Proximity dimensions and the emergence of collaboration: a HypTrails study on German AI research

Tobias Koopmann[1] · Maximilian Stubbemann[2] · Matthias Kapa[3] · Michael Paris[4] ·
Guido Buenstorf[3,5] · Tom Hanika[6] · Andreas Hotho[1] · Robert Jäschke[2,4] ·
Gerd Stumme[2]

Abstract

Creation and exchange of knowledge depends on collaboration. Recent work has suggested that the emergence of collaboration frequently relies on geographic proximity. However, being co-located tends to be associated with other dimensions of proximity, such as social ties or a shared organizational environment. To account for such factors, multiple dimensions of proximity have been proposed, including cognitive, institutional, organizational, social and geographical proximity. Since they strongly interrelate, disentangling these dimensions and their respective impact on collaboration is challenging. To address this issue, we propose various methods for measuring different dimensions of proximity. We then present an approach to compare and rank them with respect to the extent to which they indicate co-publications and co-inventions. We adapt the *HypTrails* approach, which was originally developed to explain human navigation, to co-author and co-inventor graphs. We evaluate this approach on a subset of the German research community, specifically academic authors and inventors active in research on artificial intelligence (AI). We find that social proximity and cognitive proximity are more important for the emergence of collaboration than geographic proximity.

Keywords Dimensions of proximity · Co-authorships · Co-inventorships · Embedding techniques · Collaboration

Introduction

Collaboration is a powerful tool to advance the frontier of knowledge in science and innovation. Both, the share of co-authored research articles and the average number of authors per paper increased strongly in the past decades and patents follow similar trends (Wuchty et al., 2007). These developments add to the importance of better understanding the emergence and effects of research collaboration. Numerous studies have highlighted the impact of geographic closeness for collaboration in science and innovation. However, geography is only

---

one of several dimensions of proximity upon which collaboration builds. Further proximities are cognitive, institutional, organizational and social, which have been shown to be relevant in prior research (Boschma, 2005; Broekel & Boschma, 2011). As geographic co-location is often associated with similarity in prior knowledge (cognitive proximity; Nooteboom (2001)), and also with high levels of social, institutional, and organizational proximity (Breschi & Lissoni, 2009; Heinisch et al., 2016), disentangling their impacts is challenging (Bode et al., 2019). In this paper, we explore how the various dimensions of proximity are related to the emergence of successful collaboration in research on artificial intelligence (AI). In our case study, we focus on the German AI landscape since Germany has a rapidly emerging AI community, with about 100 new professorships to be created in the near future [1] and AI is expected to have a strong impact on future technological and economic development. To identify academic collaboration in AI research, we employ the *German AI Network* (GAI), a novel data set that incorporates bibliographic information for 2131 researchers. The GAI builds upon the DBLP data set (Ley, 2009) and includes both journal publications and contributions to conferences in computer science and related fields of research. This allows us to consider conference proceedings in which the outcomes of successful collaboration in AI research are often communicated. In addition to co-authorships, we trace co-inventions of AI researchers employing the Crios-Patstat patent data set (Tarasconi, 2014). These bibliographical data sets are used to construct several similarity functions measuring how close researchers are to each other in terms of cognitive, institutional, organizational, social and geographic proximity. In constructing these proximity measures, bibliographic information is complemented by web data and information about academic genealogies. We also employ similarity measures for the text documents in our data set based on Natural Language Processing (NLP). We then adapt the Bayesian *HypTrails* approach (Singer et al., 2015), which originally was designed to compare explanations of human navigation, to the domain of co-authorship and co-inventorship. This allows to rank proximity dimensions according to how well they explain the collaboration patterns observed in our data. Our results indicate that social proximity is a key factor in the emergence of successful collaboration in German AI research.

The present paper contributes to the scientometric literature in terms of data, methods and results. First, we curate a novel data collection of German researchers from the domain of artificial intelligence and integrate rich data from various sources to obtain an encompassing view of their activities and mutual relationships. Second, regarding our contribution to research methods, we adapt the HypTrails approach to provide a new way of assessing the relative ability of various factors to explain the observable patterns in collaboration data. Third, we contribute to the discussion on proximity dimensions started by Boschma (2005). Specifically, our results suggest that social proximity is the key factor for successful collaboration between AI researchers in Germany.

Dimensions of proximity

Collaborative research activities, i.e., joint efforts to produce new knowledge, enable researchers to combine their individual knowledge base and thus to arrive at findings that they might not achieve individually (Katz and Martin, 1997). Research collaboration has been the subject of extensive scholarly attention (e.g., Newman, 2001; Newman, 2004; Lee & Bozeman, 2005). There is substantial evidence that the quantity and quality of scientific output may increase by collaboration (Glänzel & Schubert, 2005; Wuchty et al., 2007; Werker

---

[1] https://knowledge4policy.ec.europa.eu/ai-watch/germany-ai-strategy-report_en.

et al., 2019), even though too large team size may discourage creative insights (Heinze et al., 2009). In line with these empirical findings, policy makers and funding agencies encourage collaborative research activities. For collaboration among researchers to be successful, the exchange of knowledge should result in meaningful ideas, which requires suitable collaboration partners. Recent work suggests that various forms of proximity are relevant for how researchers find partners. Specifically, Boschma (2005) distinguishes five dimensions of proximity: cognitive, institutional, organizational, social and geographic proximity.

Cognitive proximity is defined by the similarity of knowledge of two actors. To exchange knowledge and jointly discover new findings, collaborating researchers need mutual understanding, which requires overlap in their knowledge bases. Successful research collaboration moreover depends on common interests in collaboration outcomes. We therefore expect that researchers are more likely to collaborate if they are cognitively close to each other. At the same time, researchers can only learn from each other if their prior knowledge is not entirely identical. This suggests that there is an optimal degree of cognitive distance between collaborating researchers which is above zero (Nooteboom, 2001). Cognitive proximity of researchers is widely investigated. For example, Liu et al. (2018) show that doctoral students are cognitively close to their advisors, and Hautala (2011) studies cognitive proximity in international research teams. Intuitively the knowledge of authors is reflected in their research. Hence, we assume that cognitive proximity can be approximated by the content of their publications. Specifically, overlaps in keywords (Xu et al., 2016) and technology classes of patents (Jaffe et al., 1993), but also similar uses of language and scientific jargon can be manifestations of cognitive proximity.

The degree of *institutional proximity* of two actors can be measured by comparing their institutional environment. Here, institutional environment refers to the routines, regulations and laws an actor is subject to (Nooteboom, 2001; Edquist & Johnson, 1996). To a large extent, the institutional environment is shaped at the societal macro level (Boschma, 2005). However, relevant institutional differences may also exist between different societal subsectors. In our empirical context, we expect that institutional differences between public research institutions and corporate R&D affect the likelihood of researchers to collaborate (Perkmann et al., 2013; Stern, 2004; Hirv, 2018).

Not only the type of organization that a given researcher is affiliated with (e.g., university or company), but also the specific individual affiliation is an important factor for collaboration. Membership in the same organization thus provides a basic measure of *organizational proximity* (Crescenzi et al., 2016). It increases the likelihood of chance encounters, but more importantly collaborative research may be based on strategic decision making within the organization. That is, researchers can be allocated to research projects by their superiors within the organization. We can further refine the measure of organizational proximity by considering the departmental structure within organizations as well as relationships between organizations (Broekel & Boschma, 2011) that are independent of respective agents.

Social proximity reflects the extent to which actors are linked by social relations. Such relations can be based on kinship, friendship, familiarity based on prior contact or other kinds of social ties (Boschma, 2005). Their relevance for the emergence of research collaboration is twofold. On the one hand, actors linked by social ties tend know each other and may be aware of each other's interests. Thus, social proximity enhances the potential to engage in collaboration. On the other hand, social ties affect the level of trust in potential collaboration partners and their competence. Accordingly, they may increase the willingness to start a collaboration. In empirical studies of science and innovation, social proximity is frequently measured by pre-existing co-authorship (Hardeman, 2015) or co-inventorship (Breschi & Lissoni, 2009) relations, including higher-degree connections (Balland,

2012). While these measures are not without limitations (Katz and Martin, 1997), it is plausible that joint work establishes social ties among the collaborators.

Finally, *geographical proximity* is defined by the proximity of two agents in physical space. The importance of geographic proximity for innovation has been discussed at least since Marshall (1890), and it was rediscovered when the interest in industry clusters re-emerged in the 1980s and 1990s (Audretsch & Feldman, 1996; Delgado et al., 2010). Fundamentally, the relevance of geographic proximity derives from the difficulty of communicating tacit knowledge (Polanyi, 1966) other than through face-to-face interaction. Recent work demonstrates its significance even within organizations and at small geographic scales (Catalini, 2018).

Geographic proximity facilitates encounters and may allow face-to-face communication and observational learning, even if agents are not characterized by high levels of proximity in any of the other dimensions (Hoekman et al., 2010). However, there is a large body of prior work indicating that geographic proximity often reflects proximity in another dimension, such as social (Breschi & Lissoni, 2009) or organizational (Buenstorf & Klepper, 2010) proximity. Since there are also overlaps between these other dimensions, our knowledge about their individual roles is limited. In the remainder of this paper, we will develop a new approach to disentangle their individual role in the emergence of collaboration in German AI research. As the first step in this endeavor, the next section provides a detailed account of our empirical measures of the individual proximity dimensions.

Measuring and quantifying proximity

In this section we propose methods to quantify and compute proximity with respect to the different dimensions covered in previous section. These methods result in similarity functions reflecting proximity of researchers given the respective dimension. An overview of the similarity functions is given in Table 1. These methods are intended for data sets of the following kind: We assume that R is a set of researchers, A a set of affiliations, P a set of publications and U a set of URLs. We then consider the following relations: is_author_of $\subseteq R \times P$, has_affiliation $\subseteq R \times A$, PhD_at $\subseteq R \times A$ and has_homepage $\subseteq R \times U$. In the section "Empirical context: the German AI community", we present concrete data providing this information.

Cognitive proximity

Since publications reflect the creation and distribution of knowledge in the academic community, we use their content to capture the research topics of authors. Cognitive proximity between authors can then be measured using the text of their respective publications. Measuring the similarity of text documents is a well studied research topic. Recently, a plenitude of methods for *representation learning* have been developed (Le & Mikolov, 2014; Sinoara et al., 2019). Representation learning, also called *embedding*, refers to the transformation of any data, for example text, into real-valued vector spaces, where the measurement of distances is well studied. Transforming text into vector representations can be accomplished by using weighted word counts (Jones, 1972), approaches based on matrix-factorization (Deerwester et al., 1990) or modern neural network architectures (Devlin et al., 2019). To utilize these methods on researchers to extract their respective research topics, we first apply them on the set of publications P. Here, we use concatenations of titles and

Table 1 Overview of similarity functions with their respective proximity dimension

Proximity Dimension	Similarity Function	Explanation
Cognitive	sim $_{LSA}$	Latent Semantic Analysis
	sim $_{NMF}$	Non Negative Matrix Factorization
	sim $_{BERT}$	BERT embedding
Institutional	sim $_{aff\_type}$	Do two researchers have both a university or non-university affiliation?
Organizational	sim $_{Affiliation}$	Amount of same affiliations
	sim $_{URL}$	URLs sharing the same hosts
	sim $_{Hyperlink}$	Distances in the Syntactic Hyperlink Graph
	sim $_{Hierachy}$	Hierarchical distance between homepages
Social	sim $_{Diss\_Loc}$	Dissertation at same location?
	sim $_{DeepWalk}$	DeepWalk embeddings
	sim $_{Node2vec-small-p}$	node2vec embeddings using $p = 0.25, q = 4$
	sim $_{Node2vec-large-p}$	node2vec embeddings using $p = 4, q = 0.25$
	sim $_{HOPE}$	HOPE embeddings
Geographic	sim $_{Geo}$	Shortest geographic distance between affiliations of two authors

abstracts of documents as input to generate a vector representation v_p for each publication $p \in P$. Then, a representation for a researcher is computed as the mean point vector of the respective publication vectors. To calculate a similarity of two researchers, we then apply cosine similarity.

In this study, we use three approaches to create vector representations for each document and hence each researcher. Two approaches use TF–IDF factorization, whereas the last one is based on neural networks. For the first two approaches, the abstracts of the documents are encoded in a TF–IDF matrix $M \in \mathbb{R}^{m \times n}$. Here, the entry $M_{i,j}$ reflects how often the j-th term occurs in the i-th document normalized by the count of occurrences in all documents. This weighting scheme incorporates the significance of a term for a given document and does not overweight words that are frequently used over all documents.

In detail, if we have documents $P = \{p_1, \ldots, p_m\}$, that are modeled as finite sequences over a set of terms $T = \{t_1, \ldots, t_n\}$, then entry $M_{i,j}$ is the product of the frequency of term t_j (the *term-frequency*) in document p_i with the inverse frequency of t_j in all documents (the *inverse-document-frequency*). Based on this, we use *Latent semantic analysis* (LSA) (Deerwester et al., 1990) and *non-negative matrix factorization* (NMF) (Lee & Seung, 1999, Lee & Seung, 2000) to generate vector representations for publications $p \in P$.

LSA computes vector representations for a set of documents by computing a singular value decomposition of the TF–IDF matrix. Building up on this, a low-rank approximation is computed. LSA is well established in the realm of text mining (Foltz, 1996; Foltz et al., 1998) and is intended to identify different words with similar meaning and to reveal the semantic structure of a given set of documents (Deerwester et al., 1990). Hence, LSA is known to compute vector representations where the measurement of distances is meaningful. In detail, LSA works as follows. It produces a factorization of the form $M = U \Sigma V^{\text{tr}}$, with $U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}, \Sigma \in \mathbb{R}^{m \times n}$ such that Σ is a diagonal matrix. Let $U_d \in \mathbb{R}^{m \times d}$ be the matrix that results by extracting only the first d columns of U and let $\Sigma_d \in \mathbb{R}^{d \times d}$ be the diagonal matrix that consists only of the first d rows and columns of Σ. Then the document vectors

with dimension d are given by the rows of the matrix $M_d := U_d \Sigma_d$. For more details, we refer the reader to (Manning et al., 2008).

Another approach to compute vector representations for a document corpus is given by NMF, which generates the vectors by factorizing the TF–IDF matrix into two non-negative matrices. Here, a specific row of the first matrix represents for the corresponding document a distribution over different topics and the rows of the second matrix represents to which extent a specific topic is connected to specific words. Hence, NMF can be used to compute understandable topic distributions for documents and therefore is commonly applied in the realm of text mining, for example for document clustering (Xu et al., 2003) or document summarization (Lee et al., 2009). In detail, NMF works as follows: To produce vector representations with dimension $d \in \mathbb{N}$ for all publications $p \in P$, NMF factorizes the TF-IDF matrix M into two non-negative matrices $W \in \mathbb{R}_{\geq 0}^{m \times d}, H \in \mathbb{R}_{\geq 0}^{d \times n}$ such that $M \approx WH$. Here, the rows of W are the vector representations of the documents.

The third approach to create vector representations is based on *BERT* (Devlin et al., 2019). BERT is designed for sentence inputs, where 'sentence' not necessarily refers to a linguistic sentence, but to an ordered sequence of words/tokens with a reasonable size. Hence, in our scenario, we are able to use the whole abstract of a given document as input. That BERT takes its input as an ordered sequence of terms is an important difference to the previous approaches, i.e. LSA and NMF. These are based on the TF-IDF matrix that is independent of the order of words. Therefore, similar use of language and scientific jargon will only be captured using BERT as an embedding approach. More specifically, BERT is a neural network model that is built upon multiple transformer layers (Vaswani et al., 2017). These models are pre-trained using a large corpus of text data, which allows them to incorporate a general language understanding. This has been shown to lead to impressive results in a variety of NLP tasks (Devlin et al., 2019). To receive a vector representation for a given document, we use the title and abstract as input into the pre-trained BERT model and extract the vector representation from the output of the neural network's last layer. While Devlin et al. (2019) provides multiple BERT models itself, we use SciBERT (Beltagy & Cohan, 2019), which is pre-trained using additional scientific texts and therefore better suited for the representation of scholarly publications.

The three approaches explained above lead to three embedding functions that map publications (i.e., their titles concatenated with their abstracts) to real-valued vectors. We name these functions $f_{\text{LSA}}, f_{\text{NMF}}, f_{\text{BERT}}$. For a given embedding function $f \in \{f_{\text{LSA}}, f_{\text{NMF}}, f_{\text{BERT}}\}$, we then compute the vector representation of a researcher $r \in R$ by $f(r) := \frac{1}{|P_r|} \sum_{p \in P_r} f(p)$, with $P_r := \{p \in P \mid (r, p) \in \text{is\_author\_of}\}$. For $m \in \{\text{LSA}, \text{NMF}, \text{BERT}\}$ we then define a similarity function sim_m via cosine similarity:

$$\text{sim}_m(r_1, r_2) := \frac{\langle f_m(r_1), f_m(r_2) \rangle}{\|f_m(r_1)\| \|f_m(r_2)\|}$$

Institutional proximity

As mentioned in previous section, institutional proximity of two actors can be understood as the proximity of the regulations and laws they are subject to. Quantifying this kind of proximity is challenging and is seldom done in related work, especially with data that is freely accessible. However, as a first step towards measuring institutional proximity, we assume that institutional circumstances differ in the private sector and the academic landscape. Hence, we state that two actors are close on the institutional level if they are both publishing from affiliations

located in the private sector or if they both are working in academic research facilities. This binary differentiation of academic and non-academic affiliations does not capture nuanced affinities that may reflect, e.g., similarities between private companies and application-oriented university departments. Nevertheless, previous work show its justification to serve as indicator for institutional proximity. (Ponds et al., 2007). Research positions in industry are sufficiently different from those at universities (Aghion et al., 2008) that scientists may prefer worse-paid university positions (Stern, 2004). The resulting similarity function looks as follows: For each affiliation $a \in A$ let $\mathrm{ac}(a) = 1$ if a is an academic affiliation and $\mathrm{ac}(a) = 0$ otherwise. Here, non-university research institutions are considered as non-academic affiliations. Then for two researchers $r_1, r_2 \in R$ we have

$$\mathrm{sim}_{\mathrm{aff\_type}}(r_1, r_2) := \begin{cases} 1 & \exists (r_1, a_1), (r_2, a_2) \in \mathrm{has\_affiliation} \\ & \mathrm{ac}(a_1) = \mathrm{ac}(a_2) \\ 0 & \mathrm{otherwise} \end{cases}$$

Here, authors with academic and non-academic affiliations are then considered institutional proximate to all other authors, since they are subject to regulations and laws from the academic and private sector.

Organizational proximity

In this work we measure organizational proximity in multiple ways. First, we argue that researchers share a relevant amount of organizational relations if they work at the same affiliation. Here, we consider affiliations from the academic and the private sector. In detail, the similarity score of two researchers is the amount of shared affiliations:

$$\mathrm{sim}_{\mathrm{Affiliation}}(r_1, r_2) := |\{a \in A \mid (r_1, a), (r_2, a) \in \mathrm{has\_affiliation}\}|$$

Second, we are using the distance of the respective web pages as representative for organizational proximity. Partnerships, internal hierarchies, individual researchers and projects of an organisation are usually reflected in their web appearance. Therefore, the hyperlink structure is used to derive a representation of organizational proximity for individual researchers. Overall three similarity functions are build upon web data.

The first one simply matches hosts: Authors are considered proximate, if they have a homepage on the same host. Since authors can have multiple homepages, we evaluate their similarity by counting joint hosts. More formally, let $\mathrm{host}(u)$ be the host for a URL $u \in U$. Then, we can compute the similarity of two researchers via:

$$\mathrm{sim}_{URL}(r_1, r_2) := |\{\mathrm{host}(u) \mid (r_1, u) \in \mathrm{has\_homepage},$$
$$\exists (r_2, v) \in \mathrm{has\_homepage} : \mathrm{host}(u) = \mathrm{host}(v)\}|$$

The next similarity function is an extension of sim_{URL} and abstracts the distance between homepages u_1 and u_2 to the distance of their respective hosts. A connection between two hosts h_1 and h_2 exists, if any page (corresponding to a URL) from host h_1 contains a hyperlink to a URL from h_2. Then distances are computed as the shortest path distance $d_{\mathrm{Hyperlink}}(u_1, u_2)$ between two URLs $u_1, u_2 \in U$ using their respective hosts. We scale these distances to be between 0 and 1. To compute the distance $d_{\mathrm{Hyperlink}}(r_1, r_2)$ between researchers $r_1, r_2 \in R$, we average the distances between their URLs. This accounts for the

fact, that authors may have multiple homepages and we capture the authors' multi-presence in the academic landscape. We define the corresponding similarity function via

$$\text{sim}_{\text{Hyperlink}}(r_1, r_2) := 1 - d_{\text{Hyperlink}}(r_1, r_2).$$

Finally, the third similarity function is based on the hierarchy of web pages. Here we assume, that the hierarchy of the web is reflect in the URL paths. The distance between two authors is expressed as the shortest path between their homepages u, whereas it is only allowed to 'climb' or 'descend' in the hierarchy. For example, the distance between the URLs *host.example.com/path1/author1* and *host.example.com/path2/author2* would be 4. After climbing to the node *host.example.com*, which are two steps, we then descend to the target node with two more steps. Overall, we compute the distance $d_{\text{hierarchy}}(r_1, r_2)$ of two researchers $r_1, r_2 \in R$ by their shortest connection. Again, we scale the distances and compute the corresponding similarities of two researchers r_1 and r_2 as mean of all distances via

$$\text{sim}_{\text{Hierachy}}(r_1, r_2) := 1 - d_{\text{Hierachy}}(r_1, r_2).$$

Assuming that universities have their own hosts and that university departments are placed as subdomains, this similarity function allows us to measure proximity on an intra-university level. Hence it can be seen as a refinement of the similarity function $\text{sim}_{\text{Affiliation}}$.

Social proximity

We present different similarity functions to approximate social proximity. As first approximation, we have a binary indicator stating two researchers as close with respect to social proximity if they have finished their PhD at the same affiliation. We expect that actors with the same roots (in terms of dissertation) will stay in contact and communicate (Burris, 2004). The similarity function is given by

$$sim_{\text{Diss\_Loc}}(r_1, r_2) := \begin{cases} 1 & \text{if } \text{Diss\_Loc}(r_1) = \text{Diss\_Loc}(r_2) \\ 0 & \text{otherwise} \end{cases},$$

where Diss_Loc maps researchers $r \in R$ to their dissertation location.

For a second approach to measure social proximity, we compute the co-author relation $\{(r_1, r_2) \in R \times R \mid \exists p \in P : (r_1, p), (r_2, p) \in \text{is\_author\_of}\}$ and build the co-author graph from this relation. Following the explanation of the relationship between social proximity and co-author graphs in section "Dimensions of proximity", we argue that the social proximity of two researchers can be approximated by closeness in the co-author graph. However, quantifying the similarity of researchers in the co-author network is not straightforward. The naive approach would be to use the shortest path distance as a measure for social distances.[2] However, since the shortest path distance produces a low amount of different values (Watts, 2003), it would provide only a very shallow insight into similarity and distances. To overcome this issue, we use embedding methods. These methods use the structure of the graph to compute vector representations of nodes which incorporate multiple aspects of the surroundings of each node, for example, overlapping neighborhoods, direct connections or similar roles of nodes. To quantify the similarity of two researchers in the co-author network, we use *node embeddings* based on the co-author graph. Having an embedding of the researchers at hand, we again compute the corresponding similarity

---

[2] And thus as a measure for similarity under the common assumption that a lower distance corresponds to a higher similarity.

via the cosine similarity of the embedding vectors. We focus on the following embedding techniques. The first embedding technique is *DeepWalk* (Perozzi et al., 2014). It was the first work that applied the skip-gram (SG) approach (Mikolov et al., 2013b, Mikolov et al., 2013a) on graphs. The SG approach was originally designed for word embeddings. Here, sentences are given as input to a two-layer network that is then trained to predict for a given word the words around it. Afterwards, the weights of the first layer matrix are used as vector representations. DeepWalk transfers this procedure to graphs in the following manner: For a given graph $G = (V, E)$, the nodes $v \in V$ are treated as the 'words' of the vocabulary. To generate embeddings, 'sentences' of nodes are generated via random walks. Today, DeepWalk is regularly used for node embeddings, since the embeddings (1) are adaptable to the emergence of new edges, (2) are capable of providing vectors where the measurement of similarity is meaningful and (3) have proven to outperform handcrafted node features in classification tasks (Perozzi et al., 2014).

The second embedding technique is *node2vec* (Grover & Leskovec, 2016), which extends Deepwalk with two parameters $p, q \in \mathbb{R}_{\geq 0}$ that allow to bias the random walk procedure. While a low p value corresponds to walks that prefer a "breadth-first behavior", low q values bias the walk towards a "depth-first behavior". The additional parameters have proven to enlighten DeepWalk in common machine learning tasks such as node classification or link prediction. As a drawback, the two additional parameters have to be chosen reasonable or have to be determined via grid-search. In our analysis, we use the p and q parameters to generate embeddings with distinctive, but reasonable properties to capture different aspects of the underlying graph. For this, we refer to (Grover & Leskovec, 2016), where the authors do parameter searching in $\{0.25, 0.5, 1, 2, 4\}$ for p and q. Following this, we generate two different embeddings and thus two corresponding similarity functions by the "extreme" choices of $p = 0.25, q = 4$ and $p = 4, q = 0.25$.

The third embedding technique is *HOPE* (Ou et al., 2016). While DeepWalk and node2vec use "sentences of nodes" as input to compute the embeddings, HOPE encodes the input graph via a similarity matrix and then computes the vector representations via factorization. Here, different similarity matrices that incorporate different information of the graphs are possible. To generate embeddings of the co-author graph, we employ the Katz–Index similarity (Katz, 1953), which is defined in the following manner. Let $G = (V, E)$ be a graph with adjacency matrix $A \in \mathbb{R}^{n \times n}$, $\beta \in \mathbb{R}_{>0}$ and let I_n be the n-th dimensional identity matrix. The similarity matrix with respect to β is then given by $S_\beta(G) := (I_n - \beta A)^{-1} \beta A$. For a given $d \in \mathbb{N}$, HOPE then computes matrices $U, V \in \mathbb{R}^{n \times d}$ such that $S_\beta \approx UV$. To get a vector representation of a node, the corresponding rows of U and V are concatenated.

Having the embedding methods at hand, we use the co-author graph to compute functions that map researchers to vectors. We denominate the resulting functions with $f_{\text{DeepWalk}}, f_{\text{Node2vec-small-p}}, f_{\text{Node2vec-large-p}}$ and f_{HOPE}. For each $m \in \{$ DeepWalk , Node2vec-small-p , Node2vec-large-p , HOPE $\}$ and for all researchers $r_1, r_2 \in R$, we define a corresponding similarity function via

$$\text{sim}_m(r_1, r_2) := \frac{\langle f_m(r_1), f_m(r_2) \rangle}{\|f_m(r_1)\| \, \|f_m(r_2)\|}$$

Geographic proximity

As explained in previous section, geographic proximity reflects the spatial separation of two actors. We simplify it by considering the longitudinal and lateral coordinates of the city associated with the affiliation of the author. As a measure of separation of two cities, we use the great-circle distance d_{Geo}. Furthermore, all similarity scores are scaled between 0 and 1. The similarity score of two researchers $r_1, r_2 \in R$ is is then calculated by

$$\text{sim}_{\text{Geo}}(r_1, r_2) = 1 - d_{geo}(r_1, r_2).$$

Ranking hypotheses about the origin of collaboration

The methods introduced in the previous section allow to quantify proximity between authors in various dimensions. Now, we present an approach, namely HypTrails, that allows to rank and compare these dimensions with respect to their impact on the emergence of cooperation in the form of joint publications and patents. In the following, we will shortly repeat the basics of HypTrails and introduce the needed adoption for our setting.

HypTrails is a descriptive approach which was originally developed to compare different hypotheses about human navigation. Human navigation can be represented by any kind of sequences, such as geographical movement (Singer et al., 2015), tagging behavior (Niebler et al., 2016b) or click trails in the web (Niebler et al., 2016a). Using the example of web trails, possible navigation behaviors could be browsing, e.g. clicking links randomly or links to items with a discount, or searching, where the clicks leads to a specific target page (Koopmann et al., 2019). Given a click sequence made by a user, HypTrails is able to compare different intuitions of navigation. More generally, HypTrails generates a ranking of hypotheses $\mathcal{H} = \{H_1, H_2, \ldots, H_n\}$ with respect to their plausibility for the observed data D. Here, D represent the users click trails, which are transformed into an adjacency matrix N. Each entry $N_{i,j}$ in the matrix expresses the amount of observed transitions between discrete states $S = \{s_1, s_2, \ldots, s_m\}$. Therefore HypTrails leverages the first order Markov Model and hence ignores second level dependencies when creating the matrix. In the web navigation example, the states represents web pages and the entries represents the normalized amount of *clicks* between pages by users. Furthermore, each hypothesis $H \in \mathcal{H}$ is expressed by its own transition matrix Q constructed on the belief of users on a specific navigation behaviour property. The "browsing" hypothesis could be a uniform distribution to express the random clicking behaviour or a matrix with high transition probability for clicks to discounted items.

For a given hypothesis H, HypTrails uses Bayesian inference to calculate the marginal likelihood $P(D|H)$, also called *evidence*, with respect to the observed data D. The input of the model are two adjacency matrices. Q is representing a hypothesis (also called prior) and N is created from the data transitions. To calculate the evidence, Hyptrails adapt the Trial-roulette method (Gore, 1987). This method incorporates a *concentration factor k*, which displays the belief in the given hypothesis. Here, k reflects the ratio of uniform distributed transitions and transitions that directly follow the hypothesis. Hence, k indicates how strongly the underlying hypothesis influences the transitions. For more detail on the computation of the evidences, we refer the reader to Singer et al. (2015).

For a given k, the marginal likelihoods $\{P(D|H) \mid H \in \mathcal{H}\}$ are used to generate an order of the hypotheses $\mathcal{H}$. More precise, hypotheses $H_1, H_2 \in \mathcal{H}$ can be compared with the

Bayes factor $B_{1,2} = \frac{P(D|H_1)}{P(D|H_2)}$. However, determining an appropriate k is challenging. Therefore, we use a range of concentration factors to compare a set of hypotheses $\mathcal{H}$.

Our setting HypTrails originally analyses sequential navigation data. Graphs can be interpreted as a generalization of sequences. Therefore, Espín-Noboa et al. (2017) showed, that the transition probabilities required for HypTrails can also be inferred from so-called attributed multi-graphs. Here an attributed multi-graph describes an undirected, weighted graph, which contains descriptive attributes for its nodes. Since co-author and co-inventor can be interpreted as such graphs, HypTrails can be applied as follows: The discrete states are given by the researchers $R = \{r_1, \ldots, r_m\}$ and the transitions $N_{i,j}$ of matrix N are collaborations (co-authorships or co-inventions) between the researchers r_i and r_j. Furthermore, the matrices corresponding to the hypotheses $\mathcal{H}$ can be extracted by leveraging attributes for author pairs, which in our case are the values from the similarity functions presented in section "Measuring and quantifying proximity". More specifically, for a similarity function sim from Table 1, we derive a hypothesis H with corresponding adjacency matrix Q via $Q_{i,j} = \text{sim}(r_i, r_j)$. Here, for a given similarity function sim, the belief of the hypothesis $H \in \mathcal{H}$ is, that a high similarity value $\text{sim}(r, s)$ for researchers $r, s \in R$ indicates the emergence of collaboration between r and s. In the following, we will not differentiate anymore between a similarity function and the corresponding hypothesis and thus use the terms interchangeably.

The result is a ranked list of hypotheses based on the extent to which they indicate an influence on collaboration. By selecting one hypotheses for each dimension of proximity, our approach allows to rank these dimensions.

Empirical context: the German AI community

To measure proximity dimensions and their impact on collaboration in the German AI landscape, we employ a data foundation that captures publication activities and inventions in this community.

The German AI network

The German AI Network (GAI) is a bibliometric data set of German AI researchers and their publications. The GAI is publicly available via Zenodo (Stubbemann & Koopmann, 2020). It is built upon the DBLP data set (Ley, 2009),[3] which contains bibliographic information of publications in the realm of computer science. DBLP is, in our experience, impressively tidy and consistently structured, especially for the amount of covered data. Furthermore, it contains information about conference proceedings, which are the predominant publishing channels in computer science. To identify the academic authors that belong to the domain of artificial intelligence, we rely on the work by Kersting et al. (2019), which provides a collection of central AI venues and the relevant venues that belong to the expanded environment of AI. Furthermore we identify German authors by using the affiliations provided by DBLP. In detail, we create the data set as follows:

---

[3] We use the DBLP dump from 2020-01-01, which can be found at https://dblp.org/xml/release/dblp-2020-01-01.xml.gz.

Table 2 Basic statistics of the German AI Network. We display, from left to right, (1) Number of German AI authors, (2) Number of publications (3) Number of publications on AI venues (4)Number of authors with at least one collaboration (5)Number of co-authorships

Authors	Publications	AI publications	Coll. authors	Co-author-ships
2131	127,780	11,344	1937	6064

- First, we extract all publications that were published in one of the venues mentioned in Kersting et al. (2019).
- Next, we identify all authors of these papers as (international) AI authors.
- Afterwards, we filter the German authors by using the given affiliations from DBLP. In detail, we search for the substring *Germany* in each affiliation. We discovered that well-formated affiliations are formed such as *University of Kassel, Germany*, for example.
- The data set consists of all authors identified in the step above and all of their respective publications, not limiting to publications of AI venues.

We optimize the construction of the data set and matching of German authors towards high precision. Accordingly, authors are included only if they are identified with a high level of confidence as German AI authors. This results in a comparatively smaller subset of the German AI community, but a small to non-existent number of false positives. While DBLP provides comparatively well disambiguated bibliographic data, it does not contain information of citations or abstracts. To enrich our data with the latter, we link it with the Semantic Scholar Corpus (Ammar et al., 2018).[4] The linking process primarily relies on titles. However, as different papers with equal titles can result in false positives, we add further linking constraints such as DOIs, years and the fact whether both publications are pre-prints or not. Basic statistics of the GAI can be found in Table 2.

The German AI inventors

We define the German AI inventors as authors who work and publish in the academic domain of artificial intelligence, which we covered with the GAI data set, and also contribute to technological change as patenting inventors. To extract these author-inventors, we use the GAI (defined above) as our starting point. We link this data set, based on names, with CIROS-Patstat[5] (Tarasconi, 2014) only considering inventors listed with a German residence.[6] This approach yields 423 possible candidates with the same name in both data sets. As before, we aim towards a high precision and hence want to ensure to have as few false positives in our sample as possible. This is achieved by analyzing the publication neighborhood. More specifically, we compare co-inventors and co-authors for each candidate. If a candidate has at least one identical co-author and co-inventor (as before based on names), we consider this candidate as a true positive. This approach leads to 212 individuals. While this is a relatively small number, given our search strategy it represents the subset of publishing AI researchers in Germany who also patent. This focus

---

[4] We use the Semantic Scholar dump from 2020-01-01.
[5] CRIOS – Patstat provides a disambiguated data set based on the European Patent Office.
[6] We prepossess the date and also respect German umlauts.

Table 3 Basic statistics of the German AI Inventors. We display, from left to right, (1) Number of German AI inventors (2) Number of inventions (3) Number of inventions with at least two inventors (4) Number of inventors with at least one collaboration (5)Number of co-inventorships

Inventors	Innovations	Co-inventions	Coll. inventors	Co-inventor-ships
212	1,007	55	72	92

on author-inventors is consistent with our interest in research collaboration, as it ensures that both collaborating partners are active researchers. Table 3 provides a summary of the resulting data. We name the resulting data set the German AI inventors (GAI-I).

The German academic web

We complement the previously mentioned data sets with web information, which allows to to capture additional information about AI researchers. In contrast to the bibliometric data, the German Academic Web (GAW) reflects an expression of the hierarchical organizational structure. At the same time, the web allows everybody to freely display their research interests and link themselves with other authors.

Overall, the GAW data set (Paris & Jäschke, 2020) has been created in an effort to establish a knowledge base of the academic landscape in Germany. It is an accumulation of semi-annual crawls since 2012, containing web pages of 150 major universities and research institutions. To match and find home pages of authors, we use URLs given by DBLP and reduce them by matching to GAW crawl seeds[7]. This allows us to remove hosts such as "gitlab" or "Orcid", which do not represent the homepage of an author. With this process we are able to find 1,163 researchers of the GAI in the GAW data set.

Analysis: ranking of proximity dimensions

Previous sections explained how we measure different forms of proximity, how we intend to compare them and which data we use. In this section, we describe our analysis which compares the proximity dimensions with respect to the extent they indicate collaboration.

Analytical setup

The aim of the analysis is to compare different hypotheses for the emergence of collaboration between different researchers. Every similarity function corresponds to a hypothesis H and the resulting pairwise similarity scores are used as entries for the prior adjacency matrix Q. The similarity functions introduced earlier depend on different hyperparameters. We display the names of the used hypotheses, the similarity functions and the used parameters in Table 4. The hyperparameters follow default choices of Scikit-Learn or from the papers where the methods were introduced. Overall, we run two different analysis. First, we study academic collaboration, where we consider co-published publications and secondly,

---

[7] We use the most recent snapshot from 2019/12.

Table 4 Overview of the hypotheses. For each hypothesis, we display the corresponding similarity function and the choice of hyperparameter, if needed. d: dimensions of the embedding, w: window size, γ: walks per node, t: walk length, p, q: node2vec bias parameters, β: parameter for HOPE embedding

Proximity dim.	Hypothesis	similarity function	parameter settings	used data
Cognitive	LSA	sim_{LSA}	$d = 100$	GAI
	NMF	sim_{NMF}	$d = 100$	GAI
	BERT	sim_{BERT}	$d = 768$	GAI
Institutional	Academic	$\text{sim}_{\text{aff\_type}}$	–	GAI
Organizational	Affiliation	$\text{sim}_{\text{Affiliation}}$	–	GAI
	URL		–	GAI,GAW
	Hyperlink	$\text{sim}_{\text{Hyperlink}}$	–	GAI,GAW
	Hierarchy	$\text{sim}_{\text{Hierachy}}$	–	GAI,GAW
Social	Diss Loc	$\text{sim}_{\text{Diss\_Loc}}$	–	GAI,DNB
	DeepWalk	$\text{sim}_{\text{DeepWalk}}$	$d = 100, w = 10, \gamma = 10,$ $t = 80$	GAI
	Node2vec Small p	$\text{sim}_{\text{Node2vec-small-p}}$	$d = 100, w = 10, \gamma = 10,$ $t = 80, p = 0.25, q = 4$	GAI
	Node2vec Large p	$\text{sim}_{\text{Node2vec-large-p}}$	$d = 100, w = 10, \gamma = 10,$ $t = 80, p = 4, q = 0.25$	GAI
	Hope	sim_{HOPE}	$d = 100, \beta = 0.1$	GAI
Geographic	Geo	sim_{Geo}	–	GAI, Wikidata

we analyze collaboration in terms of co-patentships. Some hypotheses use external information, which are not in the GAI. Here, we briefly explain the underlying data sources. The hypothesis *Diss Loc* leverages dissertation data from DNB[8]. Here, we are able to collect 1035 relevant dissertations. Hypotheses that are based on web data use the linked URls from DBLP with the GAW. Geographic distances between authors are computed using the coordinates of German cities, which we extract from Wikidata.[9]

Setup for the analysis of co-authorships. The hypotheses in the realm of social proximity, namely the graph embeddings, are build upon the co-author graphs and hence rely on the co-authorships we want to explain in our first analysis. Additionally, future collaborations should just be explained with data from previous co-authorships and publications. To tackle this problems, we split the publications and hence the co-authorships into two parts. We use all publications published until 2017 to create the hypotheses that are build upon text and graph embeddings. To create the observed transition matrix N, which contains the collaboration we want to analyse, we use the weighted co-author graph of 2018 and 2019. *Setup for the analysis of co-inventorships.* For the second analysis, we compare hypotheses with respect to the question, to which extent they describe co-patentships. Here, the matrix N of the observed transitions is given by the adjacency matrix of the weighted co-inventor graph. Due to the linking process, we have pairwise similarity scores for all inventors in the co-inventor graph. Therefore we

---

[8] The DNB is the German National Library. The data set is based on Heinisch and Buenstorf (2018), and supplemented from the DNB homepage if necessary.

[9] https://query.wikidata.org/ on 2020-05-07.

can use the same hypotheses $\mathcal{H}$ as in the first analysis with one difference. No overlap between data for the hypotheses and the observed data has to be prevented. In consequence, we omit the time split from our fist analysis. Only a small subset of German AI researchers do indeed collaborate over inventions. This can be seen when comparing GAI and GAI-I with 1, 937 AI authors which collaborate over co-publications and only 72 authors that collaborate over co-inventions. Since the aim of this work is to find key factors of collaboration, we only consider these 72 inventors for the data matrix and all hypotheses. In detail, for a given hypothesis H, represented by a adjacency matrix N, we omit the rows and columns that correspond to authors without any edges in the co-inventorship graph. *Evaluation setup* As a baseline hypothesis, we use a **Random Co-author** hypothesis, which is build by assuming that researchers choose their co-authors randomly. This works as follows. For a publication with l authors and for each of its authors we create an artificial publication with $l - 1$ randomly selected co-authors. We scale all plots with respect to this baseline. To clearly arrange the results, we first analyze each proximity dimension separately. Then, we use the hypothesis with the highest evidence as representative for each dimension of proximity and compare the different dimensions of proximity. Since institutional and geographical proximity only have one hypothesis, we only include them in the comparison over the different dimensions. When comparing the different dimensions of proximity, we also include the **True** hypothesis, which represents the actual transitions as hypothesis. Hence, it is the best possible explanation and can be understand as ground truth.

Analyzing co-authorships

Figure 1 shows the evidence scores over a selected range of k-values for different hypothesis. Figure 1a compares hypotheses that belong to cognitive proximity. For all concentration factors, the **LSA** hypothesis extracts cognitive proximity in a way, which yields the highest evidence scores. It is followed by the **NMF** and **BERT** hypotheses. Figure 1b shows the analysis of organizational proximity. For small k values, the **Affiliation** and **URL** hypothesis have the highest evidence scores. When increasing k, hence believing more in every transitions stated by the hypotheses, the **Affiliation** hypothesis drops below **URL** and **Hyperlink**. Interestingly, the **Hierachy** hypothesis has the lowest evidence scores overall. Social proximity is depicted in Fig. 1c. Here, a gap between graph based hypotheses (upper four hypotheses) and the **Diss Loc** hypothesis can be observed. The **Deepwalk** hypotheses has overall the highest evidence, followed by both node2vec approaches and **HOPE**. Finally, Fig. 1d shows the comparison of all proximity dimensions. Overall, the representative of social proximity yields the best explanation of collaboration. It is followed by the representative of cognitive proximity and organisational proximity. Institutional and geographic proximity conclude the ranking.

Analysing co-inventorships

Figure 2 shows the results for cognitive proximity. The results are similar to the previous analysis. The **LSA** hypothesis yields overall the highest evidence, followed by **NMF** and **BERT**. The organizational proximity is depicted in Fig. 2b. Here, all hypotheses lead to comparatively low evidence scores and even drop below the baseline for high k-values. For low k-values the **Affiliation** hypothesis ranks highest. On the other hand,
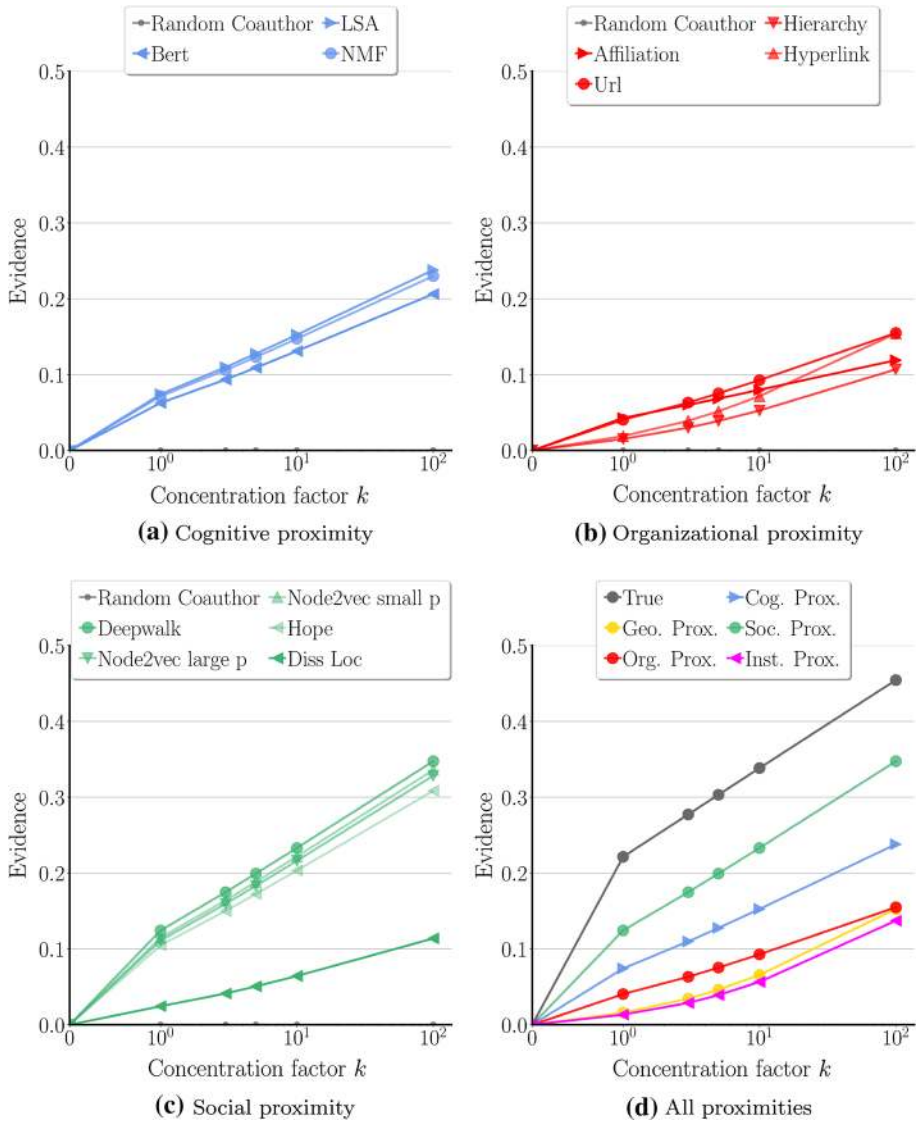
Fig. 1 HypTrail graphs for co-author graphs. For Fig. 1d we choose the **LSA** hypothesis for cognitive proximity, **URL** hypotheses as representative for organizational proximity and **Deepwalk** hypothesis for social proximity. Institutional and geographic proximity are represented by the only one hypothesis each

for high *k*-values the **Hyperlink** hypothesis has the highest evidence, followed by **URL**, **affiliation** and **Hierachy** hypothesis. The results for social proximity are depicted in Fig. 2c. Similar to the previous analysis, the **Diss Loc** is the worst indicator for collaboration. The highest evidence is achieved by the **Hope** hypothesis. Finally, Fig. 2d shows the combined view over the different proximities. As for the co-authorship, social
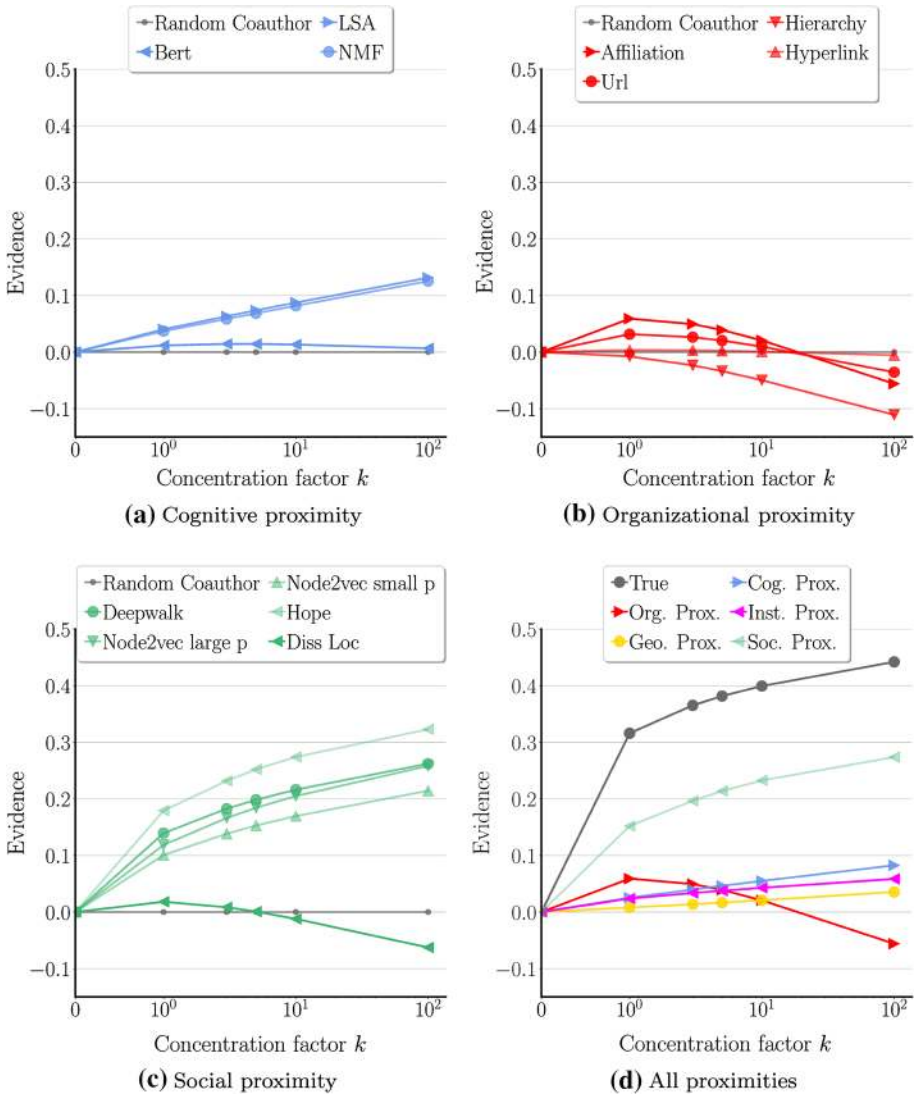
Fig. 2 HypTrail graphs for co-inventorships. For Fig. 2d we choose the **LSA** hypothesis as representative for cognitive proximity, the **affiliation** hypothesis for organizational proximity and **Hope** hypothesis for social proximity. Institutional and geographic proximity are represented by only one hypothesis each

proximity is the best indicator for collaboration. In contrast to previous analysis, for low k values organizational proximity represented by the **Affiliation** hypothesis achieves the second highest evidence. When increasing the believe (higher k factors), cognitive, institutional and geographic proximity have higher evidences.

Discussion

For both forms of collaboration, social proximity serves as the best indicator, followed by cognitive proximity. Both dimensions lead to higher results then geographic proximity, indicating that social connection between actors and similar knowledge are more important for collaboration than being co-located. This is consistent with prior findings that the role of geography mostly derives from the localization of social networks (Breschi and Lissoni, 2009). Its relevance may be further reduced by the increasing adoption of communication technologies. Our findings do not suggest that organizational proximity is an important driver of collaboration. However, our ability to analyze hypotheses based on web data is limited since we could not collect web pages for all researchers. URLs for only 1163 of 2131 researchers can be found in the GAW. This effect amplifies in the second analysis, which only considers a subset of German AI researchers. The **Diss Loc** hypothesis suffers from the same issue, where we found dissertation locations of 1035 researchers. Furthermore, we notice differences when comparing the ranking of proximities between co-authorship and co-inventorship. Organizational proximity seems to be a poor indicator of co-inventorship, finding even less support than the hypothesis that co-inventorships are chosen randomly. Furthermore, for describing co-inventorship, institutional proximity is more important than geographic proximity, which does not hold for co-authorships. This result is plausible because we expect co-inventorships to be more common in industrial contexts, where both inventors are connected to a non-academic affiliation. Another issue influencing the results are missing time-stamps for affiliations and homepages. This leads to hypotheses being built upon all data. Additionally, in the case of co-inventorships no temporal split was made because the overall number of collaborations is rather small. Our analysis primarily serves to identify associations between various forms of proximity and collaboration, whereas it is not designed to find key factors for the identification of future collaboration. Finally, various alternative representations for the different dimensions of proximity have not yet been evaluated. For example, social media data can be used to create reasonable similarity functions for social proximity.

Conclusion

In this study we presented methods to identify relationships between different dimensions of proximity and the emergence of collaboration. For each dimension, we proposed several methods to quantify the similarity of two researchers. These methods were used to create hypotheses about the origin of collaboration. To compare them with respect to their plausibility, we adapted the HypTrails approach. For our analyses, we used a novel data set of 2131 German AI researchers. By linking author data with web data, we were able to compute similarity scores between authors based on their web presence. With these linked data, we analyzed two forms of collaboration, namely co-authorship and co-inventorship. For the latter, we additionally linked our data with patent date from CRIOS Patstat. Our findings suggest that social proximity is the key factor to explain collaboration.

In future work, we plan to investigate additional forms of proximity and collaboration. For example, present-day communication and interaction often relies on social media. Hence, hypotheses based on data from such sources would lead to further interesting proximity measures. Furthermore, an interesting aspect is the influence of funded projects. Further research can study whether such projects indeed lead to more co-publications.

Additionally, joint projects can be interpreted as a form of collaboration. Hence, further investigation could tackle the question which dimensions of proximity are relevant for joint projects.

Availability of data and material The data is available on Zenodo.Code availability Upon request.

Declaration

Conflicts of interests None.

References

Aghion, P., Dewatripont, M., & Stein, J. C. (2008). Academic freedom, private-sector focus, and the process of innovation. *The RAND Journal of Economics,39*(3), 617–635.

Ammar, W. et al. (2018). "Construction of the Literature Graph in Semantic Scholar." In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan- guage Technologies, Volume 3 (Industry Papers). New Orleans - Louisiana: Association for Computational Linguistics, pp. 84–91.

Audretsch, D. B., & Feldman, M. P. (1996). R&D spillovers and the geography of innovation and production. *The American Economic Review,86*(3), 630–640.

Balland, P.-A. (2012). Proximity and the evolution of collaboration networks: Evidence from research and development projects within the global navigation satellite system (GNSS) industry. *Regional Studies,46*(6), 741–756.

Beltagy, I., K. Lo, and A. Cohan (2019). "SciBERT: Pretrained Language Model for Scientific Text." In: EMNLP.

Bode, R., G. Buenstorf, and D. P. Heinisch (2019). "Proximity and learning: evidence from a post-WW2 intellectual reparations program." In: Journal of Economic Geography. lbz023.

Boschma, R. (2005). Proximity and innovation: A critical assessment. *Regional Studies,39*(1), 61–74.

Breschi, S., & Lissoni, F. (2009). Mobility of skilled workers and co-invention networks: An anatomy of localized knowledge flows. *Journal of Economic Geography,9*(4), 439–468.

Broekel, T., & Boschma, R. (2011). Knowledge networks in the Dutch aviation industry: The proximity paradox. *Journal of Economic Geography,12*(2), 409–433.

Buenstorf, G., & Klepper, S. (2010). Why does entry cluster geographically? Evidence from the US tire industry. *Journal of Urban Economics,68*(2), 103–114.

Burris, V. (2004). The academic caste system: Prestige hierarchies in PhD exchange networks. *American Sociological Review,69*(2), 239–264.

Catalini, C. (2018). Microgeography and the direction of inventive activity. *Management Science,64*(9), 4348–4364.

Crescenzi, R., Nathan, M., & Rodríguez-Pose, A. (2016). Do inventors talk to strangers? On proximity and collaborative knowledge creation. *Research Policy,45*(1), 177–194.

Deerwester, S. C., Deerwester, Scott, Dumais, Susan T., Furnas, George W., Landauer, Thomas K., & Harshman, Richard. (1990). Indexing by latent semantic analysis. *JASIS,41*(6), 391–407.

Delgado, M., Porter, M. E., & Stern, S. (2010). Clusters and entrepreneurship. *Journal of Economic Geography,10*(4), 495–518.

Devlin, J. et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Lin- guistics: Human Language Technologies, Volume 1 (Long and Short Pa- pers). Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171-4186.

Edquist, C. and B. Johnson (1996). Institutions and organizations in systems of innovation. Univ.

Espín-Noboa, L., et al. (2017). JANUS: A hypothesis-driven Bayesian approach for understanding edge formation in attributed multigraphs. *Applied Network Science,2*(1), 16.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers,28*(2), 197–202.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes,25*(2–3), 285–307.

Glänzel, W. and A. Schubert (2005). "Analysing Scientific Networks Through Co-Authorship." In: Handbook of Quantitative Science and Technology Re- search, pp. 257-276.

Gore, S. M. (1987). Biostatistics and the medical research council. *Medical Research Council News,35*, 19–20.

Grover, A. and J. Leskovec (2016). "node2vec: Scalable Feature Learning for Networks." In: Proceedings of the 22nd ACM SIGKDD International Con- ference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM, pp. 855-864.

Hardeman, S., et al. (2015). Characterizing and comparing innovation systems by different 'modes' of knowledge production: A proximity approach. *Science and Public Policy,42*(4), 530–548.

Hautala, J. (2011). Cognitive proximity in international research groups. *Journal of Knowledge Management,15*(4), 601–624.

Heinisch, D., et al. (2016). Same place, same knowledge-same people? The geography of non-patent citations in Dutch polymer patents. *Economics of Innovation and New Technology,25*(6), 553–572.

Heinisch, D. P., & Buenstorf, G. (2018). The next generation (plus one): An analysis of doctoral students' academic fecundity based on a novel approach to advisor identification. *Scientometrics,117*(1), 351–380.

Heinze, T., et al. (2009). Organizational and institutional influences on creativity in scientific research. *Research Policy,38*(4), 610–623.

Hirv, T. (2018). Effects of European union funding and international collaboration on Estonian scientific impact. *Journal of Scientometric Research,7*, 181–188.

Hoekman, J., K. Frenken, and R. J. Tijssen (2010). "Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe." In: Research Policy 39.5. Special Section on Government as Entrepreneur, pp. 662 -673.

Jaffe, A. B., M. Trajtenberg, and R. Henderson (1993). "Geographic localization of knowledge spillovers as evidenced by patent citations." en. In: The Quarterly Journal of Economics 108.3, pp. 577-598.

Jones, K. S. (1972). "A statistical interpretation of term specificity and its application in retrieval." In: Journal of documentation.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy,26*(1), 1–18.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika,18*(1), 39–43.

Kersting, K., J. Peters, and C. A. Rothkopf (2019). "Was ist eine Professur fuer Kuenstliche Intelligenz?" In: CoRR abs/1903.09516.

Koopmann, T. et al. (2019). "On the Right Track! Analysing and Predicting Navigation Success in Wikipedia." In: Proceedings of the 30th ACM Conference on Hypertext and Social Media. HT '19. New York, NY, USA: ACM, 143-152.

Le, Q. and T. Mikolov (2014). "Distributed Representations of Sentences and Documents." In: Proceedings of the 31st International Conference on Inter- national Conference on Machine Learning - Volume 32. ICML'14. Beijing, China: JMLR.org, II-1188-II-1196.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature,401*(6755), 788–791.

Lee, D. D., & Seung, H. S. (2000). Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000* (pp. 556–562). Den- ver, CO: USA. MIT Press.

Lee, J.-H., et al. (2009). Automatic generic document summarization based on non-negative matrix factorization. *Information Processing and Management,45*(1), 20–34.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science,35*(5), 673–702.

Ley, M. (2009). DBLP: Some lessons learned. *Proceedings of the VLDB Endowment,2*(2), 1493–1500.

Liu, J., et al. (2018). Understanding the advisor-advisee relationship via scholarly data analysis. *Scientometrics,116*(1), 161–180.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to infor- mation retrieval*. Cambridge: Cambridge University Press.

Marshall, A. (1890). *The Principles of Economics*. Tech. rep: McMaster University Archive for the History of Economic Thought.

Mikolov, T., (2013a). Distributed Representations of Words, and Phrases, and their Compositionality. In: Advances in Neural Information Process- ing Systems 26: 27th Annual Conference on Neural Information Processing Systems, et al. (2013). *Proceedings of a meeting held December 5–8, 2013* (pp. 3111–3119). Nevada, United States: Lake Tahoe.

Mikolov, T. et al. (2013b). "Efficient Estimation of Word Representations in Vector Space." In: 1st International Conference on Learning Representa- tions, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences,101*(suppl 1), 5200–5205.

Newman, M. (2001). "Scientific collaboration networks. I. Network construction and fundamental results." In: Physical review. E, Statistical, nonlinear, and soft matter physics 64, p. 016131.

Niebler, T., et al. (2016a). Extracting Semantics from unconstrained navigation on wikipedia. *KI - Künstliche Intelligenz,30*(2), 163–168.

Niebler, T. et al. (2016b). "FolkTrails: Interpreting navigation behavior in a social tagging system." In: International on Conference on Information and Knowledge Management. CIKM '16. New York, NY, USA: ACM, pp. 2311- 2316.

Nooteboom, B. (2001). *Learning and innovation in organizations and economies*. Oxford: Oxford University Press.

Ou, M. et al. (2016). "Asymmetric Transitivity Preserving Graph Embedding." In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM, pp. 1105-1114.

Paris, M. and R. Jäschke (2020). Summary GAW.

Perkmann, M., et al. (2013). Academic engagement and commercialisation?: A review of the literature on university - industry relations. *Research Policy,42*(2), 423–442.

Perozzi, B., R. Al-Rfou, and S. Skiena (2014). "DeepWalk: online learning of social representations." In: The 20th ACM SIGKDD International Confer- ence on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014. ACM, pp. 701-710.

Polanyi, M. (1966). *The tacit dimension. English*. Garden City, NY: Anchor.

Ponds, R., Van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration*. *Papers in Regional Science,86*(3), 423–443.

Singer, P. et al. (2015). "HypTrails: A Bayesian Approach for Comparing Hypotheses About Human Trails on the Web." In: Proceedings of the 24th In- ternational Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015. ACM, pp. 1003-1013.

Sinoara, R. A., et al. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems,163*, 955–971.

Stern, S. (2004). Do scientists pay to be scientists? *Management science,50*(6), 835–853.

Stubbemann, M., & Koopmann, T. (2020). The German and International AI Network Data Set. *Version,2,*.

Tarasconi, G. (2014). "Crios-Patstat Database: Sources, Contents and Access Rules." In: CRIOS WP.

Vaswani, A., (2017). Attention is all you need. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, et al. (2017). *4–9 December 2017* (pp. 5998–6008). CA, USA: Long Beach.

Watts, D. J. (2003). *Six degrees: The science of a connected age*. Norton, New York: W. W.

Werker, C., Korzinov, V., & Cunningham, S. (2019). Formation and output of collaborations: The role of proximity in German nanotechnology. *Journal of Evolutionary Economics,29*(2), 697–719.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science,316*(5827), 1036–1039.

Xu, W., X. Liu, and Y. Gong (2003). "Document clustering based on nonnegative matrix factorization." In: SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Develop- ment in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada. ACM, pp. 267-273.

Xu, X., et al. (2016). A bibliographic analysis and collaboration patterns of IEEE transactions on intelligent transportation systems between 2000 and 2015. *IEEE Transactions on Intelligent Transportation Systems, 17*(8), 2238–2247.

Authors and Affiliations

Tobias Koopmann[1] · **Maximilian Stubbemann[2]** · **Matthias Kapa[3]** · **Michael Paris[4]** · **Guido Buenstorf[3,5]** · **Tom Hanika[6]** · **Andreas Hotho[1]** · **Robert Jäschke[2,4]** · **Gerd Stumme[2]**

Maximilian Stubbemann
stubbemann@l3s.de

Matthias Kapa
kapa@incher.uni-kassel.de

Michael Paris
michael.paris@hu-berlin.de

Guido Buenstorf
buenstorf@uni-kassel.de

Tom Hanika
hanika@cs.uni-kassel.de

Andreas Hotho
hotho@informatik.uni-wuerzburg.de

Robert Jäschke
robert.jaeschke@hu-berlin.de

Gerd Stumme
stumme@l3s.de

[1]   Data Science Chair, University of Würzburg, Würzburg, Germany

[2]   L3S Research Center, Hannover, Germany

[3]   INCHER and Institute of Economics, University of Kassel, University of Kassel, Germany

[4]   Humboldt-Universität zu Berlin, Berlin, Germany

[5]   Innovation and Entrepreneurship Unit, University of Gothenburg, Gothenburg, Sweden

[6]   Knowledge and Data Engineering Group, University of Kassel, Kassel, Germany