Spatial Information Research
https://doi.org/10.1007/s41324-023-00553-x

An integrated space–time framework for linkage discovery of big survey data

Xinyue Ye[1] · Xiang Lian[2] · Hongwei Xu[3] · Jiaxin Du[1] · Shuming Bao[4]

Received: 30 April 2023 / Revised: 28 September 2023 / Accepted: 29 September 2023

Abstract
In the realm of survey research, establishing connections within large datasets remains a challenge. This study aims to unveil underlying connections within extensive survey data, emphasizing the need for a more integrated approach to decipher intricate relationships among survey elements. Utilizing computational semantics, machine learning, and advanced spatiotemporal models, we developed an all-encompassing database. This novel database is adept at extracting and characterizing features from a multitude of survey studies, spotlighting relationships among metadata elements such as terms, variables, and topics. The derived relationships are systematically stored as connectivity matrices. These matrices not only quantify the degree of interconnectedness among features but also provide insights into their complex interplay. As a result, our system functions akin to a digital geographical data librarian. Beyond merely serving as a storage tool, this system facilitates interdisciplinary research. It equips researchers with the capability to discern connections between survey elements, enabling them to identify the most influential paths among features based on diverse criteria. Such a tool fosters cross-disciplinary integration and unveils potential ties between seemingly unrelated survey attributes, paving the way for breakthroughs in understanding and application.

Keywords Computational semantics · Machine learning · Spatiotemporal models · Big survey data

1 Introduction

Incorporating social survey research into the field of geographic information science is driven by two compelling reasons. Firstly, many spatially explicit challenges are interconnected with social factors, as they have both drivers and consequences within the social dimension. Secondly, community engagement and citizen response play a crucial role in shaping the research agenda of geography [1]. The tolerance and mutual understanding need to be fostered between social phenomena and their geographic context [2]. To fully comprehend the far-reaching consequences of human activities on climate and ecosystems, and the reciprocal feedback effects on human society within the framework of coupled human-earth systems (CHES), it is imperative for social scientists and earth scientists to mutually appreciate each other's perspectives [3]. This collaborative approach is vital for addressing the gaps and challenges in CHES modeling and advancing our understanding, ultimately paving the way for sustainable mitigation and adaptation strategies. Social scientists need to understand the objective of generating quantitative socio-economic projections and forecasts in

✉ Xinyue Ye
xinyue.ye@tamu.edu

Xiang Lian
xlian@kent.edu

Hongwei Xu
hongwei.xu@qc.cuny.edu

Jiaxin Du
jiaxin.du@tamu.edu

Shuming Bao
sbao@umich.edu

[1] Department of Landscape Architecture and Urban Planning & Center for Geospatial Sciences, Applications and Technology, Texas A&M University, College Station, TX 77840, USA

[2] Department of Computer Science, Kent State University, Kent, OH 44240, USA

[3] Department of Sociology, Queens College, Queens, NY 11367, USA

[4] Spatial Data Lab, Harvard University, Cambridge, MA 02138, USA

Published online: 28 October 2023

earth science, while earth scientists should recognize the unique perspectives and methodologies through which social scientists study human behavior and institutions [4]. It is essential to acknowledge the inherent challenges in predicting human activities and choices. By fostering this interdisciplinary collaboration, fruitful interactions between social and geographic phenomena can be achieved.

Since the development and application of probability sampling methods in the early twentieth century, surveys have become one of the most widely used data collection tools for empirical social science research on challenging issues on the earth [5]. Collected in a standardized form (e.g., questionnaires and structured interviews), survey data help social scientists gain new knowledge. Many survey researchers analyze secondary data collected by others because even for a relatively small sample size, high-quality survey data can be costly to collect, process, document, and curate. On the other hand, the research team that collects its own data usually designs the survey to address specific research questions and is often limited in its capacity to fully explore the potential use of the data. Therefore, the dissemination of survey data for public use not only improves transparency and replicability in social science research but also allows secondary data users to collectively exploit the full-scope scientific value of each dataset. Finding the right existing survey for secondary analysis is crucial. Secondary data users must draw on extensive literature reviews and their own prior research experiences to identify one or more candidate surveys. It may take months, if not years, for a researcher to develop a decent working knowledge about the strengths and weaknesses of a candidate survey and proceed with data analysis. This labor-intensive and time-consuming data search process can be substantially shortened with the help of large data archive centers.

However, despite many existing tools to promote effective data use at these data archive centers, data search remains largely a time-intensive process which requires researchers to choose certain search keywords or browse surveys by certain subject categories pre-defined by archivists [6]. A fundamental challenge is how to better harness the rich data from multiple surveys that cut across existing disciplinary boundaries to inform the development of new theory and hypothesis testing. A social scientist may quickly identify candidate survey data on a research topic in line with his/her research interest and expertise after a few keyword queries. However, relying on a certain data archive alone, he/she is probably limited in his/her capacity to discover other features potentially related to the same topic across surveys of various disciplinary backgrounds or to pinpoint the optimal influence chain linking two potentially related research topics across different surveys. A classic example is the social research on the association between racial segregation and racial inequality in America since the late nineteenth century. Within the discipline of sociology alone, this research inquiry can be traced back to as early as the beginning of the twentieth century when suggested that racial residential segregation could affect social interaction between whites and blacks in harmful ways. It is through more than one hundred years of constant research efforts that we are getting better at mapping out various, complicated pathways linking racial segregation to racial inequality. Residential segregation leads to a concentration of impoverished neighborhoods occupied by minorities who are faced with limited job opportunities, thereby giving rise to economic inequality between whites and racial/ethnic minorities [7]. In addition to poverty, racial residential segregation can also affect racial inequality by eroding social capital and collective efficacy which in turn increases the rates of homicide and other violent crimes [8]. Growing up and living in segregated neighborhoods expose children of racial/ethnic minority origins to an elevated level of chronic stress which can undermine their cognitive development [9] and academic performance [10]. The resulting racial inequality in human capital accumulation during childhood is likely to last, if not amplify, into racial inequality in socioeconomic status in adulthood. More importantly, these mechanisms are often intertwined with each other to either reinforce the preexisting condition of inequality or create a new source of inequality. It would be time-consuming for a researcher who only specializes in one aspect of racial segregation (say, the effect of racial segregation on concentrated poverty) to factor in chronic stress or food environment as intermediate variables in the causal chain between segregation and poverty.

Hence, a new data tool is needed to assist researchers in efficiently identifying as many logically sound pathways as possible from voluminous existing data and published studies. Most importantly, the new tool needs to have the capacity to offer, through data mining, new possible pathways for hypothesis testing. For example, neuroscience research has discovered a moderating role of social and racial contexts in the long-term effects of past event-related face recognition on affective reactions to people during subsequent encounters [11]. Such research may provide new insight into the neuroscientific basis of racial prejudice and discrimination but may not catch sociologists' attention in a timely fashion. Through its computationally efficient search of the published scientific database, the new data tool may quickly detect such an implicit connection between the neuroscience research and sociological research on race/ethnicity and provide a recommendation to its users to embark on testing the new hypothesis.

The discovery of complementary but disjointed literature is known as literature-based discovery. Literature-based discovery aims to assist researchers in generating meaningful hypotheses by mining the implicit relationships between terms in the literature. This field was pioneered by

Swanson's work starting from the late 1980s [12–14]. Generally, the literature-based discovery process starts with the extraction of terms or concepts. The terms or concepts might come from an existing knowledge base or be automatically extracted by techniques such as semantic filtering or clustering. The similarities between the terms or concepts are then computed using a variety of techniques, including lexical analysis, citation analysis, bibliographic coupling methods, clustering, or heterogeneous bibliographic information network, etc. [15]. Traditionally, the results are presented to the users with related terms. However, influence chains that represent a complex chain of intermediate terms could also be extracted. The past thirty years have seen increasingly large and diverse datasets for extracting terms and measuring similarities. At the same time, progressively more automated and complex algorithms have been investigated to deal with these data. However, most connectivity matrices, which are among one of the core components for building the influence chains, were built on the static data structure, instead of the dynamic data structure. Currently, two main approaches exist: (1) process data from third-party databases offline, which cannot incorporate dynamic updates; and (2) retrieve related datasets at query time, which cannot reflect accurate term correlations of the entire database. The semantic relatedness in the connectivity matrices could be calculated from the metadata, or other data sources including Wikipedia and Wordnet, but there is no systematic investigation on weighting different sources. Besides, it is crucial to leverage user feedback to adjust the term relatedness adaptively from user interactions through machine learning techniques. The third limitation is that most connectivity matrices were built on simple network structure, which is less efficient for computation of optimized influence chains with big data. Three challenges relate to these limitations: (1) *Weighted connectivity matrix*, semantic relatedness could be measured from different data sources, as well as user experience; (2) *Dynamic updates*, online search based on dynamically updated survey collections; and (3) *Computational efficiency*, multiscale connectivity matrices for efficient computation.

2 Literature review

2.1 Term extraction from textual data

Term extraction is the process of automatically extracting key and topical terms from documents. The central issue of term extraction is to determine correct terms from a large number of candidate terms. The candidate terms are typically selected through simple methods such as stop-list, part-of-speech tags, or *n*-grams. Researchers have investigated both supervised and unsupervised approaches to address this issue. The supervised approach relies on designed features including statistical, structural, or syntactic features [16]. Some researchers tried integrating external resources to construct features. Wikipedia is the most used reference source [17]. Unsupervised approaches include graph-based ranking, topic-based clustering, and language modeling [16]. Many tools have emerged for term extraction, but issues still exist, such as incorporating appropriate background knowledge, handling long documents, and improving evaluations [16]. In practice, a big challenge of building a user-friendly search system is the existence of synonyms. Automatically discovering synonyms has been an active topic in natural language processing tasks. The difficulty of synonym discovery mainly comes from context.

2.2 Connectivity matrix and semantic similarity

The connectivity matrix is essentially a matrix of semantic similarity measures between terms in surveys. Similar to a spatial connectivity matrix that defines the spatial connections among different addresses, a conceptual connectivity matrix defines the conceptual connections among different terms. The connectivity matrix can be formulated as a graph for calculating optimized chains between terms. The element in the matrix represents the semantic similarity between features. The most commonly used computation methods include correlation-based relevance, such as Pearson correlation, and vector cosine-based relevance [18]. The relevance calculation is highly domain-specific in terms of the features considered in the calculation and associated weights. For survey metadata, the features may include title, summary, subject, space, time, related publications, and so on. The vectors for each feature can be computed by using information retrieval methods such as TF-IDF, topic modeling methods such as Latent Dirichlet Allocation (LDA) [19].

Measuring the semantic relatedness between documents or terms has been one of the main themes in computational linguistics since the 1990s [20]. Researchers have developed various types of methods for computing the semantic relatedness, including graph-based approaches such as normalized path length, or the context-based approaches such as co-occurrence methods, or information theoretic methods. Applications of semantic similarity includes online data sources such as Wikipedia [21], knowledge graphs [22, 23], or specific domains such as biomedical science [24]. The knowledge resources used to measure the semantic relatedness include the linguistically constructed data sources such as WordNet, and collaboratively constructed sources such as Wikipedia. Researchers have also started to investigate the use of knowledge graph (such as DBpedia) to compute the semantic relatedness [23]. An evaluation of different measures is given in [25].

Although many different measures exist, choosing the right measure could be problematic in practice without

proper weighting and adjustment. Given the diversity and volume of the terms, it's not possible to ask researchers explicitly to weight different measures because it's time-consuming and expensive. Thus, it is crucial to collect the implicit feedback from users who interact with the system and adjust the weighted connectivity matrix accordingly. This process of learning from user input to adjust internal knowledge is studied well in the community of recommender systems [26]. Generally there are two categories of recommender algorithms, collaborative filtering that learn from user interest of items, and content-based filtering that break down an item in terms of its attributes. How to effectively integrate implicit user feedback and knowledge about Mixed similarity learning to solve the issue of data sparsity from relying on single data source, including heterogeneous information network [27], collaborative deep learning [28], heterogeneous-constraint item similarity model [29], and mixed similarity learning [30].

2.3 Identification of influence chains in literature-based discovery

The identification of influence chains is conceptually finding paths between two terms. Pathfinding algorithms such as the Dijkstra's algorithm have been applied to numerous studies and applications across a wide array of disciplines. For example, people have used the pathfinding functions daily on their mapping services. Pathfinding in knowledge discovery has been commonly used to measure concept similarity [31]. In literature-based discovery, researchers have envisioned the extension of the Swanson ABC model, incorporating higher-order co-occurrences to allow more than one intermediate term [32]. Recent years have seen growing interest in this direction. Wilkowski et al. (2011) suggested a discovery browsing method that guides users' search chains between two concepts through graph analysis [33]. Song, Heo, and Ding (2015) proposed a semantic path analysis method to enable the generation of possible hypotheses from biological terms [34]. Hahn-Powell, Valenzuela-Escárcega, and Surdeanu (2017) introduced a graph-based approach to identifying influence relations from publications, which allows users to explore direct and indirect influence chains [35]. Most of these studies have been focused on biomedical applications, but interest in this direction has also started to emerge among researchers in other fields. For example, researchers have investigated heuristic algorithms to identify semantic chains between concepts in Wikipedia [36].

2.4 Knowledge graph storage and search

Resource Description Framework (RDF) has been widely used for decades in the applications of the Semantic Web, which is a W3C standard to define resources and their logical relationships [37]. RDF has several data models, for example, triple store, C-store (or column store), property tables, and graphs. RDF data are usually represented by either RDF triples (in the form of $<subject, predicate, object>$) or RDF knowledge graphs. Examples of RDF graph data include DBpedia [38], YAGO [39], Wordnet [40], and so on.

Due to the large size of the multi-scale connectivity matrices, graph partitioning can be employed to reduce the space cost of large and sparse connectivity matrices. Techniques of distributed storage systems [41] and sparse matrices [42] can be leveraged in the graph partitioning process.

3 Method

An integrated approach for discovering the linkages can be suggested among big survey data (Fig. 1). A large-scale integrated database with terms extracted from survey studies can be developed by applying computational semantics and machine learning. The database can capture the relations of multiple terms embedded in the survey metadata. These relations can be stored as multi-scale connectivity matrices that measure the relatedness between terms considering complex sematic relationships. The online tool can function as a digital data librarian and help facilitate interdisciplinary research by enabling investigators to identify the optimized influence chains between terms based on different criteria. Those influence chains link terms based on the hierarchical structure of subjects. Terms are linked further to survey studies, as well as related authors and publications, in a hierarchical and network structure. This can enable the generation of a broad range of novel research questions and scientific hypotheses in a much more efficient and flexible way than previous approaches.

3.1 Term extraction from textual data

A survey usually contains the following types of features: title, summary, subject terms, space, time, related publications, and variables. Each subject term also links to a group of studies that contain the same subject. Space (geographical coverage) is represented with a list of place names and can be linked to administrative entities on different levels. Time is represented as a single year or a range of years. Related publications are normally structured in a standardized tag format such as RIS, or BIB that includes authors, organizations, and other related information. Each variable also contains the source and description.

Formally, we denote a survey s from all surveys S as a group of features $s = \{A, G, T, P, V\}$, where A is a list of subjects $A = \{a_1, a_2, \ldots, a_{n_1}\}$, G is a list of administrative units $G = \{g_1, g_2, \ldots, g_{n_2}\}$, T is a list of dates $T = \{t_1, t_2, \ldots, t_{n_3}\}$, P is a list of related publications $P = \{p_1, p_2, \ldots, p_{n_4}\}$, and V

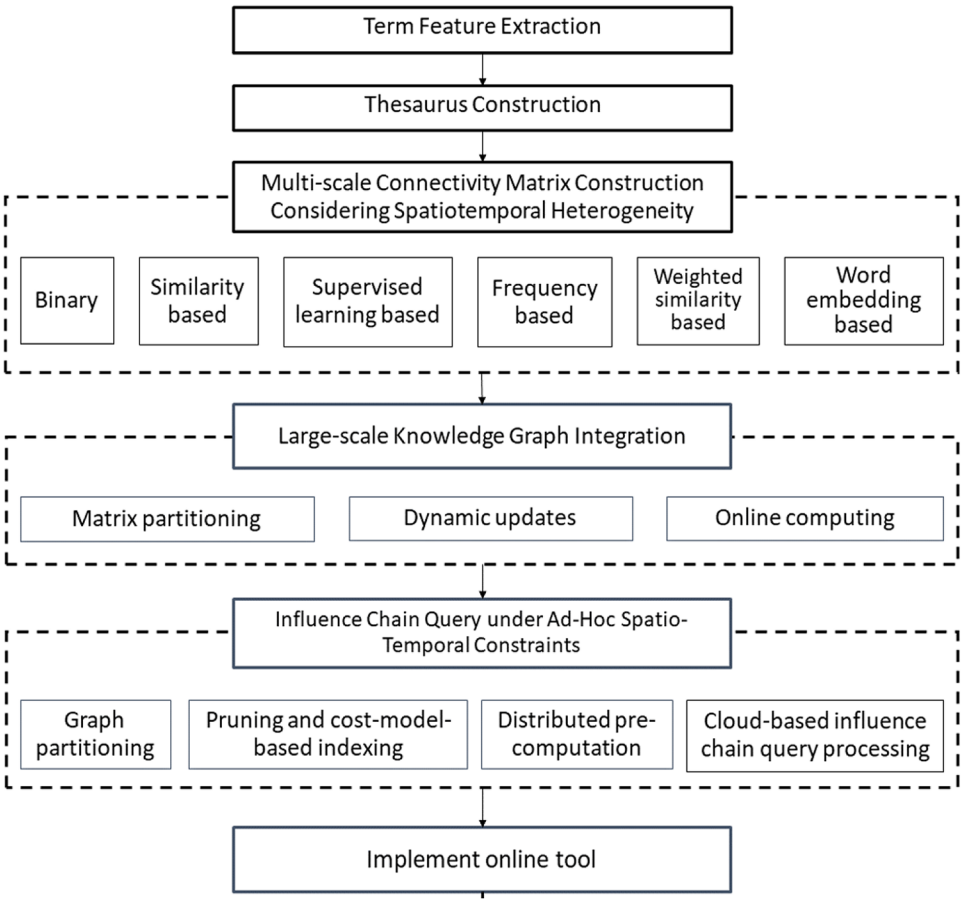

Fig. 1 The flowchart diagram

is a list of variables $V = \{v_1, v_2, \ldots, v_{n_5}\}$. All objects in a factor also belong to s. For example, a subject a_k in A belongs to s, denoted as $a_k \in s$.

We can extract terms from the variable description. This is easier than extracting terms directly from survey forms, since the latter is basically unstructured text, and can easily lead to an explosion of terms. We can combine a stop word list, part-of-speech-tags, and pre-defined lexico-syntactic patterns to select candidate terms, and then apply existing graph-based ranking methods or machine learning algorithms for term extraction. After the terms are determined, we can identify synonyms in the terms using synonym discovery algorithms adapted to the context of survey variable descriptions. Formally, each variable v_k corresponds to a set of terms C_{v_k}. The entire term set C for a survey s is decided by taking the union of C_{v_k} for all variables. A term c_k in C is said to belong to s, denoted as $c_k \in s$. The term set can be subsequently compiled into a multi-scale term structure considering the subject hierarchy.

3.1.1 Challenges

Term extraction is not trivial due to issues such as the entity resolution, deduplication, and removing the ambiguity of terms in different contexts. The multi-scale term structure needs to be validated by professional curators with domain expertise.

3.1.2 Research directions

We can combine automated tools and experience from professional curators. The human-detected analyses based on the experience provided a contextual understanding and validation of the machine-detected findings. The existing thesaurus from targeted application fields can be leveraged to build the multi-scale term structure.

3.2 Connectivity matrix construction

The construction of a connectivity matrix between different terms is a key part, which lays the foundation for graph formulation and calculation of optimized influence chains [43]. We consider the term-to-term matrix M_T. Since the terms are extracted from variable descriptions, each term reversely corresponds to a set of variables, which are related to several survey studies, and subsequently related to other terms such as subjects or publications. This chain

of mediating relations makes it possible to calculate the term-term similarities through these other features.

This proposed framework can consider six measures of the connectivity strength between two terms c_i and c_j, using various measures:

The **Binary** measure determines the relationship based on the presence of both terms in a study. if $\exists s \in S, c_i \in s \bigwedge c_j \in s$, $M_{T_{ij}} = 1$, otherwise $M_{T_{ij}} = 0$. This approach results in a sparse matrix, offering clarity to users by indicating that terms are related only if they co-occur in a study.

The **Frequency-based** measure calculates $M_{T_{ij}}$ as the count of instances where both terms c_i and c_j appear together in a study across all studies S.

For the **Similarity-based** measure, $M_{T_{ij}}$ represents the similarity between terms. This similarity is computed using cosine similarity with TF-IDF vectors. Here, each term c corresponds to a list of surveysS_c. If the term c appears k times in a survey s, s will appear k times inS_c.In essence, we treat terms as documents that contain multiple survey studies as words in information retrieval. Thus, tf(s, c) represents the number of times that survey s appears inS_c, whereas $\text{idf}(s, \text{C}) = \log \frac{|C|}{|\{c \in C, s \in S_c\}|}$. The TF-IDF vector $Q_c = \text{tf}(\text{S}, c) \cdot \text{idf}(\text{S}, c)$. The cosine similarity between c_i and c_j is then defined as $\cos(\theta) = \frac{Q_{c_1} \cdot Q_{c_2}}{Q_{c_1} Q_{c_2}}$.

The **Weighted Similarity-based** measure extends the similarity-based approach by considering a weighted average of multiple features, including time, survey studies, and related publications.

The **Supervised Approach** involves constructing multiple machine learning features from survey metadata. Initially, a subset of terms C' is randomly selected from all terms C. The similarity between terms in'C' is then rated using a fixed integer scale, such as 0 to 10. These ratings serve as training samples. Features derived from the survey metadata include statistical features like TF-IDF vectors, syntactical and lexical features, and knowledge-based features from external sources like WordNet Similarity. A machine learning model, such as logistic regression, is trained using these features, and the values in M_T are predicted based on the best-performing model.

Lastly, the **Word Embedding-based** measure employs Word2vec models trained on variable descriptions from all surveys. The model generates a similarity vector V for each word in the input text, allowing for the identification of $M_{T_{ij}}$ in the vector for either term c_i or c_j.

We can evaluate the efficiency and effectiveness of these six types of matrices in both user studies and online experiments. We can provide users with the option and flexibility to select different types of matrices, so that the similarity of the selected type between two terms (or through an optimized chain) can be calculated.

In practice, the connectivity matrices can be very sparse, but of large scale (with many terms). We can study how to partition the sparse matrix to reduce the space cost. Specifically, the connectivity matrix can be equivalently represented by a graph where vertices are terms, and edges are associated with weights, indicating the similarities between any two terms (vertices). Therefore, we can also design effective cost-model-based graph partitioning algorithms to divide a large graph (matrix) into smaller subgraphs (small partitioned matrices) and reduce the storage space of connectivity matrices.

3.2.1 Challenges

The challenges of constructing connectivity matrices lie with both the computation and storage. For the computation of the first four types of connectivity matrices, the naïve approach is to calculate the relevance or similarity between all terms sequentially. However, this method is time- and space- inefficient given a large number of terms. The machine learning models for the fifth and sixth types can not fit in a single machine either. Therefore, the problem of how to leverage cloud computing techniques to effectively compute large-scale connectivity matrices is a challenging issue. The storage of connectivity matrices is also challenging, especially when the number of terms is large. What is more, due to the sparseness or skewness of survey data, we may not be able to identify accurate similarity between any two terms (even if we apply advanced similarity measures/approaches). Therefore, it is also challenging how to effectively compute "good" similarity scores among terms.

3.2.2 Research directions

Observing the large-scale and data sparsity of connectivity matrices, we can design effective data partitioning techniques to divide the survey studies and/or publications. Then, instead of constructing a single large and sparse matrix, we can construct offline multiple small connectivity matrices (equivalently, partition a large graph from each connectivity matrix into subgraphs, based on our devised cost model) for those (highly correlated) terms over surveys/publications. In addition, we can also maintain the similarities among small connectivity matrices, which can support fast online retrieval of term similarity across surveys. When the smaller connectivity matrices are still of large scale, we can further perform the data partitioning recursively and obtain a finer resolution of matrix data, forming a hierarchical multi-scale structure. To allow the storage and processing over large scale matrix data, we can leverage cloud computing to deploy the data

partitions (e.g., partitioned connectivity matrices in different scales) to different servers, and design novel parallel and distributed algorithms (e.g., following the MapReduce framework) to efficiently compute and query connectivity matrices. To facilitate efficient data retrieval during the cloud computing, we can develop space-efficient indexes for matrices to efficiently access/retrieve the data on server nodes, and utilize data compression methods to summarize and transmit the output matrix, which can significantly reduce the communication cost over the network during the cloud computing of connectivity matrices.

In order to accurately estimate unbiased similarity scores among terms, in addition to survey data, we can explore other data sources such as knowledge bases (e.g., Wikipedia, WordNet, etc.), and consider incorporating similarity values among times in these external data sources to compute unbiased similarity measure. Specifically, given any two terms, we can obtain their different similarity values, from different similarity measures/approaches (e.g., as discussed above) and data sets (including surveys and external data sources). Then, we can use a *machine learning* approach to find appropriate weights for weighing these similarity values and calculating a final similarity score. The input of the machine learning problem is a number of similarity values from different data sources or approaches, whereas the output is a set of expected final similarity scores for pairs of terms. We can train the learning process by providing user feedback about final similarity scores and back propagate the errors to adjust parameters during the learning process. Finally, we can use this machine learning approach to derive similarity scores for pairs of terms, which can be used for data analytics over survey data.

3.3 Large-scale knowledge graph integration

We can utilize the extracted terms from survey metadata to construct knowledge graphs from connectivity matrices. Specifically, for the metadata of surveys (e.g., terms extracted from survey metadata), we obtain a knowledge graph where each vertex is a term, and each edge between any two vertices corresponds to the relationships of two terms [44]. In future research, survey contents (e.g., survey answers) can also be leveraged to construct the knowledge graph, where each answer to a survey can be represented by a knowledge graph, and each node is a survey question that is associated with both question and answers terms (keywords), and each directed edge between two nodes represents the relationship of two survey questions (e.g., if the answer is yes, then skip to the answer associated with Question 10).

3.3.1 Metadata for survey knowledge graphs

In order to integrate surveys into a single database, we can first construct a *meta-knowledge graph* for each survey study. Since different survey studies use distinct sets of terms (with different relationships among these terms), we obtain the meta-knowledge graphs of different graph topologies. Due to the heterogeneity of survey data, we then merge these meta-knowledge graphs into a single large-scale meta-graph. That is, those vertices from distinct graphs but with the same vertex labels (i.e., terms) can be merged together (each vertex is associated with term sources). If two edges from two different graphs connect the same two vertices, then edges are merged, but associated with a multi-set of weights from 2 graphs (as well as survey sources of edges). Here, the weights of the edges can be obtained from offline pre-computed multi-scale connectivity matrices, as mentioned in Sect. 2.2 (in the case that they are not stored due to the data partitioning, we can compute the weight online, that is, the similarity between two terms).

Due to dynamic updates of survey data, the metadata for survey knowledge graphs are also subject to changes. Upon the arrival of new surveys, we can first construct the meta-knowledge graph for those new surveys, and then incorporate them into the existing meta-graph.

3.3.2 Challenges

The integration of survey data is challenging. Each survey tends to have many variables. It is not time- and space- efficient to perform the data fusion over large-scale heterogeneous graphs. Most importantly, it is also challenging to guarantee the accuracy of the graph data integration. Moreover, it is non-trivial to determine how to efficiently update meta-knowledge graphs with a large amount of new survey data in a batch.

3.3.3 Research directions

We can use database techniques to handle large-scale survey data integration. Moreover, to ensure the reliability of the integrated survey data, we can also apply a probabilistic data model to the meta-knowledge graph by inferring confidences that two vertices from different meta-graphs represent the same terms in different contexts. We can design efficient batch algorithms to update meta-knowledge graphs from new survey data.

3.4 Graph query processing

We can study a number of important graph queries, as well as personalized search recommendation queries, over

large-scale integrated survey data (modeled by meta knowledge graphs) [45].

3.4.1 Survey recommendation queries

We can consider the personalized search recommendation (or prediction), with the help of our integrated big survey data. Specifically, according to our integrated meta knowledge graphs for surveys, we can infer the correlations (or co-occurrences) of a term/subject from other terms/subjects, which may potentially release some research potentials in social, behavioral, and economic research.

For example, if two terms "education" and "salary" appear frequently in surveys, it may imply that these two topics are correlated and worth studying. On the other hand, instead of obtaining the knowledge what existing surveys study, we can also recommend new knowledge about which topics have not been frequently investigated before, which might be a potential research direction, or guide the design of new surveys.

Inspired by the examples above, we can incorporate the domain knowledges of survey, social, behavioral, and economic research, and formalize novel personalized search recommendation (or prediction) queries over large-scale survey data to estimate the correlations among survey terms/subjects/topics for researchers, and predict future survey research topics. Formally, we give the definition of the *survey recommendation query* below.

Definition 3.1 (Survey Recommendation Queries). Given survey meta knowledge graphs, G, a query keyword, k, an integer parameter l for surveys, and a ranking function $r(k1, k2)$, a *survey recommendation query* obtains l keywords (terms, subjects, topics, etc.), k', that have the highest or lowest ranking scores $r(k, k')$ over surveys.

Intuitively, in Definition 3.1, we consider top-l keywords that have the highest or lowest correlations with the query keyword k in a query region Q within a period of time I, which may potentially indicate hot field research or unexplored research directions, respectively.

3.4.1.1 Challenges Due to the large-scale of graph data, it is challenging to efficiently and effectively manipulate large knowledge graphs, and estimate/predict the inherent correlations among keywords (e.g., terms/subjects) for large-scale survey data. The straightforward method is to online compute the ranking scores for every pair (k, k'), and identify the ones with highest/lowest ranks, which is however rather inefficient.

3.4.1.2 Research directions We can explore and design efficient big data techniques (e.g., the MapReduce framework) to enable intensive ranking score computations over large survey graph data. Most importantly, we can use effective pruning strategies (e.g., by using lightweight pruning methods w.r.t. lower/upper bounds of the ranking scores) to reduce the search space of the survey recommendation query.

3.4.2 Influence chain queries

Next, we study influence chain queries over large-scale integrated survey data (modeled by knowledge graphs). Intuitively, we would like to identify the optimized influence chain between two potentially related, but not directly related, terms in surveys, which is very useful for researchers to discover how important and through which chain two terms are related to each other.

As an example, given two terms, "residential segregation" and "income inequality", researchers are interested in knowing the connection chain (relationship) between these two demographic and economic phenomena across different surveys. There may be more than a single connection chain since the relationship between racial segregation and income inequality may vary in both space and time. For example, African American neighborhoods remained highly segregated until the 1990s when hypersegregation started to become less common with rising income. In contrast, the passage of the 1965 Immigration Act led to large inflows of Hispanic and Asian immigrants whose segregation levels ranged from low to high. Hispanic and Asian neighborhoods have experienced notable socioeconomic improvement since the 1970s, thereby reducing Hispanic-white and Asian-white neighborhood inequality [46]. Nevertheless, in areas with a large concentration of undocumented migrants, the level of segregation between, in particular Hispanics and whites, has actually increased [47]. One may find different chains between these two topics in our proposed survey knowledge graphs. We can provide researchers with the most convincing chains that can best describe the transitive relationships between two topics.

Definition 3.2 Given survey meta knowledge graphs G, a source keyword, ks, a destination keyword, kd, and a score function $score(P)$ of any chain P, an *influence chain query* obtains optimized chains P, between keywords ks and kd in survey knowledge graphs G, such that for any other chains P' between ks and kd, we have $score(P) > score(P')$.

In Definition 3.2, the score function $score(P)$ of chain P can be given by factors such as the summation of the relevance of edges on chain P, (the inverse of) the shortest path distance, the summation of co-occurrence frequencies on edges of chain P, and so on. The influence chain query aims to obtain the path with the highest score, which indicates the one with the strongest transitive relationships between topics

ks and *kd*. Note that, this problem can be also generalized to the one where users can select a few important topics (in addition to source and destination topics) that influence chains must pass through.

3.4.2.1 Challenges To efficiently tackle the influence chain problem, there are several major challenges. First, the connectivity matrices in Sect. 2.2 give many metrics to define meaningful and effective score functions, $score(P)$, to measure the relationship between two terms on the chain. What is more, different score functions may even conflict with each other. For example, a chain with high relevance may have low co-occurrence frequencies of edges on the chain. It is thus very challenging to determine how to select the best, or at least not the worst, score functions for research studies.

Second, on large-scale survey knowledge graphs, there are many possible influence chains between the source and destination terms. It is not efficient to enumerate all of these chains and calculate their corresponding scores.

3.4.2.2 Research directions Since it is possible to have multiple, and sometimes *conflicting*, score functions that may determine the relevance of the chain, we can alternatively consider the idea of the *skyline*. In particular, each chain P is associated with a score vector $V(P)$, in which each element $V(P)[i]$ contains a score given by a distinct score function $scorei(P)$. Then, the skylines of chains are those chains P whose vectors $V(P)$ are not *dominated by* other chains, where a chain $P1$ *dominates* another one $P2$, if (1) $V(P1)[i] \geq V(P2)[i]$ holds for all dimensions i, and (2) $V(P1)[j] > V(P2)[j]$ holds for some dimension j. In this way, we can solve the problem of conflicting score functions and return chains that are better than others in at least one dimension (with respect to a score function).

Furthermore, to tackle the influence chain query over large-scale graphs, the cloud computing (i.e., via MapReduce) can be adopted to distribute survey graphs, connectivity matrices, and queries to multiple servers and enable fast computation. In particular, effective graph partitioning algorithms can be designed to divide large-scale survey knowledge graphs into disjoint/overlapping subgraphs, based on a cost model. Then, distributed pre-computation techniques can be devised for scores of edges in graphs (e.g., pre-computing scores w.r.t. surveys in the US and in 2016), which can be used for online score retrieval and pruning false alarms while answering the query. Moreover, pruning mechanisms can be designed to reduce the search space of the problem, for example, by using lower/upper bounds of ranking scores or applying the pruning with dominances among distinct chains. Finally, cost-model-based indexing mechanisms can be used to facilitate efficient processing of cloud-based influence chain algorithms under the MapReduce framework.

3.4.3 Survey keyword search queries

Furthermore, we can investigate the keyword search queries over survey graphs. That is, given a set of query keywords, a *survey keyword search* query retrieves a number of surveys that contain all query keywords, as well as the relationships among these surveys (i.e., the graph structures), where the query keywords can be terms, subjects, topics, research fields, literature, space, and time.

Definition 3.3 (Survey Keyword Search Queries). Given survey meta knowledge graphs, G, a set of query keywords, $K = \{k1, k2, \ldots, kn\}$, and a score function $f(.)$ for evaluating the goodness of the returned subgraph answers, a *survey keyword search query* obtains a subgraph g of G (i.e., $g \subseteq G$), such that (1) vertices of *subgraph* g contains all query keywords in K, and (2) a score function $f(g)$ for subgraph g is minimized.

As given in Definition 3.3, the survey keyword search query returns a subgraph g of survey knowledge graphs G such that this subgraph contains all the n query keywords and the returned subgraph is the best one in terms of the score function $f(g)$. Here, the score function $f(g)$ can be the number of edges in subgraph g, or the summation of co-occurrences associated with edges in subgraph g, and so on. We can investigate the best score function that fits researchers' requirements in different fields.

3.4.3.1 Challenges One straightforward method is to search the entire meta knowledge graphs, and identify all subgraphs with vertices satisfying the keyword constraints. Then, we find the best subgraph with the minimal score $f(g)$. However, this method is quite inefficient, since there are an exponential number of possible subgraphs (especially, in the case where survey graphs are of large scale). The keyword search problem is usually NP-hard in the literature, which is challenging to tackle.

3.4.3.2 Research directions We can design novel indexing mechanisms for organizing terms, subjects, topics, etc. and facilitating the speed-up of fast keyword search over knowledge graphs. Most importantly, due to the hardness of the survey keyword search query, we can develop efficient approximation algorithms (e.g., greedy algorithms or sampling methods) for processing this typical keyword search problems over the integrated survey graph data.

Continuing the aforementioned research on racial segregation and inequality, a classic example is Thomas Schelling's [48] model, in which he demonstrated that individuals'

uncoordinated, modest in-group preferences would collectively result in the highly segregated residential pattern at the macro level. Such models of micro–macro linkages have been termed as "generative" social science research since they combine theoretical and computational models to explain the emergence of macro societal regularities (e.g., norms, collective actions, and epidemics) from the behavior of interdependent micro-level agents. These generative models are theoretically-driven, highly dynamic and often involve feedback loops and interdependent nonlinear processes. One recent study used an agent-based model to show that urban slums emerge as a result of interactions between multiple human agents (families, real estate developers, and politicians) and multiple aspects of the environment (housing sites, electoral wards, economic growth, and population migration) [49]. Another study also used simulation experiments to examine to what extent racial disparity in diet behavior could be reduced by improving the quality of neighborhood schools and to what extent the reduction in racial disparity might depend on the presence or absence of social influence, as well as healthy versus unhealthy social norm. We can build on the recent advance in studies of complex systems to derive generative models that link micro-level behaviors and macro-level population phenomena from data mining of the survey data archives. The tool can help identify possible influence chains between different agents and help researchers understand the complex structure of the system. Users can search optimized influence chains with related surveys and terms from these two input terms.

3.5 Visual analytics of knowledge graphs

Owing to the vast scale of the survey knowledge graphs with which we engage, comprehending and discerning the intuition behind the survey data can be challenging. A practical approach is the visualization of these knowledge graphs through interactive visual interfaces. Specifically, as each survey captures responses from individuals across various geographical regions and different years, a potent tool is the temporal and spatial mapping of survey data. This can be done by assigning different colors to represent specific statistics. Such a visual representation provides researchers with a lucid understanding of the intrinsic survey data. It can assist them in discerning the relationships between two topics within the knowledge graphs, identifying anomalies or unusual events, grasping public behavioral responses, and more. Additional features of visual analytics encompass delving deeper into the survey data, visualizing responses based on specific keywords, displaying answers to personalized survey recommendation queries, and survey path queries, as alluded to in the preceding subsection.

We can create a visual analytics module tailored for survey knowledge graphs. This will grant researchers the ability to visualize survey data across various dimensions, such as spatial, time, term, variable, topic, and many more. Our aim is to design an intuitive visual interface, allowing users to easily browse, interact with, and inquire about survey data. Additionally, our portal lets researchers save and share their impact paths. This not only facilitates learning from one another but can also spark new research ideas and pave the way for future surveys. When a user opts to share an impact path, it will be categorized based on the surveys, terms, and subjects within that path. As a result, when other users search for a specific term, subject, or survey study, the relevant impact paths will be suggested to them.

4 Discussion and conclusion

Further avenues of research could concentrate on the development and refinement of survey methodologies. In-depth exploration of search result quality assessment merits closer scrutiny. To enhance the user experience and interface design of the survey data analytics tool, research efforts should be directed toward understanding user preferences, needs, and behaviors, which can lead to the creation of more user-friendly and efficient tools that facilitate seamless data exploration and analysis. Additionally, the investigation of cloud computing platforms' potential benefits can be pursued to enhance the scalability and accessibility of survey data analytics. Optimizing the framework's performance will be essential to efficiently handle large-scale datasets. Encouraging interdisciplinary collaboration is vital for fostering partnerships and facilitating knowledge exchange.

Utilizing insights derived from survey data can lead to evidence-based decision-making, better understanding of public perceptions, and improved assessment of intervention impacts. To foster transparency, replicability, and collaboration within the research community, policymakers should advocate for open data policies that enable easy access to and sharing of survey datasets. Furthermore, policymakers can play a crucial role by providing funding for the development and implementation of cloud computing platforms and advanced technologies that facilitate survey data analytics. This investment can significantly enhance the scalability and effectiveness of survey data research. To ensure researchers, practitioners, and policymakers are well-equipped to utilize survey data analytics effectively, it is essential to invest in training programs and capacity building. Facilitating collaborations between academia and industry is vital in order to transfer survey data analytics tools and methodologies into practical applications. Such partnerships can benefit businesses, urban planning, and public services. Lastly, it is imperative to develop and uphold ethical guidelines and standards for conducting survey data research. These measures will safeguard data privacy and protect the interests of

survey participants and stakeholders, ensuring responsible and ethical research practices.

Funding National Science Foundation, 2112356, 2122054, and 2232533, Xinyue Ye.

Data availability Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Declarations

Conflict of interest The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

1. Ye, X., & Niyogi, D. (2022). Resilience of human settlements to climate change needs the convergence of urban planning and urban climate science. *Computational Urban Science, 2*(1), 1–4.
2. Liverman, D. M., & Cuesta, R. M. R. (2008). Human interactions with the Earth system: People and pixels revisited. *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group, 33*(9), 1458–1471.
3. Tan, J., Duan, Q., Xiao, C., He, C., & Yan, X. (2023). A brief review of the coupled human-Earth system modeling: Current state and challenges. *The Anthropocene Review*. https://doi.org/10.1177/20530196221149121
4. Lu, L., Li, P., Kalacska, M., & Robinson, B. E. (2023). Environmental impacts of renting rangelands: Integrating remote sensing and household surveys at the parcel level. *Environmental Research Letters, 18*(7), 074005.
5. Singh, K. K. (2022). *Research Methodology in Social Science*. KK Publications.
6. Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science, 369*(6507), 1060–1062.
7. Massey, D. S., & Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass*. Harvard University Press.
8. Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science, 277*, 918–924.
9. Massey, D. S., & Fischer, M. J. (2006). The effect of childhood segregation on minority academic performance at selective colleges. *Ethnic and Racial Studies, 29*, 1–26.
10. Charles, C. Z., Dinwiddie, G., & Massey, D. S. (2004). The continuing consequences of segregation: Family stress and college academic performance. *Social Science Quarterly, 85*, 1353–1373.
11. Cassidy, K. D., Boutsen, L., Humphreys, G. W., & Quinn, K. A. (2014). Ingroup categorization affects the structural encoding of other-race faces: Evidence from the N170 event-related potential. *Social Neuroscience, 9*, 235–248.
12. Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine, 30*, 7–18.
13. Swanson, D. R. (1986). Undiscovered public knowledge. *The Library Quarterly, 56*, 103–118.
14. Swanson, D.R., & Smalheiser, N.R. (1996). Undiscovered Public Knowledge: A Ten-Year Update. *KDD*, pp. 295–298.
15. Sebastian, Y., Siew, E.-G., & Orimaye, S. O. (2017). Emerging approaches in literature-based discovery: Techniques and performance review. *The Knowledge Engineering Review, 32*, e12.
16. Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. *ACL, 1*, 1262–1273.
17. Medelyan, O., Frank, E., & Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, *3-Vol. 3*, pp. 1318–1327.
18. Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence, 2009*, 4.
19. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.
20. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 1, pp. 448–453.
21. Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. *AAAI, 6*, 1419–1424.
22. Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2015). Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies, 8*, 1–254.
23. Zhu, G., & Iglesias, C. A. (2017). Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering, 29*, 72–85.
24. Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics, 40*, 288–299.
25. Aouicha, M. B., Taieb, M. A. H., & Hamadou, A. B. (2016). SISR: System for integrting semntic reltedness nd similrity mesures. *Soft Computing., 22*, 1855–1879.
26. Beheshti, A., Yakhchi, S., Mousaeirad, S., Ghafari, S. M., Goluguri, S. R., & Edrisi, M. A. (2020). Towards cognitive recommender systems. *Algorithms, 13*(8), 176.
27. Yu, X., Ren, X., Sun, Y., Gu, Q., Sturt, B., Khandelwal, U., Norick, B., & Han, J. (2014). Personalized entity recommendation: A heterogeneous information network approach. In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pp. 283–292.
28. Wang, H., Wang, N., & Yeung, D.-Y. (2015). Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1235–1244.
29. Chen, L., Xin, X., Wong, D., & Ding, Y. (2017). HCoM: Item-based similarity model for heterogeneous implicit feedback. In *Mobile Data Management (MDM), 2017 18th IEEE International Conference On*, pp. 40–49.
30. Liu, M., Pan, W., Liu, M., Chen, Y., Peng, X., & Ming, Z. (2017). Mixed similarity learning for recommendation with implicit feedback. *Knowledge-Based Systems, 119*, 178–185.
31. Zhu, X., Li, F., Chen, H., & Peng, Q. (2017). An efficient path computing model for measuring semantic similarity using edge and density. *Knowledge and Information Systems, 55*, 1–33.
32. Ganiz, M. C., Pottenger, W. M., & Janneck, C. D. (2005). *Recent advances in literature based discovery*. Technical report, LU-CSE-05-027 2005. Lehigh University, CSE Department.
33. Wilkowski, B., Fiszman, M., Miller, C., Hristovski, D., Arabandi, S., Rosemblat, G., & Rindflesch, T. (2011). *Discovery browsing with semantic predications and graph theory.* In *AMIA Annual Symposium Proceedings.*

34. Song, M., Heo, G. E., & Ding, Y. (2015). SemPathFinder: Semantic path analysis for discovering publicly unknown knowledge. *Journal of Informetrics, 9*, 686–703.
35. Hahn-Powell, G., Valenzuela-Escárcega, M., & Surdeanu, M. (2017). Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph. In *ACL*, 103.
36. Franzoni, V., & Milani, A. (2014). Heuristic semantic walk for concept chaining in collaborative networks. *International Journal of Web Information Systems, 10*, 85–103.
37. Hogan, A. (2020). Resource description framework. In *The Web of Data* (pp. 59–109). Springer.
38. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Kleef, P., & Auer, S. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web, 6*, 167–195.
39. Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence, 194*, 28–61.
40. Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*, 39–41.
41. Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems, 26*, 4.
42. Çatalyürek, Ü. I., Aykanat, C., & Uçar, B. (2010). On two-dimensional sparse matrix partitioning: Models, methods, and a recipe. *SIAM Journal on Scientific Computing, 32*, 656–683.
43. Shang, J., Zhang, X., Liu, L., Li, S., & Han, J. (2020). Nettaxo: Automated topic taxonomy construction from text-rich network. *Proceedings of the Web Conference, 2020*, 1908–1919.
44. Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications, 185*, 103076.
45. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys (Csur), 54*(4), 1–37.
46. Intrator, J., Tannen, J., & Massey, D. S. (2016). Segregation by race and income in the United States 1970–2010. *Social Science Research, 60*, 45–60.
47. Hall, A. (2014). Projecting regional change. *Science, 346*(6216), 1461–1462. https://doi.org/10.1126/science.aaa0629
48. Schelling, T. C. (1969). Models of segregation. *The American Economic Review, 59*, 488–493.
49. Patel, A., Crooks, A., & Koizumi, N. (2012). Slumulation: An agent-based modeling approach to slum formations. *Journal of Artificial Societies and Social Simulation, 15*(4), 2.