Applied Geography 171 (2024) 103373

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

XAI in geographic analysis of innovation: Evaluating proximity factors in the innovation networks of Chinese technology companies through web-based data

Chenxi Liu [a], Zhenghong Peng [a], Lingbo Liu [a,b,*], Hao Wu [a], Jan Kinne [c], Meng Cai [a], Shixuan Li [a]

[a] School of Urban Design, Wuhan University, Wuhan, 430072, China
[b] Center for Geographic Analysis, Harvard University, Cambridge, MA, 02138, USA
[c] ISTARI.AI, Julius-Hatry-Straße 1, 68163, Mannheim, Germany

ARTICLE INFO

Keywords:
Multidimensional proximity
Natural language processing
XGBoost
SHapley additive exPlanations (SHAP)
Innovation network

ABSTRACT

This research explores the nonlinear interactions among multidimensional proximities, including geographical, cognitive, organizational, institutional, social, and technological aspects, and their impact on innovation within networks of over three million technology firms in China. Utilizing an innovative combination of web-based hyperlink and textual data analysis, supplemented by patent information, we delve into how these proximity dimensions influence corporate innovation capabilities. Our methodology integrates text-based deep learning techniques and employs the XGBoost model along with the SHapley Additive exPlanations (SHAP) algorithm and partial dependence plots to uncover the nuanced effects of proximity on innovation. The findings reveal that while geographical distance often correlates with larger cognitive and organizational proximities, underdeveloped regions exhibit stronger technological, institutional, and social proximities compared to their developed counterparts. The study further identifies social structure and technological differences as pivotal factors impacting collaborative innovation, with both positive and negative effects fluctuating alongside changes in proximity dimensions. Notably, we uncover that geographical proximity has a pronounced boundary effect on innovation, highlighting the critical role of spatial considerations in the digital age of innovation networks. This research contributes to the understanding of urban innovation dynamics and offers valuable insights for policymakers and urban planners aiming to foster innovation ecosystems.

1. Introduction

The acceleration urbanization and rapid advancements in digital technology have established urban innovation networks as crucial drivers of urban economic growth and social progress (Gulati & Gargiulo, 1999; B. Sun et al., 2022). These networks are undergoing rapid transformation in their interaction and cooperation patterns, primarily influenced by the internet and social media (Kinne & Axenbeck, 2020). Proximity emerges as the primary concept for delineating cooperative relationships among agents, with its various dimensions significantly impacting corporate innovation outcomes. Research has demonstrated that firms strategically positioned within these networks demonstrate the highest productivity (Giuliani & Bell, 2005). Boschma has clarified the connection between cognitive, geographical, organizational, social, and institutional proximity and corporate innovation (Boschma, 2005), suggesting that the formation and effectiveness of interactions among economic entities depend on their multidimensional distances. Therefore, comprehending and assessing the effects of multidimensional proximity are vital for analyzing urban innovation dynamics (Glaeser et al., 2022; Y. Sun et al., 2022).

Nevertheless, research on multidimensional proximity presents its several challenges, particularly in data collection and methodological approaches. A notable challenge is the reliance of traditional innovation metrics on academic and patent collaborations (Abbasiharofteh & Broekel, 2021; Simensen & Abbasiharofteh, 2022), which, due to their scope limitations, insufficiently capture the breadth of firm interactions (Bailey et al., 2018) and do not promptly reflect collaboration dynamics (Nagaoka et al., 2010; Squicciarini et al., 2013). For example, compiling

* Corresponding author. Center for Geographic Analysis, Harvard University, Cambridge, MA, 02138, USA.
E-mail addresses: yololiuchenxi@whu.edu.cn (C. Liu), pengzhenghong@whu.edu.cn (Z. Peng), lingboliu@fas.harvard.edu (L. Liu), wh79@whu.edu.cn (H. Wu), jan.kinne@istari.ai (J. Kinne), mengcai@whu.edu.cn (M. Cai), lishixuan@whu.edu.cn (S. Li).

https://doi.org/10.1016/j.apgeog.2024.103373
Received 16 January 2024; Received in revised form 31 May 2024; Accepted 6 August 2024
Available online 10 August 2024

and processing official economic statistics is time-consuming, with delays frequently surpassing a year in translating real-world collaborations into accessible data, thus compromising their relevance. In contrast, the advent of information technology has broadened the scope of business operations and collaborations, moving them beyond mere physical confines. An increasing number of firms increasingly utilize websites to display their products, services, R&D efforts, and partnership dynamics, offering fresh perspectives on investigating innovative collaborations (Abbasiharofteh et al., 2023; Krüger et al., 2020; Stoehr et al., 2020). Website information typically falls into two categories: hyperlink data within the site and textual content on the web pages. Using hyperlink data, a cooperation network can be established, and the textual information from both parties can be used to analyze the differences in the cooperative relationship across various dimensions (Gök et al., 2015; Krüger et al., 2020). This information can be collected and analyzed through web scraping and natural language processing technologies, and by combining text retrieval analysis with machine learning (Kinne and Axenbeck, 2020; Stich et al., 2023). This facilitates a comprehensive discussion of corporate innovation activities on a broader scale and with greater depth.

Another challenge is the inadequacy of traditional methods to unravel the nonlinear relationships inherent in complex networks (Lotfata et al., 2023; Ma et al., 2021). Historically, linear regression has been the predominant tool for exploration (Bi et al., 2016; Wang & Liu, 2023). Yet, the intricate dynamics of innovation networks, catalyzed by corporate collaborations, defy simple linear analysis (Kinne & Axenbeck, 2020). These networks exhibit superlinear, sublinear, and threshold effects, indicating that significant innovation impacts arise only when specific factors reach predefined levels. Consequently, linear regression's limitations necessitate alternative approaches for accurate analysis (Liu et al., 2023). Machine learning emerges as a viable solution to these nonlinear complexities (Liu, 2024). However, while deep learning neural networks excel in prediction, they fall short in clarifying the influence of individual factors on outcomes. In contrast, gradient tree-based ensemble learning algorithms, such as GBDT and XGBoost, offer dual benefits: they predict outcomes and elucidate the relationships between variables and outcomes (Lotfata et al., 2023; Ma et al., 2021), providing essential support for decision-making (Grekousis & Liu, 2019). XGBoost, in particular, an enhanced version of GBDT, improves model robustness by incorporating regularization terms and column sampling, mitigating computational speed and accuracy constraints. Despite their effectiveness, machine learning models are often criticized as 'black boxes' owing to their reliance on numerous parameters for high precision, a complexity that exceeds traditional econometric models which require fewer parameters, complicating the interpretation of outputs. To overcome this, explainable artificial intelligence (XAI), such as the SHAP (Shapley additive explanations) algorithm, was developed (Goodman & Flaxman, 2017; Gunning & Aha, 2019) to demystify the contribution of feature factors within models. Although SHAP has seen application in building environments and urban transportation (Chen et al., 2023; Li, 2022), its adoption in innovation networks remains limited. Research in this domain often utilizes simpler machine learning algorithms primarily to predict technological trends (Lee et al., 2018). The integration of ensemble learning algorithms into innovation network studies is emerging, with insufficient focus on dissecting influence mechanisms and elucidating causal relationships and interpretable outcomes.

Moreover, there is a noticeable gap in comprehending the intricacies of digital collaboration, especially concerning Chinese technology firms. Additionally, the nonlinear dynamics prevalent in these contexts remain largely unexplored. While some studies have underscored the influence of geographic and technological proximities on firms' innovative performance (Liu et al., 2020), others have emphasized that innovation agents and activities are deeply ingrained in, and influenced by, the local context (Ma & Xu, 2023), resulting in significant disparities in innovation cooperation (Li et al., 2020). These uncertainties primarily arise from the aggregated data at various levels and the undiscovered interactions among the different dimensions of proximity.

To address these gaps, we conducted an analysis of over three million Chinese corporate datasets, leveraging AI techniques for feature extraction from website texts. This led to the creation of a corporate innovation index, which was validated using patent databases. Furthermore, we integrated hyperlink and website feature data to develop an index reflecting the network centrality of corporate multidimensional proximity. Subsequently, XGBoost and XAI methods were employed to analyze network characteristics and city-specific features. This approach not only enhanced the transparency and reliability of our research but also offered fresh insights into the perception and measurement of multidimensional proximity. Consequently, it promotes a deeper understanding and more effective utilization of urban innovation networks.

The remaining sections of this paper are organized as follows. Section 2 is the literature review, reviewing existing research and summarizing the main contributions of this paper. Section 3 provides detailed information on data sources and research methods. Section 4 presents the research results, including the structural analysis of urban innovation networks and the assessment of the impact of multidimensional proximity. Section 5 delves into a thorough discussion of the research results, emphasizing their significance for theory and practice. Section 6 summarizes the main findings of the study and discusses its limitations and possible directions for future research.

2. Literature review

2.1. Evolution of innovation networks and the flow of knowledge

The rapid pace of technological advancement and the escalating complexity of technology have redefined the innovation process, transitioning it from a technology-centric, demand-driven, bidirectional coupling, and interactive integration model to one predicated on collaborative networks. This paradigm is characterized by an expanding array of inter-organizational and cross-regional connections and cooperative frameworks (Ma & Xu, 2023). Such collaboration, propelled by the exchange of innovative resources such as knowledge, technology, and information, assumes a pivotal role in reshaping regional innovation landscapes. With the increased frequency of innovative resource flows among organizations, regions, and industries, knowledge sharing and technological collaboration have intensified, transforming "local spaces" into "flow spaces" and "network spaces" (Manuel, 2009). Within these networks, companies gain access to requisite knowledge and technology, thereby amplifying creative output and augmenting overall corporate performance (Hartono & Rafik, 2022).

The concept of regional collaborative innovation, rooted in Marshall's industrial district theory from the late 19th century, underscores the advantages of industrial agglomeration in terms of external economies and knowledge sharing effects (Marshall, 2013). In the 1980s, the emergence of regions like the Third Italy drew academic attention to cooperative networks and collaborative innovation among local SMEs, sparking increased research on industrial clusters and regional innovation systems (Cooke, 1992; Porter, 1990). Termed "New Regionalism," these studies highlight the importance of geographical proximity, local networks, and face-to-face communication in sustaining innovative development and competitive advantage within industrial clusters. The theory of global production networks accentuates the role of external factors, such as cross-border production organization and technological innovation networks (Coe et al., 2008). Scholars like Bathelt critique both localized and globalized perspectives, advocating for the interaction and transformation of tacit and explicit knowledge across different spatial dimensions to support local innovation. They introduced the "global pipelines-local buzz" model, emphasizing effective interaction between global knowledge flows and local adherence (Bathelt et al., 2004). The French Proximity Dynamics School challenges the notion that geographical factors alone suffice for innovation cooperation, arguing that individual geographical relationships fail to elucidate the mechanisms of interactive learning and cooperative innovation between organizations (Shaw & Gilly, 2000). Boschma and other evolutionary economic geographers advocate for a multidimensional proximity framework, providing a relational perspective to examine these dynamics (Boschma, 2005).

2.2. Multidimensional proximity and its role in innovation cooperation

Research on network growth dynamics consistently emphasizes that geographical factors alone do not determine cooperative innovation between regions. This underscores the need for a multidimensional proximity analysis to delve into the driving mechanisms of regional innovation cooperation more thoroughly and impartially (Breschi & Lissoni, 2009). Current findings increasingly downplay the role of geographical proximity, instead highlighting the growing importance of cognitive, organizational, institutional, and cultural proximities (Duan, 2018; Donaldson & Hornbeck, 2016). The concept of "multidimensional proximity," introduced by the French Proximity Dynamics School, challenges the notion that geographical closeness is the singular influencer of innovation flows (Shaw & Gilly, 2000). Building on this, Boschma developed a framework that integrates geographical, institutional, social, organizational, and cognitive proximities (Boschma, 2005). Asheim and Isaksen emphasized the importance of social, cultural, and institutional dimensions of proximity in cross-regional collaborative innovation, arguing that organizations responsible for technology transfer, such as research institutions and higher education institutions, are crucial in the innovation process (Asheim & Isaksen, 2002). Marianne Steinmo et al. argued that geographical, cognitive, organizational, and social dimensions of proximity are important driving factors for the occurrence and sustained cooperation of innovation actors across regions. The importance of different dimensions of proximity for the construction of new collaborations changes with variations in the nature of innovation actors (Steinmo & Rasmussen, 2016). Cristian Geldes et al. applied a multidimensional proximity perspective to discuss the impact of different dimensions of proximity within Chilean agribusiness clusters on the level of cooperation among enterprises across regions and the relationships between different dimensions (Geldes et al., 2015). While extensive research has validated the influence of multidimensional proximity on urban innovation networks, the focus often remains on its linear, static, and isolated impacts. However, surpassing specific proximity thresholds can trigger a lock-in effect, creating a proximity paradox where the impact of proximity on innovation may follow a U-shaped or inverted U-shaped trajectory depending on the innovation phase (Fitjar et al., 2016; Guo et al., 2021; Zhao et al., 2023).

Analyzing proximity across various dimensions mitigates uncertainty and addresses coordination challenges, thus promoting interactive learning and innovation. This paper will explore two critical questions. The first pertains to identifying the most significant dimension of proximity for innovation within the realm of online cooperation and its distinction from traditional cooperative networks. While previous research has partially examined this aspect, it predominantly assessed the significance of proximity through linear relationships, overlooking the unique features of online collaboration. The second question examines the conditions under which proximity exerts a positive or negative influence on innovation. Common belief holds that closer relationships among innovators increase interaction, thereby enhancing learning and innovation. However, this paper contends that proximity, across different dimensions, may adversely affect innovation. A dynamic approach is essential to identify the thresholds at which proximity's impact shifts from beneficial to detrimental, aiming to more efficiently boost innovation capabilities.

2.3. Measuring proximity: traditional vs. modern approaches

The traditional definition and measurement of proximity have been anchored in the realm of offline innovation networks. Geographic proximity, gauged by the spatial closeness of cooperating entities, has been a cornerstone of corporate collaboration (Hu et al., 2021; Rammer et al., 2020). Cognitive proximity, highlighting the disparities in knowledge systems among collaborators, plays a pivotal role in the efficacy of knowledge exchange, often delineated by linguistic and knowledge source differences (O' Connor et al., 2020). Organizational proximity, denoting the intensity of each collaborative tie, is intrinsically linked to the stability of cooperation and, by extension, the progression of corporate innovation, typically assessed through the diversity of cooperative endeavors (Oerlemans & Meeus, 2005). Institutional proximity captures the variances in cultural values across businesses, commonly associated with urban or global contexts (Noonan et al., 2020). Technological proximity, reflecting the degree of technological alignment and similarity between collaborators, is vital for efficient cooperation, usually inferred from patent alignments (Y. Sun et al., 2022). Social proximity, reflecting the alignment of cultural and social backgrounds among actors in innovation, influences partner selection, frequently evidenced by previous collaborative endeavors (Y. Sun et al., 2022). Conventional statistical methodologies predominantly underpin existing proximity research (Hu et al., 2021; O' Connor et al., 2020; Oerlemans and Meeus, 2005; Rammer et al., 2020; Y. Sun et al., 2022).

However, the advent of technological shifts and the proliferation of virtual information have ushered in novel paradigms for online cooperation networks, prompting the need for refreshed proximity measurement techniques. Recent scholarly endeavors have reconceptualized organizational proximity to align with the nuances of online collaboration, markedly refining its assessment through logistic regression analyses (Krüger et al., 2020). Similarly, cognitive proximity evaluation has evolved to include textual similarity analyses (Rahimi et al., 2018), reflecting the foundational knowledge base through Part-of-Speech (POS) and Term Frequency-Inverse Document Frequency (TF-IDF) metrics, optimizing the extraction of key structural insights (Du et al., 2023). Yet, these methodologies exhibit limitations in deeply understanding content, struggling to adequately harness and discern contextual structural nuances, hence compromising accuracy. Moreover, the existing body of research predominantly centers on the European context, with a notable scarcity of inquiries into developing regions like China or investigations into the nonlinear ramifications of proximity on innovation. Additionally, the conceptual overlap among various proximity dimensions complicates differentiation. Therefore, while discussions on offline collaboration have seldom integrated all six proximity types, the expansion to online collaboration allows for a more granular and non-redundant differentiation, enriching our comprehension of innovation networks.

3. Data and methodology

3.1. Data source

This study's research data consists of two parts (Fig. 1). The first part is an enterprise dataset comprising basic information and website details of Chinese technology firms. Basic information is gathered from the Qichacha platform, encompassing company-level features such as registration details, staff information, and operational data. Website details include collaboration data derived from website hyperlinks and textual data obtained through web scraping. The collaboration data entails source and target enterprise entities, with collaboration records formed through hyperlinks between them. The textual data encompasses all information found on the company's website. The data mining and collection tool utilized is ARGUS, based on the Scrapy Python framework. The second part encompasses relevant statistical data for the cities where the enterprises are situated, including patent data in the

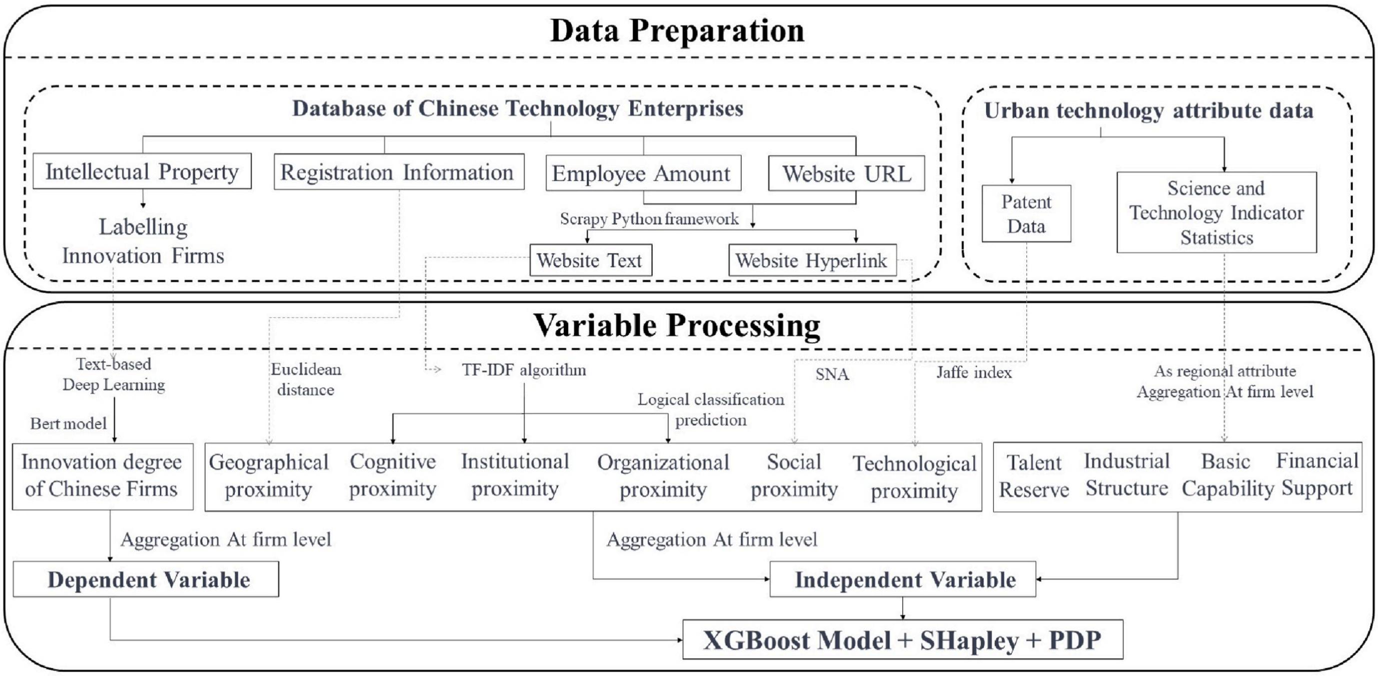

Fig. 1. Research framework.

technology field and statistical information from technology yearbooks. Patents are sourced from the patent retrieval section of the China National Intellectual Property Administration.

3.2. Variable preprocessing with NLP models

This study aims to examine the factors and mechanisms influencing the innovation probability of Chinese technology enterprises using objective data. Building on existing research and considering the characteristics of online innovation networks, six types of proximity were

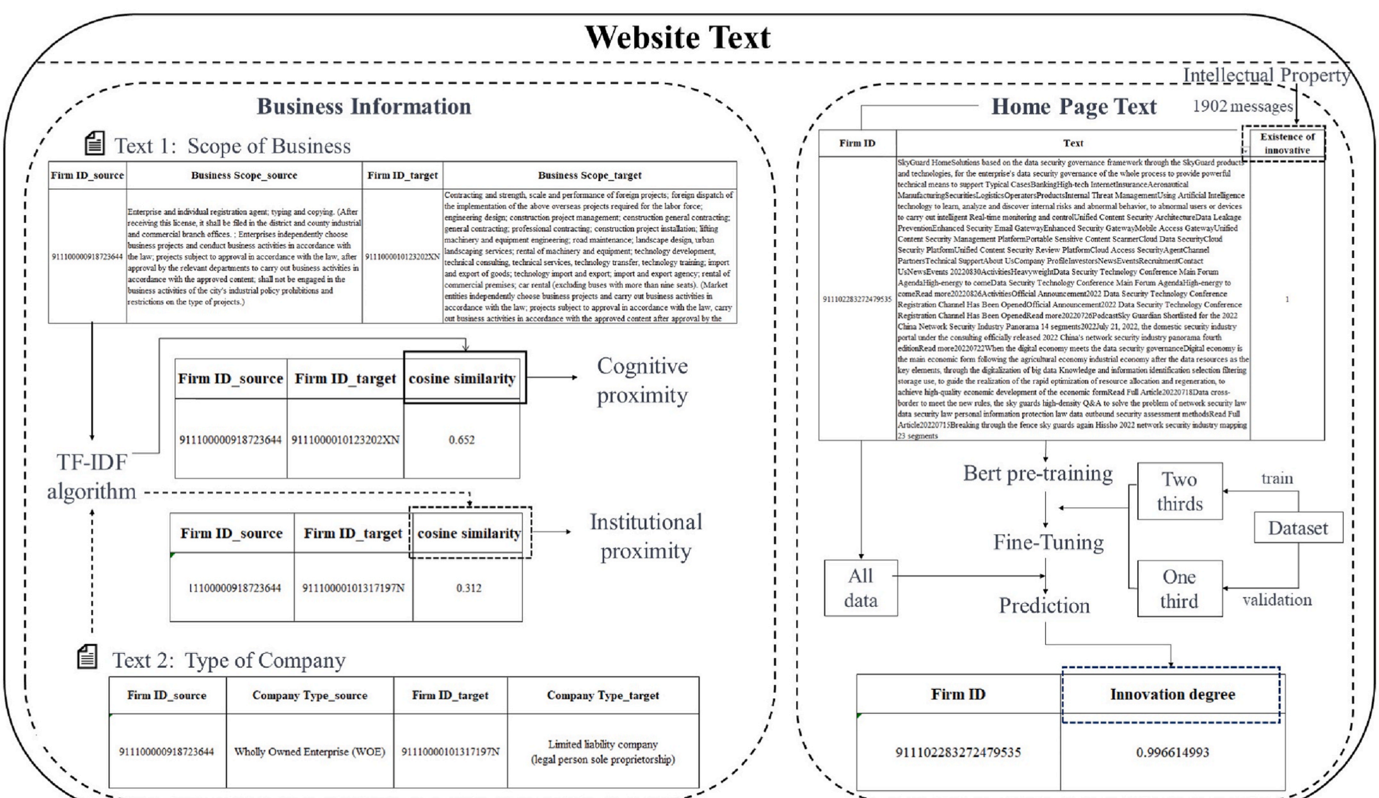

Fig. 2. Text-based deep learning process.

selected for analysis. The study analyzes their geographical patterns and explores their impact mechanisms on the probability of innovation. While past studies often relied on patent or paper collaboration data to represent innovation networks, website data, with its timeliness, granularity, and online characteristics, is more suitable for researching online innovation networks. Therefore, this paper utilizes information from enterprise websites to explore innovation networks, supplemented by collaborative patent authorization data to enhance the scientific rigor of the study. The deep learning based on text is mainly achieved by the TF-IDF algorithm and the BERT model (Fig. 2).

3.2.1. Independent variables: proximity and urban technology attribute

3.2.1.1. Cognitive proximity. Cognitive proximity refers to the similarity in knowledge, expertise, and cognitive processes among collaborators within an innovation network (O' Connor et al., 2020). In this study, cognitive proximity is assessed by measuring the similarity between the textual content of the websites of collaborating entities. Companies utilize webpage text to showcase their products and services, providing valuable insights into their offerings, credibility, achievements, key personnel decisions, and strategies. Essentially, website text serves as a representation of the company's knowledge base (Krüger et al., 2020). The cosine similarity of the vectors representing the textual content of their websites is calculated to quantify cognitive differences between collaborating companies. This approach is widely employed in natural language processing research (Gentzkow et al., 2019; Rahimi et al., 2018). The TF-IDF algorithm is utilized to transform each document into a fixed-size vector of size V, where V represents the size of the dictionary comprising all the words from the entire text corpus. Subsequently, the features in the vector are converted into weighted values, facilitating the calculation of the similarity between these vectors and the input data. During the execution of the TF-IDF algorithm, we set the minimum document frequency to 1.5% and the maximum document frequency to 65% (based on popularity filtering). Cosine similarity is then employed to determine the similarity between two different document vectors, as represented in formula 1:

$$similarity = \frac{V_{qi} \times V_{dj}}{|V_{qi}| \times |V_{dj}|} \tag{1}$$

In the equation, V_{qi} and V_{dj} represent the vectors corresponding to query documents i and j, respectively.

The final calculation result ranges from 0 (text completely different, indicating low cognitive proximity) to 1 (text completely the same, indicating high cognitive proximity).

3.2.1.2. Organizational proximity. Organizational proximity pertains to the structural and relational similarities or linkages between collaborators within an innovation network, influencing the ease of coordination and resource sharing (Oerlemans and Meeus, 2005). The relationships between collaborating companies are classified into two categories: inter-industry collaboration and intra-industry collaboration, with organizational proximity treated as a binary variable. Inter-industry collaboration tends to be more stable and intimate than intra-industry collaboration, often representing cooperation that is less substitutable. We determine whether companies are in the same industry based on the similarity of textual content related to their business scope displayed on their respective webpages (Krüger et al., 2020). The greater the similarity in text, the closer the industry affiliation.

Using this criterion, we quantify the nature of each hyperlink relationship between two companies as either 0 (indicating weak intra-industry collaboration) or 1 (indicating strong inter-industry collaboration). This facilitates binary machine learning classification tasks (Formula 1). Initially, we randomly select 1700 pairs of collaborating companies from the dataset to create the training dataset for this classification task, labeling each pair as either intra-industry or inter-industry collaboration. Subsequently, we encode the textual content of each pair of hyperlinked companies using TF-IDF and train a logistic regression classifier with the provided data. Although artificial neural networks were also considered during testing, experimental results favored the logistic regression classifier for achieving higher accuracy. Finally, two-thirds of the data are used for training the classifier, with the remaining third for testing to evaluate the model's performance. The trained model demonstrates an accuracy of 0.780, indicating robust performance and reliable results.

The predicted results range from 0 (indicating intra-industry collaboration, weak organizational relationship) to 1 (indicating inter-industry collaboration, strong organizational relationship).

3.2.1.3. Geographical proximity. Geographic proximity is defined as the extent of physical closeness among entities within an innovation network (Rammer et al., 2020). To quantify proximity, the Euclidean distance between each pair of companies is computed using their latitude and longitude coordinates. These distances are then normalized and subtracted from 1, converting "distance" into "proximity" for comparative analysis. The final results range from 0 (indicating far distance, low geographical proximity) to 1 (indicating close distance, high geographical proximity).

3.2.1.4. Technological proximity. Technical proximity refers to the similarity or compatibility in technological capabilities, structures, or domains among entities within an innovation network. Recognizing the crucial role of cities as facilitators of corporate innovation, the urban technological structure is considered the foundational environment for corporate innovation (Crowley & Jordan, 2022). The code information associated with each patent allows us to gain insights into the technological development of cities, with each code representing a distinct type of technology. Building upon prior research (Aldieri, 2013), the Jaffe index (Jaffe, 1986) was employed to characterize the technological structure of cities based on three primary categories of patent activities in the technological field of cities (according to the Chinese patent classification published by the National Intellectual Property Administration) (Yingcheng et al., 2023), thereby quantifying technological proximity. This can be represented as formula 2:

$$T_{ij} = \frac{\sum_{k=1}^{m} F_{ki} F_{kj}}{\sqrt{\sum_{k=1}^{m} F_{ki}^2 \sum_{k=1}^{m} F_{kj}^2}} \tag{2}$$

In the computation, we consider the data of invention patents, utility model patents, and design patents, resulting in a total of m = 3 technological domains. $F_{ki(j)}$ represents the number of patents produced by city i(j) in the technological domain k. The larger the Jaffe index, the more similar the technological structures of the two regions.

The final calculated results range from 0 (indicating significant structural differences and low technological proximity) to 1 (indicating small structural differences and high technological proximity).

3.2.1.5. Social proximity. Social proximity refers to the level of social relationships or trust among entities within an innovation network, indicating the degree of collaboration (Errico et al., 2022). Drawing on social network analysis principles for expressing social relationships (Otte & Rousseau, 2002; Wasserman & Galaskiewicz, 1994), social proximity is measured as the reciprocal of the shortest path between entities in the collaboration network (Opsahl et al., 2010). The expression is as follows:

$$SP_{i,j} = \ln \frac{1}{SD_{i,j}};\ SD_{i,j} = \min\left(\left(\frac{1}{CO_{i,h}}\right) + \ldots + \left(\frac{1}{CO_{h,j}}\right)\right) \tag{3}$$

In the equation, SP_{ij} represents the social proximity between institution *i* and *j*. SD_{ij} denotes the shortest path between institution *i* and *j* in the innovation collaboration network. stands for the number of collaborations between institution *i* and *j*, and *h* is the number of intermediaries (intermediating institutions) between institution *i* and *j* in the innovation collaboration network.

The final result ranges from 0 (indicating significant structural differences and low technological proximity) to 1 (suggesting minimal structural differences and high technological proximity).

3.2.1.6. Institutional proximity. Institutional proximity refers to the consistency or similarity among entities within an innovation network in terms of operating system types and governance structures (Chen et al., 2023). In China, keywords such as "limited liability," "sole proprietorship," and "collective operation" represent different institutional characteristics in company types. Within each type, the differences in these keywords can reflect variations in their operating systems. For example, the difference in keywords between "limited liability company" and "joint-stock company" implies distinctions in the internal operational mechanisms and business forms of these enterprises. Therefore, in this study, the similarity of company type keywords is used to indicate the institutional proximity between collaborating parties. Similar to the calculation method for cognitive proximity, TF-IDF recognition is applied to the textual data of company types involved in collaboration, and cosine similarity is computed to represent the level of institutional differences between the two enterprises (Krüger et al., 2020). The final result ranges from 0 (indicating significant differences in institutional systems and low institutional proximity) to 1 (suggesting minimal differences in institutional systems and high institutional proximity).

3.2.1.7. Urban technology attributes as control variables. Urban attributes serve as a fertile ground for collaborative endeavors, significantly influencing the flow of knowledge and innovation activities (Rammer et al., 2020). The breadth of the knowledge base, along with socioeconomic performance, and the cultural and social milieu, fundamentally impacts the capacity of innovators to navigate challenges within the collaboration process more effectively (Yao et al., 2020). Extensive research on innovation has explored various urban and regional features, including R&D investment, economic development levels, talent availability, industrialization, scientific and technological infrastructure, and openness (Dennis Wei et al., 2011; Wei, 2015). Consequently, it is imperative to consider not only the disparities in knowledge among innovation agents but also the variation in their urban innovation environments. In alignment with the research objectives of this study, four urban attributes have been identified and incorporated as independent variables into the model, detailed in Table 1. Due to data accessibility limitations, city-level data were used for certain indicators instead of firm-level data. While this approach has some limitations, it also allows for the capture of regional differences, providing valuable insights into the spatial variation of the studied phenomena.

Table 1
Description and processing of urban technology attribute.

Attribute	Description	Calculation method
Talent reserve	The greater the reserve of technology talent, the stronger the advantage in innovation and development.	This is expressed in terms of the number of individuals with a university education per one hundred thousand people.
Industrial structure	It can reflect the technological and economic connections between industries in that region.	This is based on the proportion of the second and third industries in the GDP.
Technological competitiveness	It can provide a visual representation of the current level of technological development and capabilities.	This is based on the number of technology institutions in the region.
Financial support	It reflects the policy inclination towards technological development.	This is based on the proportion of local financial investment in science and technology to the local general public budget expenditure.

3.2.2. Dependent variable: innovation probability values as innovation index

Understanding innovation activities is complex and profound, requiring a contextual analysis of the entire textual information on the websites. Therefore, this paper chooses to train and predict innovation probabilities using the BERT model. The BERT model constitutes a large-scale, pre-trained language model built upon the Transformer architecture, particularly well-suited for classification tasks demanding an in-depth comprehension of textual content (Pota et al., 2021; Tomihira et al., 2020). BERT's innovation lies in its utilization of the bidirectional Transformer for language modeling. Empirical evidence demonstrates that bidirectional training of language models results in a more profound contextual understanding in comparison to unidirectional models (F. Zhao et al., 2022; T. Zhao et al., 2022).

BERT's application in text classification tasks primarily involves two steps: initially, training the model on an unlabeled, extensive corpus to acquire linguistic proficiency, facilitating subsequent text classification tasks. Secondly, the model undergoes fine-tuning for text classification tasks through supervised training on labeled datasets. BERT employs two pre-training methods. The first method is the Masked Language Model, where random placeholders replace 15% of the text characters. Through rigorous training on a substantial text corpus, the model minimizes loss and adeptly predicts the specific characters concealed by these placeholders. The second method is Next Sentence Prediction, primarily concentrating on learning sentence relationships while taking into account individual words. In this approach, sentence segments are randomly substituted, and the model's objective is to predict their suitability as the following sentence after the original one.

In the small sample used for training the innovation probability prediction model, the determination of innovation activities is not obtained through supervised classification. Instead, it is based on official data obtained from the China National Intellectual Property Administration. Enterprises with new knowledge achievements within the past three years are designated as having innovation activities, labeled with "1". Those without new achievements are considered to have no innovation activities, labeled with "0". This approach avoids errors introduced by human expertise, ensuring the accuracy of the small sample and greatly enhancing the reliability of the results. Following the collection and organization of data, the small sample comprises a total of 1902 companies, of which 869 are categorized as innovative companies, while 1033 are classified as non-innovative companies. Using this labeled small sample as a foundation, we link these labels with the descriptive textual content found on the companies' websites. Subsequently, the Bert model is employed to train and predict the textual data set from the companies' websites, using a learning rate of 0.00003 and a maximum sequence length of 512. We allocate two-thirds of the data for training and reserve one-third for validation testing. The model attains a notable accuracy of 0.97 (Fig. 3), and the details are shown in Table 2. Thereby it ensures the reliability and scientific validity of the dependent variable. Ultimately, we derive the innovation probability values for all companies with websites.

3.3. Methodology

3.3.1. eXtreme gradient boosting (XGBoost)

XGBoost is a gradient boosting method that sequentially integrates decision trees using the gradient descent optimization algorithm to minimize the model's error (Chen et al., 2023). In practice, the XGBoost algorithm can be seen as an optimized version of the gradient boosting decision tree algorithm (GBDT). GBDT is commonly used to solve classification and regression problems, as it can accurately detect complex nonlinear relationships between dependent and independent variables.

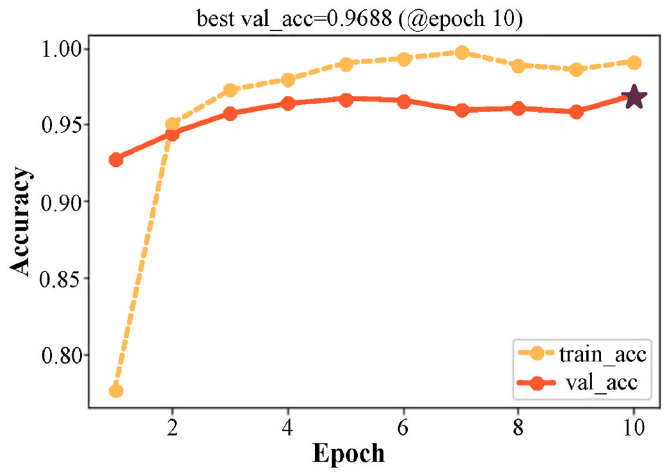

Fig. 3. Accuracy of Bert's model training and validation sets.

Table 2
Bert model results.

Epoch	Train_loss	Train_acc	lr	Val_loss	Val_acc
1	0.432227	0.776902	0.00003	0.221754	0.927155
2	0.141243	0.950487	0.00003	0.153806	0.943966
3	0.082603	0.972403	0.00003	0.140695	0.956897
4	0.060580	0.979437	0.00003	0.109454	0.963362
5	0.029251	0.989989	0.00003	0.134777	0.966595
6	0.024269	0.992965	0.00003	0.127809	0.965517
7	0.011462	0.997024	0.00003	0.156501	0.959052
8	0.034375	0.988636	0.00003	0.165017	0.960129
9	0.035929	0.985660	0.00003	0.200773	0.957974
10	0.029066	0.991071	0.00003	0.160705	0.968750

It can also compute and visualize the relative importance of each factor. XGBoost, built upon distributed computing, further significantly improves the computation speed and performance of the original model, making it the fastest and most integrated decision tree algorithm currently available. Assuming the XGBoost model consists of K CARTs, it can be represented as formula 4:

$$\widehat{y}_i = \varnothing(X_i) = \sum_{k=1}^{K} f_k(X_i), f_k \in F \tag{4}$$

Here, $\widehat{y}_i$ represents the predicted value, f_k represents the k-th tree, $f_k(X_i)$ represents the value of the i-th sample in the k-th tree, K is the total number of samples, X_i is the input value of the i-th sample, and F represents all possible K CARTs. In this study, we utilized the "XGBoost" package by calling it in R version 4.2.2 and RStudio version 4.1.3 to use this model.

3.3.2. Shapley and partial dependence plots

The partial dependence plot (PDP) (Friedman, 2001) visualizes how a single independent variable affects the dependent variable, revealing the marginal effect of the independent variable. It addresses the problem of machine learning being a black box where only predictions are possible without analysis. Shapley, derived from a concept in cooperative game theory, was originally used to fairly distribute the contribution values of players in a collective achievement (Shapley, 2016). Through extensions, this algorithm can quantify and visualize the contribution of each feature in the model. Building upon the Partial Dependence Plot, Shapley further opens up the black box of machine learning (Lundberg and Lee, 2017; Lundberg et al., 2019). The Shapley values of features are calculated using the following formula 5:

$$Shapley(X_j) = \sum_{S \subseteq N \setminus \{j\}} \frac{k!(p-k-1)!}{p!} \left(f\left(S \bigcup \{j\}\right) - f(S)\right) \tag{5}$$

In formula 5, p represents the number of features, $N \setminus \{j\}$ denotes the set of all possible feature combinations excluding feature X, j. S represents a feature set from $N \setminus \{j\}$, $f(S)$ represents the model prediction for the feature set S, $f(S \bigcup \{j\})$ represents the model prediction for the feature set S with the additional feature X. Current research results demonstrate that the integration of XGBoost and SHAP works well and can efficiently compute the SHAP values of features (Lundberg and Lee, 2017). In this study, both PDP and SHAP calculations were performed by calling the "gbm" package and the "SHAPforxgboost" library in R version 4.2.2 and RStudio version 4.1.3.

4. Result

4.1. Statistics of multidimensional proximity and innovation probability

4.1.1. Preliminary descriptive statistics

Table 4 presents the initial descriptive statistics for multidimensional proximity and innovation probability values at the enterprise level, following z-score standardization. A notably strong positive skewness in cognitive proximity (6.606) indicates a tendency among firms to seek partnerships with entities displaying considerably diverse cognitive structures. Organizational proximity (1.095), with a moderate rightward skew, suggests prevalent cross-industry collaboration, where firms form associations leveraging functional complementarities to enhance capabilities. Despite interactions occurring in digital environments, geographical proximity remains substantial in partnership formations. Lower average levels of institutional and social proximity suggest a current inclination towards forming online collaborative networks with partners exhibiting significant differences in institutional structures and engaging in more limited social interactions.

This trend may signify a pursuit of novel institutional frameworks and the cultivation of less conventional social ties to diversify innovation influences. Additionally, the right-skewed kurtosis for institutional and social proximity might indicate a pronounced inclination towards collaboration across different backgrounds, underscoring an underlying complexity not fully captured by mean levels. Finally, the negative skewness for technological proximity (−7.574) reflects a preference for collaborations with entities of similar technological standing, emphasizing the importance of compatible technical capabilities and shared knowledge bases critical for joint innovation endeavors. These findings collectively depict a nuanced spectrum of collaboration preferences, as firms strategically position themselves within the innovation landscape.

4.1.2. Spatial distribution of innovation probability

Fig. 4 illustrates the spatial distribution of innovation probability among Chinese technology firms, revealing a non-uniform distribution across the country characterized by geographic clustering. Higher innovation probabilities are clustered in the southeast, encompassing Beijing and its surrounding areas, the southeastern coastal region, Wuhan-Changsha, and the Sichuan-Chongqing areas, indicating a concentration of innovative activities in economically developed and urbanized regions. This clustering pattern may be attributed to several factors, including the presence of advanced infrastructure, the concentration of human capital, and the establishment of higher education and

Table 3
Multiple model results.

Model	RMSE(Root mean squared error)
Linear regression	0.181
Geographically Weighted Regression	0.274
eXtreme Gradient Boosting	0.102

Table 4
Firm cooperation statistics.

Variable	Mean	Std.	Skewness	Kurtosis
Independent variables				
Cognitive proximity	0.021	0.060	6.606	64.509
Organizational proximity	0.209	0.086	1.095	1.786
Geographical proximity	0.824	0.131	−0.224	−0.640
Technological proximity	0.974	0.121	−7.574	57.940
Social proximity	0.118	0.138	3.560	19.146
Institutional proximity	0.287	0.378	1.234	−0.252
Talent reserve	0.190	0.155	1.986	4.636
Industrial structure	0.731	0.188	−1.258	2.536
Technological competitiveness	0.031	0.098	6.532	53.566
Financial support	0.134	0.153	2.290	6.845
Dependent variable				
Predicted innovation possibility	0.680	0.079	0.122	−0.180

research institutions conducive to innovation. Clusters around Guangdong and Fujian may be indicative of regional innovation systems supported by policies such as Special Economic Zones (SEZs) and substantial levels of foreign direct investment (FDI). Conversely, lower innovation probabilities observed in the Yunnan-Guizhou region, Inner Mongolia, and Xinjiang may result from factors such as lower economic development, reduced investment in research and development, and potentially fewer collaborative networks stimulating innovation. These spatial disparities underscore the significance of region-specific policies aimed at leveraging local strengths and addressing constraints to promote balanced innovation capacities across regions.

4.1.3. Spatial distribution of multidimensional proximity

The analysis of multidimensional proximity, depicted in Fig. 5, unveils the intricate dynamics shaping the innovation landscape. The rhombus-shaped structure of the online collaboration network, emphasized by geographical proximity, signifies robust inter-regional connections among major economic centers – Beijing, Shanghai, Guangzhou, Chongqing, and Wuhan – pivotal nodes for innovation diffusion. The generally low cognitive proximity, with sporadic spikes between specific cities, implies potential barriers to widespread knowledge transfer. This could stem from regional specialization in particular technologies, creating cognitive gaps where firms lack shared languages and understanding beyond their respective niches. Despite the prevalent low cognitive proximity, the existence of high-value connections between peripheral and more developed regions suggests selective yet significant knowledge exchanges, possibly driven by targeted partnerships or government policies fostering inter-regional collaboration. The pattern of organizational proximity, characterized by high-value connections concentrated between developed regions and specific interior and northeastern areas, likely reflects the presence of branch offices or shared corporate cultures facilitating collaboration.

Meanwhile, technical proximity, evidenced by its high-value linkages, underscores a robust exchange of technical knowledge and skills, particularly within the main collaborative network, essential for high-tech industries. This clustering effect illustrates how similar businesses congregate and benefit from shared technological advancements. Institutional and social proximity, marked by low-value connections in developed regions but high-value connections in less developed areas, suggests that while formal institutions and social structures in developed regions may not drive innovation directly, in less developed areas, these factors could compensate for other deficiencies, such as economic or infrastructural deficits. This multifaceted exploration of proximity dimensions reveals the nuanced and interdependent nature of innovation networks, where different forms of proximity interact to shape the innovation capabilities and potential of regions across China. A comparison with innovation networks constructed using patent data indicates a largely consistent basic framework (Abbasiharofteh et al., 2023; Cantner & Meder, 2007; Cao et al., 2019; Zhao et al., 2023). However, the network nodes and connections developed in this study exhibit greater richness, highlighting specific differences that will be discussed further.

4.2. Result of XGBoost model and SHAP explainer

4.2.1. Model comparison

To highlight the suitability and effectiveness of XGBoost in this study, we also utilized linear regression and geographically weighted regression models for comparison purposes (Table 3). The analysis resulted in an RMSE value of 0.102, indicating minimal model error and ensuring strong scientific validity. As shown in Table 3, the XGBoost model demonstrated lower errors in comparison to both the linear regression and geographically weighted regression models, confirming

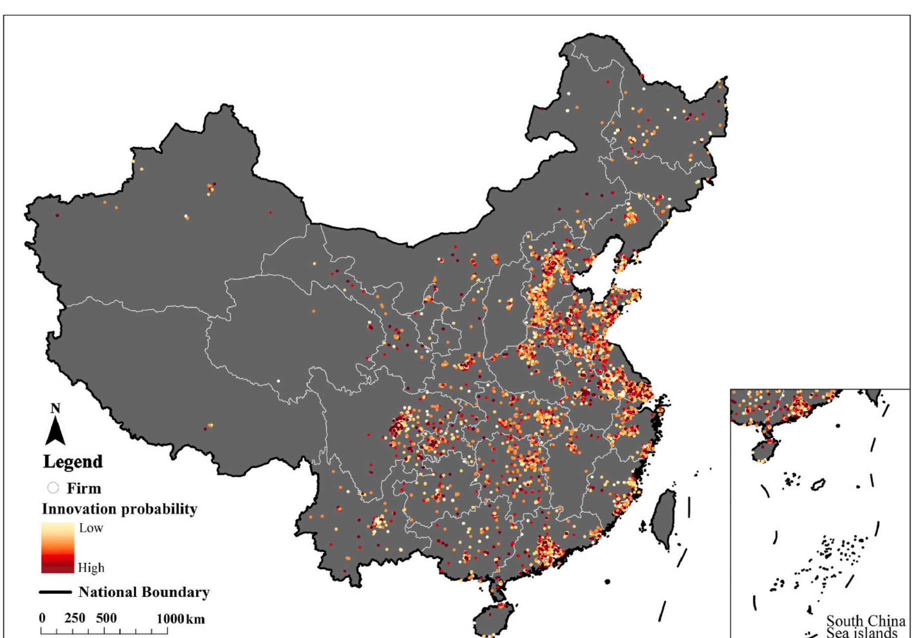

Fig. 4. Spatial statistics of innovation probability for Chinese technology enterprises.

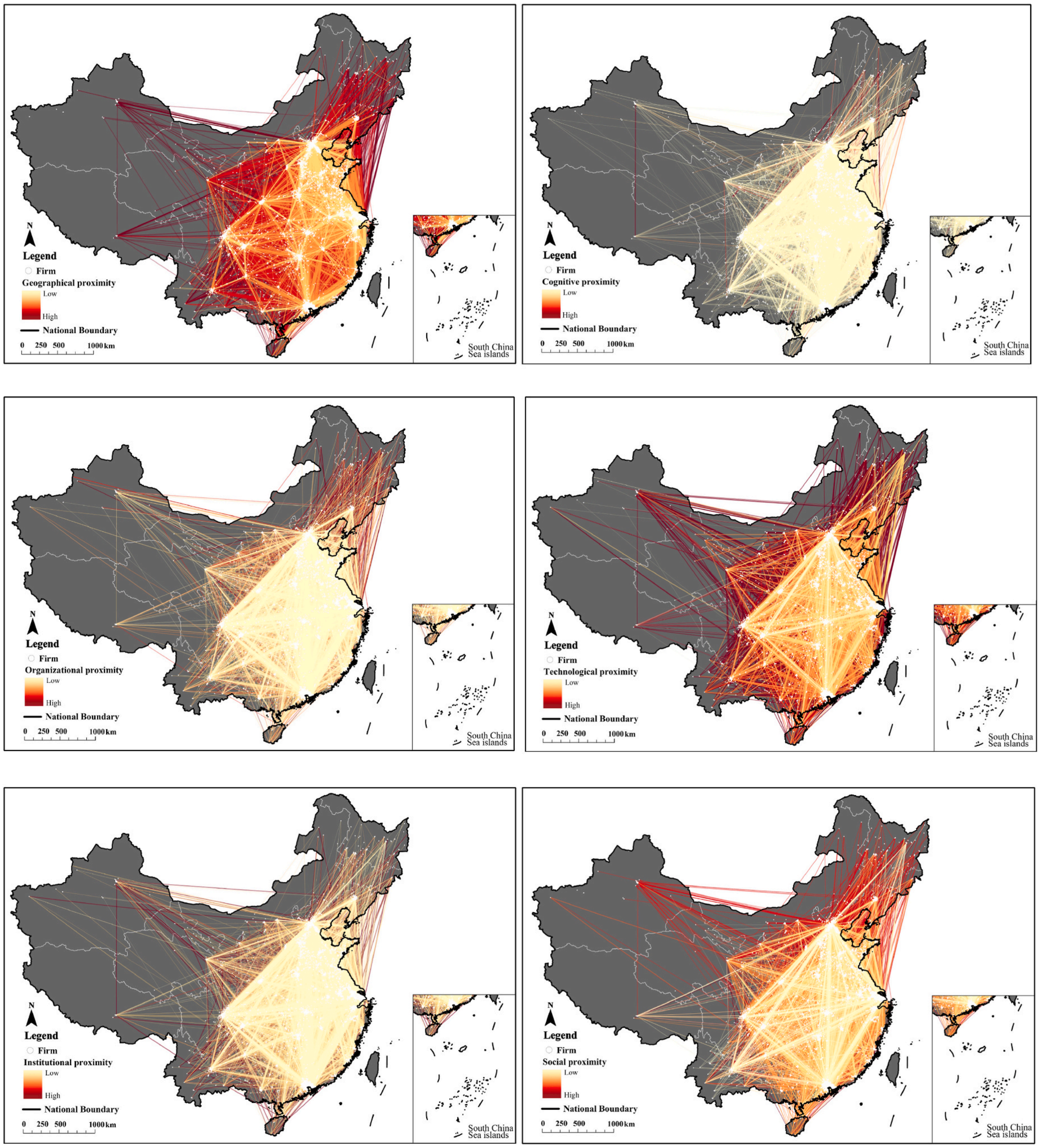

Fig. 5. Spatial distribution of enterprise proximity.

the superiority of XGBoost for our research.

4.2.2. The relative importance of factors explained by SHAP

When thoroughly evaluating the combined influence of various numerical values representing multi-class proximities on the likelihood of corporate innovation, their significance varies throughout the process. Illustrated in Fig. 6, based on SHAP, are the calculated importance results. Social proximity emerges as the most influential factor, while organizational proximity exerts the least impact. The horizontal axis represents the calculated SHAP values, with higher values indicating a more substantial contribution of the corresponding feature. Each point on the graph represents a data point from the dataset, vertically aligned to illustrate density, with a color strip below indicating the values of each point. Higher feature values are depicted in purple, and lower values in yellow. For example, a purple point on the social proximity factor indicates that a higher level of technical proximity (as shown in the color strip below) has a more significant positive impact on the probability of innovation (corresponding to larger positive values on the x-axis).

Fig. 7 elucidates the interactive process of variation in the six types of proximity values and their changes in importance based on the SHAP algorithm. For social proximity, a larger value initially exhibits an inverted U-shaped negative impact, followed by a logarithmic-linear relationship in the positive impact phase when the value exceeds 1. In the case of technical proximity, a larger value shifts the impact on the probability of innovation from negative to positive, with the degree of influence gradually increasing, demonstrating a logarithmic-linear relationship. The impact of geographic proximity changes from positive to negative, forming a logarithmic-linear relationship. In contrast, institutional proximity undergoes a gentle sine non-linear relationship between its value and importance, indicating a negative impact. As for cognitive proximity, as the value increases, it overall forms an inverted

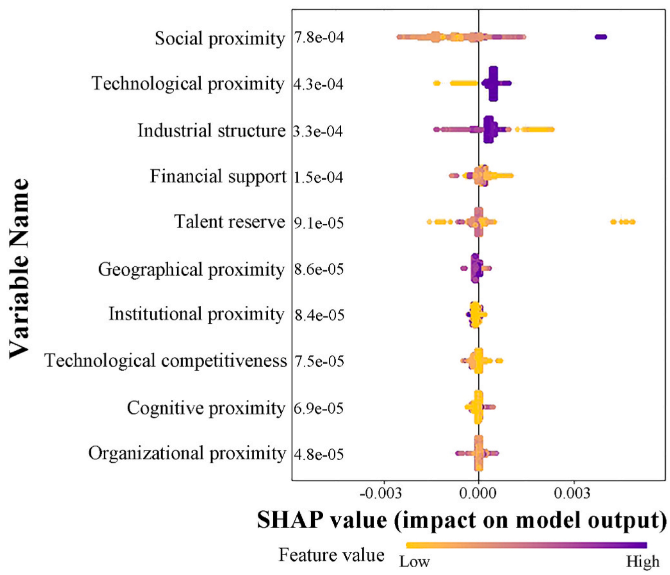

Fig. 6. SHAP summary plot for the XGBoost model.

U-shaped relationship, with a negative impact in the initial and final segments and a positive impact in the middle segment. Organizational proximity forms a very gentle positive exponential relationship.

Regarding urban attribute factors, the relationship between industrial structure and the probability of innovation approximately follows a power function, exhibiting a negative impact initially. For financial support, a larger value initially experiences a wavelike decline, transitioning to a negative impact phase, then rising to form a small peak in the wave, and finally declining again, with the peak level below the initial level. Talent reserve and technological competitiveness both exhibit very gentle logarithmic relationships.

4.2.3. The relationship between factors and innovation probabilities

For a more comprehensive understanding of how independent variables influence the dependent variable, Fig. 8 illustrates the impact process of changes in six proximity types on innovation probability values.

The impact of various proximity measures on innovation probability exhibits distinct patterns. Social proximity emerges as the most influential, initially increasing, showing fluctuations, and then sharply rising at higher levels. Technical proximity decreases at lower levels, experiences moderate growth at intermediate levels, and declines again at higher levels. Initially, geographic proximity positively correlates with innovation probability, but this relationship diminishes beyond a certain threshold. Institutional proximity demonstrates fluctuating effects, while cognitive proximity initially increases before subsequently declining. Organizational proximity undergoes minor fluctuations initially, followed by a sharp decrease at higher levels.

In the realm of urban attributes, industrial structure emerges as the primary influencer on innovation probability. As the share of secondary and tertiary industries increases, innovation probability rises, fluctuates, and generally declines after reaching a certain level. Financial support boosts innovation probability as it increases, maintaining a consistently high influence beyond a certain threshold. Talent reserve influences innovation probability through several stages: initially increasing with additional reserves, then sharply declining, followed by another rise before stabilizing at a higher level. Regarding technological competitiveness, low levels trigger a decline in innovation probability, which then rises and stabilizes as competitiveness improves. At the lowest levels of competitiveness, innovation probability reaches its peak.

5. Discussion

This research aims to unravel the intricate nonlinear dynamics between proximity dimensions within corporate webpage hyperlink networks and their impact on innovation. The findings underscore the pivotal role of these digital networks as hubs for innovation, where proximity emerges not as a static metric but as a dynamic interplay of factors that can either hinder or spur firms' innovative capacities. The nuanced insights derived from this study, illuminated by Explainable Artificial Intelligence (XAI), elucidate the complex and nonlinear relationship between virtual representations of proximity and the propensity for innovation, a perspective that is becoming increasingly relevant as the physical and digital realms intertwine more closely.

Among the proximity dimensions, technological proximity at the city level emerges as the most influential variable for the likelihood of innovation. The findings suggest that current trends drive enterprises to forge collaborations primarily with entities sharing similar technological structures (Cantner and Meder, 2007; Yu et al., 2014), particularly in underdeveloped regions, aligning with prior research on patent innovation networks (Griffith et al., 2006). This inclination is attributed to the heightened costs associated with significant disparities in research foundations, a major hindrance to innovation in these areas (Bunduchi et al., 2011; Yang & Cai, 2009). However, the study observes a noteworthy phenomenon: beyond a certain threshold, excessive technological similarity may actually impede innovation. This paradox arises due to the negative consequences of homogenization, including constraints on new technology development and limited exploration of new fields (De Noni et al., 2018). Conversely, when the level of similarity surpasses a critical threshold, profound technological congruence significantly enhances entities' ability to assimilate and leverage each other's technologies. In such cases, the benefits of efficient communication outweigh the drawbacks of homogenization, fostering and sustaining more innovative collaborations (Cantner and Meder, 2007; Jaffe, 1986). This phase holds particular relevance for nurturing innovation partnerships in underdeveloped regions.

Organizational proximity, as the second most important variable for the probability of innovation, underscores the prevalence of cross-sector collaboration where firms establish robust, functionally complementary partnerships. This observation is consistent with the results of the existing Chinese patent cooperation, the German website cooperation network (Cao et al., 2019; Krüger et al., 2020). Much like cognitive proximity, its spatial distribution aims to counteract adverse geographical effects by leveraging compensatory advantages (Mascia et al., 2017; Pouder & StJohn, 1996). Cross-industry cooperation notably fosters the exchange and synergy of organizational knowledge across diverse sectors (Markovic et al., 2020). However, as collaboration intensity escalates, the benefits of complementarity diminish, overshadowed by pronounced industry disparities that hinder the circulation of innovative resources (Gao et al., 2023). Upon surpassing a certain threshold, these obstacles predominate, significantly reducing innovation potential. Nevertheless, once equilibrium is achieved, innovation levels exceed their initial baseline, illustrating the superior impact of cross-sector collaboration on corporate innovation compared to within-industry partnerships. This underscores the importance of balanced cross-disciplinary complementarity and resource exchange in fostering innovation.

Observations on social proximity suggest a preference among enterprises for online collaborations with less interactive partners, a trend diverging sharply from traditional patent innovation networks (Errico et al., 2022; Zhang et al., 2023). In contrast, economically developed areas exhibit a stronger reliance on established social relationships for collaborative innovation, consistent with geoeconomic studies (Baru, 2012; Beeson, 2018). Analyses by SHAP and PDP validate that social proximity negatively impacts innovation probability. This challenges previous research advocating for the path-dependent nature of innovation cooperation (Eiriz et al., 2013; Heringa et al., 2014), suggesting

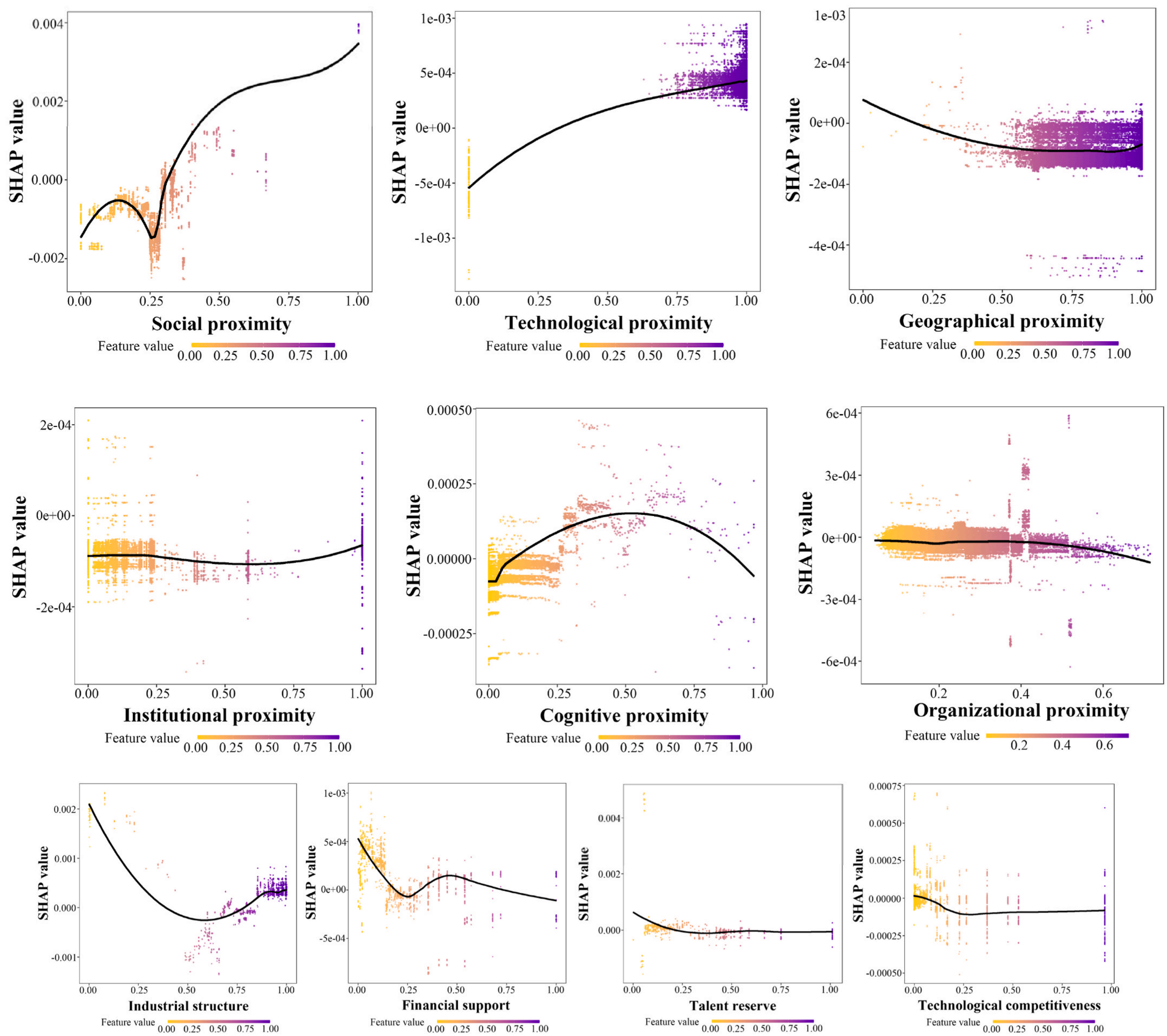

Fig. 7. SHAP plots for each factor.

instead that online collaboration with previously unaffiliated enterprises could potentially enhance innovation chances, as suitable diversification fosters the production of novel outputs. Furthermore, as proximity approaches a threshold value, the likelihood of innovation declines and remains suppressed. This implies that excessively homogeneous cultural backgrounds and social connections impede the pursuit of new innovations in online collaboration contexts, obstructing the generation of new technological discoveries.

Geographical proximity exhibits a notable boundary effect, with contemporary enterprise collaborations maintaining an elevated average level of geographical proximity, emphasizing the enduring importance of geographical costs even in online collaborations (Abbasiharofteh et al., 2023; T. Zhao, et al., 2022). This finding is consistent with those of studies conducted on patent networks (Lim & Han, 2023). In underdeveloped regions, locational disadvantages necessitate compensatory measures through enhanced forms of other proximity, aligning with additional research findings (Santamaría et al., 2021). Moreover, analyses by Shapely and PDP confirm that moderate geographical proximity enhances innovation cooperation (Abbasiharofteh et al., 2023; T. Zhao, et al., 2022). Intriguingly, our analysis reveals a reduced innovation probability within China's metropolitan radii as geographical closeness decreases, attributed to heightened homogeneity in knowledge and technology, which fosters competitive uniformity among enterprises, thereby adversely affecting innovation (Mascia et al., 2017). However, surpassing a threshold in geographical proximity significantly mitigates the adverse effects of competition, notably boosting innovation probabilities.

Additionally, findings regarding institutional proximity indicate that businesses are increasingly forming online partnerships with entities exhibiting significant systemic variances. This analysis underscores that a moderate degree of systemic disparity initially promotes complementary benefits among parties (Chen et al., n.d.) and ensures uninterrupted knowledge exchange across different organizations. Excessive institutional resemblance between entities can foster exploitative innovation, significantly hampering innovative endeavors (Zhong et al., 2023). This highlights a fundamental difference in the collaborative orientation of online innovation networks compared to the collaborative characteristics of patent and scholarly paper innovation networks (Eiriz

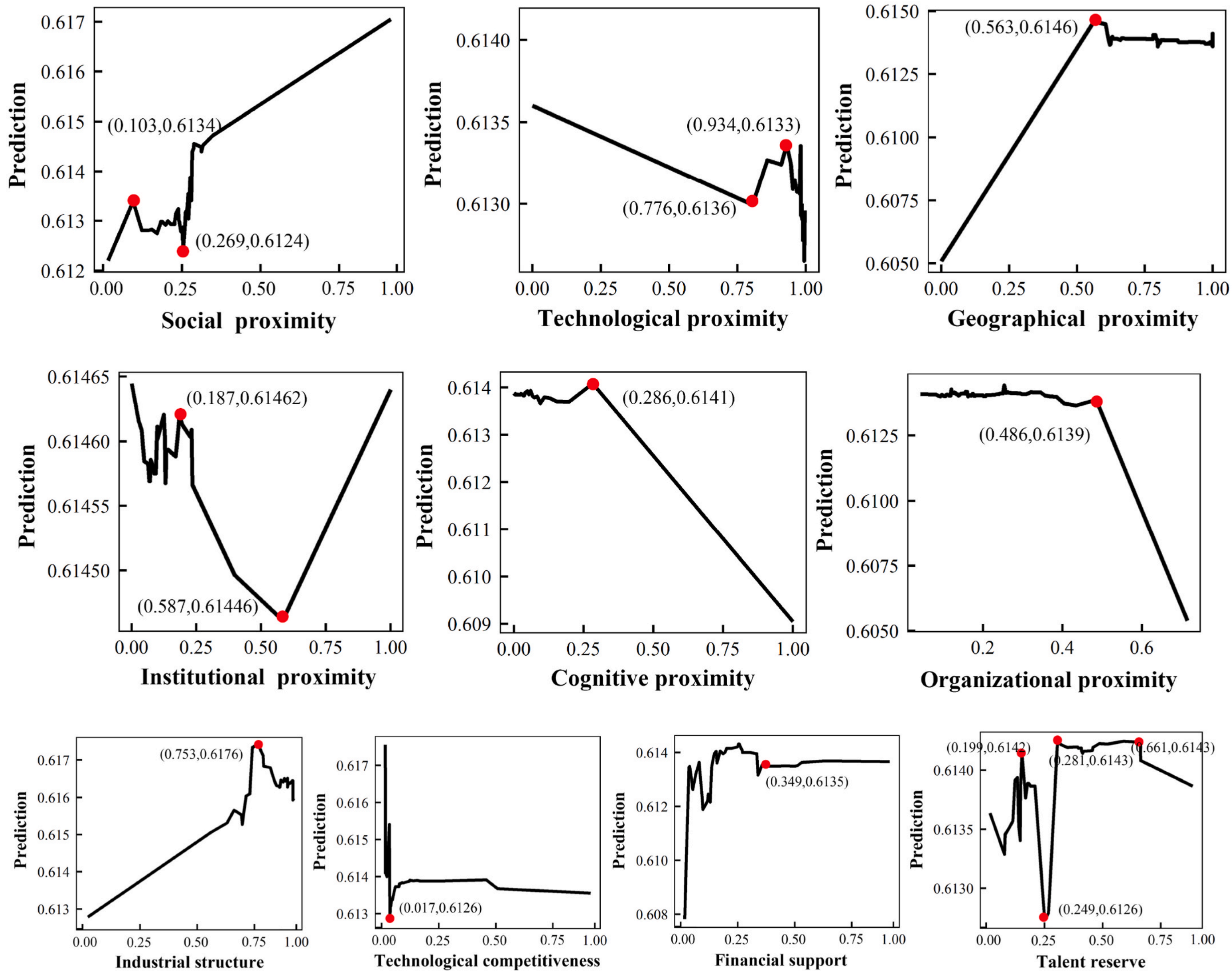

Fig. 8. PDP based on XGBoost modeling.

et al., 2013; Heringa et al., 2014).

The findings suggest that businesses currently prefer to establish collaborations with partners who exhibit significant differences in textual cognition. This contradicts not only prior studies suggesting that innovation networks based on patents tend to cooperate more with entities having high cognitive similarity (Criscuolo et al., 2018), but also findings from the analysis of German website cooperation data (Abbasiharofteh et al., 2023; Krüger et al., 2020). It is important to acknowledge that our assessment of cognitive proximity represents a simplified interpretation of a complex process. Companies from various sectors, such as cultural technology innovators and chip developers, operating in the same marketplace (e.g., AI applications in real-world scenarios), contribute to the reported statistical outcomes. Additionally, spatial analysis reveals that businesses in less developed areas increasingly rely on greater cognitive proximity to mitigate their disadvantages. This finding supports the theory proposed by Miriam Krüger, suggesting that cognitive and geographical proximities can mutually compensate (Krüger et al., 2020). Results from Shapley and PDP further validate this conclusion, albeit to a limited extent. Once cognitive similarity between partners exceeds a certain threshold, there is a significant surge in innovation potential. This is because similar knowledge frameworks enhance the efficiency of knowledge and technology exchanges among entities, consistent with previous findings (O' Connor et al., 2020).

Overall, while previous studies may have assumed a monotonous or linear impact of multidimensional proximities on innovation, the clarification offered by SHAP values in our study unveils an intricate network of subtle nonlinear relationships. These intricate dynamics become especially apparent when examining threshold effects and non-monotonic patterns, including peak responses or V-shaped correlations, providing alternative perspectives in the ongoing discussions on interpretations of multidimensional proximity. This complexity illuminates the drawbacks of linear models and aggregated data approaches that have dominated previous research, indicating the necessity for a more sophisticated analytical framework.

6. Conclusion

This study utilizes data from collaborative activities on Chinese technology firms' websites, supplemented by patent data, to meticulously construct the innovation network of enterprises. It considers both traditional and online collaboration models, enhancing the robustness of the research foundation. A text-based deep learning approach refines and optimizes the connotations of proximity and innovation probability related to informatization. Finally, the XGBoost algorithm, combined with SHAP and PDP, examines the nonlinear relationship between proximity and corporate innovation probability, reducing potential errors inherent in conventional linear regression. This methodology reveals the intricacies of machine learning models, providing profound insight into the influence of online collaboration on innovation networks. The research findings lay the groundwork for future investigations into the nonlinear implications of multidimensional proximity for innovation.

The precision of the text-based deep learning model is remarkably high. The results confirm that a database primarily based on website data, complemented by patent data, is highly suitable for the comprehensive analysis of enterprise innovation networks, without constraints related to specific regions or levels of geographical analysis. Concurrently, proximity within their innovation networks can be meaningfully assessed by combining the features of informatization in enterprise cooperation. Upon considering online features, the innovation networks exhibit numerous contrasting or significantly complementary features compared to traditional networks. Furthermore, enterprises can enhance their innovation probability through compensatory mechanisms across various proximities.

The high accuracy of the XGBoost algorithm results in identifying the influence mechanisms, demonstrating the reliability of the model. The findings highlight that, globally, social proximity has the most significant impact on innovation probability, followed by technological proximity, whereas organizational proximity has the least influence. Importantly, close relationships between companies do not automatically lead to higher innovation probabilities, due to the complex nonlinear relationship between proximity and innovation. This suggests that achieving enhanced enterprise innovation efficiency requires a multifaceted approach, involving a comprehensive balance among these various perspectives to identify the optimal solution.

Building on previous discussions, online collaboration has brought about significant changes in the innovation network, characterized by a complex influencing mechanism. These conclusions are significant as digital transformation rapidly reshapes the operational and collaborative paradigms of enterprises. A more accurate understanding of the transformed innovation network is essential for enhancing business efficiency and the urban innovation foundation. However, this study has limitations; primarily, it relies on only one year of website and patent data for spatial pattern analysis, within a relatively short timeframe. This limitation prevents the definitive exclusion of occasional results and fails to capture the dynamic process of the "online" transformation of the innovation network. Future research endeavors will employ network mining methods to acquire longer-term data, incorporating a broader time dimension for a nuanced discussion of the spatiotemporal evolution of online collaboration.

Conflict of interest statement

The authors declare no conflict of interest

CRediT authorship contribution statement

Chenxi Liu: Writing – original draft, Methodology, Conceptualization. **Zhenghong Peng:** Writing – review & editing, Funding acquisition, Conceptualization. **Lingbo Liu:** Writing – review & editing, Methodology, Conceptualization. **Hao Wu:** Methodology, Funding acquisition. **Jan Kinne:** Writing – review & editing, Methodology, Data curation. **Meng Cai:** Conceptualization. **Shixuan Li:** Methodology, Data curation.

Acknowledgement

This research was funded by National Science Foundation grant #1841403 and National Natural Science Foundation of China grant #52078390 and #51978535.

References

Abbasiharofteh, M., & Broekel, T. (2021). Still in the shadow of the wall? The case of the Berlin biotechnology cluster. *Environment & Planning A, 53*, 73–94. https://doi.org/10.1177/0308518X20933904

Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2023). The digital layer: Alternative data for regional and innovation studies. *Spatial Economic Analysis*, 1–23. https://doi.org/10.1080/17421772.2023.2193222

Aldieri, L. (2013). Knowledge technological proximity: Evidence from US and European patents. *Economics of Innovation and New Technology, 22*, 807–819. https://doi.org/10.1080/10438599.2013.788838

Asheim, B. T., & Isaksen, A. (2002). *Regional innovation systems: The integration of local "sticky" and global "ubiquitous"*. Knowledge.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *The Journal of Economic Perspectives, 32*, 259–280. https://doi.org/10.1257/jep.32.3.259

Baru, S. (2012). Geo-economics and strategy. *Survival, 54*, 47–58. https://doi.org/10.1080/00396338.2012.690978

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*, 31–56. https://doi.org/10.1191/0309132504ph469oa

Beeson, M. (2018). Geoeconomics with Chinese characteristics: The BRI and China's evolving grand strategy. *Economic and Political Studies, 6*, 240–256. https://doi.org/10.1080/20954816.2018.1498988

Bi, K., Huang, P., & Wang, X. (2016). Innovation performance and influencing factors of low-carbon technological innovation under the global value chain: A case of Chinese manufacturing industry. *Technological Forecasting and Social Change, 111*, 275–284. https://doi.org/10.1016/j.techfore.2016.07.024

Boschma, R. (2005). Proximity and innovation: A critical assessment. *Regional Studies, 39*, 61–74. https://doi.org/10.1080/0034340052000320887

Breschi, S., & Lissoni, F. (2009). Mobility of skilled workers and co-invention networks: An anatomy of localized knowledge flows. *Journal of Economic Geography, 9*, 439–468. https://doi.org/10.1093/jeg/lbp008

Bunduchi, R., Weisshaar, C., & Smart, A. U. (2011). Mapping the benefits and costs associated with process innovation: The case of RFID adoption. *Technovation, 31*, 505–521. https://doi.org/10.1016/j.technovation.2011.04.001

Cantner, U., & Meder, A. (2007). Technological proximity and the choice of cooperation partner. *J Econ Interac Coord, 2*, 45–65. https://doi.org/10.1007/s11403-007-0018-y

Cao, X., Zeng, G., & Ye, L. (2019). The structure and proximity mechanism of formal innovation networks: Evidence from Shanghai high-tech ITISAs. *Growth and Change, 50*, 569–586. https://doi.org/10.1111/grow.12294

Chen, Y., Zhang, X., Grekousis, G., Huang, Y., Hua, F., Pan, Z., & Liu, Y. (2023). Examining the importance of built and natural environment factors in predicting self-rated health in older adults: An extreme gradient boosting (XGBoost) approach. *Journal of Cleaner Production, 413*, Article 137432. https://doi.org/10.1016/j.jclepro.2023.137432

Chen, Y., Zhu, Q., & Sarkis, J. (n.d.). Heterogeneity in corporate green supply chain practice adoption: Insights from institutional fields. Business Strategy and the Environment n/a. https://doi.org/10.1002/bse.3499.

Coe, N. M., Dicken, P., & Hess, M. (2008). Global production networks: Realizing the potential. *Journal of Economic Geography, 8*, 271–295. https://doi.org/10.1093/jeg/lbn002

Cooke, P. (1992). Regional innovation systems: Competitive regulation in the new Europe. *Geoforum, 23*, 365–382. https://doi.org/10.1016/0016-7185(92)90048-9

Criscuolo, P., Laursen, K., Reichstein, T., & Salter, A. (2018). Winning combinations: Search strategies and innovativeness in the UK. *Industry & Innovation, 25*, 115–143. https://doi.org/10.1080/13662716.2017.1286462

Crowley, F., & Jordan, D. (2022). Do local start-ups and knowledge spillovers matter for firm-level R&D investment? *Urban Studies, 59*, 1085–1102. https://doi.org/10.1177/0042098021995105

De Noni, I., Orsi, L., & Belussi, F. (2018). The role of collaborative networks in supporting the innovation performances of lagging-behind European regions. *Research Policy, 47*, 1–13. https://doi.org/10.1016/j.respol.2017.09.006

Dennis Wei, Y. H., Liefner, I., & Miao, C.-H. (2011). Network configurations and R&D activities of the ICT industry in Suzhou municipality, China. *Geoforum, 42*, 484–495. https://doi.org/10.1016/j.geoforum.2011.03.005

Dezhong Duan, D. D. (2018). Spatial-temporal complexity and growth mechanism of city innovation network in China. *Scientia Geographica Sinica, 38*, 1759–1768. https://doi.org/10.13249/j.cnki.sgs.2018.11.003

Donaldson, D., & Hornbeck, R. (2016). Railroads and American economic growth: A "market access" approach. *Quarterly Journal of Economics, 131*, 799–858. https://doi.org/10.1093/qje/qjw002

Du, W., Ge, C., Yao, S., Chen, N., & Xu, L. (2023). Applicability analysis and ensemble application of BERT with TF-IDF, TextRank, MMR, and LDA for topic classification based on flood-related VGI. *ISPRS International Journal of Geo-Information, 12*, 240. https://doi.org/10.3390/ijgi12060240

Eiriz, V., Faria, A., & Barbosa, N. (2013). Firm growth and innovation: Towards a typology of innovation strategy. *Innovation, 15*, 97–111. https://doi.org/10.5172/impp.2013.15.1.97

Errico, F., Corallo, A., Spennato, A., & Berlingerio, G. E. (2022). *Spatial proximity versus social distance: Partnership development in the cross-border cooperation*. J Knowl Econ. https://doi.org/10.1007/s13132-022-01077-9

Fitjar, R. D., Huber, F., & Rodríguez-Pose, A. (2016). Not too close, not too far: Testing the goldilocks principle of 'optimal' distance in innovation networks. *Industry & Innovation, 23*, 465–487. https://doi.org/10.1080/13662716.2016.1184562

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*. https://doi.org/10.1214/aos/1013203451

Gao, B., Hong, J., Guo, H., Dong, S., & Lan, Z.-Z. (2023). Cooperative evolution and symmetry breaking in interdependent networks based on alliance mechanisms. *Physica A: Statistical Mechanics and Its Applications, 609*, Article 128320. https://doi.org/10.1016/j.physa.2022.128320

Geldes, C., Felzensztein, C., Turkina, E., & Durand, A. (2015). How does proximity affect interfirm marketing cooperation? A study of an agribusiness cluster. *Journal of Business Research, 68*, 263–272. https://doi.org/10.1016/j.jbusres.2014.09.034

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*. https://doi.org/10.1257/jel.20181020

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

Glaeser, C. K., Glaeser, S., & Labro, E. (2022). Proximity and the management of innovation. *Management Science*. https://doi.org/10.1287/mnsc.2022.4469

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics, 102*, 653–671. https://doi.org/10.1007/s11192-014-1434-0

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation.". *AI Magazine, 38*, 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Grekousis, G., & Liu, Y. (2019). Where will the next emergency event occur? Predicting ambulance demand in emergency medical services using artificial intelligence. *Computers, Environment and Urban Systems, 76*, 110–122. https://doi.org/10.1016/j.compenvurbsys.2019.04.006

Griffith, R., Harrison, R., & Van Reenen, J. (2006). How special is the special relationship? Using the impact of U.S. R&D spillovers on U.K. Firms as a test of technology sourcing. *The American Economic Review, 96*, 1859–1875. https://doi.org/10.1257/aer.96.5.1859

Gulati, R., & Gargiulo, M. (1999). Where do interorganizational networks come from? *American Journal of Sociology, 104*, 1439–1493. https://doi.org/10.1086/210179

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine, 40*, 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Guo, M., Yang, N., Wang, J., & Zhang, Y. (2021). Multi-dimensional proximity and network stability: The moderating role of network cohesion. *Scientometrics, 126*, 3471–3499. https://doi.org/10.1007/s11192-021-03882-6

Hartono, A., & Rafik, A. (2022). Linking open innovation, innovation barriers and performance of Indonesian firms. *International Journal of Innovation Science, 14*, 713–732. https://doi.org/10.1108/IJIS-10-2020-0218

Heringa, P. W., Horlings, E., van der Zouwen, M., van den Besselaar, P., & van Vierssen, W. (2014). How do dimensions of proximity relate to the outcomes of collaboration? A survey of knowledge-intensive networks in the Dutch water sector. *Economics of Innovation and New Technology, 23*, 689–716. https://doi.org/10.1080/10438599.2014.882139

Hu, C., Mao, J., Tian, M., Wei, Y., Guo, L., & Wang, Z. (2021). Distance matters: Investigating how geographic proximity to ENGOs triggers green innovation of heavy-polluting firms in China. *Journal of Environmental Management, 279*, Article 111542. https://doi.org/10.1016/j.jenvman.2020.111542

Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits and market value. Working Paper Series https://doi.org/10.3386/w1815.

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics, 125*, 2011–2041. https://doi.org/10.1007/s11192-020-03726-9

Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2020). *The digital layer: How innovative firms relate on the web*. ZEW Discussion Papers.

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change, 127*, 291–303. https://doi.org/10.1016/j.techfore.2017.10.002

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems, 96*, Article 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845

Li, D., Wei, Y. D., Miao, C., & Chen, W. (2020). Innovation, innovation policies, and regional development in China. *Geographical Review, 110*, 505–535. https://doi.org/10.1080/00167428.2019.1684194

Lim, H., & Han, C. (2023). National borders transcended: The impact of geographical proximity on the growth of global innovation networks among cities in east asia. *International Journal on the Unity of the Sciences, 27*, 570–598. https://doi.org/10.1080/12265934.2021.1915854

Liu, C., Peng, Z., Liu, L., & Li, S. (2023). Innovation networks of science and technology firms: Evidence from China. *Land, 12*(7), 1283.

Liu, L. (2024). An ensemble framework for explainable geospatial machine learning models. *International Journal of Applied Earth Observation and Geoinformation, 132*, Article 104036.

Lotfata, A., Grekousis, G., & Wang, R. (2023). Using geographical random forest models to explore spatial patterns in the neighborhood determinants of hypertension prevalence across chicago, Illinois, USA. *Environment and Planning B: Urban Analytics and City Science 23998083231153401*. https://doi.org/10.1177/23998083231153401

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). Consistent individualized feature attribution for tree ensembles. https://doi.org/10.48550/arXiv.1802.03888.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. https://doi.org/10.48550/arXiv.1705.07874.

Ma, H., & Xu, X. (2023). The effects of proximities on the evolving structure of intercity innovation networks in the Guangdong–Hong Kong–Macao greater bay area: Comparison between scientific and technology knowledge. *International Journal on the Unity of the Sciences, 27*, 390–413. https://doi.org/10.1080/12265934.2022.2085154

Ma, M., Zhao, G., He, B., Li, Q., Dong, H., Wang, S., & Wang, Z. (2021). XGBoost-based method for flash flood risk assessment. *Journal of Hydrology, 598*, Article 126382. https://doi.org/10.1016/j.jhydrol.2021.126382

Manuel, C. (2009). The new economy: Informationalism, globalization, networking. In *The rise of the network society* (pp. 77–162). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444319514.ch2.

Markovic, S., Jovanovic, M., Bagherzadeh, M., Sancha, C., Sarafinovska, M., & Qiu, Y. (2020). Priorities when selecting business partners for service innovation: The contingency role of product innovation. *Industrial Marketing Management, 88*, 378–388. https://doi.org/10.1016/j.indmarman.2020.06.001

Marshall, A. (2013). *Principles of economics*. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137375261

Mascia, D., Pallotti, F., & Angeli, F. (2017). Don't stand so close to me: Competitive pressures, proximity and inter-organizational collaboration. *Regional Studies, 51*, 1348–1361. https://doi.org/10.1080/00343404.2016.1185517

Nagaoka, S., Motohashi, K., & Goto, A. (2010). Patent statistics as an innovation indicator. In B. H. Hall, & N. Rosenberg (Eds.), *Handbook of the economics of innovation, handbook of the economics of innovation* (Vol. 2, pp. 1083–1127). North-Holland. https://doi.org/10.1016/S0169-7218(10)02009-5.

Noonan, L., O'Leary, E., & Doran, J. (2020). The impact of institutional proximity, cognitive proximity and agglomeration economies on firm-level productivity. *Journal of Economics Studies, 48*, 257–274. https://doi.org/10.1108/JES-07-2019-0345

O' Connor, M., Doran, J., & McCarthy, N. (2020). Cognitive proximity and innovation performance: Are collaborators equal? *European Journal of Innovation Management, 24*, 637–654. https://doi.org/10.1108/EJIM-11-2019-0347

Oerlemans, L., & Meeus, M. (2005). Do organizational and spatial proximity impact on firm performance? *Regional Studies, 39*, 89–104. https://doi.org/10.1080/0034340052000320896

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks, 32*, 245–251. https://doi.org/10.1016/j.socnet.2010.03.006

Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science, 28*, 441–453. https://doi.org/10.1177/016555150202800601

Porter, M. (1990). Competitive advantage of nations. *Competitive Intelligence Review, 1*, 14. https://doi.org/10.1002/cir.3880010112

Pota, M., Ventura, M., Fujita, H., & Esposito, M. (2021). Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. *Expert Systems with Applications, 181*, Article 115119. https://doi.org/10.1016/j.eswa.2021.115119

Pouder, R., & StJohn, C. H. (1996). Hot spots and blind spots: Geographical clusters of firms and innovation. *Academy of Management Review, 21*, 1192–1225. https://doi.org/10.2307/259168

Rahimi, S., Mottahedi, S., & Liu, X. (2018). *The geography of taste: Using yelp to study urban culture*. Multidisciplinary Digital Publishing Institute. https://doi.org/10.20944/preprints201806.0389.v1

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

Santamaría, L., Nieto, M. J., & Rodríguez, A. (2021). Failed and successful innovations: The role of geographic proximity and international diversity of partners in technological collaboration. *Technological Forecasting and Social Change, 166*, Article 120575. https://doi.org/10.1016/j.techfore.2021.120575

Shapley, L. S. (2016). 17. A value for n-person games. In *17. A value for n-person games* (pp. 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018.

Shaw, A. T., & Gilly, J.-P. (2000). On the analytical dimension of proximity dynamics. *Regional Studies, 34*, 169–180. https://doi.org/10.1080/00343400050006087

Simensen, E. O., & Abbasiharofteh, M. (2022). Sectoral patterns of collaborative tie formation: Investigating geographic, cognitive, and technological dimensions. *Industrial and Corporate Change, 31*, 1223–1258. https://doi.org/10.1093/icc/dtac021

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). *Measuring patent quality: Indicators of technological and economic value*. Paris: OECD. https://doi.org/10.1787/5k4522wkw1r8-en

Steinmo, M., & Rasmussen, E. (2016). How firms collaborate with public research organizations: The evolution of proximity dimensions in successful innovation projects. *Journal of Business Research, 69*, 1250–1259. https://doi.org/10.1016/j.jbusres.2015.09.006

Stich, C., Tranos, E., & Nathan, M. (2023). Modeling clusters from the ground up: A web data approach. *Environment and Planning B: Urban Analytics and City Science, 50*, 244–267. https://doi.org/10.1177/23998083221108185

Stoehr, N., Braesemann, F., Frommelt, M., & Zhou, S. (2020). Mining the automotive industry: A network analysis of corporate positioning and technological trends. In H. Barbosa, J. Gomez-Gardenes, B. Gonçalves, G. Mangioni, R. Menezes, & M. Oliveira (Eds.), *Complex networks XI, springer proceedings in complexity* (pp. 297–308). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-40943-2_25.

Sun, B., Ruan, A., Peng, B., & Lu, W. (2022a). Talent flow network, the life cycle of firms, and their innovations. *Frontiers in Psychology, 13*, Article 788515. https://doi.org/10.3389/fpsyg.2022.788515

Sun, Y., Sun, Y., & Zhang, J. (2022). How does proximity affect the dual innovation of alliance partner? The role of knowledge coupling. *IEEE Access, 10*, 19149–19161. https://doi.org/10.1109/ACCESS.2022.3150930

Tomihira, T., Otsuka, A., Yamashita, A., & Satoh, T. (2020). Multilingual emoji prediction using BERT for sentiment analysis. *International Journal of Web Information Systems, 16*, 265–280. https://doi.org/10.1108/IJWIS-09-2019-0042

Wang, F., & Liu, L. (2023). *Computational methods and GIS applications in social science.* CRC Press.

Wasserman, S., & Galaskiewicz, J. (1994). *Advances in social network analysis: Research in the social and behavioral sciences.* SAGE Publications.

Wei, Y. H. D. (2015). Network linkages and local embeddedness of foreign ventures in China: The case of suzhou municipality. *Regional Studies, 49*, 287–299. https://doi.org/10.1080/00343404.2013.770139

Yang, Y., & Cai, N. (2009). Study on the selection of firm technological innovation model based on game analysis. In *2009 16th international conference on industrial engineering and engineering management. Presented at the 2009 16th international conference on industrial engineering and engineering management* (pp. 279–283). https://doi.org/10.1109/ICIEEM.2009.5344590

Yao, L., Li, J., & Li, J. (2020). Urban innovation and intercity patent collaboration: A network analysis of China's national innovation system. *Technological Forecasting and Social Change, 160*, Article 120185. https://doi.org/10.1016/j.techfore.2020.120185

Yingcheng, L., Weiting, X., & Xiaowu, H. (2023). The geography of intercity technological proximity: Evidence from China. *International Journal on the Unity of the Sciences, 27*, 355–370. https://doi.org/10.1080/12265934.2021.1938641

Yu, X., Chen, Y., & Nguyen, B. (2014). Knowledge management, learning behavior from failure and new product development in new technology ventures. *Systems Research and Behavioral Science, 31*, 405–423. https://doi.org/10.1002/sres.2273

Zhang, C., Wang, B., & Ye, J. (2023). Social proximity and urban innovation: A megalopolis perspective. *Applied Economics Letters*, 1–5. https://doi.org/10.1080/13504851.2023.2276363, 0.

Zhao, F., Li, X., Gao, Y., Li, Y., Feng, Z., & Zhang, C. (2022). Multi-layer features ablation of BERT model and its application in stock trend prediction. *Expert Systems with Applications, 207*, Article 117958. https://doi.org/10.1016/j.eswa.2022.117958

Zhao, C., Wang, K., & Dong, K. (2023). How does innovative city policy break carbon lock-in? A spatial difference-in-differences analysis for China. *Cities, 136*, Article 104249. https://doi.org/10.1016/j.cities.2023.104249

Zhao, T., Yang, M., Cao, Z., & Wang, X. (2022). Understanding the joint impacts of cognitive, social, and geographic proximities on the performance of innovation collaboration between knowledge-intensive business services and the manufacturing industry: Empirical evidence from China. *Frontiers in Psychology, 13*.

Zhong, X., Ren, G., & Wu, X. (2023). Not all innovation is prioritized: Economic policy uncertainty, industry environmental, and firms' relative exploitative innovation emphasis. *European Journal of Innovation Management*. https://doi.org/10.1108/EJIM-01-2023-0006. ahead-of-print.