

University of Groningen

The digital layer

Abbasiharofteh, Milad; Krüger, Miriam ; Kinne, Jan; Lenz, David; Resch, Bernd

Published in:
Spatial Economic Analysis

DOI:
[10.1080/17421772.2023.2193222](https://doi.org/10.1080/17421772.2023.2193222)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2023

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2023). The digital layer: alternative data for regional and innovation studies. *Spatial Economic Analysis*, 18(4), 507-529. Advance online publication. <https://doi.org/10.1080/17421772.2023.2193222>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

The digital layer: alternative data for regional and innovation studies

Milad Abbasiharofteh, Miriam Krüger, Jan Kinne, David Lenz & Bernd Resch

To cite this article: Milad Abbasiharofteh, Miriam Krüger, Jan Kinne, David Lenz & Bernd Resch (2023) The digital layer: alternative data for regional and innovation studies, *Spatial Economic Analysis*, 18:4, 507-529, DOI: [10.1080/17421772.2023.2193222](https://doi.org/10.1080/17421772.2023.2193222)

To link to this article: <https://doi.org/10.1080/17421772.2023.2193222>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



View supplementary material [↗](#)



Published online: 14 Apr 2023.



Submit your article to this journal [↗](#)



Article views: 1782



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)

The digital layer: alternative data for regional and innovation studies

Milad Abbasiharofteh ^a, Miriam Krüger^b, Jan Kinne^{c,d,e,f}, David Lenz^{c,g} and Bernd Resch^{e,f}

ABSTRACT

The lack of large-scale data revealing the interactions among firms has constrained empirical studies. Utilizing relational web data has remained unexplored as a remedy for 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 the innovation capabilities of firms 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 for 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

HISTORY Received 17 March 2022; in revised form 7 March 2023

1. INTRODUCTION

Innovation and its impact on economic growth have been of great interest in the past decades (Marshall, 1890; Schumpeter, 1911). 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. Often, 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 (Strumsky & Lobo, 2015; Vedres, 2021).

CONTACT Milad Abbasiharofteh  m.abbasiharofteh@rug.nl

^aEconomic Geography, University of Groningen, Groningen, the Netherlands

^bChair of Innovation Economics, Technical University of Berlin, Berlin, Germany


^cistari.ai, Mannheim, Germany

^dZEW Centre for European Economic Research, Mannheim, Germany

^eZ_GIS, University of Salzburg, Salzburg, Austria

^fCenter for Geographic Analysis, Harvard University, Cambridge, MA, USA

^gDepartment of Econometrics and Statistics, Justus-Liebig-University, Gießen, Germany

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/17421772.2023.2193222>

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 (Abbasiharofteh & Broekel, 2020; Breschi & Lissoni, 2009; Owen-Smith & Powell, 2004; Simensen & 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.

Scholars argue that unresolved research questions in regional and innovation studies call for alternative data sources to be used and linked with traditional data sources (Bottai et al., 2022; Duranton & Kerr, 2018). In a business context, websites serve as a showcase for firms' products, services, credibility, achievements, critical decisions, strategies, and relationships with other firms (Gök et al., 2015). This information is usually encoded using text. In recent years, techniques to retrieve and analyse textual data coupled with high-performance machine learning enabled researchers to harvest and analyse this information by employing web scraping and natural language processing (NLP) techniques (Gök et al., 2015; Kinne & Axenbeck, 2020; Stich et al., 2022).

In this study, we focus on relational web data (also known as hyperlinks), which have 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). Hyperlink data promises a particularly up-to-date and extensive view of the digital reflection of real-world company networks (Park, 2003). 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.

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. 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.

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 Section 3, we present the data and methodology used to construct the Digital Layer as well as the estimation of the innovation capabilities of firms. Then, 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 finally 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.

2. 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.

Scholars acknowledge that a knowledge transfer network is one of the main ways whereby companies access complementary resources and improve their innovation capabilities (Gulati & Gargiulo, 1999). Brusoni et al. (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. Bell (2005) finds that taking a central position in a manager network positively correlates with the increase in the innovation capabilities of Canadian firms. 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.

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 & Albert, 1999). Thus, seeing inter-firm 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).

Furthermore, several scholars interpret the relevance of taking a central position concerning the structural holes and receiving good ideas (Burt, 2004). 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 (Mazzola et al., 2018). 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 attention to the relevance of the structure of inter-firm networks, sociologists argue that researchers should not remain agnostic about the content of exchanged knowledge (Moody, 2011). Evidence for this argument has been found in management and organization studies. Kobarg et al. (2019) analyse 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. Studies in economic geography show that firms can excel in different innovation modes based on the type of exchanged information through their relations (Fitjar & Rodríguez-Pose, 2013; Jensen et al., 2007).

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 et al., 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. 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.

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 enabling factor in knowledge creation in an uncertain environment. Firms' status and innovation capabilities may be positively related. 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. It is well established that the likelihood of forming social and advice tie relations decreases substantially if the geographic distance exceeds a certain threshold (Kabirigi et al., 2022; Sonn & Storper, 2008).

Around the turn of the twentieth century, Marshall (1890) argued that the availability of specialized suppliers (sharing), the availability of 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 et al., 2004). The geographic collocation also enables firms to benefit from non-interactive learning through observing other firms (Glückler, 2013). There is a large body of literature studying industrial clusters and the geography of innovation (Audretsch & Feldman, 1996).

Graevenitz et al. (2022) show that the diffusion of innovation is still spatially bounded despite recent advances in telecommunication and transport systems. Studies of related diversification also provide evidence of the comparative advantage for the collocation of workers with similar skills and local inter-industry matching driven by skill-relatedness. Skill-relatedness mimics the rationale behind Marshallian externalities (Boschma et al., 2014).

While recent studies suggest the importance of collocation 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 et al., 2004). Empirical results supporting this conjecture are mixed. For instance, while Berg (2018) shows that innovation capabilities benefit from both short and long-distance relations, the study of Aarstad et al. (2016) suggests that only local interactions contribute to the innovativeness of small and medium-sized enterprises. 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 dimensions¹ (Boschma, 2005).

The evolutionary economic geography approach suggests that the cognitive dimension of relations plays a critical role in firms' learning and innovation capabilities. Cognitive proximity seems even more relevant when considering that innovation increasingly requires larger teams that consist of experts specialized in similar or related fields (van der Wouden, 2020). In other words, 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.

Among numerous studies, 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.

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 (Boschma, 2005). The notion of 'optimal' proximity builds on Nooteboom's (1999) 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 does not contribute to firms' innovative performance (Boschma & Frenken, 2010).

Empirical evidence for this has been presented by Wuyts et al. (2005), Hagedoorn and Cloudt (2003), 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.

3. 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.

3.1. 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 & 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 & Axenbeck, 2020). We removed firms without address information from our dataset and geocoded the remaining firms using street-level geocoding.

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

3.2. 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 & 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 1.4 if the hyperlink connection between a pair of firms is unidirectional or weight 2.1 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 2.1 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.

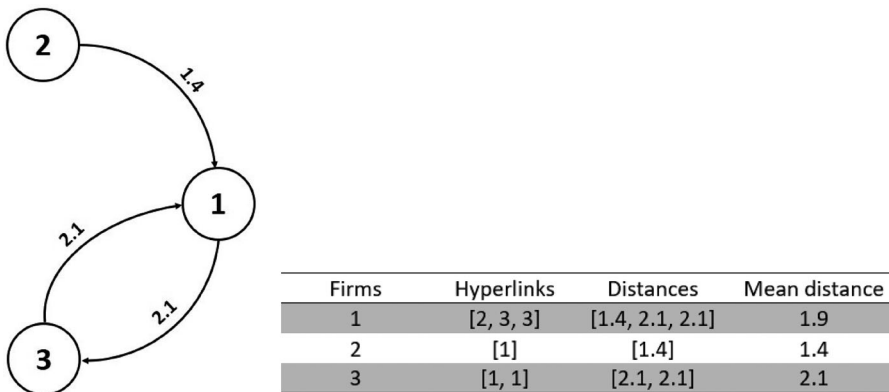


Figure 1. Schematic representation of a firm hyperlink network. Network of three firms with hyperlink connections and a corresponding exemplary distance measure.

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.

3.3. 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 labelled (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. 2020). 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 Appendix A in the supplemental data online). Casted to a binary variable using a classification threshold of 0.4 (see Kinne & 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 & Axenbeck, 2020 for details).

4. 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.

4.1. 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 *LinkCount* 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.

4.2. 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.

4.3. 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, and their products and services. This information is usually codified as text and can be extracted and analysed to assess firms' products and services (Gök et al., 2015). In their 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. The tf-idf algorithm assigns each document to a fixed-size sparse vector of size V , where 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*). We 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 NLP studies (Rahimi et al., 2018).

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.

4.4. 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 doing 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). If 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).²

First, we created a training dataset for that classification task by sampling 5,000 random pairs of hyperlinked firms from our dataset. Subsequently, we labelled 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.

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 labelled dataset and used one-third (952 firms) as a test set to evaluate the model's performance. Table 1 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.

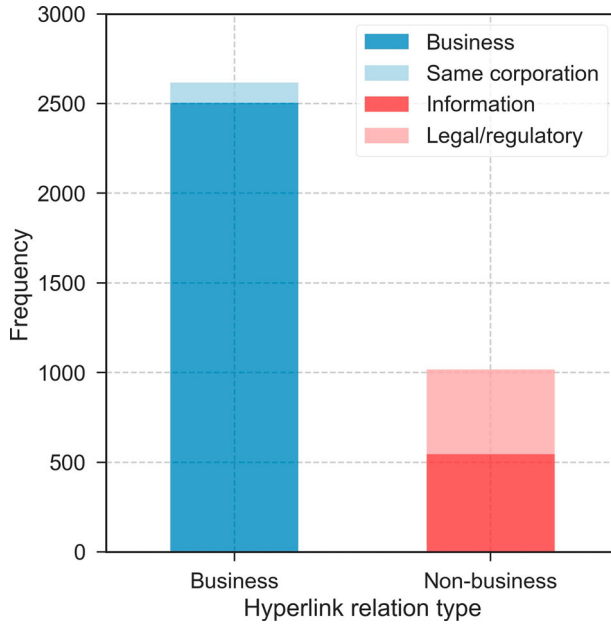


Figure 2. Manually labelled training dataset of hyperlinked firm pairs.

Table 1. 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			

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 *Non-BusinessRelation*) to 1.0 (high probability of non-business relation; large *NonBusinessRelation*).

Appendix B in the supplemental data online provides the pairwise Pearson correlation coefficients of variables. Given the dependent variable is bounded between zero and one, we opted for a set of beta regression models (see Appendix C in the supplemental data online, providing detailed information on the estimation strategy). **Figure 3** compares normal and beta distributions with the distribution of the dependent variable (*InnoProb*).

5. 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 colouring gives the mean innovation probability for the

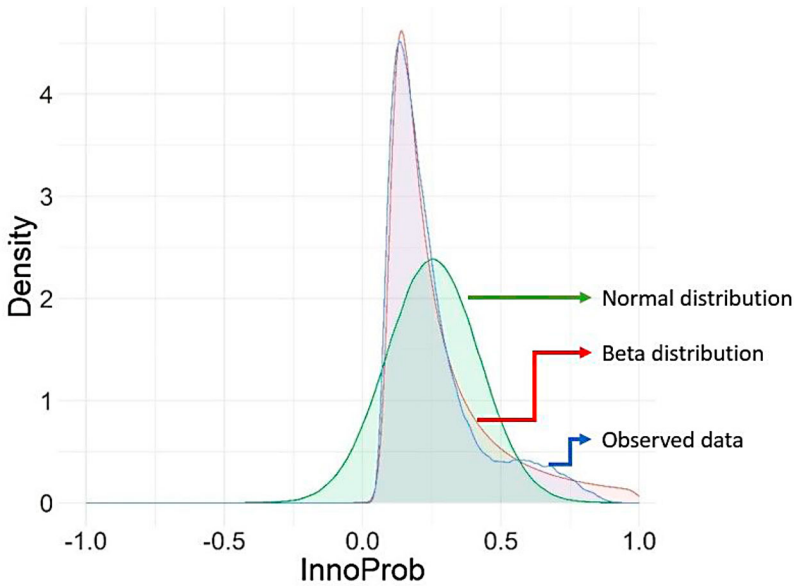


Figure 3. The *InnoProb* (the dependent variable), normal and beta distributions (p : 1.099 and q : 0.131). Note: the normal distribution is estimated based on the mean and standard deviation of *InnoProb* (mean: 0.25, sd: 0.17). For this figure, the beta distribution is estimated based on randomly selected 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.

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 et al., 2012). Unsurprisingly, the density of hyperlink connections between any two areas seems highly dependent on population.³

Figure 5 illustrates the distribution of *InnoProb* stratified by sector (also see Appendix D illustrating the distribution of the four variables of interest stratified by firms’ innovation capabilities). 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 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). Appendix A in the supplemental data online provides descriptive statistics for the variables.

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.

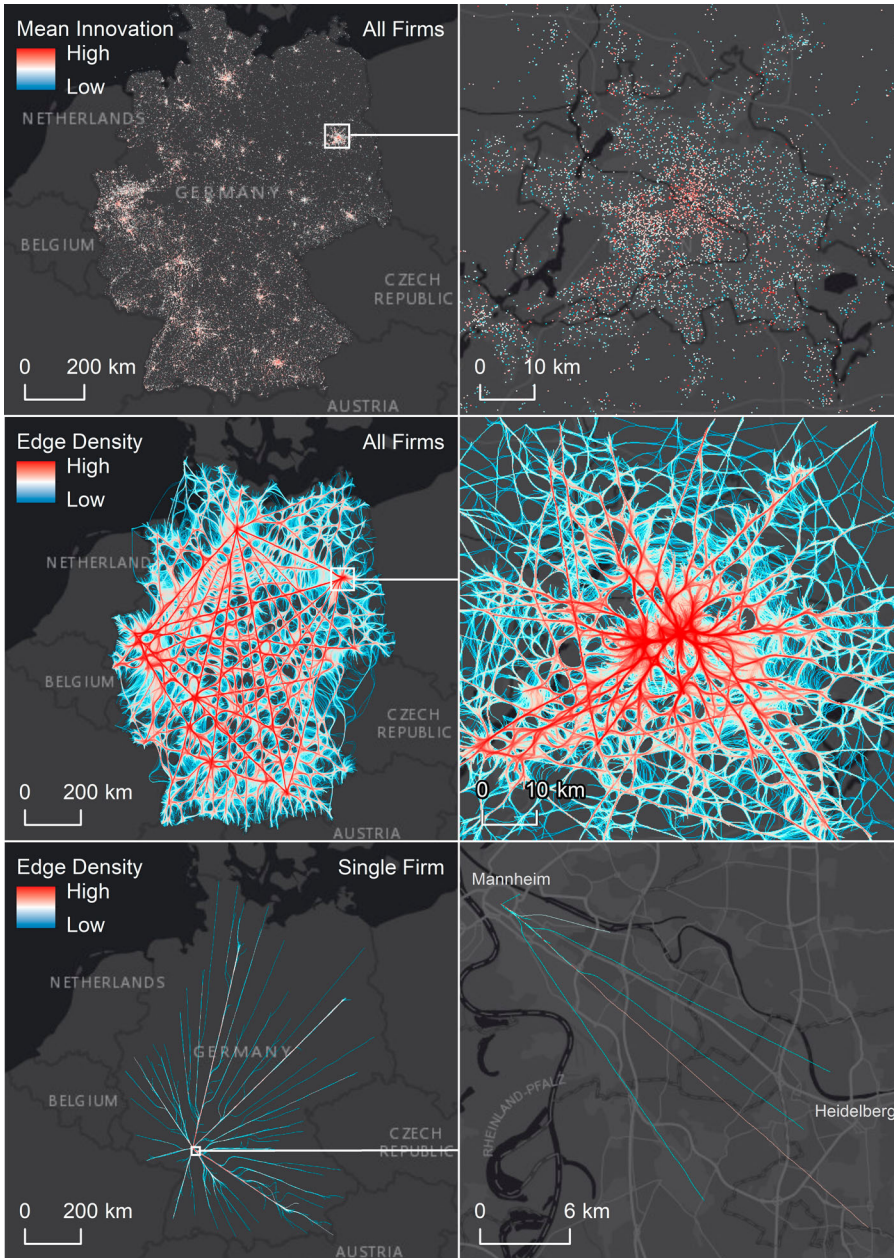


Figure 4. The Digital Layer of Germany. Top row: Mean product innovator probability for Germany (left) and Berlin (right). Middle row: Hyperlink connections between firms in Germany (left) and Berlin (right). Bottom row: Hyperlink connections of a single firm observation in Germany (left) and the Rhine-Neckar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

5.1. 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 Hünernmund and Louw (2022) that estimated effect sizes of control variables

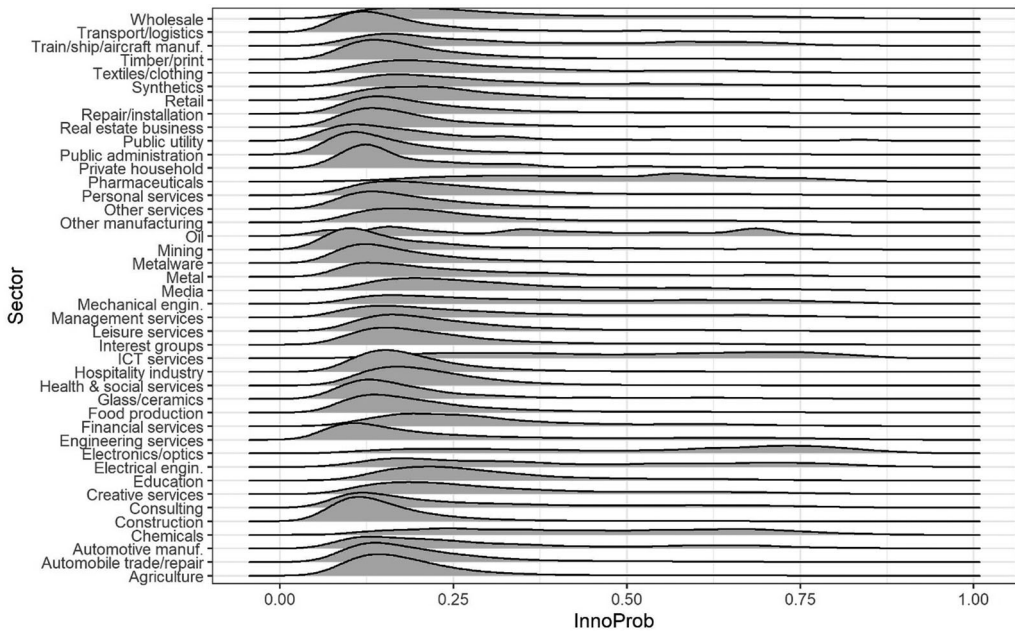


Figure 5. The distribution of the dependent variable stratified by sector.

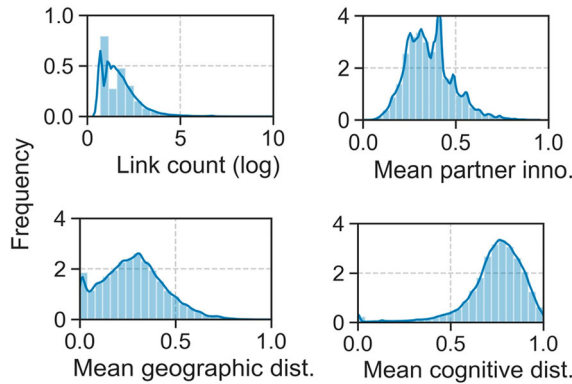


Figure 6. Kernel density estimations for variables of interest.

(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.⁴ 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 & Zeileis, 2010). 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 2). 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

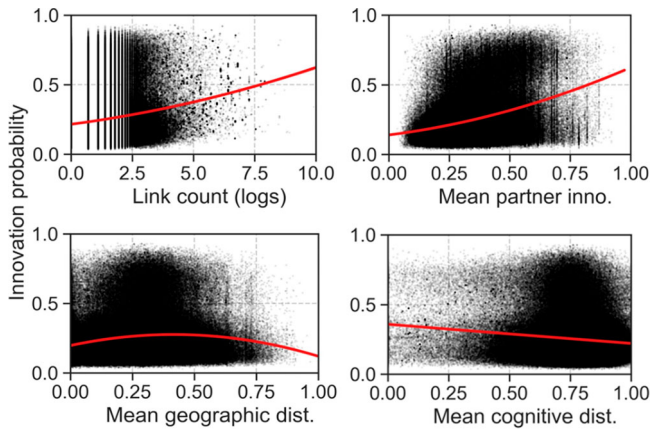


Figure 7. Scatter plots for firm-level predicted innovation probability and variables of interest.

results suggest that the number of hyperlinks is positively associated with firms’ innovation capabilities. More specifically, increasing *LinkCount* by one standard deviation increases the likelihood of the innovation capability of firms by 15%. Similarly, Giuliani and Bell (2005)

Table 2. Results of the beta regressions.

	Dependent variable: <i>InnoProb</i>				
	(1)	(2)	(3)	(4)	(5)
LinkCount (z-score)	0.12*** (0.001)	0.14*** (0.001)	0.14*** (0.001)	0.13*** (0.001)	0.14*** (0.001)
InnoPartner (z-score)		0.21*** (0.001)	0.21*** (0.001)	0.21*** (0.001)	0.21*** (0.001)
GeoDist (z-score)			0.01*** (0.001)	0.01*** (0.001)	0.01*** (0.001)
CogDist (z-score)				-0.07*** (0.001)	-0.06*** (0.001)
CogDistSquared					0.01*** (0.0005)
(phi)	9.42*** (0.02)	10.06*** (0.02)	10.06*** (0.02)	10.18*** (0.02)	10.19*** (0.02)
Constant	-0.76*** (0.01)	-0.88*** (0.01)	-0.88*** (0.01)	-0.89*** (0.01)	-0.90*** (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: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

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 21% increase in likelihood of firms' innovation capability by one unit increase in *InnoPartner*. These findings align with those in the literature that emphasize the relevance of inter-firm relations as knowledge transfer channels (Kobarg et al., 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 et al., 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, and cognitively distant firms encounter problems interpreting exchanged knowledge beyond their absorptive capacity.

By including a quadratic term of *CogDist* (i.e., *CogDistSquared*) a potential inverted U-shaped relationship can be investigated between the cognitive distance of linked firms and their innovation capabilities. Figure 8 shows the relation between *CogDist* and its quadratic 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 are 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 (Balland et al., 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.

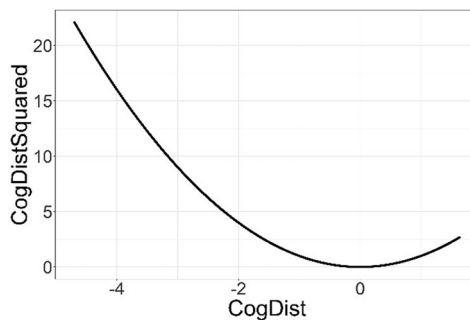


Figure 8. Cognitive distance variable (*CogDist*) and its quadratic term (*CogDist Squared*, mean: 0.99, sd: 2.66). Note: *CogDist* and its quadratic term are negatively but not strongly correlated (Pearson correlation: -0.66).

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 3) is positive and statistically significant. Since including an interaction term based on two continuous variables may lead to a biased estimation of interaction effects, 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.

The descriptive statistics (see Appendix A in the supplemental data online) suggest no significant difference between the four variables of interest among firms with low, average and

Table 3. Results of the beta regressions with and without interaction terms.

	Dependent variable: <i>InnoProb</i>			
	(1)	(2)	(3)	(4)
LinkCount (z-score)	0.14*** (0.001)	0.14*** (0.001)		
InnoPartner (z-score)	0.21*** (0.001)	0.22*** (0.001)		
LinkCount × InnoPartner		0.10*** (0.001)		
LinkCount (dummy)			0.21*** (0.002)	0.16*** (0.003)
InnoPartner (dummy)			0.32*** (0.002)	0.27*** (0.003)
LinkCount (dummy) × InnoPartner (dummy)				0.19*** (0.005)
GeoDist (z-score)	0.01*** (0.001)	0.01*** (0.001)	0.04*** (0.001)	0.04*** (0.001)
CogDist (z-score)	-0.06*** (0.001)	-0.06*** (0.001)	-0.06*** (0.001)	-0.06*** (0.001)
CogDistSquared	0.01*** (0.0005)	0.01*** (0.0005)	0.01*** (0.0005)	0.01*** (0.0005)
(phi)	10.19*** (0.02)	10.33*** (0.02)	9.90*** (0.02)	9.93*** (0.02)
Constant	-0.90*** (0.01)	-0.89*** (0.01)	-0.96*** (0.01)	-0.94*** (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: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

high degrees of innovation capabilities. Table 4 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 (Appendix E in the supplemental data online includes further robustness checks).

Table 4. Results of the beta regressions on the full and a sample of innovative firms.

	Dependent variable: <i>InnoProb</i>			
	Full sample	Full sample	InnovativeFirms*	InnovativeFirms*
LinkCount (z-score)	0.14*** (0.001)	0.14*** (0.001)	0.07*** (0.001)	0.06*** (0.001)
InnoPartner (z-score)	0.21*** (0.001)	0.22*** (0.001)	0.15*** (0.002)	0.14*** (0.002)
LinkCount × InnoPartner		0.10*** (0.001)		0.06*** (0.002)
GeoDist (z-score)	0.01*** (0.001)	0.01*** (0.001)	0.005*** (0.002)	0.01*** (0.002)
CogDist (z-score)	-0.06*** (0.001)	-0.06*** (0.001)	-0.03*** (0.002)	-0.03*** (0.002)
CogDistSquared	0.01*** (0.0005)	0.01*** (0.0005)	-0.004*** (0.001)	-0.002*** (0.001)
(phi)	10.19*** (0.02)	10.33*** (0.02)	13.04*** (0.05)	13.21*** (0.05)
Constant	-0.90*** (0.01)	-0.89*** (0.01)	-0.07*** (0.02)	-0.06*** (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.

6. 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 a 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 analyse 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 2020). 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 NLP methods have enabled researchers to train algorithms based on hyperlinks' ambient texts (texts surrounding hyperlinks) to classify hyperlinks. For instance, researchers can utilize the descriptors of goods and services provided by the trademark data to investigate whether ambient texts can capture inter-firm transactional relationships (Abbasiharofteh et al., 2022). It is important to note that 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 analysing the nodal and dyadic attributes of hyperlink portfolios (i.e., link count and geographic and cognitive distances on hyperlinked firms). We did not consider other structural network measures at the triadic level (e.g., triadic closure) and meso level (e.g., community membership). Several studies have shown the relevance of these measures in the innovation capabilities of individuals and firms (Abbasiharofteh, 2020; Abbasiharofteh et al., 2020; Lobo & Strumsky, 2008; Strumsky & Lobo, 2015). 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 analysing a hyperlink network are promising avenues for future work.

The Digital Layer approach has 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 NLP methods, company relationships can thus not only be surveyed quantitatively but also be evaluated in terms of quality and scope.

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 (Janssen & Abbasiharofteh, 2022). The Digital Layer approach provides the possibility to assess the impact of mission-oriented policies by analysing 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. 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 analyse 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.

ACKNOWLEDGEMENT

This article is based on an earlier working paper 'The Digital Layer: How Innovative Firms Relate on the Web' by Miriam Krüger, Jan Kinne, David Lenz und Bernd Resch published as a ZEW Discussion Paper (no. 20-003) available at: <https://www.zew.de/publikationen/the-digital-layer-how-innovative-firms-relate-on-the-web>

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

FUNDING

This work was supported by Bundesministerium für Bildung und Forschung [grant number 16IF1001].

NOTES

¹ Boschma (2005) formulates five proximity dimensions (geographical, cognitive, organizational, social, institutional). We however focus on geographical and cognitive dimensions

that have attracted the most attraction in the literature, whereas other dimensions are less studied or used mainly as control variables in empirical studies.

² 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 $(x - \bar{x})/\text{sd}(x)$, where \bar{x} and $\text{sd}(x)$ are the mean and standard deviation of x , respectively.

³ Figure 4 is not intended to be of high analytical value but rather to give an overview of the dataset and its granularity.

⁴ The coefficients of control variables and corresponding standard errors are available upon request from authors.

ORCID

Milad Abbasiharofteh  <http://orcid.org/0000-0001-9694-4193>

REFERENCES

- Aarstad, J., Kvitastein, O. A., & Jakobsen, S.-E. (2016). Local buzz, global pipelines, or simply too much buzz? A critical study. *Geoforum; Journal of Physical, Human, and Regional Geosciences*, 75, 129–133. <https://doi.org/10.1016/j.geoforum.2016.07.009>
- Abbasiharofteh, M. (2020). Endogenous effects and cluster transition: A conceptual framework for cluster policy. *European Planning Studies*, 28(12), 1–24. <https://doi.org/10.1080/09654313.2020.1724266>
- Abbasiharofteh, M., & 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). <https://doi.org/10.1177/0308518X20933904>
- Abbasiharofteh, M., Castaldi, C., & Petralia, S. G. (2022). From patents to trademarks: Towards a concordance map (Final report): European Patent Office (EPO).
- Abbasiharofteh, M., Kogler, D. F., & Lengyel, B. (2020). Atypical combination of technologies in regional co-inventor networks. *Papers in Evolutionary Economic Geography (PEEG)*, 20.55, from <http://econ.geo.uu.nl/peeg/peeg2055.pdf>.
- Audretsch, D., & Feldman, M. P. (1996). R&D spillovers and the geography of innovation and production. *American Economic Review*, 86(3), 630–640.
- Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3), 259–280. <https://doi.org/10.1257/jep.32.3.259>
- Balland, P.-A., Boschma, R., & Frenken, K. (2022). Proximity, innovation and networks: A concise review and some next steps. In A. Torre, & D. Gallaud (Eds.), *Handbook of proximity relations* (pp. 70–80). Edward Elgar Publishing.
- Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512. <https://doi.org/10.1126/science.286.5439.509>
- Bathelt, H., Malmberg, A., & Maskell, P. (2004). Clusters and knowledge: Local buzz, global pipelines and the process of knowledge creation. *Progress in Human Geography*, 28(1), 31–56. <https://doi.org/10.1191/0309132504ph4690a>
- Bell, G. G. (2005). Clusters, networks, and firm innovativeness. *Strategic Management Journal*, 26(3), 287–295. <https://doi.org/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. <https://doi.org/10.1177/2393957517749072>
- Bersch, J., Gottschalk, S., Müller, B., & Niefert, M. (2014). The Mannheim Enterprise Panel (MUP) and firm statistics for Germany. Mannheim: ZEW Discussion Paper 14–104.
- Blomqvist, K., & 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. <https://doi.org/10.1504/IJMCP.2006.009645>

- Boschma, R. (2005). Proximity and innovation: A critical assessment. *Regional Studies*, 39(1), 61–74. <https://doi.org/10.1080/0034340052000320887>
- Boschma, R., Eriksson, R. H., & Lindgren, U. (2014). Labour market externalities and regional growth in Sweden: The importance of labour mobility between skill-related industries. *Regional Studies*, 48(10), 1669–1690. <https://doi.org/10.1080/00343404.2013.867429>
- Boschma, R., & Frenken, K. (2010). The spatial evolution of innovation networks. A proximity perspective. In R. Boschma, & R. Martin (Eds.), *The handbook of evolutionary economic geography* (pp. 120–135). Edward Elgar Pub.
- Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., & Liberati, C. (2022, September 7–9). Unconventional data for policy: Using Big Data for detecting Italian innovative SMEs. In *Conference on information technology for social good (GoodIT'22)* (pp. 338–344). New York, NY: ACM.
- Breschi, S., & Lissoni, F. (2009). Mobility of skilled workers and co-invention networks: An anatomy of localized knowledge flows. *Journal of Economic Geography*, 9(4), 439–468. <https://doi.org/10.1093/jeg/lbp008>
- Brusoni, S., Prencipe, A., & 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. <https://doi.org/10.2307/3094825>
- Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, 110(2), 349–399. <https://doi.org/10.1086/421787>
- Cantner, U., & Meder, A. (2008). Regional and technological effects on cooperation behavior. Jena Economic Research Papers, #14-2008.
- Chandler, D., Haunschild, P. R., Rhee, M., & Beckman, C. M. (2013). The effects of firm reputation and status on interorganizational network structure. *Strategic Organization*, 11(3), 217–244. <https://doi.org/10.1177/1476127013478693>
- Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, 34(2). <https://doi.org/10.18637/jss.v034.i02>
- Duranton, G., & Kerr, W. (2018). The logic of agglomeration. In G. L. Clark, M. P. Feldman, M. S. Gertler, & D. Wójcik (Eds.), *The new Oxford handbook of economic geography* (pp. 347–365). Oxford University Press.
- Eriksson, R., & Lindgren, U. (2008). Localized mobility clusters: Impacts of labour market externalities on firm performance. *Journal of Economic Geography*, 9(1), 33–53. <https://doi.org/10.1093/jeg/lbn025>
- Fitjar, R. D., & Rodríguez-Pose, A. (2013). Firm collaboration and modes of innovation in Norway. *Research Policy*, 42(1), 128–138. <https://doi.org/10.1016/j.respol.2012.05.009>
- Gault, F. (ed.). (2013). *Handbook of innovation indicators and measurement*. Edward Elgar Publishing.
- Giuliani, E. (2007). The selective nature of knowledge networks in clusters: Evidence from the wine industry. *Journal of Economic Geography*, 7(2), 139–168. <https://doi.org/10.1093/jeg/lbl014>
- Giuliani, E., & Bell, M. (2005). The micro-determinants of meso-level learning and innovation: Evidence from a Chilean wine cluster. *Research Policy*, 34(1), 47–68. <https://doi.org/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), 880–894. <https://doi.org/10.1080/00343404.2013.779659>
- Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, 102(1), 653–671. <https://doi.org/10.1007/s11192-014-1434-0>
- Graevenitz, G. v., Graham, S. J. H., & Myers, A. F. (2022). Distance (still) hampers diffusion of innovations. *Regional Studies*, 56(2), 227–241. <https://doi.org/10.1080/00343404.2021.1918334>
- Gulati, R. (1999). Network location and learning: The influence of network resources and firm capabilities on alliance formation. *Strategic Management Journal*, 20(5), 397–420. [https://doi.org/10.1002/\(SICI\)1097-0266\(199905\)20:5<397::AID-SMJ35>3.0.CO;2-K](https://doi.org/10.1002/(SICI)1097-0266(199905)20:5<397::AID-SMJ35>3.0.CO;2-K)
- Gulati, R., & Gargiulo, M. (1999). Where do interorganizational networks come from? *American Journal of Sociology*, 104(5), 1439–1493. <https://doi.org/10.1086/210179>
- Hagedoorn, J., & Cloudt, M. (2003). Measuring innovative performance: Is there an advantage in using multiple indicators. *Research Policy*, 32(8), 1365–1379. [https://doi.org/10.1016/S0048-7333\(02\)00137-3](https://doi.org/10.1016/S0048-7333(02)00137-3)

- Haus-Reve, S., Fitjar, R. D., & Rodríguez-Pose, A. (2019). Does combining different types of collaboration always benefit firms?: Collaboration, complementarity and product innovation in Norway. *Research Policy*, 48(6), 1476–1486. <https://doi.org/10.1016/j.respol.2019.02.008>
- Hünermund, P., & Louw, B. (2022). On the Nuisance of Control Variables in Regression Analysis. arXiv pre-print, 1–22.
- Hurter, C., Ersoy, O., & Telea, A. (2012). Graph bundling by kernel density estimation. *Computer Graphics Forum*, 31(3pt1), 865–874. <https://doi.org/10.1111/j.1467-8659.2012.03079.x>
- Janssen, M. J., & 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. <https://doi.org/10.1016/j.techfore.2022.121689>
- Jensen, M. B., Johnson, B., Lorenz, E., & Lundvall, BÅ. (2007). Forms of knowledge and modes of innovation. *Research Policy*, 36(5), 680–693. <https://doi.org/10.1016/j.respol.2007.01.006>
- Kabirigi, M., Abbasiharofteh, M., Sun, Z., & 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. <https://doi.org/10.1016/j.agsy.2022.103465>
- Kinne, J. (2018). ARGUS - An automated robot for generic universal scraping, from <https://github.com/datawizard1337/ARGUS>.
- Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, 28(4), 443. <https://doi.org/10.1007/s11192-020-03726-9>
- Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PloS one*, 16(4), e0249071. <https://doi.org/10.1371/journal.pone.0249071>
- Kobarg, S., Stumpf-Wollersheim, J., & Welp, I. M. (2019). More is not always better: Effects of collaboration breadth and depth on radical and incremental innovation performance at the project level. *Research Policy*, 48(1), 1–10. <https://doi.org/10.1016/j.respol.2018.07.014>
- Lazzeretti, L., & Capone, F. (2016). How proximity matters in innovation networks dynamics along the cluster evolution. A study of the high technology applied to cultural goods. *Journal of Business Research*, 69(12), 5855–5865. <https://doi.org/10.1016/j.jbusres.2016.04.068>
- Lee, Y., Cho, I., & Park, H. (2015). The effect of collaboration quality on collaboration performance: Empirical evidence from manufacturing SMEs in the Republic of Korea. *Total Quality Management & Business Excellence*, 26(9–10), 986–1001. <https://doi.org/10.1080/14783363.2015.1050169>
- Lin, H.-F. (2014). The impact of socialization mechanisms and technological innovation capabilities on partnership quality and supply chain integration. *Information Systems and e-Business Management*, 12(2), 285–306. <https://doi.org/10.1007/s10257-013-0226-z>
- Lobo, J., & Strumsky, D. (2008). Metropolitan patenting, inventor agglomeration and social networks: A tale of two effects. *Journal of Urban Economics*, 63(3), 871–884. <https://doi.org/10.1016/j.jue.2007.07.005>
- Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An introduction to information retrieval*. Cambridge University Press.
- Marshall, A. (1890). *Principles of economics*. MacMillan.
- Mazzola, E., Perrone, G., & 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), 960–982. <https://doi.org/10.1111/jpim.12438>
- Moody, J. (2011). Network dynamics. In P. Bearman, P. Hedström, P. Dodds, & D. J. Watts (Eds.), *The Oxford handbook of analytical sociology* (pp. 447–474). Oxford University Press.
- Nooteboom, B. (1999). Innovation, learning and industrial organisation. *Cambridge Journal of Economics*, 23(2), 127–150. <https://doi.org/10.1093/cje/23.2.127>
- Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., & Van Den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research Policy*, 36(7), 1016–1034. <https://doi.org/10.1016/j.respol.2007.04.003>
- OECD. (2018). *Oslo manual: Guidelines for collecting, reporting and using data on innovation, 4th edition, The measurement of scientific, technological and innovation activities*. OECD Publishing.

- Owen-Smith, J., & Powell, W. W. (2004). Knowledge networks as channels and conduits: The effects of spillovers in the Boston biotechnology community. *Organization Science*, 15(1), 5–21. <https://doi.org/10.1287/orsc.1030.0054>
- Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections*, 25(1), 49–61.
- Rahimi, S., Mottahedi, S., & Liu, X. (2018). The geography of taste: Using yelp to study urban culture. *ISPRS International Journal of Geo-Information*, 7(9), 376. <https://doi.org/10.3390/ijgi7090376>
- Rammer, C., Kinne, J., & Blind, K. (2020). Knowledge proximity and firm innovation: A microgeographic analysis for Berlin. *Urban Studies*, 57(5), 996–1014. <https://doi.org/10.1177/0042098018820241>
- Schumpeter, J. A. (1911). *Theorie der wirtschaftlichen Entwicklung (Theory of economic development)*. Duncker und Humblot.
- Simensen, E. O., & Abbasiharofteh, M. (2022). Sectoral patterns of collaborative tie formation: Investigating geographic, cognitive, and technological dimensions. *Industrial and Corporate Change*, 31(5), 1–36. <https://doi.org/10.1093/icc/dtac021>
- Sonn, J. W., & Storper, M. (2008). The increasing importance of geographical proximity in knowledge production: An analysis of US patent citations, 1975–1997. *Environment and Planning A*, 40(5), 1020–1039. <https://doi.org/10.1068/a3930>
- Stich, C., Tranos, E., & Nathan, M. (2022). Modeling clusters from the ground up: A web data approach. *Environment and Planning B: Urban Analytics and City Science*, 50(1), 244–267. <https://doi.org/10.1177/23998083221108185>
- Strumsky, D., & Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy*, 44(8), 1445–1461. <https://doi.org/10.1016/j.respol.2015.05.008>
- 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. <https://doi.org/10.1093/icc/dtz058>
- Vedres, B. (2021). Network mechanisms in innovation: Borrowing and sparking ideas around structural holes. *SSRN Electronic Journal*, 45(3), 425. <https://doi.org/10.2139/ssrn.3878902>
- Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics*, 113(2), 331–360. <https://doi.org/10.1162/003355398555595>
- Wuyts, S., Colombo, M. G., Dutta, S., & Nooteboom, B. (2005). Empirical tests of optimal cognitive distance. *Journal of Economic Behavior & Organization*, 58(2), 277–302. <https://doi.org/10.1016/j.jebo.2004.03.019>