Journal of Informetrics 17 (2023) 101455

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Quantifying knowledge synchronization with the network-driven approach

Jisung Yoon [a,b,c,d], Jinseo Park [e], Jinhyuk Yun [f,*], Woo-Sung Jung [c,g,*]

[a] Kellogg School of Management at Northwestern University, Evanston, IL, 60208, USA
[b] Northwestern Institute on Complex Systems, Evanston, IL, 60208, USA
[c] Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang 37673, Republic of Korea
[d] Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408, USA
[e] Center for Global R&D Data Analysis, Korea Institute of Science and Technology Information, Seoul 02456, Republic of Korea
[f] School of AI Convergence, Soongsil University, Seoul 06978, Republic of Korea
[g] Department of Physics, Pohang University of Science and Technology, Pohang 37673, Republic of Korea

A B S T R A C T

Humans acquire and accumulate knowledge through language usage and eagerly exchange their knowledge for advancement. Although geographical barriers had previously limited communication, the emergence of information technology has opened new avenues for knowledge exchange. However, it is unclear which communication pathway is dominant in the 21st century. Here, we explore the dominant path of knowledge diffusion in the 21st century using Wikipedia, the largest communal dataset. We evaluate the similarity of shared knowledge between population groups, distinguished based on their language usage. When population groups are more engaged with each other, their knowledge structure is more similar, where engagement is indicated by socio-economic connections, such as cultural, linguistic, and historical features. Moreover, geographical proximity is no longer a critical requirement for knowledge dissemination. Furthermore, we integrate our data into a mechanistic model to better understand the underlying mechanism and suggest that the main channel of information distribution in the 21st century is based online.

1. Introduction

Human language and knowledge are fundamentally intertwined and influence one another (Code, 1980). Knowledge, which is defined as the ability to perceive and comprehend a subject, can be obtained through various sources, including memory, education, and practice (Grimm, 2014). Epistemologists traditionally investigated the nature and origins of knowledge. For instance, Immanuel Kant, a prominent epistemologist, claims that human perception is the basis of the general rules of nature that structure all our experiences (Kant, 2000). Because the experience could be different depending on the environments of population groups, knowledge structure can vary based on personalities, the country one lives in, or language profile based on a person's social structure and education system. Humans conventionally acquire information through language, synthesizing knowledge from a flow of sensory experience (Schieffelin & Ochs, 1986). Thus, language profiles may influence the knowledge structure.

Researchers have considered that information can be spread through the mobility of people. For example, until the 16th century, the *Silk Road* had played an important role in the transmission of knowledge between Europe and Asia (Andrea, 2014, Lu et al., 2016). Beyond physical contact between groups, modern information technology offers interactive online resources, such as user comments on a web page, social networks, and internet messengers, which allow knowledge to be transferred. Therefore, the emergence

* Corresponding authors.
E-mail addresses: jinhyuk.yun@ssu.ac.kr (J. Yun), wsjung@postech.ac.kr (W.-S. Jung).

https://doi.org/10.1016/j.joi.2023.101455
Received 24 October 2022; Received in revised form 10 July 2023; Accepted 1 September 2023

Available online 27 September 2023

of information society raises intriguing questions: do social interactions influence the organization of human knowledge in the 21st century? If yes, what is the main contemporary channel of information distribution that geographical boundaries gradually diminished thanks to the advance of technology?

In this study, we attempt to answer the aforementioned questions using the knowledge structure for users of each language, who share their *habitus* inherited from their antecedents. We compare the knowledge structures between languages. The way in which language is used reflects the innate knowledge structure of its users. Thus, we consider the usage of each language as a proxy for its users' *habitus*, and hereafter, we use the term "language" to indicate the collective usage of the language users, unless otherwise specified. Researchers commonly use large scholarly databases as fundamental sources to explore knowledge structure for investigating the mechanisms of scientific innovations. However, these databases are suitable for examining shared knowledge within research communities rather than covering the society in general (Qian et al., 2009, Song & Kim, 2013, Su & Lee, 2010, Hu et al., 2014, Sakata et al., 2012, Fortunato et al., 2018). Although previous studies have achieved the quest of understanding human innovation to some degree, it also necessitates complementary data with more general coverage, including non-scholars. Wikipedia, on the other hand, enables us to construct knowledge structures encompassing general society of specific language users and compare the knowledge structures across different groups.

Here, we use Wikipedia's multi-lingual linkage to evaluate the similarity of knowledge structures among different language editions to track the dominant pathway of contemporary information distribution. First, we construct 59 hierarchical knowledge networks based on the relationship between the pages and categories of each Wikipedia language edition, where each page or category — known as a rich proxy for the knowledge structure (Nastase & Strube, 2008, Schönhofen, 2009, Yoon et al., 2018) — is regarded as a scientific concept. Using a personalized page rank algorithm (Jeh & Widom, 2003), we build *genealogy vectors* for each subject in the knowledge network. Then, using Wikipedia's multi-lingual linkage, we determine subject similarity by comparing the genealogical vectors of each subject among the knowledge networks from different language usage groups. We discover a plausible modular structure of languages comprising multidimensional factors, such as geographical, cultural, linguistic, and historical factors, by aggregating multiple topic similarities between languages into a knowledge structure similarity.

Using this massive knowledge graph, we also discover geographically disassociated interactions, such as cooperative scientific research and social ties, by comparing with other socio-economic data, thereby explaining the synchronization of knowledge structures rather than geographical proximity. Furthermore, we successfully regenerate the similarity of empirical knowledge structures based on various socio-economic ties, supporting our previous observations, and uncover the potential mechanism underlying the synchronization of knowledge structures with the mechanistic model, inspired by the simple synchronization model (Kuramoto, 1975). This study enables us to understand the contemporary *Silk Road* of knowledge dissemination and that virtual social interactions shape the structure of human knowledge, as indicated by the massive records of online-based collaborative knowledge in the form of Wikipedia.

2. Related work

2.1. Knowledge diffusion originated from interactions and knowledge synchronization

Knowledge diffusion necessitates interactions. For instance, the Silk Road was historically one of the most important channels for the transmission of knowledge (Andrea, 2014, Lu et al., 2016). In addition to direct physical interaction, modern information technology enables online interactive materials. Scientific collaboration contributes to the production and dissemination of knowledge (Lambiotte & Panzarasa, 2009). Global student exchanges have played a vital role in the dissemination of contemporary knowledge (Bhandari & Blumenthal, 2011). Active social interactions can facilitate the efficient transfer of information between team or company members (Inkpen & Tsang, 2005, Wu et al., 2007, Ringberg & Reihlen, 2008). Dyadic social capital can promote collaborative knowledge creation (Tu, 2020), and city-level collaboration networks and knowledge networks have a favorable impact on its creativity (Ba et al., 2021). Language is also an essential aspect of the process of knowledge diffusion (Welch & Welch, 2008, Ambos & Ambos, 2009).

One way to understand the mechanism of information spreading (or knowledge diffusion) is a synchronization process (Pluchino et al., 2005, Arenas et al., 2008, Jalili, 2013, Del Chiappa & Baggio, 2015). As an illustrative example, opinion formation is a fascinating phenomenon where individuals' viewpoints can evolve through interactions with others, leading to collective behavior (Boccaletti et al., 2006). This phenomenon can be interpreted as a process of synchronization (Pluchino et al., 2005). In this context, they consider a model comprising N agents, each holding an opinion denoted as x_i (an integer or real number), and these agents are connected through contact networks. The probability of an agent assimilating an opinion is higher if they are connected in the network. This model has proven valuable in comprehending how the underlying network topology influences opinion formation. Similarly, a consensus phenomenon can also be understood as a synchronization process enabling the capture of many general features of social systems (Pluchino et al., 2006). The most popular and classic model of synchronization is the Kuramoto model (Kuramoto, 1975) to model how the entire system synchronizes from their interactions (Jalili, 2013). Using the synchronization framework, the influence of interaction structure on the information-spreading process was also investigated (Kirst et al., 2016, Zhang et al., 2016). Empirically, in the field of health, it is known that knowledge cannot be transferred without persistent interactions and synchronizing processes (Cernada, 2019, Havelock, 1979).

3. Data

3.1. Wikipedia data set

We used Wikipedia SQL dump of 59 different language editions on February 1, 2019. A list of language editions and their abbreviation are provided in Table S1. Wikipedia is considered a representative example of collaborative knowledge, growing through collaboration and competition of contributors; and has been studied to understand the dynamics of collective intelligence (Yasseri, Sumi, & Kertész, 2012, Yasseri, Sumi, Rung et al., 2012, Yun et al., 2019). Specifically, we used two collections of the Wikipedia dump: category membership link records (`*-categorylinks.sql.gz`) and interlanguage link records (`*-langlinks.sql.gz`). First, category membership link records contained directed linkage between a category and other items (e.g., page and category) in Wikipedia. We filtered `page` → `category` and `category` → `category` relationship (e.g., `page:Complex system` → `category:System theory`) to extract the reference relationship between scientific concepts. Second, interlanguage link records contained information of items in other language editions that were identical or reasonably similar to the source article. For instance, `page:Complex system` in English Wikipedia has a language linkage with `page:Système complexe` in French Wikipedia, indicating that the two documents on this topic are identical.

3.2. Socio-economic data sets

To verify our hypothesis that social connection yields the similarity in knowledge structure, we collected additional country-to-country socio-economic datasets. Although we collected 59 language editions, only 52 languages exist in the country-to-country socio-economic data because it is difficult to find the usage statistics for languages such as Bosnian, Welsh, Croatian, Norwegian Nynorsk, Scottish, Serbian, and Cantonese. The export data were extracted from two different sources: IMF Data in December 2019 and UN Comtrade export data in January 2020. We obtained the statistics of scientific papers (citations and collaborations) from SCOPUS's April 2019 data; patent information was retrieved from PATSTAT's Spring 2019 data. The international student count was collected from OECD in December 2019. Please note that only inbound international student numbers in OECD countries were collected, and thus, the statistics are highly asymmetric and incomplete because it does not provide a number for international students in non-OECD countries. Facebook Social Connected Index (SCI) is the index indicating the degree of the social connection between the two regions, which has been used in various disciplines recently (Bailey et al., 2018, 2020, Vahedi et al., 2021, Du et al., 2021).

All socio-economic data was directed, except the paper/patent collaboration and Facebook SCI. Because socio-economic data fluctuates and has a wide range of year-to-year variations, we aggregate the available data for compensating issues. For each data set, SCOPUS data covers 2000 to 2018, PATSTAT data covers 2000 to 2018, IMF Data covers 2000 to 2018, UN Comtrade covers 2014 to 2018, and international student count covers 2013 to 2016. We use the socio-economic data from the country that our language usage data covers (178 countries). For each data set, SCOPUS data covers 178 countries, PATSTAT data covers 178 countries, IMF Data covers 166 countries, UN Comtrade covers 169 countries, and international student count covers 168 countries.

4. Methods

4.1. Knowledge structure of each language edition

From the linkage between Wikipedia pages and categories, we extracted a hierarchical *knowledge network* of each language edition. Categories are generally located at the end of a Wikipedia page and are designed to link related entries under a shared topic to make navigation easier. For the knowledge network, we designated a node as a category or page, treated as a proxy of a subject or scientific notion, and if there was a hyperlink from a subject to another subject, we assigned a directed link (Yoon et al., 2018). We considered categories and pages to be identical when they shared an identical name (e.g., `category:science` and `page:science`); thus, we merged them into a single node with inheriting connected links. As an illustrative example displayed in Fig. 1a, `page:Complex system` hyperlinks `category:System science`, and therefore, we assigned a directed link from the node `Complex system` to `System science`.

Our primary interest was to understand the dissemination of scientific and technological knowledge. We thus sampled a subnetwork derived from an artificial root node named science & technology, which is assigned as a common parent node of "science" and "technology" for each language edition. Note that "Science" covers all branches of science, such as applied sciences, formal sciences, natural sciences, and social sciences. A constructed knowledge network is directed and unweighted with several cycles from complex connections among nodes. Then, we obtained 59 knowledge networks based on their written language. The basic statistics of the derived networks are listed Table S1.

4.2. Genealogy vector of scientific concept

The obtained knowledge network represents the relationships between the subjects. As shown in Fig. 1b, English Wikipedia users identify complex systems with complex dynamics and complex system theory, and other language users may consider different associations. For example, French Wikipedia users identify the complex system with distinct topics such as concept sociologique (sociological concepts) and analyse des réseaux sociaux (social network analysis). To investigate such differences systematically, we

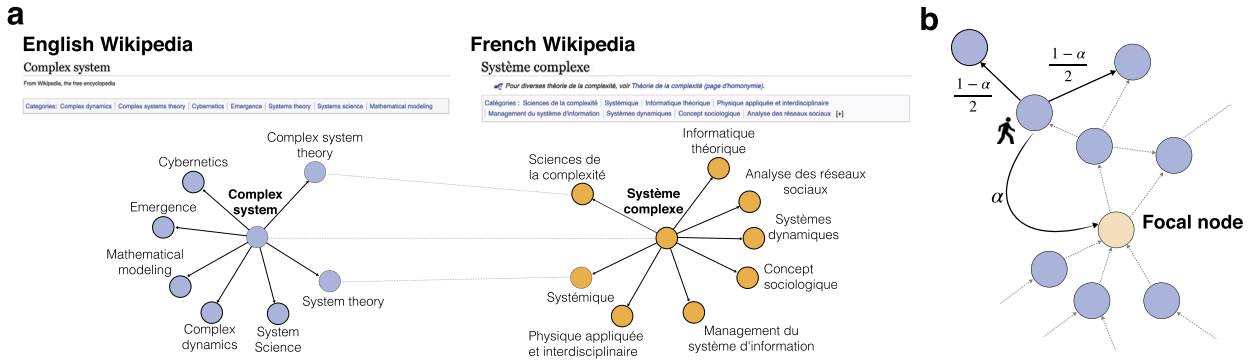

Fig. 1. Wikipedia knowledge network and genealogy vector of a subject. a. Example of a page-category hyperlink in English Wikipedia (left). The `page:Complex system` page has hyperlinks to several categories to which the page belongs. One can express such relations using the network (graph), where the node represents an entity (which can be a page, a category, or their union) with the links representing the hyperlink relationship between them. The identical `page:Complex system` page is titled `page:système complexe` in French Wikipedia (right). Note that the hyperlink structures of two language editions are different, even for the identical entities. Between these two language editions, there are dotted and gray lines denoting the existence of language links between them, indicating that the two subjects are identical. **b.** Method for calculating the genealogy vector of a focal node. We obtain the genealogy vector using the personalized Page Rank algorithm, which calculates the probability that a random walker starting from the focal node visits other nodes. The random walker starts at the focal node and traverses with probability $1-\alpha$ to its nearest neighbors. The walker also occasionally returns to the focal node with a teleport probability, α. The focal genealogy vector of the focal node is the random walker's stationary distribution for the visited nodes.

introduced the concept of a *genealogy vector* using personalized PageRank, which depicts how people correlate different subjects with the focal node regarding both nearest and non-nearest neighbors in the network.

We calculated genealogy vectors of a given subject for each edition using our variant of personalized page rank (PPR) algorithm (Jeh & Widom, 2003). The PPR is a node ranking algorithm with respect to a specific source node using the random walker on networks. For every timestep, the random walker moved to a nearby node chosen randomly with a probability proportionate to the edge weight between them, while the walker could return to the starting node, with a chance of fixed probability α. Thus, the stationary distribution of the random walker starting from node i, denoted by $p_i = (p_{ik})$ was given by

$$p_i = (1-\alpha)Wp_i + \alpha v_i, \tag{1}$$

where W is an adjacency matrix for a given network and v_i is a column vector of length N (number of nodes in the network) whose elements are zero except ith element equals to one. The teleport probability α is a tunable parameter, for which we used $\alpha = 0.3$ in this study.

For hierarchies of the structure of knowledge itself, we introduced a hierarchical bias on the transition matrix, W. First, we defined l_i as the shortest path of node i from the root node, practically interpreted as the level of node i. Therefore, a given starting node, i, transition matrix $W = (W_{ij})$ was defined as follows:

$$W_{ij} = \frac{A_{ij} * k^{l_i}}{\sum_j A_{ij} * k^{l_j}}, \tag{2}$$

where A_{ij} is the adjacency matrix and k is the tunable hyperparameter that controls the behavior of the random walker toward the hierarchy. If $k > 1$, the random walker was more likely to visit lower-level nodes, whereas the random walker tended to visit the higher-level nodes when $k < 1$. In this study, we used $k = 0.5$ considering the hierarchy from the root node for a given subject. The PPR value of the source node is $p_{ii} \simeq \alpha$ by definition, although we forcibly assigned $p_{ii} = 0$ to remove the self-preference of the genealogy vectors. Then, we normalized genealogy vectors so that the sum of the vector is 1.

4.3. Modeling multi-lingual linkage data

In Wikipedia, interlanguage link records are a relationship between two items in other language editions that were identical or reasonably similar to the source article. As shown in Fig. 1b (dotted line), We model the interlanguage link records with a bipartite network. A bipartite network is a network whose nodes can be divided into two disjoint and independent sets a and b; that is, every link connects a node in a to one in b. In our definition, set a and b is each language edition and $T_{ij}^{a\to b} > 0$ indicates interlanguage link between scientific concepts i from language edition a and scientific concepts j in language edition b. In the most simple case, `page:Complex system` in English Wikipedia has a language linkage with `page:Système complexe` in French Wikipedia, indicating that the two documents on this topic are identical. Here, $T^{\text{English}\to\text{French}}_{\texttt{page:Complex system},\texttt{page:Système complexe}}$ is one and otherwise zero. If one subject in the language edition a is connected k multiple subjects in the language edition b (many-to-many case, Fig. S7), we set $T_{ij}^{a\to b} = \frac{1}{k}$. Intuitively, one can understand $T^{a\to b}$ as the translation matrix between two different knowledge networks from interlanguage link records.

4.4. Calculation of subject similarity and knowledge structure similarity

We hypothesized that the interlanguage similarities between genealogy vectors of the same subject could be used as cognitive similarities between them. We presented a simple example to depict the computation of similarity when one subject is solely connected to another subject, and we present more complex cases (*e.g.* many-to-many) in Supporting Information (Fig. S6). For the most straightforward and common case, we first defined this similarity as follows:

$$d_x^{a\to b} = d(p_x^a T^{a\to b}, p_x^b), \tag{3}$$

where p_x^a and p_x^b are genealogy vectors of subject x in the knowledge network of a language edition and b language edition, respectively, and $T^{a\to b}$ is translation matrix between two different knowledge networks from interlanguage link records. For the distance function d, we used the l_2 Euclidean distance.

We then defined subject similarities between the same subject in different language editions as

$$s_x^{a\to b} = \frac{\sqrt{2} - d_x^{\alpha\to\beta}}{\sqrt{2}}, \tag{4}$$

where $\sqrt{2}$ is the theoretical maximum value of the distance between vectors whose elements are positive, and their sum is one.

Finally, we defined the knowledge structure similarity by aggregating the similarities between two language editions by averaging the similarity of all co-existing concepts as follows:

$$s^{a\to b} = \frac{\sum_{x\in\mathcal{A}} s_x^{a\to b}}{|\mathcal{A}|}, \tag{5}$$

where $\mathcal{A}$ is the subject that co-exists in both language editions.

4.5. Community detection

We employed Leiden algorithms (Traag et al., 2019) to find community structure from the knowledge structure similarity network, which is a refined version of the well-known Louvain algorithm. We used a quality function Q of the Potts model with the configuration null mode (Leicht & Newman, 2008) as follows:

$$Q = \sum_{ij}\left(A_{ij} - \gamma\frac{k_i^{out}k_j^{in}}{m}\right)\delta(\sigma_i,\sigma_j), \tag{6}$$

where k_i^{out} and k_i^{in} are the out-strength and in-strength of node i, respectively, A is the weighted adjacency matrix, m is the total edge weight, and σ_i denotes the membership of node i. Here, γ is the resolution parameter, where $\delta(\sigma_i,\sigma_j) = 1$ if $\sigma_i = \sigma_j$ and 0 otherwise. We may control resolution parameter γ to vary the number of clusters, and we use γ with a default value of 1. Q quantifies the modularity of the networks, which is high when having dense connections between the nodes within the community but sparse connections between nodes in different modules. A modularity value of the obtained community is 167.29.

4.6. Mapping the country-level statistics onto the language

The socio-economic data described earlier were county-level data, whereas our similarity measure was language-level statistics. For our analysis, we projected country-to-country data to language-to-language data using the language profile for each country.

Consider a country-level-statistics, $X \in \mathcal{R}^{N_c * N_c}$, where N_c is the total number of unique countries in the dataset and X_{ij} denotes the socio-economic quantity between countries i and j. To map this matrix onto the language space, we constructed a country to a language projection matrix, $A \in \mathcal{R}^{N_c * N_l}$, where N_l is the total number of unique languages. The elements of matrix A were obtained from the language usage profile of countries. We assigned the proportion of language a in country i onto $A_{i\alpha}$. For instance, $A_{\text{English, United States}}$ was 0.821 because the usage share of English in the United States is 82.1%. Using this metric, we constructed the language-to-language socio-economic data, Y, with simple projection $Y = A^T X A$, where Y_{ab} means projected socio-economic quantity between languages a and b.

5. Result

5.1. Geographical proximity still influences, but socio-economic interaction shape the knowledge structure

A natural step forward was to find the possible sectors of languages whose members are more closely associated with each other. For this purpose, we constructed the *similarity network* from the pairwise knowledge structure similarity, where nodes represent the language of Wikipedia, and the link's weight indicates similarity between languages. Considering constructed similarity network is densely connected, we extracted the backbone of the networks by calculating the ego-centric importance of each link (Waltman et al., 2020) as follows

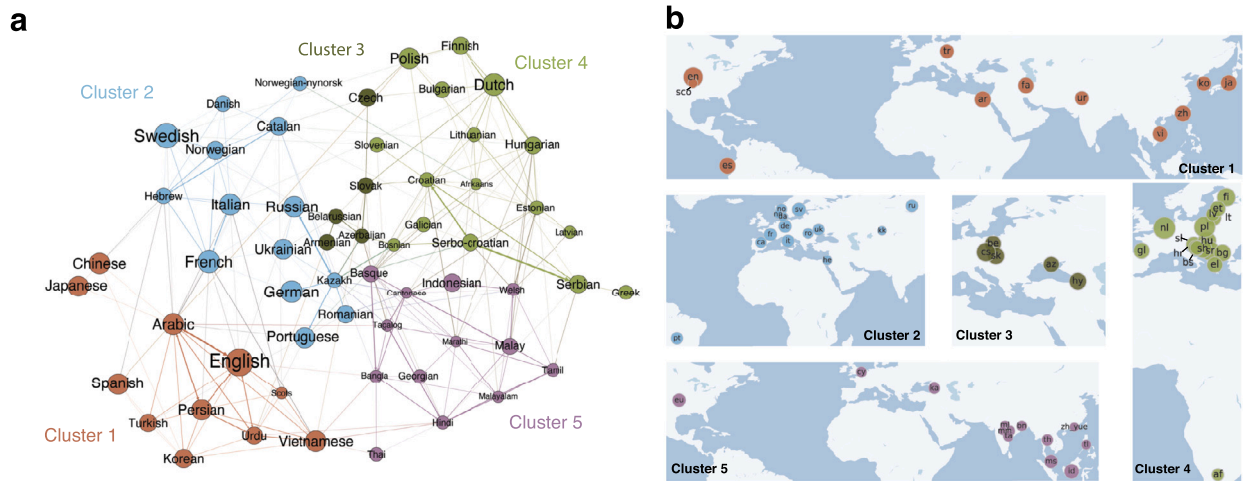

Fig. 2. Geographical proximity affects the similarity of knowledge structure across language usage groups. a. We find five language communities from similarity networks using the Leiden algorithm (Traag et al., 2019) (see Methods). The node colors indicate community memberships, whereas size indicates the number of documents of the corresponding Wikipedia edition on the log scale. **b.** Geographical dispersion of the languages for each community. The location of each language is estimated from the Wikipedia pageview data with geotags (see S2 Text).

$$r_{ij} = \sum_{m,k} s_{mk} \frac{s_{ij}}{\sum_k s_{ik} \sum_k s_{kj}}. \tag{7}$$

We then removed links whose normalized weight is under a certain threshold, t, and chose the minimum value of t that network remains in a single component. As presented in Fig. 2a, five distinct communities are identified by the Leiden algorithm (Traag et al., 2019) (see Methods) that indicate that the clusters seem to be affected by geographical proximity (Fig. 2b), which is similar to a previous study on Wikipedia bilateral ties. In this instance, geography best explains the formation of the cluster (Karimi et al., 2015). English is in the center and serves as a hub node, while intermediate hub languages such as Spanish, German, French, Russian, Portuguese, Chinese, and Dutch also function as cluster centroids (Ronen et al., 2014). Four identified clusters (Cluster 2–5) show close geographical proximity within each cluster, implying geographical proximity affects the knowledge structure similarity.

However, cultural and historical backgrounds also play an important factor in the cluster, particularly for those of Cluster 4 (light green). For example, Afrikaans, a language mostly spoken in South Africa, Namibia, and Botswana, evolved from European Dutch dialects (Pithouse et al., 2009, Heese, 1971) during the era of imperialism. One may note that geographic proximity does not appear to be a key determinant for Cluster 1 (orange), which spans the globe from the Far East to the Americas, and includes English as the de facto international language. The result above echoes with the user's geo-location distribution of each language edition (Yasseri, Sumi, & Kertész, 2012). For example, English or Spanish Wikipedia shows widespread geographical distributions and page-view data in Fig. S2 and Fig. S3. On the other hand, the French Wikipedia displays a distribution concentrated in Europe and Western Africa in Fig. S4. In Fig. 2, English and Spanish Wikipedia belong to Cluster 1, which is the most globally dispersed cluster, whereas French Wikipedia belongs to Cluster 2 (light blue), which is agglomerated toward Europe. These findings imply that knowledge distribution is still influenced by geographical proximity, which impacts the synchronization of knowledge structures between languages, while knowledge dissemination could also be influenced by other factors.

Nowadays, advances in technology provide new channels for interaction. For instance, modern information technology enables us to communicate with thousands, and even millions, of people in real time. One can also physically reach distant countries faster than ever before, with high-speed trains and air transportation being widely available. The cost of travel has also reduced significantly over time, owing to globalization (Hummels, 2007). Thus, such new routes can be new pathways for knowledge dissemination. Accordingly, we expand our analysis to include various language socio-economic connections to verify these new knowledge dissemination pathways. Because most socio-economic data focus on the interaction between countries, we first extract the language usages statistics of each country from the language database (Ethnologue global dataset, 2019). Then, we compare the projected socio-economic connection to a paired knowledge structure similarity, to identify potential contemporary Silk Roads for knowledge dissemination. First, we find that geographical distance no longer plays a central role in knowledge dissemination in the 21st century (Fig. 3a). The geographical distance shows a weak and insignificant correlation with the knowledge similarities between countries (coefficient of determination is $R^2 = 0.01$, and regression coefficient $\beta = -0.005$). This implication is consistent with a previous study that the influence of geographic distance on information flows was minimal at the continental level and even irrelevant at the inter-continental level (Abramo et al., 2020a, 2020b). As we expect, there might be a new route, and the importance of the geographical proximity diminished in the knowledge exchange (Murray et al., 2020).

We observe positive and more significant correlations from the non-geographical interactions (Fig. 3). For example, the scientific interaction reflected in paper collaboration shows a higher coefficient of determination (Fig. 3e; $R^2 = 0.16$, $\beta = 0.009$) than those with geographical proximity (Fig. 3a; $R^2 = 0.01$, $\beta = -0.005$). Indirect scientific interactions also show a positive correlation, although comparably lower than direct interactions (Fig. 3c; $R^2 = 0.13$, $\beta = 0.007$). Because we consider the structure of knowledge denoted

in Wikipedia under the science and technology topic, a strong connection in scientific collaboration will reasonably result in countries to have similar knowledge structures. This result also supports the previous evidence, which suggests internationally mobile scientists could help knowledge transmission (Aman, 2022). Similarly, the soft power movement, which is counted as the number of international students, shows a high coefficient of determination because more than half of international students return to their homelands (OECD, 2011) (Fig. 3d; $R^2 = 0.15$, $\beta = 0.008$). Furthermore, we find non-intellectual interactions positively correlate with knowledge similarity. For instance, the amount of export values, which are not directly related with the knowledge interchange, also correlates with knowledge structure to some degree (Fig. 3b, $R^2 = 0.10$, $\beta = 0.005$, IMF), and the result is robust for export values from different data sources (Fig. S10 $R^2 = 0.09$, $\beta = 0.005$, UN Comtrade). In other words, two language usage groups with strong socio-economic linkages are more likely to have similar knowledge structures. It is, nonetheless, a natural phenomenon because all these linkages are somehow related to knowledge exchange, which ultimately entails knowledge synchronization.

By contrast, personal friendship is not necessarily associated with knowledge structure because it is not directly related to knowledge exchange; instead, it may be related to knowledge similarity through their information exchanges. Therefore, one might expect a weak or non-existent connection between friendship and knowledge. However, we find unanticipated significant and strong correlations between knowledge similarity and personal friendships, measured by the number of mutual friends in social media (Facebook social connectedness index (SCI); see Fig. 3f, where their coefficient of determination $R^2 = 0.17$). This social link, which is reflected by the number of mutual friends in social networks, is the leading candidate for the Silk Road of the twenty-first century, which encompasses several levels of direct and indirect links among people on the web, although it is not widely considered the main channel of knowledge dissemination today.

Furthermore, our regression analysis demonstrates significant overall effect sizes for socio-economic indicators (Table S4). Specifically, Cohen's f^2 shows marginal but significant effects for variables like an international student, paper collaboration, and Facebook SCI, while Cohen's d_1 and d_2 consistently indicate medium effects across all socio-economic indicators, supporting the notion of their substantial impact on knowledge structure similarity. Note that the effect size increases in a comparable order as R^2, confirming our observations. We further examined the robustness of the degree of association to support our findings statistically (see S6 Text). We estimate R^2 using out-of-sample prediction and demonstrate that our finding is not simply a correlation between observed pairs (Fig. S8). These findings corroborate and elucidate our primary findings and claims.

We also perform a multivariate analysis to examine the relationship between socio-economic characteristics and knowledge similarity (see S7 Text). The results consistently indicate that while other socio-economic interactions contribute marginally to knowledge synchronization, the social link (specifically, Facebook SCI) is the most significant correlate. Additionally, we examine the effect of language relatedness on knowledge synchronization. The result indicates that language relatedness is also associated with the knowledge structure similarity ($R^2 = 0.023$), yet the degree of association is insignificant compared to other socio-economic interactions. Moreover, we prove that international students or Facebook SCI partially encodes the language relatedness already shown by the explanatory power gain of the model (see Fig. S9).

In summary, we find the degree of association between socio-economic interactions and knowledge structure to occur in the following order: geographical distance (Fig. 3a, $R^2 = 0.01$) $\ll$ export (Fig. 3b, $R^2 = 0.10$) $<$ weak knowledge dissemination—paper citation (Fig. 3c, $R^2 = 0.03$) $<$ soft-power movement—international students (Fig. 3d, $R^2 = 0.15$) $<$ strong knowledge dissemination—paper collaboration (Fig. 3e, $R^2 = 0.16$) $<$ mutual friendship on the web—Facebook SCI (Fig. 3f, $R^2 = 0.17$). Taken together, the results demonstrate that social connections shape the collective knowledge structures of language users, regardless of whether it is explicitly connected to the knowledge transmission process. The possible mechanism behind the transmission will be discussed below, through our stochastic modeling.

5.2. Mechanistic model for the knowledge dissemination

Our empirical analysis described in the previous sections reveals that i) knowledge structures are more likely to be similar if interactions exist between language usage groups and ii) the degree of association in knowledge structures varies based on the types of interactions. To understand the hidden mechanism of the observed correlation patterns, we identify the key factors driving the synchronization of knowledge structures. First, we assume that people are more likely to be similar when they interact more frequently and *vice versa* (Guéguen et al., 2011). Second, the channel of interaction is progressively moving from a physical route to an online media space, which enables people to interact with overseas countries in real-time (Wasko & Faraj, 2008). We consider a subject as a vector representation, which can be viewed as similar when they are close to each other. This is a similar concept to neural embedding (Peng et al., 2021); however, we avoid declaring the embedding explicitly. Instead, we develop a mechanistic model to reproduce the synchronization of the knowledge structure with proximity among the language usage groups, using randomized initial vectors, motivated by the classic model to elucidate synchronization phenomena (Kuramoto, 1975).

By incorporating the aforementioned factors, we build a mechanistic model of knowledge spreading and synchronization. A simple mechanistic model has been widely used to understand microscopic dynamics which produce macroscopic observations and has achieved prominent achievements across a wide range of the topics such as network science (Barabási & Albert, 1999), public policy (Lempert, 2002), financial market (Feng et al., 2012), consumer energy choice (Rai & Henry, 2016), inequalities in Wikipedia (Yun et al., 2019), information seeking process (Lydon-Staley et al., 2021), and mobility related to COVID-19 (Chang et al., 2021). For simplicity, we only simulate the synchronization of a single subject's genealogy vector. The model comprises N_l agents representing artificial language usage groups. Every agent has the capacity to store d different subjects, and each digit represents their knowledge perception toward a target subject. Thus, each agent has a knowledge structure of d dimensional vector similar to the genealogy

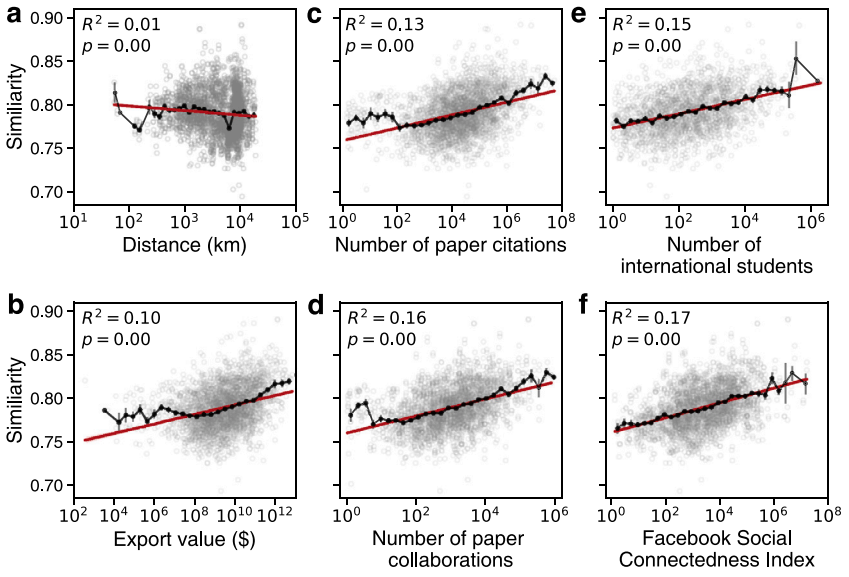

Fig. 3. Interrelationship of knowledge structure across language usage groups reveals the impact of socio-economic interactions. The correlation between structural similarity of knowledge and socio-economic factors: **a.** Geographical distance for the centroids of language pairs, **b.** Amount of exported goods for language pairs (IMF), **c.** Number of citations on papers for language pairs (SCOPUS), **d.** Number of co-authorship on paper for language pairs (SCOPUS), **e.** Number of the international students for language pairs (OECD), and **f.** Facebook Social Connected Index, the strength of connectedness between areas by represented by Facebook friendship ties, for language pairs. An increasing pattern of association is observed in the result. The red line is the line of regression results. Black dots are the mean similarity for each bin where the error bar denotes the standard error.

vector discussed in the previous section. We then define the genealogy matrix, $V \in \mathcal{R}^{N_l \times d}$, by stacking the genealogy vectors of agents so that its row, $V_i \in \mathcal{R}^d$, represents the genealogy vector of agent i.

We further assume that the initial status for agents is independent (thus, each row is orthogonal to the others), aiming to get insight into the situation in which the agent adjusts their differences from the most radical status. We assigned the dimension of the vectors to be a multiple of the number of artificial user groups. Otherwise, the simulation result was biased toward a set of vectors that were not orthogonal or had more nonzero elements. Accordingly, we set each row with an equal number of equally weighted nonzero values (*e.g.*, 1). From the orthogonal condition, each column had only one nonzero value so that $Rank(V)$ was equal to the number of user groups. Then, we normalized each row similar to the empirical genealogy vectors. We simulated the model with 52 language usage groups and using 520 dimensions, resulting genealogy matrix $V = \mathcal{R}^{52\times520}$ for the results.

We additionally introduce proximity $p(i,j)$ from agent i to agent j to describe the degree of interaction between the agents. We use log-transformed empirical data (*e.g.*, Facebook SCI and paper citations) as the proximity $p(i,j)$ and normalize them by dividing proximity $p(i,j)$ by the total sum of proximity for agent i, to use the proximity as the selection probability. Hence, the normalized proximity weight between agents i to j is given by $\hat{p}(i,j) = \frac{p(i,j)}{\sum_k p(i,k)}$.

For every simulation step, the genealogy vector of agent i is updated as the following process. First, agent i chooses its neighbor agent, j, with the probability of $\hat{p}(i,j)$, considering proximity. Then, the genealogy vector of agent i, V_i is updated as follows:

$$V_i(t+1) = V_i(t) + lr \cdot [V_j(t) - V_i(t)], \tag{8}$$

where $V_j(t)$ and $V_i(t)$ is the genealogy vector of group i and chosen reference neighbor, j, at time t, respectively. lr is the fixed learning rate for updating, and we use $lr = 0.001$ for the results displayed in the main text. In this process, an agent's pair with high proximity has a higher chance of being influenced and is more likely to become similar genealogy vectors as a consequence.

As more iteration t passed by, genealogy vectors were synchronized according to the proximity matrix. In the real world, new concepts are consistently introduced to society, and thus the synchronization is hard to reach, yet we neglected the introduction of a new concept. As a result, the simulation ended with an identical vector after all. Our motivation was to obtain a general insight from various proximities. Hence, we captured the most optimized synchronization case by calculating the modeled similarity $S^{model}(t) \in \mathcal{R}^{N_l \times N_l}$ with a pairwise Euclidean distance with normalization for every iteration t (the distance is expressed by eq. (4)). Then, we chose the final simulation result for the given proximity $S^{model}(t^*)$ as

$$S^{model}(t^*) = \underset{t}{\operatorname{argmin}} ||S^{model}(t) - S^{empirical}||_F, \tag{9}$$

where $S^{empirical}$ is the empirical knowledge structure similarity from Wikipedia's knowledge network, and $||\cdot||_F$ is the Frobenius norm of a vector. For example, a Frobenius norm between the reconstructed similarity and actual similarity has a minimum value at $t = 399$, as depicted in the snapshot of Facebook SCI proximity in Fig. 4b. Therefore, we choose $S^{model}(399)$ as the final state of the simulation.

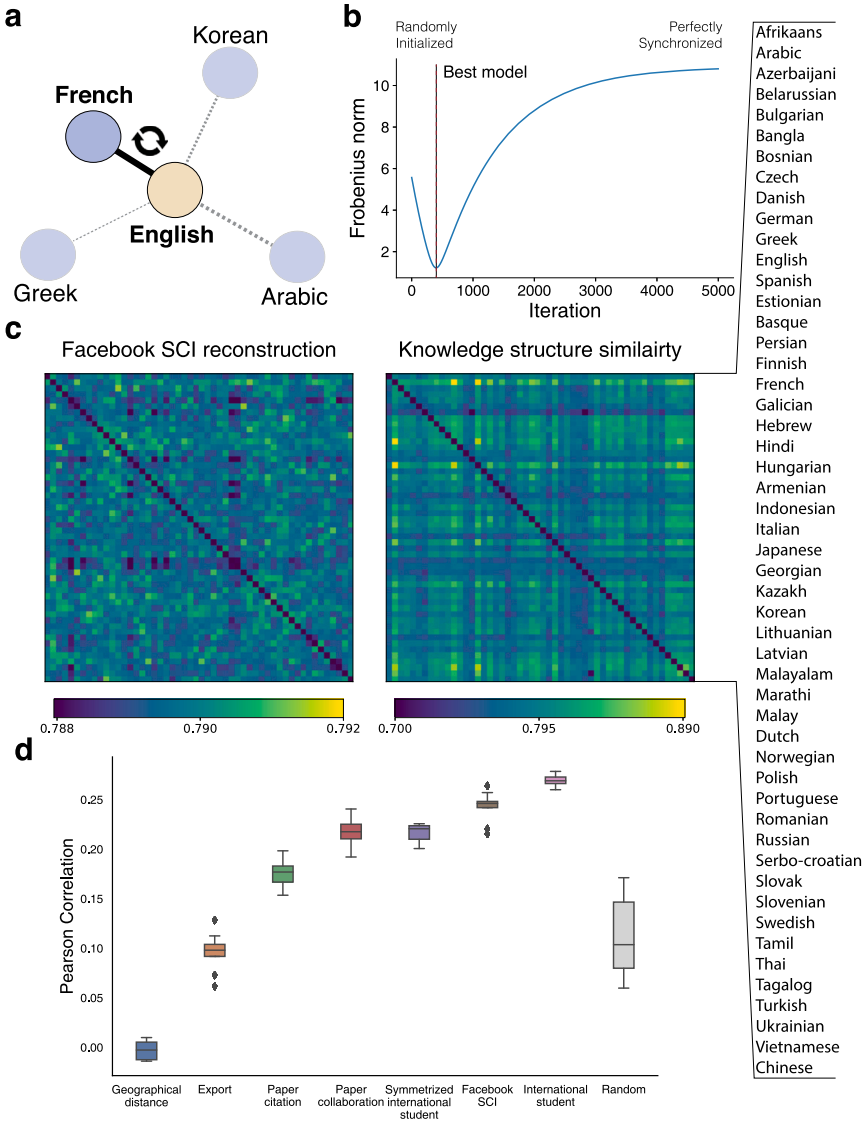

Fig. 4. Simple synchronization model for the knowledge similarity with various proximity indicators a. Schematic diagram describing the model. For each time step, t, each agent selects one of their neighbors based on the probability proportional to the given proximity. As shown in the example, the English agent chooses the French agent as a neighbor for synchronization. **b.** Criteria for choosing the best model. The model begins with the randomly initialized orthogonal vectors, which are synchronized as more iterations occur. For each proximity, we choose the best model similarity at time step t that shows the minimum Frobenius distance from the empirical knowledge similarity matrix. **c.** As an illustrative example, we present the snapshot of the best model similarity matrix for Facebook Social Connected Index ($r = 0.257$, left) and empirical knowledge structure similarity from Wikipedia (right). **d.** The Pearson correlation between the best model similarity and empirical knowledge similarity indicates how well each proximity can reproduce the empirical knowledge similarity. To compensate for any randomness impact, we test 100 different initializations, repeated ten times for each.

We test the model with the six socio-econmic empirical proximities (geographical distance, export, paper citations, paper collaboration, international student, and Facebook SCI, See Data), along with one random proximity as a null model. As the geographical distance is not a proximity measure, we use its log-transformed reciprocal as the geographical proximity. For all other cases, we use log-transformed proximity, similar to the empirical analysis. As an illustrative example of our model results, we present the knowledge similarity matrix of the Facebook SCI from our model and empirical data in Fig. 4c, which shows similar structures.

Previously, we displayed the association between the knowledge similarity and the socio-economic proximities, which is high for the countries that exchange more. The results of our simple synchronization model are consistent with the empirical observations. We show a pairwise Pearson correlation r between the similarity matrix of the model and empirical observation, as an indicator of how well the association pattern has been reproduced in our model. We found a similar increasing pattern of the association with our empirical observations above (Fig. 4d). Specifically, we observed the lowest Pearson correlation for the geographical distance ($r \simeq 0$), followed by the amount of export ($r = 0.091$), the paper citation number ($r = 0.174$), and the paper collaboration number ($r = 0.225$).

One exception was the order between the number of international students and Facebook SCI, which showed the third highest as well as the highest coefficient of determination respectively, in our empirical observations (Fig. 3a). From our model, Facebook SCI showed a Pearson correlation $r = 0.239$, while the international student proximity number showed a Pearson correlation of $r = 0.269$. Note that the international student numbers are highly asymmetric originating, because it has been collected for only the inbound international students studying in OECD countries. To compensate for this asymmetry impact, we also tested the model with the symmetrized count of international students, by averaging the number of inbound and outbound students. The symmetrized model showed a significantly lower correlation than that of the asymmetry model ($r = 0.216$), which is in-between the paper citation and the paper collaboration as similar as the empirical observation.

Because of the knowledge structure synchronization through multiple channels, the current state of similarity is the result of an accumulated exchange process through many routes, which even include factors we have neglected. Our model is the simplest replica, using only a single route of exchange, but implies that social interaction can shape the structure of human knowledge. Moreover, the observed similarity was more robust for virtual connections, which may overcome geographical barriers in contemporary society.

6. Discussion

Humankind has accumulated and exchanged knowledge through various channels over time, facilitated by the technology of the time. Society has gradually progressed toward more efficient commutation, from physically proximate communication routes to virtual online interactions. In this study, we explore the similarity of knowledge structures between users of different languages. We compare the similarities with socio-economic proximities, to identify the main route of contemporary knowledge exchange. Our results indicate the importance of both scientific and social connectedness in knowledge exchange, which shows the significant association between knowledge structures. Thus, this observation indicates that the changes in the main channel of knowledge exchange are from physical contact to online interactions. Our mechanistic model was motivated by the synchronization phenomena proposed to investigate the hidden mechanism behind the current state of knowledge similarity. The model replicated the interactions between language usage groups and reproduced the trend of knowledge similarities. Both the empirical data and model revealed a key factor of knowledge exchange: that is, socio-economic interaction led to a synchronization of the knowledge structures between different cultural areas. Our approach has important implications for science studies, as online collaborative knowledge can provide non-experts with valuable insight into knowledge dissemination. This is a difficult assignment for traditional data set (e.g., papers or patents), which focus primarily on the knowledge structure of professionals, who make up just a small part of society overall.

Our study has several limitations, mainly from the restrictions on data collection. First, one may argue that Wikipedia's accessibility (or inaccessibility in some regions) can affect results. The Chinese people living in mainland China have been unable to access Wikipedia since 2015 (it remains inaccessible). Nonetheless, the Chinese version of Wikipedia is one of the website's most active language editions, and it plays an important role in the similarity network. These findings make it difficult to attribute knowledge structure to a specific geographical region. Second, the quality of Wikipedia is not flawless. Despite the platform's extensive efforts to maintain accuracy and reliability, it remains susceptible to occasional inaccuracies, bias, and vandalism due to the open-editing nature of Wikipedia. To address this issue, our study exclusively utilizes category link data, which offers a more robust and consistent approach compared to relying solely on raw texts. Third, OECD international student data did not include data for non-OECD countries. Fourth, various data dimensions between similarities in knowledge structure and socio-economic data restricted a direct comparison between the data and the model. Our derived knowledge structure from Wikipedia was also focused on language users rather than a specific nation because of the nature of the Wikipedia dataset. Nonetheless, the majority of socio-economic data is collected at the national level. In our study, we projected languages onto countries based on language usage statistics, yet it may be imperfect statistics. If a direct comparison were possible, it would provide additional information; however, we decided to leave this for future research. Fifth, the co-editing behavior of multi-lingual users plays an essential role in information sharing of Wikipedia articles and pathways of the cultural transfers (Karimi et al., 2015). Hence, investigating individual-level co-editing behavior across different language editions could be an interesting subject, yet we left it for further study.

We believe that Wikipedia data have considerable potential for future research. We investigated only the relationships between categories, pages, and language editions; nevertheless, there are billions of records with article content or user statistics that could be explored, interlanguage link records and Wikidata's curated collection provide well-structured, high-quality multilingual linkages that connect semantically similar objects. We show that language usage groups have diverse knowledge structures, indicating that, even if people face the same issue, they may have different perspectives on it. Quantifying the differences in interest changes based on their spoken language may be beneficial. Furthermore, unprecedented contemporary global problems, such as the COVID-19 pandemic, threaten to cause significant changes in how people work worldwide (Brynjolfsson et al., 2020, Yang et al., 2021) and collaborate (Lee & Haupt, 2021). Such changes may accelerate non-physical interactions for knowledge exchange. By pinpointing the primary paths of knowledge diffusion, we want to shed light on the unknown mechanism of general rules of knowledge evolution. Therefore, we would like to emphasize that our study is not simply restricted to Wikipedia, but has the potential for broader applications in contemporary society.

CRediT authorship contribution statement

Jisung Yoon: Conceptualization, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. **Jinseo Park:** Conceptualization, Data curation, Methodology, Writing – review & editing. **Jinhyuk Yun:** Conceptualization, Data

curation, Methodology, Writing – original draft, Writing – review & editing. **Woo-Sung Jung:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Data availability

Wikipedia data are available at wiki-dumps, https://dumps.wikimedia.org/, export data are available at https://data.imf.org/?sk=9d6028d4-f14a-464c-a2f2-59b2cd424b85 (IMF) and https://comtrade.un.org/ (UN), and Facebook Social Connected Index is available at https://dataforgood.facebook.com/dfg/tools/social-connectedness-index. Paper (SCOPUS) and patent (PATSTAT) data can be accessed under a license agreement, which cannot share publicly due to the data's copyright.

Additional information

Supporting Information is available for this paper. Correspondence and requests for materials should be addressed to Dr. Jinhuk Yun and Dr. Woo-Sung Jung.

Code availability

The code used in this analysis can be found at https://github.com/jisungyoon/Structure-of-Science.

Acknowledgements

We thank M. Ahn, I. Hong, H. Kim, L. Miao, and Y.-Y. Ahn for their helpful discussions. This work was supported by the National Research Foundation of Korea (NRF) with grant number NRF-2021R1F1A10630301 (J.Y.; W.S.J.) and NRF-2022R1A2C1091324 (J.Y.). The Korea Institute of Science and Technology Information (KISTI) also offered institutional support for this work (K-23-L03-C01; J.P.) and provided KREONET, our high-speed internet connection. J.Y. would like to acknowledge the support of the National Science Foundation Grant Award Number EF-2133863.

This research was also supported by the MSIT (Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-RS-2022-00156360) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

We would also like to thank Facebook Inc., for making the Social Connectedness Index dataset available to us.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.joi.2023.101455.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2020a). The role of geographical proximity in knowledge diffusion, measured by citations to scientific literature. *Journal of Informetrics*, *14*, Article 101010.

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2020b). Knowledge spillovers: Does the geographic proximity effect decay over time? A discipline-level analysis, accounting for cognitive proximity, with and without self-citations. *Journal of Informetrics*, *14*, Article 101072.

Aman, V. (2022). Internationally mobile scientists as knowledge transmitters: A lexical-based approach to detect knowledge transfer. *The Journal of the Association for Information Science and Technology*.

Ambos, T. C., & Ambos, B. (2009). The impact of distance on knowledge transfer effectiveness in multinational corporations. *Journal of International Management*, *15*, 1–14.

Andrea, A. J. (2014). The silk road in world history: A review essay. *Asian Review of World Histories*, *2*, 105–127.

Arenas, A., Díaz-Guilera, A., Kurths, J., Moreno, Y., & Zhou, C. (2008). Synchronization in complex networks. *Physics Reports*, *469*, 93–153.

Ba, Z., Mao, J., Ma, Y., & Liang, Z. (2021). Exploring the effect of city-level collaboration and knowledge networks on innovation: Evidence from energy conservation field. *Journal of Informetrics*, *15*, Article 101198.

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.

Bailey, M., Farrell, P., Kuchler, T., & Stroebel, J. (2020). Social connectedness in urban areas. *Journal of Urban Economics*, *118*, Article 103264.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Bhandari, R., & Blumenthal, P. (2011). Global student mobility and the twenty-first century silk road: National trends and new directions. In *International students and global mobility in higher education* (pp. 1–23). Springer.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, *424*, 175–308.

Brynjolfsson, E., et al. (2020). Covid-19 and remote work: An early look at us data, Tech. Rep., National Bureau of Economic Research.

Cernada, G. P. (2019). Knowledge into action: A guide to research utilization. Routledge.

Chang, S., et al. (2021). Mobility network models of Covid-19 explain inequities and inform reopening. *Nature*, *589*, 82–87.

Code, L. (1980). Language and knowledge. *Word*, *31*, 245–258.

Del Chiappa, G., & Baggio, R. (2015). Knowledge transfer in smart tourism destinations: Analyzing the effects of a network structure. *Journal of Destination Marketing & Management*, *4*, 145–150.

Du, Z., et al. (2021). International risk of the new variant Covid-19 importations originating in the United Kingdom. *MedRxiv*.

Ethnologue global dataset, https://www.ethnologue.com/. (Accessed 30 October 2019).

Feng, L., Li, B., Podobnik, B., Preis, T., & Stanley, H. E. (2012). Linking agent-based models and stochastic models of financial markets. *Proceedings of the National Academy of Sciences*, *109*, 8388–8393.

Fortunato, S., et al. (2018). Science of science. *Science*, *359*.
Grimm, S. R. (2014). Understanding as knowledge of causes. In *Virtue epistemology naturalized* (pp. 329–345). Springer.
Guéguen, N., Martin, A., & Meineri, S. (2011). Similarity and social interaction: When similarity fosters implicit behavior toward a stranger. *The Journal of Social Psychology*, *151*, 671–673.
Havelock, R. G. (1979). Planning for innovation through dissemination and utilization of knowledge. Center for Research on Utilization of Scientific Knowledge.
Heese, J. A. (1971). *Die Herkoms van die Afrikaner, 1657-1867*. A. A. Balkema.
Hu, Z., Fang, S., & Liang, T. (2014). Empirical study of constructing a knowledge organization system of patent documents using topic modeling. *Scientometrics*, *100*, 787–799.
Hummels, D. (2007). Transportation costs and international trade in the second era of globalization. *The Journal of Economic Perspectives*, *21*, 131–154.
Inkpen, A. C., & Tsang, E. W. (2005). Social capitalinformation spreading, networks, and knowledge transfer. *The Academy of Management Review*, *30*, 146–165.
Jalili, M. (2013). Social power and opinion formation in complex networks. *Physica A: Statistical Mechanics and its Applications*, *392*, 959–966.
Jeh, G., & Widom, J. (2003). Scaling personalized web search. In *Proceedings of the 12th international conference on world wide web* (pp. 271–279).
Kant, I. (2000). *Critique of the power of judgment*. Cambridge University Press.
Karimi, F., Bohlin, L., Samoilenko, A., Rosvall, M., & Lancichinetti, A. (2015). Mapping bilateral information interests using the activity of Wikipedia editors. *Palgrave Communications*, *1*, 1–7.
Kirst, C., Timme, M., & Battaglia, D. (2016). Dynamic information routing in complex networks. *Nature Communications*, *7*, Article 11061.
Kuramoto, Y. (1975). International symposium on mathematical problems in theoretical physics. *Lecture Notes in Physics*, *30*, 420.
Lambiotte, R., & Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, *3*, 180–190.
Lee, J. J., & Haupt, J. P. (2021). Scientific collaboration on Covid-19 amidst geopolitical tensions between the US and China. *The Journal of Higher Education*, *92*, 303–329.
Leicht, E. A., & Newman, M. E. (2008). Community structure in directed networks. *Physical Review Letters*, *100*, Article 118703.
Lempert, R. (2002). Agent-based modeling as organizational and public policy simulators. *Proceedings of the National Academy of Sciences*, *99*, 7195–7196.
Lu, H., et al. (2016). Earliest tea as evidence for one branch of the silk road across the Tibetan Plateau. *Scientific Reports*, *6*, 1–8.
Lydon-Staley, D. M., Zhou, D., Blevins, A. S., Zurn, P., & Bassett, D. S. (2021). Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature Human Behaviour*, *5*, 327–336.
Murray, D., et al. (2020). Unsupervised embedding of trajectories captures the latent structure of mobility. arXiv preprint, arXiv:2012.02785.
Nastase, V., & Strube, M. (2008). Decoding Wikipedia categories for knowledge acquisition. In *AAAI: Vol. 8* (pp. 1219–1224).
OECD (2011). How many international students stay on in the host country? https://www.oecd-ilibrary.org/content/component/eag_highlights-2011-14-en.
Peng, H., Ke, Q., Budak, C., Romero, D. M., & Ahn, Y.-Y. (2021). Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances*, *7*, Article eabb9004.
Pithouse, K., Mitchell, C., & Moletsane, R. (2009). *Making connections: Self-study & social action, Vol. 357*. Peter Lang.
Pluchino, A., Latora, V., & Rapisarda, A. (2005). Changing opinions in a changing world: A new perspective in sociophysics. *International Journal of Modern Physics C*, *16*, 515–531.
Pluchino, A., Latora, V., & Rapisarda, A. (2006). Compromise and synchronization in opinion dynamics. *The European Physical Journal B-Condensed Matter and Complex Systems*, *50*, 169–176.
Qian, Y., Liang, J., & Dang, C. (2009). Knowledge structure, knowledge granulation and knowledge distance in a knowledge base. *International Journal of Approximate Reasoning*, *50*, 174–188.
Rai, V., & Henry, A. D. (2016). Agent-based modelling of consumer energy choices. *Nature Climate Change*, *6*, 556–562.
Ringberg, T., & Reihlen, M. (2008). Towards a socio-cognitive approach to knowledge transfer. *Journal of Management Studies*, *45*, 912–935.
Ronen, S., et al. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, *111*, E5616–E5622.
Sakata, I., Sasaki, H., & Kajikawa, Y. (2012). Identifying knowledge structure of patent and innovation research. *Journal of Intellectual Property Association of Japan*, *8*, 56–67.
Schieffelin, B. B., & Ochs, E. (1986). Language socialization. *Annual Review of Anthropology*, *15*, 163–191.
Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. *Web Intelligence and Agent Systems: An International Journal*, *7*, 195–207.
Song, M., & Kim, S. Y. (2013). Detecting the knowledge structure of bioinformatics by mining full-text collections. *Scientometrics*, *96*, 183–201.
Su, H.-N., & Lee, P.-C. (2010). Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in technology foresight. *Scientometrics*, *85*, 65–79.
Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*, 1–12.
Tu, J. (2020). The role of dyadic social capital in enhancing collaborative knowledge creation. *Journal of Informetrics*, *14*, Article 101034.
Vahedi, B., Karimzadeh, M., & Zoraghein, H. (2021). Spatiotemporal prediction of Covid-19 cases using inter-and intra-county proxies of human interactions. *Nature Communications*, *12*, 6440.
Waltman, L., Boyack, K. W., Colavizza, G., & van Eck, N. J. (2020). A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies*, *1*, 691–713.
Wasko, M., & Faraj, S. (2008). The web of knowledge: An investigation of knowledge exchange in networks of practice. *Academy of Management Journal*.
Welch, D. E., & Welch, L. S. (2008). The importance of language in international knowledge transfer. *Management International Review*, *48*, 339–360.
Wu, W.-L., Hsu, B.-F., & Yeh, R.-S. (2007). Fostering the determinants of knowledge transfer: A team-level analysis. *Journal of Information Science*, *33*, 326–339.
Yang, L., et al. (2021). The effects of remote work on collaboration among information workers. *Nature Human Behaviour*, 1–12.
Yasseri, T., Sumi, R., & Kertész, J. (2012). Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PLoS ONE*, *7*, Article e30091.
Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PLoS ONE*, *7*, Article e38869.
Yoon, J., Yun, J., & Jung, W.-S. (2018). Build Up of a subject classification system from collective intelligence. *New Physics: Sae Mulli*, *68*, 647–654. http://www.npsm-kps.org/journal/DOIx.php?id=10.3938/NPSM.68.647.
Yun, J., Lee, S. H., & Jeong, H. (2019). Early onset of structural inequality in the formation of collaborative knowledge in all wikimedia projects. *Nature Human Behaviour*, *3*, 155–163.
Zhang, Z.-K., et al. (2016). Dynamics of information diffusion and its applications on complex networks. *Physics Reports*, *651*, 1–34.