread for all sectors. Other advantages are the timeliness of the web indicator and its low collection costs. The described web-based indicator is available for all 534,627 firms in our dataset.

Due to the sampling scheme of the MIP, the survey dataset includes larger and older firms on average, and certain sectors are over-represented (for more information, see Rammer et al. 2019). Even though the web dataset is closer to the overall German firm population, the results of Kinne and Axenbeck (2020) show that it is not unbiased. More extensive and older firms from certain sectors are more likely to have a website and thus are over-represented in the web dataset. On average, firms in the survey dataset are located in more densely populated areas. All these differences are statistically significant according to a t-test.

On the other hand, the number of hyperlinks per firm is not significantly different, but the distribution is highly skewed, especially for the *web dataset*. The maximum *link count* in the web dataset is about 169,000 and corresponds to the German branch of a well-known Silicon Valley-based tech company.

The mean product innovator probability (hereafter, *InnoProb*) in the web dataset is 25% (see Table 1). Casted to a binary variable using a classification threshold of 0.4 (see Kinne and Lenz 2021) results in only 16% predicted product innovators compared to 25% in the survey dataset. Given that the latter dataset intentionally over-samples innovative firm types due to the sampling procedures outlined in OECD (2018) while the web dataset is closer to the overall firm population, these values are credible (see also Kinne and Axenbeck 2020 for details).

Variables

In this section, we outline how we operationalize the network position of each firm, mean partner innovation, geographical and cognitive distances to firm's link partners, and the type of each hyperlink. We calculate the mean for all these measures as outlined in Figure 1. We also calculated standard deviations to capture the heterogeneity of each firm's network. Still, we found that a simple hyperlink count per firm sufficiently predicts network heterogeneity.

Link count and mean partner innovation

Link count *(LinkCount)* is a count of all the hyperlinks a firm maintains to other firms. In Figure 1, *firm 1* has a link count of 3, and *firm 3* has a link count of 2, for example. As such, the link count variable is analogous to the degree centrality measure in social network analysis. Alternatively, we counted the number of firms' hyperlinks to innovative firms (*InnoProb* greater than the 75th percentile) to distinguish between high- and low-quality hyperlinks regarding knowledge exchange and learning *(InnoLinkCount)*. The result suggests that *InnoLink* and *InnoLinkCount* strongly correlate (the Pearson correlation coefficient: 0.96). Therefore, we refrain from including *InnoLinkCount* in our analysis.

The mean partner innovation *(InnoPartner)* reflects the innovativeness of the hyperlinked partners that a firm has in the Digital Layer. It is calculated by taking the mean of the firm-level web-based innovation indicator (see the Data section) of the hyperlinked partners of a firm.

Geographic distance

We measure geographic distance (*GeoDist*) by calculating the Euclidean distance between firms that are hyperlinked. For each firm, we calculated the mean Euclidean distance to its partners.

Cognitive distance

The cognitive distance (*CogDist*) between hyperlinked firms is operationalized by calculating the cosine similarity between their website texts. We know that firms use their websites to present themselves, their products, and services. This information is usually codified as text and can be extracted and analyzed to assess firms' products, and services (Gök, Waterworth, and Shapira 2015). In its entirety, website texts describe a firm's knowledge base, and we use them to calculate the cognitive distance between the firm and its hyperlinked partners.

We represent the firms' website texts in a high-dimensional vector space by transferring them using a term frequency-inverse document frequency (tf-idf) scheme (Manning et al. 2009). The tf-idf algorithm assigns each document to a fixed-size sparse vector of size $\boldsymbol{V}$, where $\boldsymbol{V}$ is the size of a dictionary composed of all words found in the overall text corpus. We restricted our dictionary to words with a minimum document frequency of 1.5% and a maximum document frequency of 65% (*popularity-based filtering*). Each entry in the tf-idf vector of a document corresponds to one word in the dictionary, representing the relative

importance of this word in the document. A 0 value represents words that do not appear in a given document.

Specifically, in the first step (the tf step), the number of appearances per word in a single document is counted. In the second step, the inverse document frequency (idf) is used as a weighting scheme to adjust the tf counts. Conceptually, the idf weights determine how much information a specific word provides by means of how frequently a word appears in the overall document collection. The intuition is that very frequent words that appear in many documents should be given less weight than less frequent words, as infrequent words are more valuable as a distinguishing feature.

We then use the tf-idf vector of a firm to calculate its similarity to the website texts of other firms, which have a hyperlink to the firm under consideration. We quantify the similarity between the two website texts by computing the cosine similarity of their vector representations (Manning et al. 2009), an approach widely adopted in natural language processing studies (Mikolov et al. 2013; Rahimi, Mottahedi, and Liu 2018; Gentzkow, Kelly, and Taddy 2019). For the sake of consistency, by multiplying the similarity values by minus one, we transform the calculated cosine similarities to cosine distances, which range from -1 (identical texts) to 0 (maximal dissimilar texts). Again, we then calculate the mean of the cognitive distances between a firm and its hyperlinked partners.

It is important to note that we use the z-score of the four variables described above to ease the interpretation of the regression results. A z-score corresponds to $(x- \bar{x})/sd(x)$, where $\bar{x}$ and $sd(x)$ are the mean and standard deviation of x, respectively.

Hyperlink type

We operationalize hyperlink type as a binary variable by classifying the nature of each relation between hyperlinked firms as one of the following two classes. First, non-business relations are between firms that are not directly related to making business with each other and are non-monetary. Such relations primarily include membership in (industrial) associations or chambers of commerce and references to regulatory or legal bodies (e.g., commercial courts and commercial registries). Hyperlinks to purely informative web content are also part of this class. Such references may include, for example, hyperlinks from a pharmacy to an external website that informs about healthy diets or a hyperlink from a firm to the website of a local news outlet that reports about the firm's latest achievements. Second, business relation includes all hyperlinks between firms that do or did business together. Frequently, firms include hyperlinks to other companies' websites to present them as testimonials or because they have an ongoing business relationship (e.g., web hosting, web design, web mail providers,

certification services). Suppose a firm hyperlinks to its own social media profiles, the firm that operates the social media platform is a business partner of that firm (because they provide the platform and make money from it). Hyperlinks between entities of the same corporate group or between personal websites of employees and their employer (e.g., professor to university) are also part of this class.

The business relation is closer than the non-business relation as the ties represented by it are usually more formal and reoccurring. In that sense, we quantify the nature of each hyperlink connection between two firms as either value 0.0 (weak non-business relation) or 1.0 (strong business relation) that can be predicted in a binary machine learning classification task. We again use the firms' website texts for this classification and relate them in the tf-idf vector space (see cognitive distance section above).

First, we created a training dataset for that classification task by sampling 5,000 random pairs of hyperlinked firms from our dataset. Subsequently, we labeled each hyperlink as representing either a business or non-business relation. We were able to label 3,632 hyperlink connections unambiguously. Figure 2 shows that more than two-thirds of the hyperlinks were labelled as business relations, with only a few being hyperlinks between firms of the same corporate group. *Non-business relations*, on the other hand, are of information only and legal/regulatory nature to about equal shares.

[Figure 2 about here]

We then created numerical vectors for each hyperlinked firm pair by concatenating their respective tf-idf vectors. The resulting vectors have two times the dimension of our initial dictionary and effectively encode the texts of both firms. We tested several binary classifiers with these vectors and their corresponding labels from the training data and decided on a primary logistic regression classifier with balance class weights. For our classification task, the performance of the logistic regression classifier was overall superior in terms of accuracy and more balanced compared to more sophisticated binary classifiers we tested (e.g., artificial neural networks and random forest). We trained the logistic regression classifier on two-thirds of the labeled dataset and used one-third (952 firms) as a test set to evaluate the model's performance. Table 2 reports precision, recall, f1-score, and accuracy of the trained model in the test set. The overall accuracy of 0.92 and an f1-score of 0.92 indicate outstanding performance.

We used the trained model to predict the type of each of the 7,076,560 hyperlink connections in our dataset. The predictions range from 0.0 (high probability of business relation; small *NonBusinessRelation*) to 1.0 (high probability of non-business relation; large *NonBusinessRelation*).

[Table 1 about here]

[Table 2 about here]

Estimation strategy

Using linear regression models when the dependent variable is bounded between zero and one strongly violates the critical assumptions of linear modeling. It brings about untrustworthy p-values (also known as the efficiency issue of regression coefficients). Statisticians developed beta regression models to remedy this situation, in which dependent variables are rates, proportions, or concentration indices (Ferrari and Cribari-Neto 2004). The essential advantage of the beta regression model is that it can assume, among others, left- or right-skewed density shapes based on a combination of parameter values (Cribari-Neto and Zeileis 2010).

Figure 3 compares normal and beta distributions with the distribution of the dependent variable (*InnoProb*). To ensure the robustness of our estimated coefficients, we opted for employing a set of beta regressions to investigate the association between independent variables and firms' innovation capabilities. Cribari-Neto and Zeileis (2010) formally express the beta density as:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{(p-1)}(1-y)^{(q-1)} \quad , \quad 0 < y < 1, \tag{1}$$

where $p, q > 0$ and $\Gamma(.)$ is the gamma function. For a regression model, Ferrari and Cribari-Neto (2004) suggest an alternative parameterization by setting $\mu = p/(p+q)$ and $\Phi = p+q$:

$$f(y;\mu,\Phi) = \frac{\Gamma(\Phi)}{\Gamma(\mu\Phi)\Gamma((1-\mu)\Phi)} y^{\mu\Phi-1}(1-y)^{(1-\mu)\Phi-1} \quad , \quad 0 < y < 1, \tag{2}$$

with $0 < \mu < 1$ and $\Phi > 0$. We conduct beta regressions using the betareg R-package developed by Grün, Kosmidis, and Zeileis (2012). Table 3 provides the pairwise Pearson correlation coefficients of variables.

[Table 3 about here]

Results and discussion

We created the Digital Layer of Germany according to the procedure described in the previous section. The top panel of Figure 4 shows the distribution of product innovator firms in Germany (left) and Berlin (right), where each cell's coloring gives the mean innovation probability for the companies in the respective cell. The middle panel shows the distribution of hyperlink connections in Germany (left) and Berlin (right). The lower panel shows the ego network of an exemplary firm (the Centre for European Economic Research) both for overall Germany (left) and for the Rhine-Neckar region (right) where the firm is located. The networks shown in Figure 4 were created using a graph bundling method based on kernel density estimation (Hurter, Ersoy, and Telea 2012). Unsurprisingly, the density of hyperlink connections between any two areas seems highly dependent on population[2].

[Figure 4 about here]

Figure 5 illustrates the distribution of *Innoprob* stratified by sector. We observe a similar distribution pattern of the dependent variable across industries, with a peak reached before the *Innoprob* value of 0.25. A more careful investigation of these distributions by a set of Kolmogorov-Smirnov tests reveals that only a few sectors (e.g., wholesale and oil sectors) have statistically similar *Innoprob* distributions.

[Figure 5 about here]

Figure 6 shows kernel density estimations of the four variables of interest. The normalized mean geographic distance distribution has a mean and a median of 0.28 (235 km). It follows a normal distribution with an over-proportional accumulation of observations at a mean distance of 0.0 (i.e., companies that maintain hyperlinks to other companies located in the same street). Considering the mean cognitive distance, a value of of 0.0 corresponds to firms that share identical texts with their hyperlink partners. The mean link count is 13.01, and the median is 4, while the maximum link count in our dataset is 168,961 (the German branch of a major tech company from Silicon Valley). Mean partner innovation is again somewhat normally distributed with a mean of 0.36 and a median of 0.34. The local firm density variable (Density) distribution is highly skewed. On average, firms in the dataset have 176.8 other firms within one kilometer of their geographic location. The median is 53, and the maximum is 3,930

[2] Figure 4 is not intended to be of high analytical value but rather to give an overview of the dataset and its granularity.

(downtown Hamburg). We do not observe any difference in the distribution of four variables among firms with different degrees of innovation capability.

[Figure 6 about here]

Figure 7 shows scatterplots and fitted regression lines of second order between innovation and several variables. We also tested regressions of the third order, which yielded only slightly different results. The number of firm partners (*LinkCount*) and the mean innovation probability of these partners (*InnoPartner*) show a strong positive and linear relation to the firm's innovation probability. The relation between a firm's innovation probability and the mean cognitive distance to its hyperlink partners is negative but less distinct.

[Figure 7 about here]

Discussion of regression results

In the remainder of this section, we discuss the results of the beta regression models and robustness checks. All estimated models include control variables and sector fixed effects. Following the argument of Hünermund and Louw (2022) that estimated effect sizes of control variables (i.e., *Size*, *Age*, *Density*, and *NonBusinessRelation*) might represent a mix of multiple causal mechanisms, we refrain from reporting and interpreting the coefficients of control variables[3]. Instead, we focus on reporting and discussing the coefficients of the main variables of interest. In addition, we have used the heteroskedasticity-consistent estimation of standard errors due to the heteroskedasticity inherent in the beta models (Cribari-Neto and Zeileis 2010). As discussed before, the four variables of interest are included in the models as z-scores (i.e., having the same scale), whereby we can more easily interpret and compare the effect sizes.

First, we conducted beta regressions and added variables of interest stepwise (Table 4). The values of the Akaike information criterion (AIC) suggest that the full model provides the best goodness of fit. Since the sign, the degree of significance, and the effect size of variables do not substantially change, and we discuss the results of the full mode (Model 5). Our results suggest that the number of hyperlinks is positively associated with firms' innovation capabilities. More specifically, increasing *LinkCount* by one standard deviation increases the odds of the innovation capability of firms by 15 percent. Similarly, Uzzi (1996) and Giuliani

[3] The coefficients of control variables and corresponding standard errors are available upon requests from authors.

and Bell (2005) suggest a positive relationship between the degree centrality of inter-firm network and their innovative performance.

The quality of hyperlinks captured by the average innovation capabilities of hyperlinked firms (*InnoPartner*) is positively related to the dependent variable. More interestingly, this variable has a greater effect size compared to the one of *LinkCount.* That is a 23 percent increase in odds of firms' innovation capability by one unit increase in *InnoPartner*. These findings align with the ones in the literature that emphasize the relevance of inter-firm relations as knowledge transfer channels (Kobarg, Stumpf-Wollersheim, and Welpe 2019).

Contrary to our expectations, the reported results suggest that geographic distance negatively correlates with the dependent variable, and the effect size is about one order of magnitude smaller than the first two variables. This result comes as a surprise because this is contradictory to recent empirical evidence suggesting geographic distance still hampers innovations (Graevenitz, Graham, and Myers 2022). It is important to note that this finding needs to be interpreted in relation to the nature of hyperlink data and the relatively low cost of creating a hyperlink relation compared to formal collaborative ties (e.g., joint patenting).

In line with the theoretical arguments, cognitive distance (*CogDist*) between linked firms negatively correlates with firms' innovation capabilities. It is plausible to argue that firms innovate in areas close to their knowledge base (Nelson and Winter 1982), and cognitively distant firms encounter problems interpreting exchanged knowledge beyond their absorptive capacity (Cohen and Levinthal 1990).

By including a quadratic term of *CogDist* (i.e., *CogDistSquared*) to investigate a potential inverted U-shape relationship between the cognitive distance of linked firms and their innovation capabilities. Figure 8 shows the relation between *CogDist* and its quadradic term, suggesting that smaller values of *CogDist* have considerably greater weight in *CogDistSquared*. Interestingly, a change in the sign of the quadratic term suggests that lower values of cognitive distance among hyperlinked firms positively related to their innovation capabilities. This finding resonates with the 'optimal' cognitive distance argument that two firms benefit from interaction if their technological and cognitive backgrounds do not fully overlap. However, at the same time, they are cognitively close enough to be capable of absorbing and exploiting each other's knowledge (Cohen and Levinthal 1990; Nooteboom 1999; Balland, Boschma, and Frenken 2022). It is important to note that our cognitive proximity measure must be understood as a one-dimensional mapping of a high-dimensional relationship. There may be companies with entirely different backgrounds (e.g., a software and a mechanical engineering company) that both participate in the same market (e.g., internet-of-things) and consequently share a

similar knowledge base according to our text-based measure for the cognitive distance variable. Our results could, therefore, also indicate that cognitively close hyperlinked firms share similar target markets rather than similar technologies.

[Table 4 about here]

[Figure 8 about here]

It is plausible to expect that the number and quality of firms' hyperlinks variables have positive joint effects on firms' innovation capabilities. While there is no statistically significant difference between the reported coefficients of variables of interest across models with and without interaction terms, the interaction term in Model 2 (Table 5) is positive and statistically significant. Since including an interaction term based on two continuous variables may lead to a biased estimation of interaction effects (Juhász, Tóth, and Lengyel 2020), Models 3 and 4 are based on a dichotomized version of *LinkCount* and *InnoPartner*. More precisely, *LinkCount (dummy)* and *InnoPartner (dummy)* take the value of one if their original values are greater than the 75th percentile of *LinkCount* and *InnoPartner*, respectively, and they take the value of zero otherwise. The result does not substantially change after this specification.

[Table 5 about here]

The descriptive statistics suggests no significant difference between the four variables of interest among firms with low, average, and high degrees of innovation capabilities. Table 6 shows the results of beta regression models on the full and a sample of innovative firms. In the smaller sample, we included firms with *InnoProb* greater than the 75th percentile corresponding to a threshold of 0.3, which is close to what Kinne and Lenz (2021) also suggest in their study as an innovation classification threshold. The findings indicate no statistical difference between the reported coefficients of the four variables of interest and the interaction terms between the full model and the one of a smaller sample. However, the only difference with the full model is that the sign of the coefficient of *CogDistSquared* remains negative.

Given that the relationship between *CogDist* and *CogDistSquared* in the smaller sample is similar to the one in the full sample (Figure 8), this finding implies a negative association holds for any degree of cognitive distance among hyperlinked firms with a higher degree of innovation capabilities. One reason for this may be that firms with a higher degree of innovation

capabilities are also very specialized and interact with firms with the same knowledge bases. Since we do not have a measure of specialization for firms in this study, we cannot disentangle the effects of these two factors and leave it to further empirical investigations.

[Table 6 about here]

Robustness checks[4]

We conducted several robustness checks to test the reliability of our main findings under alternative specifications. First, we ran similar beta regression models for smaller samples of firms in Berlin (N: 20,290) and Frankfurt am Main (N: 4,316) to ensure that the high degree of significance is not driven by a large number of observations (N: 509,165). The results suggest that the sign, significance, and effect sizes do not substantially change and align with the results of the original model.

Second, we tested the correlation between independent variables and an alternative dependent variable. That is, a dummy variable that takes the value of one if the company filed for at least one patent in the past ten years before the time of web scraping, and it takes the value of zero otherwise. We created a dummy variable because the distribution of the number of filed patents among firms is highly skewed (ranging between 0 and 1904), and about 2% of firms have at least one patent. We ran two logit models using the alternative dependent variable with and without an interaction term. The results suggest a positive association between the number and quality of hyperlinks and filling for a patent with relatively large effect sizes. Similar to our main finding, this robustness check suggests a weaker negative association between the cognitive distance of linked firms and their patenting activities. More interestingly, in line with the theoretical arguments developed based on the results of empirical studies primarily using formal collaboration data (Abbasiharofteh and Broekel 2020; Graevenitz, Graham, and Myers 2022), we found that geographic distance between hyperlinked firms negatively correlates with patenting. This result may indicate the potential difference between hyperlink data and formal collaborative ties. We will underline this difference in the next section and motivate further research on this issue.

Third, we also checked the association between independent variables and firms' innovation capabilities using the German Community Innovation Survey. As described in the data section, this dataset includes a much smaller number of firms that responded to the innovation survey and declared whether they introduced new or significantly improved products or services. We

[4] All robustness checks are available upon request.

utilized this information and created a dummy variable that takes the value of one if a given firm introduced an innovation. It takes the value of zero otherwise. Using two logit models with and without an interaction term, this robustness check suggests the positive correlation of the number and average innovation capabilities of linked firms with the dependent variable. In contrast, the coefficients of *GeoDist* and *CogDist* are not statistically significant (see Appendix D).

All in all, robustness checks strongly support a positive association between the number of hyperlinks, innovative partners, and firms' innovation capabilities. We found mixed results concerning the correlation between the geographic and cognitive distance of hyperlinked firms and innovation capabilities. It is important to note that the datasets used for robustness checks have limitations. Only 2% of firms can be classified as innovative by the patent data, and this data does not perfectly reflect the innovation capability of a broad population of firms. Similarly, the German Community Innovation Survey data includes a much smaller share of German firms.

Conclusion

In this study, we have introduced the Digital Layer, a novel, web-based approach to exploring innovation systems. The Digital Layer contains the geographic locations of German companies with a website and the hyperlink connections between them. In addition, each company in the Digital Layer is described by the textual content of its website, which serves as the basis to assess the firm's innovation capability and the distance to its hyperlink partners. In addition to geographic distance, we have operationalized text-based measures for cognitive distance. Next, we have showcased the use of this alternative data in the context of economic geography and innovation studies. Our empirical results suggest that firms' innovation capabilities are indeed positively associated with the quantity and quality of their hyperlinks and, to a lesser extent, with hyperlink relations to geographically distant and cognitively close firms. Thus, this study shows that a theoretically informed analysis of firms' hyperlink portfolios can reveal firms' innovation capabilities. Our work contributes to developing a new methodological tool set for research in multiple fields ranging from economic geography to regional and innovation studies and management and economics. Therefore, we encourage researchers to take this study as a point of departure for future research that was previously constrained by the lack of micro-data and analytical tools.

We acknowledge several limitations of our work that open up new opportunities for future research. First and foremost, we have observed and reported correlations and cannot infer any strict causality. For instance, we do not provide statistical proof on whether companies are more

innovative because they have more hyperlinks, or innovative firms tend to connect to more firms on the web. That is the potential reverse causality between hyperlink portfolios and innovation capabilities. Access to a Digital Layer panel dataset can pave the way for a causal analysis of firms' hyperlink portfolios and innovation capabilities. Soon, such a dataset would be comparatively easy to generate by applying a consistent web scraping strategy at different points in time to an up-to-date sample of companies. One advantage of our presented approach is that in a future dynamic analysis, we can observe whether certain hyperlinks persist or disappear.

Second, we have approximated the effects of geographical and cognitive distances among linked firms in the Digital Layer but have not accounted for institutional distance. One should account for the institutional distance when going a step further and expanding our proposed analysis approach to an international scope. Given that we only analyze the network of firms located in Germany and additionally control for sectors, we assume that the macro-level institutional setting is sufficiently uniform and does not affect our analysis too much. We should note, though, that there is, in fact, evidence of relevant city-level effects of socio-cultural settings on firms' relationships (Abbasiharofteh and Broekel 2021). Similarly, the social closeness between firms is mainly established by personal ties (like friendship or kinship between employees) and is assumed to increase trust and more effective communication. We can barely gain insights into employee relations depending on firm websites as our primary data source. Therefore, data from job-related social networks (e.g., LinkedIn) promises immense potential for future studies, especially if such data can be integrated into the Digital Layer of company websites.

Third, it is not too far-fetched (and backed by our manual classification of hyperlink relations) to assume that a hyperlink between two firms is associated with a kind of knowledge exchange between these two. We have not, however, distinguished between the hyperlinks based on their type and intensity. Recent advances in Natural Language processing (NLP) methods have enabled researchers to train algorithms based on hyperlinks' ambient texts (texts surrounding hyperlinks) to classify hyperlinks (Vaswani et al. 2017). For instance, researchers can build on these techniques to classify inter-firm hyperlinks into the supply chain, a joint venture (e.g., joint research), and outsourcing (e.g., training, advice-seeking, and marketing) (Tsamenyi et al. 2010). This aspect is of critical importance because firms can create an inter-firm hyperlink at a relatively low cost compared to getting involved in a joint research project with other firms. Thus, future research on the techniques mentioned above should investigate the extent to which hyperlinks represent mutual learning and knowledge exchange.

Finally, we have only focused on analyzing the nodal and dyadic attributes of hyperlink portfolios (i.e., link count and geographic and cognitive distances on hyperlinked firms). However, we did not consider other structural network measures at the triadic- (e.g., triadic closure) and Meso levels (e.g., community membership). Several studies have shown the relevance of these measures in the innovation capabilities of individuals and firms (Lobo and Strumsky 2008; Strumsky and Lobo 2015; Abbasiharofteh 2020; Abbasiharofteh, Kogler, and Lengyel 2020). Similarly, due to data limitations, we did not include a control variable for firms' R&D expenditures, which could lead to an omitted variable bias. Going beyond the nodal and dyadic levels and adding more control variables in analyzing a hyperlink network are promising avenues for future work.

The Digital Layer approach promises the excellent potential for evidence-based assessment of sectoral and place-based innovation policies. As we have shown, the Digital Layer can be created for any regional unit in a sector-independent and cost-effective manner to provide up-to-date insight into the interconnectedness of the firm population represented on the Internet. Combined with modern Natural Language Processing (NLP) methods, company relationships can thus not only be surveyed quantitatively and evaluated in terms of quality and scope. For instance, one of the main aims of innovation mission-oriented policy is to bring stakeholders from different fields to trigger innovative ideas for tackling grant societal challenges (Mazzucato 2018; Wanzenböck et al. 2019; Janssen and Abbasiharofteh 2022). The Digital Layer approach provides the possibility to assess the impact of mission-oriented policies by analyzing the cognitive distance of hyperlinked firms before and after implementing such policies. Our suggested method also contributes to recent transition policy efforts to create directionalities for a joint green and digital transition ('twin transition') of European economies (Muench et al. 2022). The implication of our approach, iteratively coupled with NLP methods to identify firms' green and digital goods and services based on the web data, offers an unprecedented ability to identify and analyze how firms diversify into new green and digital capabilities and inter-firm relations. This twin transition observatory provides much-needed inputs to investigate firms and place-based diversification trajectories and to assess the impact of transition policies.

References

Aarstad, J., Kvitastein, O. A., and Jakobsen, S.-E. 2016. Local buzz, global pipelines, or simply too much buzz? A critical study. Geoforum 75:129–33.

Abbasiharofteh, M. 2020. Endogenous effects and cluster transition: A conceptual framework for cluster policy. European Planning Studies 30:1–24.

Abbasiharofteh, M., and Broekel, T. 2020. Still in the shadow of the wall? The case of the Berlin biotechnology cluster. Environment and Planning A: Economy and Space 46 (3). doi:10.1177/0308518X20933904.

Abbasiharofteh, M., Kogler, D. F., and Lengyel, B. 2020. Atypical Combination of Technologies in Regional Co-inventor Networks. Papers in Evolutionary Economic Geography (PEEG) 20.55. http://econ.geo.uu.nl/peeg/peeg2055.pdf.

Anderson, N., Potočnik, K., and Zhou, J. 2014. Innovation and Creativity in Organizations. Journal of Management 40 (5): 1297–1333. doi:10.1177/0149206314527128.

Andersson, M., and Larsson, J. P. 2022. Mysteries of the trade? Skill-specific local agglomeration economies. Regional Studies 56 (9): 1538–53. doi:10.1080/00343404.2021.1954611.

Aral, S. 2016. The Future of Weak Ties. American Journal of Sociology 121 (6): 1931–39. doi:10.1086/686293.

Archibugi, D., and Planta, M. 1996. Measuring technological change through patents and innovation surveys. Technovation 16 (9): 451–519. doi:10.1016/0166-4972(96)00031-4.

Axenbeck, J., and Breithaupt, P. 2021. Innovation indicators based on firm websites-Which website characteristics predict firm-level innovation activity? PloS one 16 (4): e0249583. doi:10.1371/journal.pone.0249583.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., and Wong, A. 2018. Social Connectedness: Measurement, Determinants, and Effects. Journal of Economic Perspectives 32 (3): 259–80. doi:10.1257/jep.32.3.259.

Balland, P.-A., Boschma, R., and Frenken, K. 2022. Proximity, innovation and networks: A concise review and some next steps. In Handbook of Proximity Relations, ed. A. Torre and D. Gallaud, 70–80: Edward Elgar Publishing.

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., and Hidalgo, C. A. 2020. Complex economic activities concentrate in large cities. Nature human behaviour. doi:10.1038/s41562-019-0803-3.

Barabási, A.-L., and Albert, R. 1999. Emergence of Scaling in Random Networks. Science 286 (5439): 509–12. doi:10.1126/science.286.5439.509.

Basole, R. C., Russell, M. G., Huhtamäki, J., Rubens, N., Still, K., and Park, H. 2015. Understanding Business Ecosystem Dynamics. ACM Transactions on Management Information Systems 6 (2): 1–32. doi:10.1145/2724730.

Bathelt, H., Malmberg, A., and Maskell, P. 2004. Clusters and knowledge: Local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography 28 (1): 31–56. doi:10.1191/0309132504ph469oa.

Bell, G. G. 2005. Clusters, networks, and firm innovativeness. Strategic Management Journal 26 (3): 287–95. doi:10.1002/smj.448.

Berg, S.-H. 2018. Local Buzz, Global Pipelines and Hallyu: The Case of the Film and TV Industry in South Korea. Journal of Entrepreneurship and Innovation in Emerging Economies 4 (1): 33–52. doi:10.1177/2393957517749072.

Bersch, J., Gottschalk, S., Müller, B., and Niefert, M. 2014. The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany. Mannheim: ZEW Discussion Paper 14–104.

Blomqvist, K., and Levy, J. 2006. Collaboration capability a focal concept in knowledge creation and collaborative innovation in networks. International Journal of Management Concepts and Philosophy 2 (1): 31. doi:10.1504/IJMCP.2006.009645.

Bloom, N., Jones, C. I., van Reenen, J., and Webb, M. 2020. Are Ideas Getting Harder to Find? American Economic Review 110 (4): 1104–44. doi:10.1257/aer.20180338.

Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. 2009. Network Analysis in the Social Sciences. Science 323 (5916): 892–95. doi:10.1126/science.1165821.

Boschma, R. 2005. Proximity and Innovation: A Critical Assessment. Regional Studies 39 (1): 61–74. doi:10.1080/0034340052000320887.

Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., and Liberati, C. 2022. Unconventional data for policy. In Conference on Information Technology for Social Good, 338–44. New York, NY, USA: ACM.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30 (1-7): 107–17. doi:10.1016/S0169-7552(98)00110-X.

Broekel, T. 2019. Using structural diversity to measure the complexity of technologies. PloS one 14 (5): e0216856. doi:10.1371/journal.pone.0216856.

Broekel, T., and Bednarz, M. 2019. Disentangling link formation and dissolution in spatial networks: An Application of a Two-Mode STERGM to a Project-Based R&D Network in the German Biotechnology Industry. Networks and Spatial Economics 46:677–704.

Brusoni, S., Prencipe, A., and Pavitt, K. 2001. Knowledge Specialization, Organizational Coupling, and the Boundaries of the Firm: Why Do Firms Know More than They Make? Administrative Science Quarterly 46 (4): 597–621. doi:10.2307/3094825.

Burt, R. S. 2004. Structural Holes and Good Ideas. American Journal of Sociology 110 (2): 349–99. doi:10.1086/421787.

Crespo, J., Suire, R., and Vicente, J. 2016. Network structural properties for cluster long-run dynamics: Evidence from collaborative R&D networks in the European mobile phone industry. Industrial and Corporate Change 25 (2): 261–82. doi:10.1093/icc/dtv032.

Cribari-Neto, F., and Zeileis, A. 2010. Beta Regression in R. Journal of Statistical Software 34 (2). doi:10.18637/jss.v034.i02.

Crossan, M. M., and Apaydin, M. 2010. A Multi-Dimensional Framework of Organizational Innovation: A Systematic Review of the Literature. Journal of Management Studies 47 (6): 1154–91. doi:10.1111/j.1467-6486.2009.00880.x.

Daas, P. J.H., and van der Doef, S. 2020. Detecting innovative companies via their website. Statistical Journal of the IAOS 36 (4): 1239–51. doi:10.3233/SJI-200627.

Ferrari, S., and Cribari-Neto, F. 2004. Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics 31 (7): 799–815. doi:10.1080/0266476042000214501.

Fitjar, R. D., and Rodríguez-Pose, A. 2013. Firm collaboration and modes of innovation in Norway. Research Policy 42 (1): 128–38. doi:10.1016/j.respol.2012.05.009.

Friedman, T. L. 2005. The world is flat: A brief history of the twenty-first century. 1. ed. New York, NY: Farrar Straus and Giroux. http://www.loc.gov/catdir/enhancements/fy0617/2004028685-b.html.

Fritsch, M., Titze, M., and Piontek, M. 2020. Identifying cooperation for innovation—a comparison of data sources. Industry & Innovation 27 (6): 630–59. doi:10.1080/13662716.2019.1650253.

Gault, F., ed. 2013. Handbook of Innovation Indicators and Measurement: Edward Elgar Publishing.

Giuliani, E., and Bell, M. 2005. The micro-determinants of meso-level learning and innovation: Evidence from a Chilean wine cluster. Research Policy 34 (1): 47–68. doi:10.1016/j.respol.2004.10.008.

Glückler, J. 2013. Knowledge, Networks and Space: Connectivity and the Problem of Non-Interactive Learning. Regional Studies 47 (6): 880–94. doi:10.1080/00343404.2013.779659.

Gök, A., Waterworth, A., and Shapira, P. 2015. Use of web mining in studying innovation. Scientometrics 102:653–71.

Gulati, R., and Gargiulo, M. 1999. Where Do Interorganizational Networks Come From? American Journal of Sociology 104 (5): 1439–93. doi:10.1086/210179.

Haus-Reve, S., Fitjar, R. D., and Rodríguez-Pose, A. 2019. Does combining different types of collaboration always benefit firms? Collaboration, complementarity and product innovation in Norway. Research Policy 48 (6): 1476–86. doi:10.1016/j.respol.2019.02.008.

Jacobs, J. 1970. The economy of cities. Vintage books 584. New York: Vintage Books.

Jaffe, A. M. T. H. R. 1993. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. Quarterly Journal of Economics 108 (3): 577–98.

Janssen, M. J., and Abbasiharofteh, M. 2022. Boundary spanning R&D collaboration: Key enabling technologies and missions as alleviators of proximity effects? Technological Forecasting and Social Change 180 (7): 121689. doi:10.1016/j.techfore.2022.121689.

Jensen, M. B., Johnson, B., Lorenz, E., and Lundvall, B. Å. 2007. Forms of knowledge and modes of innovation. Research Policy 36 (5): 680–93. doi:10.1016/j.respol.2007.01.006.

Juhász, S., and Lengyel, B. 2018. Creation and persistence of ties in cluster knowledge networks. Journal of Economic Geography 121: 1203–26. 10.1093/jeg/lbx039.

Juhász, S., Tóth, G., and Lengyel, B. 2020. Brokering the core and the periphery: Creative success and collaboration networks in the film industry. PloS one 15 (2): e0229436. doi:10.1371/journal.pone.0229436.

Kabirigi, M., Abbasiharofteh, M., Sun, Z., and Hermans, F. 2022. The importance of proximity dimensions in agricultural knowledge and innovation systems: The case of banana disease management in Rwanda. Agricultural Systems 202 (3): 103465. doi:10.1016/j.agsy.2022.103465.

Kinne, J., and Axenbeck, J. 2020. Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. Scientometrics 28 (4): 443. doi:10.1007/s11192-020-03726-9.

Kinne, J., and Lenz, D. 2021. Predicting innovative firms using web mining and deep learning. PloS one 16 (4): e0249071. doi:10.1371/journal.pone.0249071.

Kogut, B., and Zander, U. 1992. Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology. Organization Science 3 (3): 383–97. doi:10.1287/orsc.3.3.383.

Lazega, E., Mounier, L., Snijders, T., and Tubaro, P. 2012. Norms, status and the dynamics of advice networks: A case study. Social Networks 34 (3): 323–32. doi:10.1016/j.socnet.2009.12.001.

Lorenzen, M. 2018. The Geography of the Creative Industries: Theoretical Stocktaking and Empirical Illustration. In Clark, Feldman, Gertler, and Wójcik, The new Oxford handbook of economic geography, 305–23.

Manning, C. D., Raghavan, P, and Schutze, H. 2009. An Introduction to Information Retrieval. Cambridge (Online edition): Cambridge University Press.

Marmaros, D., and Bruce, S. 2006. How Do Friendships Form? The Quarterly Journal of Economics 121 (1): 79–119.

Marshall, A. 1890. Principles of Economics. London: MacMillan.

Mazzola, E., Perrone, G., and Handfield, R. 2018. Change Is Good, But Not Too Much: Dynamic Positioning in the Interfirm Network and New Product Development. Journal of Product Innovation Management 35 (6): 960–82. doi:10.1111/jpim.12438.

Mazzucato, M. 2018. Mission-oriented innovation policies: Challenges and opportunities. Industrial and Corporate Change 27 (5): 803–15. doi:10.1093/icc/dty034.

Micek, G. 2018. Attempt at Summarising Past Studies on Geographic Proximity. kraków: Wydawnictwo Naukowe Uniwersytetu Pedagogicznego w Krakowie.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Moody, J. 2011. Network Dynamics. In The Oxford Handbook of Analytical Sociology, ed. P. Bearman, P. Hedström, P. Dodds, and D. J. Watts, 447–74: Oxford University Press.

Morrison, A. 2008. Gatekeepers of Knowledge within Industrial Districts: Who They Are, How They Interact. Regional Studies 42 (6): 817–35. doi:10.1080/00343400701654178.

Muench, S., Stoermer, E., Jensen, K., Asikainen, T., Salvi, M., and Scapolo, F. 2022. Towards a green and digital future. Luxembourg: Publications Office of the European Union.

Nathan, M., and Rosso, A. 2022. Innovative events: Product launches, innovation and firm performance. Research Policy 51 (1): 104373. doi:10.1016/j.respol.2021.104373.

Nelson, R. R., and Winter, S. 1982. An Evolutionary Theory of Economic Change. Cambridge, MA: Harvard University Press.

Nooteboom, B. 1999. Innovation, learning and industrial organisation. Cambridge Journal of Economics 23 (2): 127–50. doi:10.1093/cje/23.2.127.

OECD. 2018. Oslo Manual: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition, The Measurement of Scientific, Technological and Innovation Activities. Paris: OECD Publishing.

Park, H. W. 2003. Hyperlink network analysis: A new method for the study of social structure on the web. Connections 25: 49–61.

Ponds, R., van Oort, F., and Frenken, K. 2007. The geographical and institutional proximity of research collaboration. Papers in Regional Science 86 (3): 423–43. doi:10.1111/j.1435-5957.2007.00126.x.

Rammer, C., Kinne, J., and Blind, K. 2020. Knowledge proximity and firm innovation: A microgeographic analysis for Berlin. Urban Studies 57 (5): 996–1014. doi:10.1177/0042098018820241.

Romer, P. 1990. Endogenous Technological Change. Journal of Political Economy 98: 71–102.

Schumpeter, J. A. 1911. Theorie der wirtschaftlichen Entwicklung (Theory of economic development). Berlin: Duncker und Humblot.

Simensen, E. O., and Abbasiharofteh, M. 2022. Sectoral patterns of collaborative tie formation: investigating geographic, cognitive, and technological dimensions. Industrial and Corporate Change 00:1–36.

Strumsky, D., and Lobo, J. 2015. Identifying the sources of technological novelty in the process of invention. Research Policy 44 (8): 1445–61. doi:10.1016/j.respol.2015.05.008.

Ter Wal, A. 2014. The dynamics of the inventor network in German biotechnology: Geographic proximity versus triadic closure. Journal of Economic Geography 14 (3): 589–620. doi:10.1093/jeg/lbs063.

Tranos, E., Carrascal-Incera, A., and Willis, G. 2022. Using the Web to Predict Regional Trade Flows: Data Extraction, Modeling, and Validation. Annals of the American Association of Geographers 13 (1): 1–23. doi:10.1080/24694452.2022.2109577.

van der Wouden, F. 2020. A history of collaboration in US invention: Changing patterns of co-invention, complexity and geography. Industrial and Corporate Change 29 (3): 599–619. doi:10.1093/icc/dtz058.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017. Attention Is All You Need. http://arxiv.org/pdf/1706.03762v5.

Vaughan, L., Gao, Y., and Kipp, M. 2006. Why are hyperlinks to business Websites created? A content analysis. Scientometrics 67 (2): 291–300. doi:10.1007/s11192-006-0100-6.

Vaughan, L., and Wu, G. 2004. Links to commercial websites as a source of business information. Scientometrics 60 (3): 487–96. doi:10.1023/B:SCIE.0000034389.14825.bc.

Vedres, B. 2021. Network mechanisms in innovation: Borrowing and sparking ideas around structural holes. SSRN Electronic Journal 45 (3): 425. doi:10.2139/ssrn.3878902.

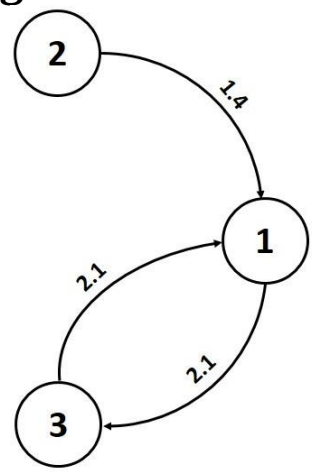
Verbrugge, L. M. 1983. A research note on adult friendship contact: a dyadic perspective. Soc. F. 62 (78).

Wanzenböck, I., Wesseling, J., Frenken, K., Hekkert, M., and Weber, M. 2019. A framework for mission-oriented innovation policy: Alternative pathways through the problem-solution space.

Weitzman, M. L. 1998. Recombinant Growth. The Quarterly Journal of Economics 113 (2): 331–60. doi:10.1162/003355398555595.

Zipf, G. K. 1949. Human behaviour and the principle of least effort. Cambridge: Addison–Wesley.

Figures



Firms	Hyperlinks	Distances	Mean distance
1	[2, 3, 3]	[1.4, 2.1, 2.1]	1.9
2	[1]	[1.4]	1.4
3	[1, 1]	[2.1, 2.1]	2.1

Figure 1. Schematic representation of a firm hyperlink network. Network of three firms with hyperlink connections and a corresponding exemplary distance measure.

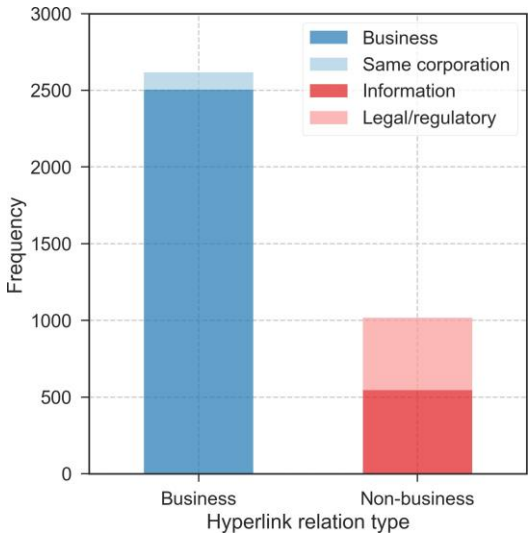


Figure 2. Manually labelled training dataset of hyperlinked firm pairs.

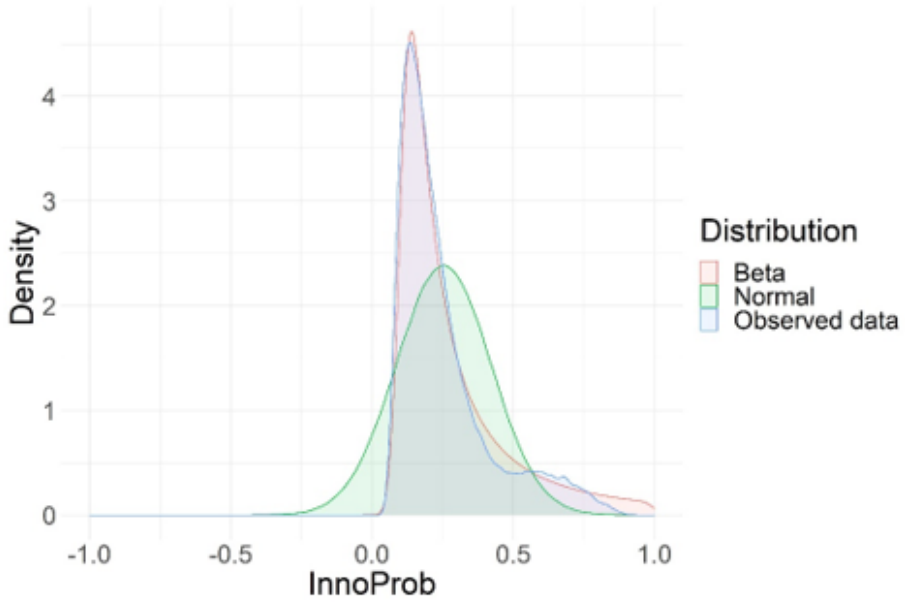


Figure 3. The *InnoProb* (the dependent variable), normal, and beta distributions (p: 1.099 and q: 0.131).

Note: the normal distribution is estimated based on the mean and standard deviation of *InnoProb* (mean: 0.25, sd: 0.17). For this figure, the beta distribution is estimated based on randomly selected p and q parameters. The illustrated beta distribution is selected based on its similarity to the one of *InnoProb* suggested by the Kolmogorov–Smirnov test.

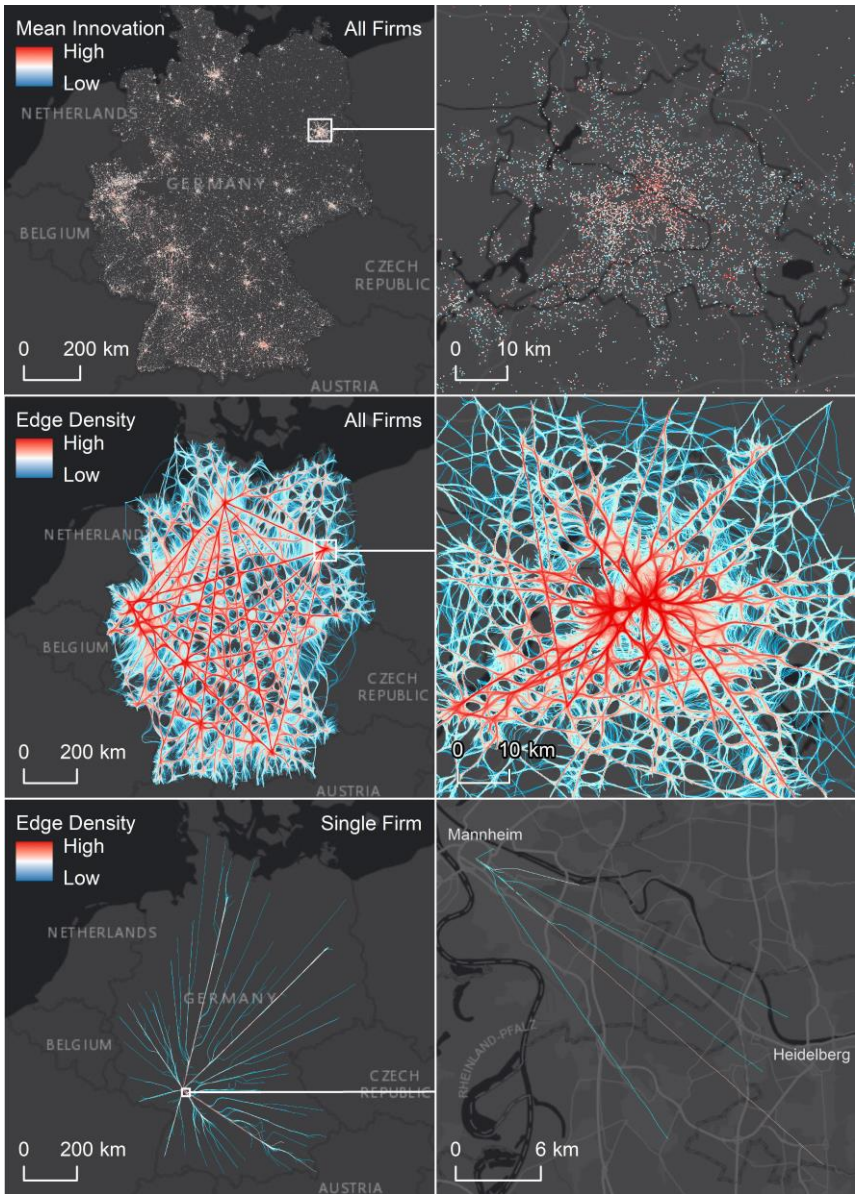


Figure 4. The Digital Layer of Germany. Top row: Mean product innovator probability for Germany (left) and Berlin (right). Middle row: Hyperlink connections between firms in Germany (left) and Berlin (right). Bottom row: Hyperlink connections of a single firm observation in Germany (left) and the Rhine- Neckar region (right).

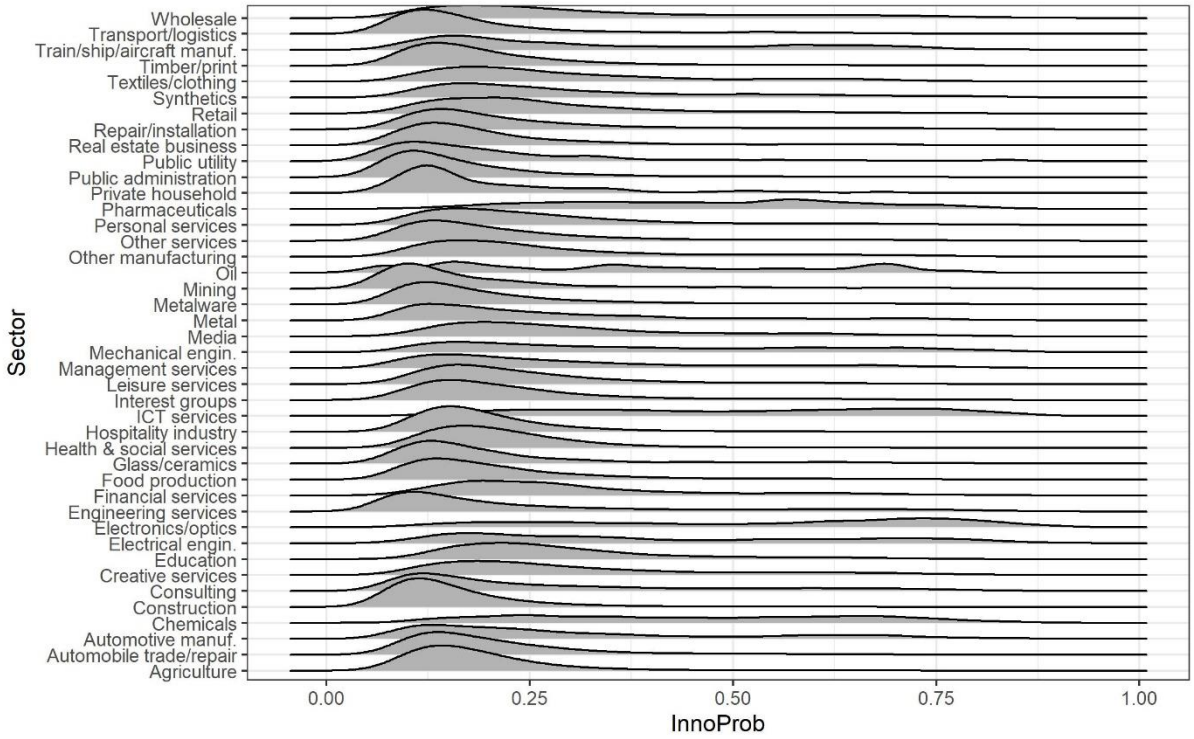


Figure 5. The distribution of the dependent variable stratified by sector.

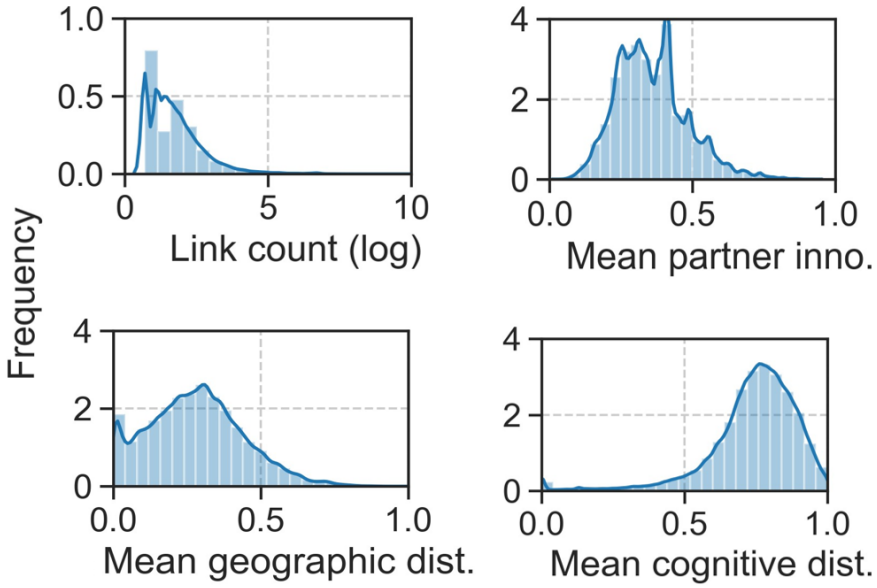


Figure 6. Kernel density estimations for variables of interest.

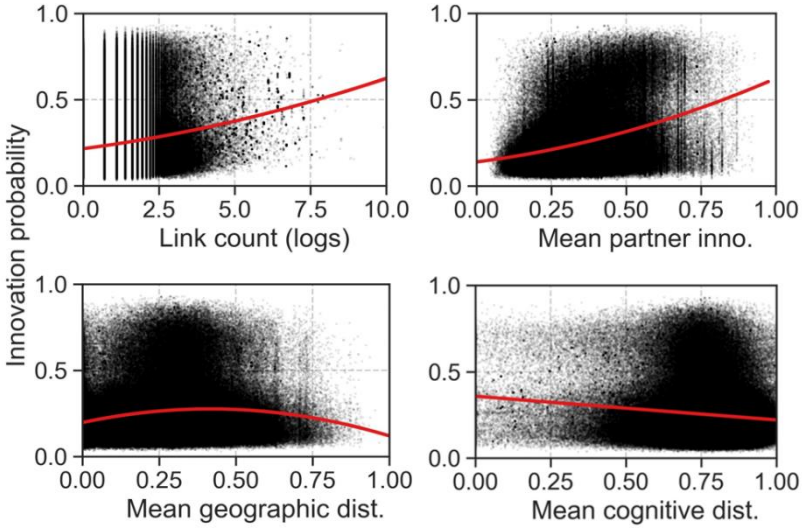


Figure 7. Scatter plots for firm-level predicted innovation probability and variables of interest.

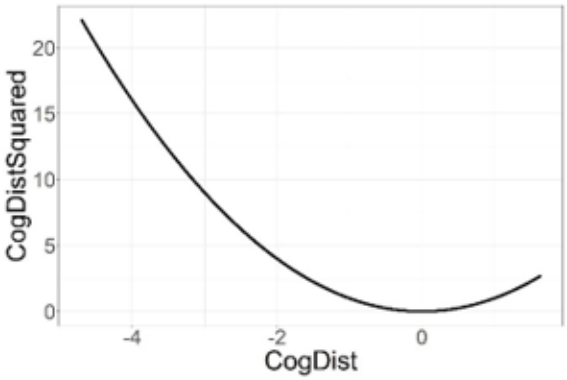


Figure 8. Cognitive distance variable (*CogDist*) and its quadradic term (*CogDist Squared*, mean: 0.99, sd: 2.66).

Note: *CogDist* and its quadradic term are negatively but not strongly correlated (Pearson correlation: -0.66)

Tables

Table 1. Descriptive statistics.

Statistic	N	Mean	St. Dev.	Min	Max	Share of missing values (%)
InnoProb	534,627	0.25	0.17	0.03	0.93	0.00
LinkCount (z-score)	534,627	0.00	1.00	-1.32	9.78	0.00
InnoPartner (z-score)	534,627	0.00	1.00	-2.41	4.17	0.00
GeoDist (z-score)	512,924	0.00	1.00	-1.75	4.50	4.05
CogDist (z-score)	509,165	0.00	1.00	-4.70	1.64	4.76
Size	534,627	3.26	1.75	1	5	0.00
Age	534,627	3.22	0.85	1	5	0.00
Density	534,627	1.73	3.66	0.00	39.30	0.00
NonBusinessRelation	509,205	0.25	0.18	0.01	0.94	4.76

Table 2: Classification report for hyperlink type (*NonBusinessRelation*) prediction in the test set.

Label	Precision	Recall	f1-score	Support
Non-business	0.86	0.88	0.87	271
Business	0.95	0.94	0.95	681
Macro average	0.90	0.91	0.91	952
Weighted average	0.92	0.92	0.92	952
	Accuracy			
Overall	0.92			

Table 3. Correlation table.

	1	2	3	4	5	6	7	8	9
1) InnoProb	1.00								
2) LinkCount (z-score)	0.20	1							
3) InnoPartner (z-score)	0.31	-0.08	1						
4) GeoDist (z-score)	0.09	-0.11	0.33	1					
5) CogDist (z-score)	-0.13	-0.2	0.05	0.14	1				
6) Size	-0.02	0.01	-0.04	-0.02	-0.01	1			
7) Age	-0.12	0.04	-0.12	-0.09	0	0.11	1		
8) Density	0.14	0.09	0.05	0.02	-0.09	0.01	-0.06	1	
9) NonBusinessRelation	-0.14	-0.04	-0.3	-0.08	-0.02	0.06	0.12	-0.01	1

Table 4. Results of the beta regressions.

	Dependent variable: InnoProb				
	(1)	(2)	(3)	(4)	(5)
LinkCount (z-score)	0.12***	0.14***	0.14***	0.13***	0.14***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
InnoPartner (z-score)		0.21***	0.21***	0.21***	0.21***
		(0.001)	(0.001)	(0.001)	(0.001)
GeoDist (z-score)			0.01***	0.01***	0.01***
			(0.001)	(0.001)	(0.001)
CogDist (z-score)				-0.07***	-0.06***
				(0.001)	(0.001)
CogDistSquared					0.01***
					(0.0005)
(phi)	9.42***	10.06***	10.06***	10.18***	10.19***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	-0.76***	-0.88***	-0.88***	-0.89***	-0.90***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Controls	Yes	Yes	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes	Yes	Yes
Observations	509,205	509,205	509,205	509,165	509,165
AIC	-676551.6	-707798.8	-707821.7	-713650.8	-713916.0
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 5. Results of the beta regressions with and without interaction terms.

	Dependent variable: InnoProb			
	(1)	(2)	(3)	(4)
LinkCount (z-score)	0.14***	0.14***		
	(0.001)	(0.001)		
InnoPartner (z-score)	0.21***	0.22***		
	(0.001)	(0.001)		
LinkCount × InnoPartner		0.10***		
		(0.001)		
LinkCount (dummy)			0.21***	0.16***
			(0.002)	(0.003)
InnoPartner (dummy)			0.32***	0.27***
			(0.002)	(0.003)
LinkCount (dummy) × InnoPartner (dummy)				0.19***
				(0.005)
GeoDist (z-score)	0.01***	0.01***	0.04***	0.04***
	(0.001)	(0.001)	(0.001)	(0.001)
CogDist (z-score)	-0.06***	-0.06***	-0.06***	-0.06***
	(0.001)	(0.001)	(0.001)	(0.001)
CogDistSquared	0.01***	0.01***	0.01***	0.01***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)
(phi)	10.19***	10.33***	9.90***	9.93***
	(0.02)	(0.02)	(0.02)	(0.02)
Constant	-0.90***	-0.89***	-0.96***	-0.94***
	(0.01)	(0.01)	(0.01)	(0.01)
Controls	Yes	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes	Yes
Observations	509,165	509,165	509,165	509,165
AIC	-713916.0	-721027.3	-700311.9	-701976.5
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6. Results of the beta regressions on the full and a sample of innovative firms.

	Dependent variable: InnoProb			
	Full sample	Full sample	Innovative Firms*	Innovative Firms*
LinkCount (z-score)	0.14***	0.14***	0.07***	0.06***
	(0.001)	(0.001)	(0.001)	(0.001)
InnoPartner (z-score)	0.21***	0.22***	0.15***	0.14***
	(0.001)	(0.001)	(0.002)	(0.002)
LinkCount × InnoPartner		0.10***		0.06***
		(0.001)		(0.002)
GeoDist (z-score)	0.01***	0.01***	0.005***	0.01***
	(0.001)	(0.001)	(0.002)	(0.002)
CogDist (z-score)	-0.06***	-0.06***	-0.03***	-0.03***
	(0.001)	(0.001)	(0.002)	(0.002)
CogDistSquared	0.01***	0.01***	-0.004***	-0.002***
	(0.0005)	(0.0005)	(0.001)	(0.001)
(phi)	10.19***	10.33***	13.04***	13.21***
	(0.02)	(0.02)	(0.05)	(0.05)
Constant	-0.90***	-0.89***	-0.07***	-0.06***
	(0.01)	(0.01)	(0.02)	(0.02)
Controls	Yes	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes	Yes
Observations	509,165	509,165	129,259	129,259
AIC	-713916.0	-721027.3	-157872.6	-159579.3

Note: *p<0.1; **p<0.05; ***p<0.01

*Innovative firms are the ones with *InnoProb* values greater than the 75th percentile.