// NO.20-003 | 01/2020

DISCUSSION PAPER

// MIRIAM KRÜGER, JAN KINNE,
DAVID LENZ, AND BERND RESCH

The Digital Layer: How Innovative Firms Relate on the Web

Leibniz Association

ZEW

The Digital Layer: How innovative firms relate on the Web

Miriam Krüger[a,1], **Jan Kinne**[b,c,d,e], **David Lenz**[b,f], **and Bernd Resch**[d,e]

[a]Technical University of Berlin, Berlin, Germany; [b]istari.ai, Mannheim, Germany; [c]Department of Economics of Innovation and Industrial Dynamics, ZEW Centre for European Economic Research, Mannheim, Germany; [d]Department of Geoinformatics - Z_GIS, University of Salzburg, Salzburg, Austria; [e]Center for Geographic Analysis, Harvard University, Cambridge, Massachusetts, USA; [f]Department of Econometrics and Statistics, Justus-Liebig-University, Gießen, Germany

Abstract

In this paper, we introduce the concept of a Digital Layer to empirically investigate inter-firm relations at any geographical scale of analysis. The Digital Layer is created from large-scale, structured web scraping of firm websites, their textual content and the hyperlinks among them. Using text-based machine learning models, we show that this Digital Layer can be used to derive meaningful characteristics for the over seven million firm-to-firm relations, which we analyze in this case study of 500,000 firms based in Germany. Among others, we explore three dimensions of relational proximity: (1) Cognitive proximity is measured by the similarity between firms' website texts. (2) Organizational proximity is measured by classifying the nature of the firms' relationships (business vs. non-business) using a text-based machine learning classification model. (3) Geographical proximity is calculated using the exact geographic location of the firms. Finally, we use these variables to explore the differences between innovative and non-innovative firms with regard to their location and relations within the Digital Layer. The firm-level innovation indicators in this study come from traditional sources (survey and patent data) and from a novel deep learning-based approach that harnesses firm website texts. We find that, after controlling for a range of firm-level characteristics, innovative firms compared to non-innovative firms maintain more numerous relationships and that their partners are more innovative than partners of non-innovative firms. Innovative firms are located in dense areas and still maintain relationships that are geographically farther away. Their partners share a common knowledge base and their relationships are business-focused. We conclude that the Digital Layer is a suitable and highly cost-efficient method to conduct large-scale analyses of firm networks that are not constrained to specific sectors, regions, or a particular geographical level of analysis. As such, our approach complements other relational datasets like patents or survey data nicely.

Keywords: Web Mining | Innovation | Proximity | Network | Natural Language Processing

JEL Classification: O30, R10, C80

1. Introduction

Since Schumpeter (1) innovation has been recognized as the key element driving economic growth (2). As a consequence, for decades both researchers and policy makers have focused on understanding innovation dynamics in networks of firms and the drivers behind them. One of the well researched aspects thereby is the impact of proximity on learning, knowledge creation and innovation. Boschma (3) conceptualized five dimensions of proximity that are related to the innovativeness of a firm in a network of firms: cognitive, geographical, organizational, institutional and social proximity. The theoretical approach of (3) found wide adaption in economic geography but has proven to be difficult to operationalize in large-scale empirical studies (see Literature review section). In this paper, we introduce a novel approach based on web mining to map firm networks and to analyze the characteristics of innovative firms in them. For that, we create a so-called Digital Layer of the network of firms located in Germany from large scale web scraping of firm websites, their textual content and the hyperlinks among them. This allows us to analyze firm-to-firm relations and firm characteristics at a larger scale and higher granularity compared to studies using traditional data based on questionnaire-based surveys or patents.

This way, we are able to investigate the characteristics of over half a million firms located in Germany and over seven million relations among them. Using text-based classification and text similarity models from machine learning, we create quantitative measures that describe the position and relationships of each firm in the Digital Layer. We demonstrate that these measures offer meaningful insights on firm-level innovativeness. These measures include the number of partners that a firm has in the network, the innovativeness of its partners, as well as several proximity measures describing the relation to the link partners of each firm.

We then relate these measures (and several firm-level control variables) to the innovativeness of firms in a regression analysis. In this regression analysis, we use two different firm-level innovation indicators as the dependent variable. First, we use a traditional indicator from the questionnaire-based German Community Innovation Survey (CIS) which includes information for about 2,500 firms in our dataset. Second, we use a web-based firm-level innovation indicator developed by (4) which is based on an artificial neural network classification model trained on website texts of firms surveyed in the CIS. The latter indicator is available for all 513,026 firms in our dataset.

With this study we aim to answer the following research

Miriam Krüger and Jan Kinne designed the study and wrote the paper. Jan Kinne gathered the data. Jan Kinne, Miriam Krüger and David Lenz analyzed the data. Bernd Resch supervised the study, discussed the results and proof-read the paper.

The authors declare no conflict of interest. The authors would like to thank the German Federal Ministry of Education and Research for providing funding for the research project (TOBI - Text Data Based Output Indicators as Base of a New Innovation Metric and NETINU - Networks of Innovative Firms; funding ids 16IFI001 and 16IFI106) of which this study is a part.

[1]To whom correspondence should be addressed. E-mail: jan.kinne@zew.de

January 23, 2020

questions:

1. **Research Question 1**: Is our approach to create a *Digital Layer* of interrelated and textually described firms suitable for a large scale web-based analysis of firm networks?
2. **Research Question 2**: How do innovative and non-innovative firms differ concerning their relationships in the Digital Layer and are the observed statistical relations between the different dimensions of proximity and firm innovation in line with the established theory?

The remainder of this paper is structured as follows: First, we give an overview of the literature related to this study. We then present the datasets used to create the Digital Layer and to assess firm-level innovation. In the following methodology section, we outline how we developed measures of firm-to-firm proximity and firm-level embeddedness. We then present our results and discuss them in the following two sections. We finalize this paper with our conclusions and an outlook to potential future research.

2. Literature review

Firm networks, proximity and innovation. More than two decades ago, (5) pointed out that technological change has brought into existence a new type of economy where "information is the key ingredient of social organization and flows of messages and images between networks constitute the basic thread of social structure." According to his reasoning, it is now networks that form the social morphology of our societies and "the extent to which a network has access to technological know-how is at the roots of productivity and competitiveness". In his book "Why information grows" (6) further builds upon this concept of our economy as a social construct of connected firms. Firms again are regarded as networks of individuals and the degree to which firms and networks of firms are capable of producing and crystalizing information lies at the core of why some places are economically successful and others are not. This paradigm differs from the previous view on innovative places and competitive firms as summarized by (7):

> "For a long time, a fundamental debate existed in economic geography about the question whether places are more relevant for the competitiveness of firms, or whether networks matter more (Castells 1996). While the concept "space of places" expresses the idea that the location matters for learning and innovation (being in the right place is what counts), the concept of "space of flows" focuses more on the idea that networks are important vehicles of knowledge transfer and diffusion (meaning that being part of a network is crucial). In a nutshell, the cluster literature claimed that regions are drivers of innovation and economic development: firms in clusters benefit almost automatically from knowledge externalities that are "in the air", as Marshall once put it. [. . .] This is not to say that the cluster literature overlooked the importance of networks. The problem was, however, that the cluster literature suggested that the space of place and the space of flows showed a great deal of overlap (Boschma and Ter Wal 2007). [. . .] Knowledge networks are not territorial, [though], but social constructs that may cross the boundaries of regions. Knowledge diffuses through social networks which may be dense between local agents, but may also span across the world."

And it is not only geography that matters for effective knowledge flows, learning and innovation. (3) conceptualized five dimensions of proximity that play a crucial role for inter-organizational interaction and innovation: cognitive, institutional, social, organizational and geographical proximity. (8) wrote:

> "In short, cognitive proximity indicates the extent to which two organizations share the same knowledge base; organizational proximity the extent to which two organizations are under common hierarchical control, social proximity the extent to which members of two organizations have friendly relationships, institutional proximity the extent to which two organizations operate under the same institutions, and geographical proximity the physical distance or travel time separating two organizations."

Traditional relational data for innovation networks. To empirically assess these different dimensions of proximity and their relation to innovation in firms, relational data is needed. So far relational data has been obtained from either primary survey data or secondary data sources such as patent data. Even though other sources of secondary network data exist, e.g. strategic alliance databases or co-publications, patent data is the most widely used. (7) review and assess the advantages and drawbacks of primary survey data and secondary patent data as relational datasets:

Primary survey data is obtained through interviews and/or questionnaires either by means of the roster-recall methodology or the snowball method (for more information on these methods see (7)). As this is very costly and time-intensive, primary survey data generally fails to capture an entire firm population and is thus regionally or sectorally bounded. Moreover, the quality of the obtained data is very dependent on the response rate of firms. Most datasets represent a static network at one point in time, as the conduction of longitudinal surveys for a potential dynamic analysis of firm networks is even more costly and time-intensive. They therefore conclude that "network analysis on the basis of primary data is most appropriate for small clusters of firms or relatively small sectors within a region." An advantage of survey data is that it can record different dimensions of relationships across the same set of actors. An example for that is (9) study, in which a network of business relations and a network of knowledge-based relationships is identified.

Secondary patent data provides relational links based on the information about the patent applicant or the inventors. The node in the network is hence either the firm or the inventor. A link between firms or inventors exists in case of co-patenting or multi-applicant inventorship. Patent data therefore only reveals relatively formal cooperative links that resulted in a filed patent. Many other forms of inter-firm cooperation and more informal inter-firm interaction are not captured. Moreover, there are only some sectors that strongly rely on patents to protect their innovations, such as the pharmaceutical or the semiconductor industries. Other sectors, such as software industries and services, protect their innovations more likely via secrecy or trademarks. Network studies based on patent data are thus more appropriate for certain sectors than for others. An advantage is, however, that one can construct and analyze networks back in time, as patent data is available for a long time-series. This allows for dynamic analyses of inter-firm networks.

The Digital Layer as a new generation of web-based relational data. In this study, we introduce the concept of a Digital Layer created from large-scale web scraping of geolocated firm websites. The relations among firms in the Digital Layer are constructed from the hyperlinks between their websites, enriched with quantitative measures based on the websites' textual content. (10) identified hyperlinks as the "basic structural element of the internet". He points to hyperlinks as a new social or communication channel and as a means for organizations to exchange information and sustain cooperative relationships. According to him, a hyperlink system is comprised of organizations that are linked together around a common background, interest, or project. In this sense, we expect the Digital Layer to reveal relationships among firms which are of cooperative rather than competitive nature. We explore how the position and relationships of each firm in the Digital Layer relate to firm innovation based on quantitative measures, which operationalize the cognitive, organizational, and geographical proximity to link partners. This way, our dataset allows us to empirically investigate the characteristics of inter-firm interaction and innovation at a larger scale and higher granularity than with previous datasets available. Our dataset does not constrain us to specific sectors (see data section) and bears great potential for a dynamic network analysis of inter-firm relationships (see future work section).

Innovative and non-innovative firms in the Digital Layer. Based on the findings of previous studies using patent and survey data (11–13), we expect that innovative and non-innovative firms differ with regard to their position and relationships in the Digital Layer. (7), for example, reference the study of (12) which found "empirical evidence that firms with cutting-edge technology are usually positioned in the core of inter-firm collaboration networks." Moreover, (13) and (11) found a positive relationship between network centrality of firms and their innovative performance. We hence expect innovative firms to have a higher degree centrality, meaning more hyperlinks, than non-innovative firms in the Digital Layer. Based on the concept of *homophily* (14, 15), meaning that actors link to actors that are similar to them, we also expect that innovative firms especially link to firms of their own sector and to other innovative firms.

Proximity and innovation in the Digital Layer. (8) claim that "it depends on the optimal level of proximity between agents whether their connection will lead to a higher level of innovative performance or not". This means that both too much and too little proximity to partners can hamper interactive learning and innovation. Hence, we expect that close relationships between firms in terms of their cognitive, geographical, and organizational proximity (social and institutional proximity are not assessed in this study), are not necessarily related to higher innovativeness.

Concerning geographical proximity, it is argued that remotely located firms with merely distant partners will not be able to catch the *local buzz* and knowledge spillovers that firms in densely urban locations can grasp from more frequent and sometimes serendipitous face-to-face interactions with other economic actors. On the other hand, local over-embeddedness without any *global pipelines* might lead to missing the next crucial development from another place (16). Some trans-regional linkages are considered crucial to protect from so-called *technological lock-ins* (7, 17). In this sense, we expect innovative firms to have a mixture of local and trans-regional links.

In the case of cognitive proximity, (8) argue that a firm's cognitive base needs to be "close enough to new knowledge in order to communicate, understand and process it successfully". If the cognitive distance between actors becomes too large, learning and knowledge flows are hampered. (18) found that "firms innovate in areas close to their current cognitive capabilities along well-defined technological trajectories". (19) showed that cognitive proximity may enable RD alliances and (20) identified cognitive proximity of actors via patent citations. We thus expect innovative firms in the Digital Layer to be linked to firms that are in close cognitive proximity.

In the case of organizational proximity, "a continuum is assumed ranging from one extreme of 'on the spot' market, to informal relations between firms [...] to the other extreme of a hierarchically organized firm" (8). (11) found a positive relationship between firm survival and a mixture of embedded trust-based ties and arm's length market based ties of a firm. In this sense, we expect innovative firms to be linked with both organizationally close and distant firms.

3. Data

In this section, we first present the base firm dataset of this study. We then outline how web scraping was used 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) that are used in this study.

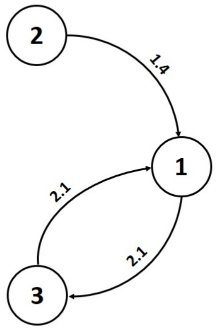
Firm base data. We use the Mannheim Enterprise Panel (MUP) of 2019 as our base dataset. The MUP is a firm panel database that covers the entire population of firms in Germany. It is updated on a semi-annual basis (21). In addition to firm-level characteristics, such as firm size, age, and location, the 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 (22) 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 well as medium sized and larger firms from sectors like manufacturing and ICT (information and communication technology) services. The MUP, nonetheless, represents a comprehensive dataset with a very high URL coverage in those firm groups that are the most relevant for the development of innovation (22, 23). We removed firms without address information from our dataset and geocoded the remaining firms using street-level geocoding (without house numbers; see e.g. (24)).

The geocoded firms were also used to calculate a firm-level location control variable by counting the number of other firms within one kilometer around each individual 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 (25) who showed that spillovers from local knowledge sources decay within a few hundred meters.

Constructing the Digital Layer. For the web scraping of the firm websites, we used ARGUS (26), 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 (see (22)) from the data. 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 an schematic representation). At this, 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 corresponding exemplary proximity value (say, the geographical distance between *firm 1* and *firm 3*) is weighted by 2.0 when calculating the mean proximity of *firm 1*.



Firm	Hyperlinks	Proximities	Mean Proximity
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

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

After constructing the network, we excluded 150,116 firms (21.9%) without any hyperlink connections to other firms. Firms without any links have considerably fewer employees (11.9 vs. 27.7) than firms with hyperlinks and are younger (23.0 vs. 24.8 years) as well. 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 though (see below). Overall, there are 7,076,560 hyperlink connections in our dataset.

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 (4).

The MIP survey is the German contribution to the Community Innovation Survey (CIS), which is conducted every two years in the European Union, and has been used in an array of innovation studies (27). The survey methodology and the definition of innovation follows the Oslo Manual (28) and covers firms with five or more employees from the sectors of manufacturing and business-oriented services. In the survey, firms are asked whether they introduced new or significantly improved products or services (*product innovations*) during the three years prior to the survey, as well as 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 is able to predict the product innovator probability of any out-of-sample firm with a website. (4) have shown that this approach can be used to generate reliable firm-level innovation indicators even in industrial sectors and size groups that are not covered in the training data (i.e. in the MIP survey). This web-based indicator is available for all 534,757 firms in our dataset.

Table 1 presents key descriptive statistics for both innovation datasets (i.e. the MIP *survey dataset* and the deep learning based *web 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 (23)). Even though the web dataset is closer to the overall German firm population, the results of (22) showed that it is not unbiased. Larger and older firms from certain sectors are more likely to have a website and thus are over-represented in the web dataset. Firms in the survey dataset are located in more densely populated areas on average. All these differences are statistically significant according to a t-test. The number of hyperlinks per firms, on the other hand, are not significantly different, but the distribution is extremly skewed especially for the *web dataset*. As a consequence, we use logs of this variable for the further analysis.

We report both the original continuous (C in Table 1) web-based innovation indicator and a binary (B) recast to make it comparable to the binary MIP survey indicator. The mean product innovator probability in the web dataset is 25%. Casted to a binary variable using a classification threshold of 0.4 (see (4)) 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 (28)) while the web dataset is closer to the overall firm population, these values are credible (see also (4) for details).

4. Methodology

In this section, we outline how we operationalize the network position of each individual firm. Geographical, cognitive, and organizational proximity to each firm's link partners reflect the distances between firms with values of 0.0 indicating closest proximity and values of 1.0 indicating farthest distance. We

Table 1. Firm characteristics.

Variable	Mean	Median	Min	Max	Filled
		Survey dataset (n=2,463)			
Link count	11.36	5	1	992	1.00
Employees	81.97	39	1	5,060	0.80
Age	42.85	28.99	2.95	908	0.99
Firm density	879.50	79	0	3,879	1.00
Surveyed inno.	0.25	0	0	1	1.00
Pred. inno. (C)	0.30	0.23	0.37	0.93	1.00
Pred. inno. (B)	0.24	0	0	1	1.00
		Web dataset (n=543,825)			
Link count	13.01	4	1	168,961	1.00
Employees	27.65	6	1	244,038	0.52
Age	24.79	17.03	0.91	1019	0.94
Firm density	176.80	53	0	3,930	1.00
Surveyed inno.	-	-	-	-	-
Pred. inno. (C)	0.25	0.20	0.03	0.93	1.00
Pred. inno. (B)	0.16	0	0	1	1.00

also create firm-level measures that grasp the innovativeness of hyperlinked partners and the overall number of partners a firm is hyperlinked to. For all these measures we calculate the mean as it was outlined in Figure 1. In an earlier version of this paper, we also calculated standard deviations to capture the heterogeneity of each individual firm's network but found that a simple hyperlink count per firm sufficiently predicts for network heterogeneity.

Link count and mean partner innovation. *Link count* is a simple 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* measure in social network analysis.

The *mean partner innovation* is a simple measure that 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 Data section) of the hyperlinked partners of a firm.

Geographical proximity. We measure geographical proximity by calculating the euclidean distance between firms that are hyperlinked. For each firm, we then calculate the mean euclidean distance to its partners. We normalize the resulting distances to values between 0.0 (0.0 meters) and 1.0 (840,858 meters, the maximum value in our dataset) to make it easier to compare geographical proximity with the other two dimensions of proximity, which naturally range from 0.0 to 1.0.

Cognitive proximity. The cognitive proximity between hyperlinked firms is operationalized by calculating the similarity between their website texts. We know that firms use their websites to present themselves, their products and services. These information are usually codified as text and can be extracted and analyzed to assess a firms' products, services, credibility, achievements, key personnel decisions, and strategies (29). In its entirety, website texts are a description of a firm's knowledge base and we use it to calculate the cognitive proximities 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 (see e.g. (30)). The tf-idf algorithm transfers 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*). 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. Words that do not appear in a given document are represented by a 0 value.

Specifically, in a first step (the tf step) the number of appearances per word in a single document are counted. In a 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 is provided by a specific word by means of how frequently a word appears in the overall document collection. The intuition is that very frequent words that appear in a lot of documents, should be given less weight compared to less frequent words, as infrequent words are more useful 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 (see e.g. (30)), an approach widely adopted in natural language processing studies (see e.g. (31–33)). For the sake of consistency, we transform the calculated cosine similarities to cosine distances, which range from 0.0 (identical texts) to 1.0 (maximal dissimilar texts). Again, we then calculate the mean of the cognitive distances between a firm and its hyperlinked partners.

Organizational proximity. We operationalize organizational proximity as a binary variable by classifying the nature of each relation between hyperlinked firms as one of the following two classes:

- **Non-business relation**: Non-business relations are relations between firms that are not directly related to making business with each other and are of non-monetary nature. Such relations primarily include the membership in (industrial) associations or chambers of commerce, and references to regulatory or legal bodies (e.g. commercial courts, commercial registries). Hyperlinks to purely informative web contents 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.

- **Business relation**: This class includes all hyperlinks between firms that do or did business together. Oftentimes, firms include hyperlinks to the websites of other companies to present them as testimonials or because they have an ongoing business relation (e.g. web hosting, web design, web mail providers, certification services). If a firm hyperlinks to its own social media profiles, the company that operates the social media platform is a business partner of that firm as well (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.

In terms of the degree of organizational proximity, 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. For this classification, we again use the firms' website texts and relate them in the tf-idf vector space (see cognitive proximity 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 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 few of them being hyperlinks between firm of the same corporate group. *Non-business relations* on the other hand are of information only and legal/regulatory nature to about equal shares.

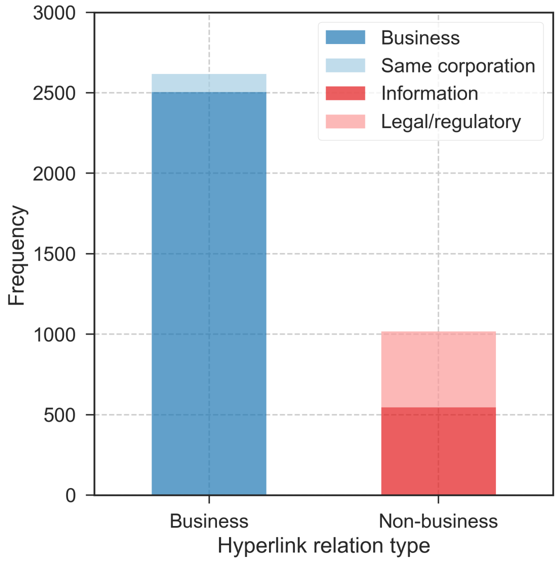


Fig. 2. Organizational proximity classes in training dataset. Manually labelled training dataset of hyperlinked firm pairs.

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 for a basic 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 which 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 performance of the model.

Table 2. Classification report for organizational proximity type 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 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 a very good 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 organizational distance) to 1.0 (high probability of non-business relation; large organizational distance). We summarized each firm's network by calculating the mean organizational distance over all its hyperlink connections.

5. Results

Figure 3 maps the Digital Layer of Germany which we created according to the procedure described in the previous section. The top panel of Figure 3 shows the distribution of product innovator firms in Germany (left) and Berlin (right) where the coloring of each cell gives the mean innovation probability for the companies contained 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 3 were created using a graph bundling method based on kernel density estimation (34). Unsurprisingly, the density of hyperlink connections between any two areas seems to be highly dependent on population. However, Figure 3 is not intended to be of high analytical value but rather to give an overview of the dataset and its granularity.

Figure 4 shows kernel density estimations for all three types of firm-level proximity as well as for link count, mean partner innovation, and local firm density. The distribution of the normalized mean geographic proximity has a mean and a median of 0.28 (235 km) and follows a normal distribution with an over-proportional accumulation of observations at mean distance 0.0 (i.e. companies that maintain hyperlinks to other companies located in the same street). Mean cognitive distance and organizational distance follow a similar normal-like distribution with higher means (0.74 and 0.75) and medians (0.75 and 0.75). Considering mean cognitive distance, an over-proportional frequency of 0.0 observations can be seen (i.e. firms that share identical texts with their hyperlink partners). In the case of mean organizational distance, on the other hand, a high frequency of 1.0 values can be seen (i.e. a lot of of firms have partner networks that consist of only non-business relations). In table 1 we already saw that the distribution of link count is highly skewed. 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 com-

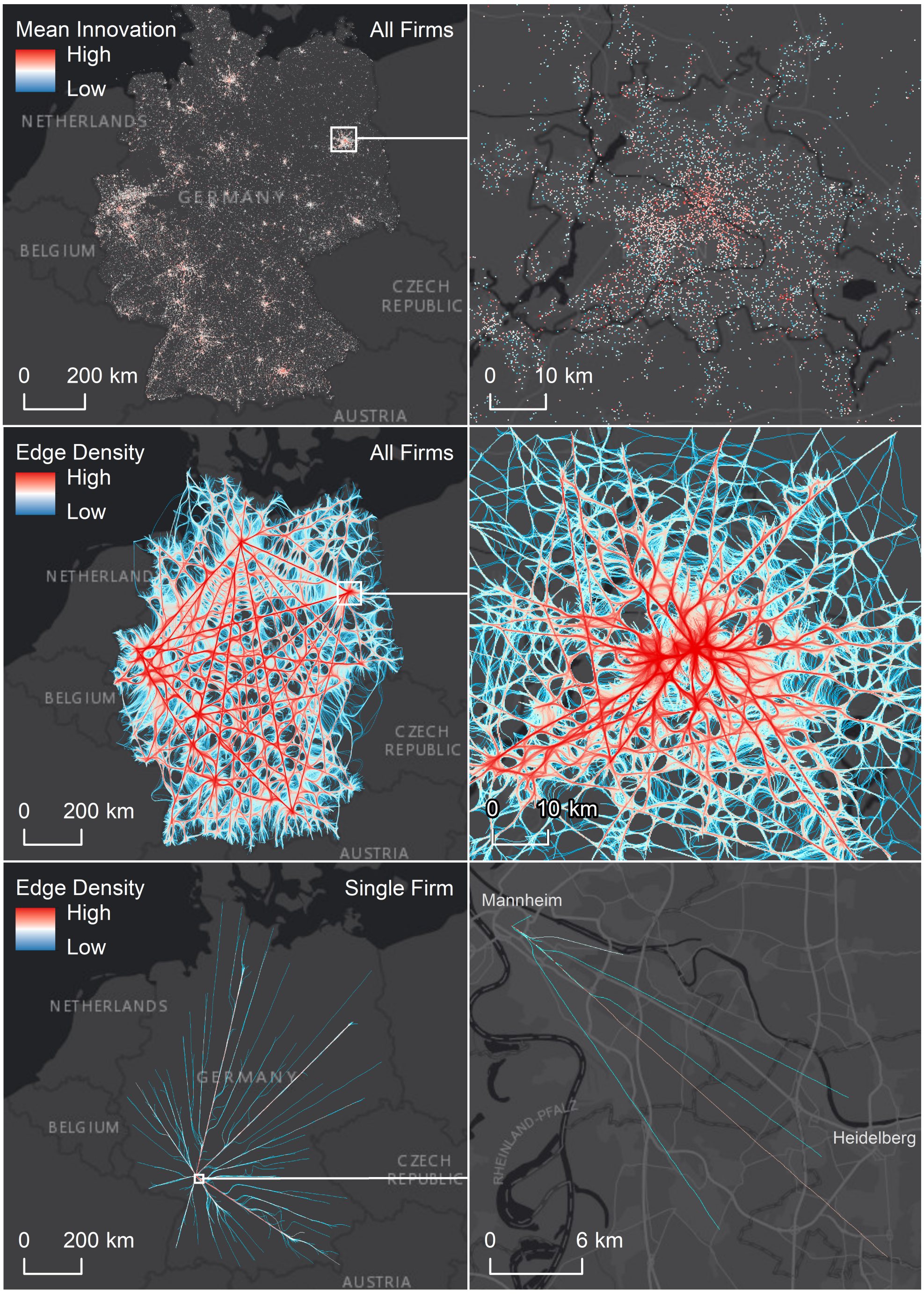


Fig. 3. The Digital Layer of Germany. Top row: Mean product innovator probability for Germany (left) and Berlin (right). Middle row: Hyperlink connections between firms in Germany (left) and Berlin (right). Bottom row: Hyperlink connections of a single firm observation in Germany (left) and the Rhine-Neckar region (right).

pany from the Silicon Valley). Mean partner innovation is again somewhat normal distributed with a mean of 0.36 and a median of 0.34. The distribution of the local firm density variable is very skewed again. On average, firms in our dataset have 176.8 other firms within one kilometer of their geographic location. The median is at 53 and the maximum value is 3,930 (downtown Hamburg).

Figure 5 shows the correlation table for all variables except for the *sector* variable which is categorical. The high correlation between the size of a company (*employees*) and its *age* is well known. However, there is also a strong positive correlation between firm size and the number of hyperlinked partners (*link count*) that a firm has. The *innovation* of firms shows a strong positive correlation to their hyperlinked partners' innovation (*mean partner innovation*). Having many partners (*link count*) is strongly negative correlated to *mean cognitive distance* (i.e. firms with many partners usually have similar partners). We also see a strong positive correlation between *mean geographic*

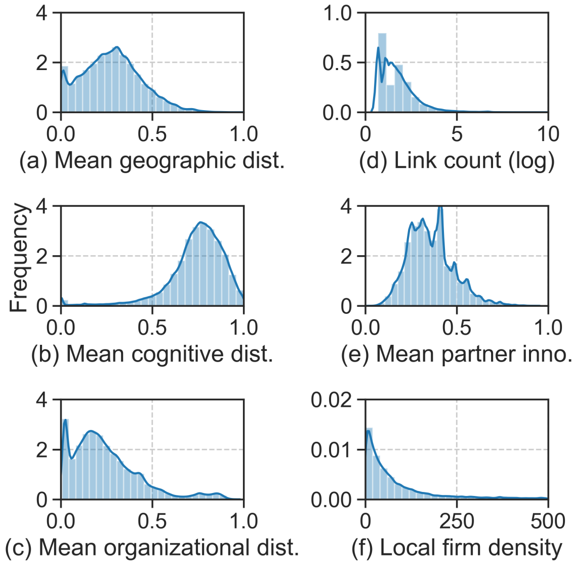
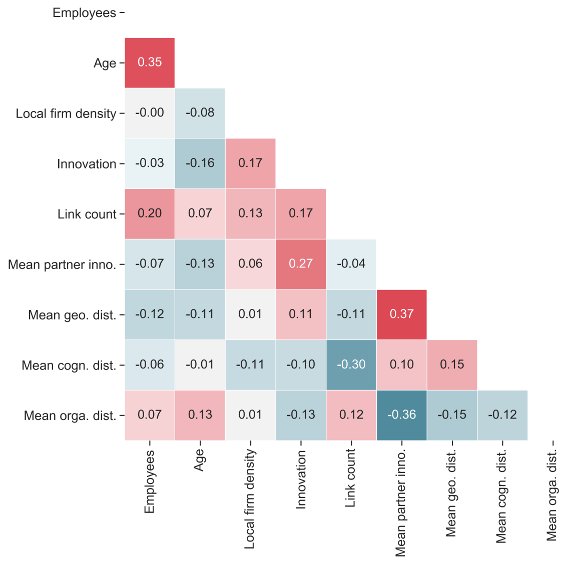


Fig. 4. Kernel density estimations for variables of interest. Geographical (a), cognitive (b), and organizational (c) distances. Link count (d), mean partner innovation (e), local firm density (f).

distance and *mean partner innovation* (i.e. firms with innovative partners maintain long-distance relationships). The strong negative correlation between *mean partner innovation* and *mean organizational distance* indicates that firms with innovative partners usually maintain stronger organizational ties.



	Employees	Age	Local firm density	Innovation	Link count	Mean partner inno.	Mean geo. dist.	Mean cogn. dist.	Mean orga. dist.
Employees									
Age	0.35								
Local firm density	-0.00	-0.08							
Innovation	-0.03	-0.16	0.17						
Link count	0.20	0.07	0.13	0.17					
Mean partner inno.	-0.07	-0.13	0.06	0.27	-0.04				
Mean geo. dist.	-0.12	-0.11	0.01	0.11	-0.11	0.37			
Mean cogn. dist.	-0.06	-0.01	-0.11	-0.10	-0.30	0.10	0.15		
Mean orga. dist.	0.07	0.13	0.01	-0.13	0.12	-0.36	-0.15	-0.12	

Fig. 5. Correlation table. Correlation table with Spearman's correlation coefficients.

Figure 6 shows scatterplots and fitted regression lines of second order between innovation and our main variables of interest. We also tested regressions of third order which yielded only slightly different results. Both the number of partners of a firm (*link count*) and the mean innovation probability of these partners (*mean partner innovation*) show a strong positive and linear relation to the firm's own innovation probability. The relations between a firm's innovation probability and the mean cognitive and organizational distance to its hyperlink partners are both negative but less distinct. The mean geographical distance to a firm's partners as well as the local firm density show inverse-U shaped relationships to the firm's innovation probability.

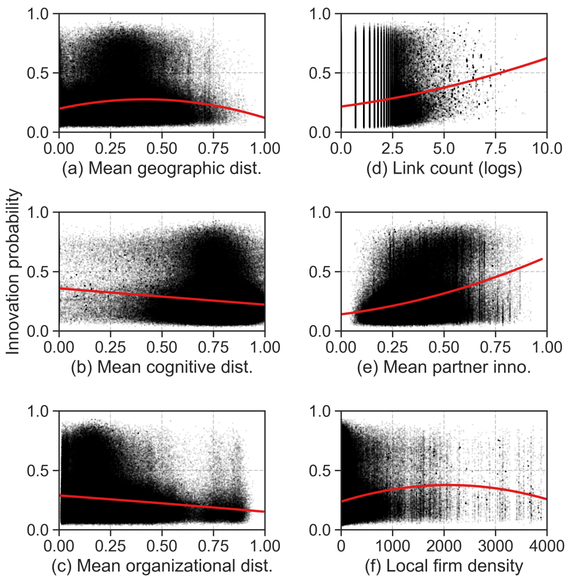


Fig. 6. Scatter plots for firm-level predicted innovation probability and variables of interest. Scatter plots and fitted regression lines of second order for geographical (a), cognitive (b), and organizational (c) distances, link count (d), mean partner innovation (e), and local firm density (f).

In Table 3, we present the results of an ordinary least square (OLS) regression with firm-level innovation as the dependent variable, control variables, and our variables of interest. The results are presented for both the innovation indicators from our *web dataset* (513,026 observations) and the *survey dataset* (2,405 observations) as a robustness check. Additionally, we also report the results of a fourth regression where we used the firms' statuses as patent-holders (1) or non-patent-holders (0). Concerning the web dataset, we run the regression for both the original continuous *product innovator probability* indicator and a binary recast (classification threshold 0.4). We use dummy variables to control for firms' sectors (*mechanical engineering* as baseline sector), size (number of employees; *missing* category as baseline), and age (*missing* category as baseline). We further control for the local firm density at the location of each firm.

6. Discussion

Our analysis revealed that innovative and non-innovative firms differ in terms of their network positions and hyperlink relations in the Digital Layer. Thereby our results are consistent as raw correlations (see Figure 5) and in our regression setting (see Table 3), which additionally controls for several firm characteristics. We find that innovative firms compared to

Table 3. Regression results

Variable	Web dataset (continuous y)	Web dataset (binary y)	Survey dataset (binary y)	Patent dataset (binary y)
	Constant			
Constant	0.2053***	-3.4465***	-2.6743	-5.5191***
	Firm-level controls			
Sector	Yes	Yes	Yes	Yes
Size	Yes	Yes	Yes	Yes
Age	Yes	Yes	Yes	Yes
Firm density (in 100)	0.0072***	0.0099***	0.0068	0.0021
Firm density (in 100) sq.	-0.0001***	-0.0002***	-0.0005**	-0.0002
	Hyperlink partners			
Link count (log)	0.0270***	0.0404***	0.0294***	0.0435***
Mean partner inno.	0.3036***	0.4602***	0.3803***	0.1745***
	Proximity			
Mean geo. distance	0.2404***	0.2688***	-0.1916	0.2455***
Mean geo. distance sq.	0.0260***	-0.0490**	-0.0966	-0.2399**
Mean cogn. distance	-0.1972***	-0.2045***	-0.0953	0.0284
Mean cogn. distance sq.	0.0733***	-0.0084	0.0398	0.0975*
Mean orga. distance	-0.4267***	-0.8022***	0.2706	0.0360
Mean orga. distance sq.	0.1151***	0.0994***	-0.5607	0.0652
	Proximity interactions			
Geo. dist. * orga. dist.	-0.0863***	-0.0377*	0.5660	0.0962
Geo. dist. * cogn. dist.	-0.2965***	-0.2566***	0.2122	-0.2845***
Cogn. dist. * orga. dist.	0.4326***	0.8583***	-0.1679	-0.1782
	Model statistics			
Model type	Robust OLS	Robust logit (average marginal effects)		
Observations	513,026	513,026	2,384	29,772
(Pseudo) R-squared	0.32	0.24	0.25	0.24
F-test/Wald chi2	3,187***	73,299***	379***	4,225***

non-innovative firms:

1. Have more hyperlinked partners.
2. Have partners that are more innovative.
3. Use geographic proximity to overcome cognitive distance to hyperlinked partners or use cognitive proximity to overcome geographic distance to their partners.

These findings are consistent for all of our used innovation datasets that we included as a robustness check. Finding *3* is consistent for the web dataset and the patent dataset but not for the survey dataset. Due to the comparatively small number of observations in the survey dataset, we were not able to identify statistically significant coefficients for our proximity measures in this dataset.

Link count. Previous studies like (13) and (11) found a positive relationship between the network centrality of firms and their innovation performance. Network centrality is equivalent to the number of hyperlink relations of each firm (*degree centrality*) in our study setup. In the scatter plots in Figure 6, we identified a strong positive and linear relationship between a firm's innovation probability and the number of hyperlinked partners. This positive relationship holds true when controlling for firm characteristics and other explanatory variables (see Table 3) and is consistent for all four datasets.

Mean partner innovation. All our results reveal a strong and positive relationship between a firm's innovation status and the mean product innovator probabilities of its hyperlinked partners, indicating that innovative firms are linked to other innovative firms. This is very much in line with the concept of *homophily* (14, 15), meaning that actors connect to other actors that are similar to them.

Firm density. (11) suggested that locally embedded firms have higher survival chances, but that the positive effect of embeddedness can reach a turning point, after which it reverses into a negative effect. If we assume that local firm density is a valid proxy for local firm embeddedness, our results in Figure 6 confirm the findings of (11) on a much larger scale. The existence of an optimal level of firm density is also revealed in the regression results for our web datasets. As the survey dataset is governed by a different sampling procedure and has very different descriptive statistics in terms of firm density (see Data section), we found no significant relationship between firm density and innovation for the survey dataset. On the basis of (16, 17)'s concept of “local buzz and global pipelines”, we expected that successful firms are able to catch local knowledge flows (high local firm density) but maintain global pipelines (high mean geographical distance) to other innovative firms.

Mean geographical distance. Looking at the scatter plots in 6, we find that the mean geographical distance to hyperlinked partners has an optimum in its relation to a firm's innovation probability. As (35) explain, an optimum does not indicate that there is an optimal geographical distance but rather that a balanced level of local and non-local linkages to other companies generates an average distance that is most conducive to innovation. Concerning the regression results, we can confirm this for the patent dataset only, while we find a monotonically positive relationship between mean geographical distance and firm innovation in the web dataset.

Mean organizational distance. The raw correlations in Table 5 show a negative relation between mean organizational distance and innovation (i.e. innovative firms tend to form business instead of non-business relations). This negative relationship is also revealed in our regression results for both web datasets, where an increase in a firm's innovation probability is associated with a decline of its mean organizational distance in the variable range from 0.0 to 1.0. We assume business relationships to be closer than non-business relationships, because they are generally more formal and reoccurring. In this sense, it appears reasonable that knowledge flows and learning are more effective among organizationally close firms and go along with a higher innovativeness in the focal firm.

Mean cognitive distance. Both the raw correlations (see 6) and the regression results for our two web datasets reveal a negative relationship between cognitive distance and innovation within the value range of our dependent variable (0.0 to 1.0). This indicates that innovative companies connect to other companies that have a similar knowledge base (i.e. small *cognitive distance*). These findings are in in line with theory of (18) who argued that firms innovate in areas close to their own knowledge base. However, our measure for cognitive proximity has to be understood as a one-dimensional mapping of a high-dimensional process. There may be companies with quite 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 measure for cognitive proximity. So our results might indicate that innovative firms and their partners share similar target markets rather that they are from the same sector.

We also found a negative relationship between innovation and an interaction term of cognitive and geographical distance for both our web datasets and the patent dataset. This may indicate that cooperation with cognitively distant companies can be successful (i.e. relate positively to firm innovation) when such partners are geographically close. It seems reasonable that geographic proximity helps to bridge knowledge gaps between dissimilar companies, for example by allowing for frequent face-to-face contact and the communication of tacit knowledge. Similarly, large geographical distances may not be hampering knowledge flows between partners if they share a common knowledge base which eases mutual understanding.

7. Conclusion

The Digital Layer. The aim of this study was to introduce a new approach to generate a web-based dataset of interrelated and textually described firms, the so-called *Digital Layer*. We constructed this Digital Layer by web mining the content of over half a million websites of German firms, resulting in a geolocated network with over seven million hyperlink relations. Making use of text-based machine learning models, we were able to operationalize proximity concepts that were difficult to analyze in large-scale empirical studies using other data sources. In a second step, we were able to empirically assess the relationship of these proximity measures and innovation in firms. For this, we used three different firm-level innovation indicators: a traditional indicator from the questionnaire-based German Community Innovation Survey (CIS), a novel indicator generated from deep learning of website texts (4), and firm-level patent statistics. Our results showed that the Digital Layer is suitable for conducting large-scale analyses of firm networks that are not constrained to specific sectors, regions, or geographical levels of analysis.

Proximity and innovation. Our case study revealed that innovative firms are differently connected within the Digital Layer compared to non-innovative firms. We were able to confirm the results of previous studies, showing that innovative firms have more (hyperlinked) partners and that their partners are on average more innovative compared to the partners of non-innovative firms. Analogous to the theory of "local buzz and global pipelines" (16, 17), we found that innovative firms are located in high density areas and still maintain relations to firms that are geographically farther away. We were able to operationalize meaningful and convenient measures of geographical, organizational, and cognitive proximity from the Digital Layer. Our results indicate that close relationships are not necessarily related to higher firm innovativeness but that it rather depends "on the optimal level of proximity between agents" (8). We also found that the relation between innovation and proximity may be indeed rather complex and that different dimensions of proximity interact with each other.

Future research. We believe that the Digital Layer approach bears great potential for the empirical analysis of firm networks. As of now, we only have gathered data for one year, but we plan to reconstruct the hyperlink networks of previous years on the basis of web archive data and also to collect data in future years by continuing to gather web data using our presented approach. Such time series data would allow researchers to investigate innovation dynamics such as firm-to-firm knowledge spillovers and the diffusion of technology between firms, industrial sectors, and regions. The high level of granularity of the Digital Layer also allows for further analyses of microgeographical intraurban firm networks as well as the analysis of networks of cities. We also expect that the depth of the Digital Layer allows for many more studies in other economic or social science settings. Moreover, the Digital Layer can add meaningful insights to the research on the multilayered structure of corporate networks (36).

1. Schumpeter J (1942) *Capitalism, Socialism and Democracy.* (Harper & Brothers).
2. van Egeraat C, Kogler DF (2013) Global and regional dynamics in knowledge flows and innovation networks. *European Planning Studies* 21(9):1317–1322.
3. Boschma RA (2005) Proximity and innovation: A critical assessment. *Regional Studies* 39(1):61–74.
4. Kinne J, Lenz D (2019) Predicting Innovative Firms Using Web Mining and Deep Learning.
5. Castells M (1996) *The rise of the network society.* (Blackwell Publishers Cambridge, MA).
6. Hidalgo C (2015) *Why information grows: The evolution of order, from atoms to economies.* (Basic Books).
7. Ter Wal AL, Boschma RA (2009) Applying social network analysis in economic geography: Framing some key analytic issues. *The Annals of Regional Science* 43(3):739–756.
8. Boschma R, Frenken K (2009) The spatial evolution of innovation networks: A proximity perspective.
9. Giuliani E (2005) The structure of cluster knowledge networks: uneven and selective, not pervasive and collective in *DRUID Tenth Anniversary Summer Conference.* pp. 27–29.
10. Park HW (2003) Hyperlink network analysis: A new method for the study of social structure on the web. *Connections* 25:49–61.
11. Uzzi B (1996) The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American sociological review* pp. 674–698.
12. Gay B, Dousset B (2005) Innovation and network structural dynamics: Study of the alliance network of a major sector of the biotechnology industry. *Research policy* 34(10):1457–1475.
13. 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.
14. Skvoretz J (1991) Theoretical and methodological models of networks and relations. *Social networks* 13(3):275–300.
15. Powell WW, White DR, Koput KW, Owen-Smith J (2005) Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American journal of sociology* 110(4):1132–1205.
16. 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.
17. Asheim BT, Isaksen A (2002) Regional innovation systems: the integration of local 'sticky' and global 'ubiquitous' knowledge. *The Journal of Technology Transfer* 27(1):77–86.
18. Winter SG, Nelson RR (1982) An evolutionary theory of economic change. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.*
19. 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.
20. Breschi S, Lissoni F (2006) Mobility of inventors and the geography of knowledge spillovers: new evidence on us data. *KITeS Working Papers* (184).
21. Bersch J, Gottschalk S, Müller B, Niefert M (2014) The Mannheim Enterprise Panel (MUP) and firm statistics for Germany.
22. Kinne J, Axenbeck J (2018) Web Mining of Firm Websites : A Framework for Web Scraping and a Pilot Study for Germany.
23. Rammer C, et al. (2019) Innovationen in der deutschen Wirtschaft, (ZEW Centre for European Economic Research, Mannheim), Technical report.
24. Zandbergen PA (2008) A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems* 32(3):214–232.
25. Rammer C, Kinne J, Blind K (2019) Knowledge Proximity and Firm Innovation: A Microgeographic Analysis for Berlin. *Urban Studies* forthcomin.
26. Kinne J (2018) ARGUS - An Automated Robot for Generic Universal Scraping.
27. Gault F, Aho E, Alkio M, Arundel A, Bloch C (2013) *Handbook of Innovation Indicators and Measurement* ed. Gault F. (Edward Elgar Publishing Ltd, Glos, UK), p. 486.
28. OECD, Eurostat (2018) *Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation.* (OECD/eurostat, Luxembourg, Paris), 4th edition, p. 258.
29. Gök A, Waterworth A, Shapira P (2015) Use of web mining in studying innovation. *Scientometrics* 102(1):653–671.
30. Manning CD, Raghavan P, Schutze H (2009) *An Introduction to Information Retrieval.* (Cambridge University Press, Cambridge, England), Online edi edition, p. 569.
31. 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.
32. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space.
33. Grentzkow M, Kelly BT, Taddy M (2017) Text as Data.
34. Hurter C, Ersoy O, Telea A (2012) Graph Bundling by Kernel Density Estimation. *Computer Graphics Forum* 31(3pt1):865–874.
35. Boschma R, Frenken K (2010) The spatial evolution of innovation networks: A proximity perspective in *The Handbook of Evolutionary Economic Geography.* (Edward Elgar Publishing).
36. de Jeude JvL, Aste T, Caldarelli G (2019) The multilayer structure of corporate networks. *New Journal of Physics* 21(2):025002.

Download ZEW Discussion Papers from our ftp server:

http://ftp.zew.de/pub/zew-docs/dp/

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

//

IMPRINT

ZEW – Leibniz-Zentrum für Europäische
Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European
Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de