







ARTICLE



<https://doi.org/10.1057/s41599-023-01774-z>

OPEN

Climbing up the ladder of abstraction: how to span the boundaries of knowledge space in the online knowledge market?

Haochuan Cui ¹, Tiewei Li²  & Cheng-Jun Wang ² 

The challenge of raising a creative question exists in recombining different categories of knowledge. However, the impact of recombination remains controversial. Drawing on the theories of knowledge recombination and category-spanning, we claim that the impact of knowledge spanning on the appeal of questions is contingent upon questions' knowledge hierarchy in the knowledge space. Using word embedding models and network analysis to quantify knowledge spanning and knowledge hierarchy respectively, we test our hypotheses with the data collected from a large online knowledge market ($N = 463,545$). Knowledge spanning has an inverted U-shaped influence on the appeal of questions: the appeal of questions increases up to a threshold, after which point the positive effect reverses. However, with the increase in knowledge hierarchy, the inverted U-shape is weakened and disappears quickly. We fill the research gap by conceptualizing question-asking as knowledge-spanning and highlighting the theoretical underpinnings of knowledge hierarchy. The theoretical and practical implications for future research on knowledge recombination are discussed.

¹School of Systems Science, Beijing Normal University, Beijing, China. ²School of Journalism and Communication, Nanjing University, Nanjing, China.
email: litiewei219@163.com; wangchengjun@nju.edu.cn

Introduction

The theories of knowledge recombination suggest that knowledge spanning plays a crucial role in asking novel questions (Schumpeter, 1934; Nelson and Winter, 1982; Moaniba et al., 2018; Guan et al., 2018; Zhang et al., 2019). Yet, the role played by knowledge recombination is controversial (Nelson and Winter, 1982; Kuhn, 1977; Ferguson and Carnabuci, 2017). On the one hand, knowledge recombination can bring a broader niche market and increase cultural products' potential for extraordinary achievements (Keuschnigg and Wimmer, 2017; Hsu et al., 2012; Ordanini et al., 2018). On the other hand, spanning multiple categories is very challenging, and it faces an even more considerable risk of failure (Ferguson and Carnabuci, 2017; Hsu et al., 2009). The recombination of these competing ideas suggests that there exists an inverted U-shaped impact of knowledge recombination (Uzzi and Spiro, 2005; Foster et al., 2015; Askin and Mauskopf, 2017; Shi et al., 2021). Thus, those questions of higher-level knowledge recombination would be punished, which will hinder asking novel questions. How to encourage people to ask more novel questions becomes an important puzzle.

To address this research puzzle and extend this line of research, we assert that the hierarchy of knowledge (i.e., the level of abstraction) can help loosen the tension between knowledge recombination and tradition. Yet, existing research largely ignores the essential role played by the hierarchy of knowledge (Cole, 1983; Hayakawa and Hayakawa, 1939; Lerner and Lomi, 2018; Tang, et al., 2019). Knowledge is not evenly distributed, and everyone is at a certain level of knowledge hierarchy. Individual actors can only see and understand the knowledge below their own knowledge hierarchy. If individual actors want to expand their scope of knowledge recombination, they need to increase their knowledge hierarchy. In short, the hierarchy of knowledge limits the scope of knowledge recombination. Taken together, we argue that the impact of knowledge spanning on the appeal of questions is contingent upon questions' knowledge hierarchy in the knowledge space.

In addition, there is a lack of research about the knowledge recombination existing in the daily life of ordinary people (Badilescu-Buga, 2013; Shin et al., 2001; Shi et al., 2021). Prior research primarily focuses on knowledge recombination in scientific or technological innovations. Knowledge begins with asking questions (Gazan, 2010; Ravi et al., 2014; Shi et al., 2021). In the academic community, there is a closely connected social network. Thus, the penalties for research that spans the boundary are even more severe. In the online knowledge market, whether the essential tension between knowledge recombination and tradition still exists deserves more attention. Knowledge is one kind of public good. To avoid the tragedy of the commons (Hardin, 1968; Ostrom, 1990) and enable, support, and facilitate knowledge sharing (Stewart, 1997), the knowledge market is established. The question-and-answer (i.e., Q&A) website is one form of knowledge market. According to the types of payments, the Q&A websites can be classified into two categories: fee-based knowledge markets (e.g., Google Answers) and free knowledge markets (e.g., Quora, Zhihu). The free knowledge exchange model works by increasing the reputation of answers and questioners. How to ask creative questions plays an essential role in attracting high-quality answers in the online knowledge market. Thus, we posit that free knowledge markets are more suitable to study the impact of knowledge spanning in asking questions, compared with the fee-based knowledge markets.

By synthesizing the theories of recombination (Schumpeter, 1934; Nelson and Winter, 1982) and category spanning (Hannan and Freeman, 1977; Hannan et al., 2007; Hannan, 2010; Hsu et al., 2009), we conceptualize the act of knowledge

recombination on the Q&A website as *knowledge spanning*. Knowledge spanning is one particular kind of knowledge recombination. It reflects the distance of boundary spanning the knowledge space. This study aims to investigate the roles played by knowledge spanning and knowledge hierarchy in asking attractive questions. Against the backdrop of prior research, the contributions of this study are summarized as follows. First, we contribute to the theories of knowledge recombination by formulating the hypothesis of knowledge spanning in the context of the online knowledge market. There exists an inverted U-shaped relationship between knowledge spanning and the appeal of questions. Second, and more interestingly, we discover that knowledge hierarchy moderates the nonlinear impact of knowledge spanning. To be specific, the inverted U-shape gets more prominent when the knowledge hierarchy (i.e., the level of abstraction) of questions is smaller. In contrast, with the increase of knowledge hierarchy, the inverted-U-shaped influence is weakened and disappears quickly. Third, and practically, our findings suggest that maintaining a hierarchical tree of categories in the knowledge market can better encourage knowledge spanning by empowering people to ask abstract questions. Fourth, we show the perspectives of geometry and network provide a powerful lens to investigate knowledge recombination with the aid of word embedding models (Mikolov et al., 2013a, 2013b; Kozłowski et al., 2019; Lee and Sohn, 2021). In all, our findings suggest that knowledge spanning, knowledge hierarchy, and their interaction shape the appeal of questions on the Q&A website.

In the following sections, we first outline the theories of knowledge recombination, reformulate our research puzzlement, and propose operational hypotheses. Second, we introduce the research methods and the findings of this present study. Although prior research highlights the significance of capturing the extent of recombination in the knowledge space (Moaniba et al., 2018), there is a lack of methods to explicitly model the knowledge space (Morita et al., 2004; Min et al., 2021). Fortunately, word embedding models provide a powerful lens to represent the knowledge space and measure the distance of knowledge spanning (Mikolov et al., 2013a, 2013b; Tang et al., 2019). We address the challenges as mentioned above by using the word embedding methods and network analysis to quantify knowledge spanning and knowledge hierarchy respectively. Third, we summarize the findings and discuss the implications.

Theoretical framework

Theories of knowledge recombination. Creativity is stimulated when diverse ideas are mingled (Schumpeter, 1934; Nelson and Winter, 1982). According to the seminal work of Nelson and Winter (1982), "The creation of any sort of novelty in art, science, or practical life—consists to a substantial extent of a *recombination* of conceptual and physical materials that were previously in existence" (Nelson and Winter, 1982, p. 130). Similarly, Koestler coined the term "bisociation" to elucidate the logic of recombination and the nature of creativity (Koestler, 1964). Koestler (1964) asserts that human beings are used to thinking within a single frame of thought. Nevertheless, creative ideas lie in the intersection area of two frames of thought. The act of creation operates through blending the ingredients from two seemingly incompatible frames of thought. Based on the concept of *bisociation*, Koestler (1964) examines the creative activity in humor, art, and scientific research. In all, innovation often means reassessing and rearranging the existing knowledge, techniques, concepts, and materials. Therefore, the freedom to explore new directions or divergent thinking is necessary.

Knowledge recombination can be viewed as a reshuffling process of the knowledge network. Uzzi and Spiro (2005) argue that the small-world network affects human behaviors by influencing connectivity and cohesion. In a small-world network, people interact within local clusters, and long-ranged ties connect these local clusters (Watts and Strogatz, 1998). These properties of the small-world network promote the diffusion of ideas from one cluster to another in the social system. As one form of recombination, the small-world network can be realized by randomly rewiring the links with a probability that adds long-ranged ties while preserves local clustering. However, if the probability of reshuffling is larger than a threshold, the local clustering would drop dramatically, which will hinder knowledge recombination. Thus, there is a tradeoff in choosing the optimal degree of recombination.

In general, the factors affecting knowledge recombination (e.g., how people ask questions) can be classified into at least three categories. At the individual level, it is related to a person's interest, taste, education (e.g., training), career trajectory, new resources (e.g., mentor, collaborators, expertise, information), and the expenditure of effort. At the local level, it is influenced by institutional factors and disciplinary culture. At the social level, it is also influenced by the social structure and policies. According to Foster et al. (2015), all these factors can be well considered within Bourdieu's field theory of science (Bourdieu, 1975, 2004; Foster et al., 2015).

Bourdieu's field theory of science provides an organizing framework to understand knowledge recombination. Bourdieu (1975) asserts that agents' habitus and capital in the field shape their practice and lifestyles. The habitus is an acquired system of habits, tastes, dispositions, skills, and expectations. It determines how individuals perceive and react to the social world. As the structured structure (cognitive schemes), habitus is the principle organizing the perception of the social world; As the structuring structure (generating schemes), habitus organizes practices and perceptions. The interplay between agents' positions and their habitus guides their actions. Taking scientific research as an example, the strategic choice of research question continually re-creates tradition and punishes deviance in the field (Bourdieu, 1975). Researchers strive to accumulate scientific capital and occupy dominant positions in specific fields (Guan et al., 2017, 2018; Zhang et al., 2019).

Yet, Kuhn (1977) argues that both convergent thinking and divergent thinking are essential to innovation. The essential tension between these two modes of thought generates the best scientific research (Kuhn, 1977). In the stage of normal science, people are not innovators but problem solvers. The essential tension between productive tradition and risky innovation affects researchers' choice of research problems. Most problems can be answered using the existing theories and methods. "Only investigations firmly rooted in the contemporary scientific tradition are likely to break that tradition and give rise to a new one" (Kuhn, 1977, p. 227). Kuhn asserts that the most fundamental advance in science depends on recognizing the causes of the crisis and the traditional tools and standards developed by well-defined traditions help people better identify the loci of troubles. The collision between theoretical expectations and unexpected observations brings breakthrough discoveries (Peirce and Bisanz, 2016).

Knowledge creation and recombination on the Q&A website.

The Q&A website constitutes an informal social field of questions asking for the public. With the accelerated updating and iteration of knowledge, individuals and organizations cannot deal with their problems relying on their own knowledge stock

(Ostrom, 1990; Gazan, 2010; Qi et al., 2020). Although scientists and elites can also take part in the Q&A process, most participants are ordinary people. The users of the knowledge market can not only share knowledge but also integrate existing knowledge and develop new knowledge through online interaction (Qi et al., 2020). Everyone can act as the creator, disseminator, and consumer of knowledge. With the aid of the knowledge market (e.g., Q&A website), ordinary people's individual questions are aggregated into a global knowledge space. Consequently, the hidden individual knowledge becomes visible social knowledge which transforms the way to acquire knowledge.

In contrast with scientific research and technical innovations, the Q&A website is characterized by creating and spreading social knowledge. When people ask a question on the Q&A website, they may not expect to get a technical answer (Liu, Jansen, 2017; Liu and Jansen, 2018). Just as Tilly (2006) has illustrated, technical explanations are only one of four types of reasons adopted by people in daily life. The other three types of reasons are conventions, stories, and codes (e.g., legal judgments). When people ask and answer questions, they confirm, repair, claim, or deny the social relations among them (Tilly, 2006). Therefore, social factors also matter for questions asked on the Q&A website (Fu and Oh, 2019). The social factors extensively investigated in prior research can be classified into three categories:

First, users' attributes shape the appeal of questions. Users' online profile, posting behavior, language style, and social activities reflect their intention of question asking. Liu and Jansen (2017) find that these non-Q&A features can achieve a 70% success rate. Further, Guan et al. (2018) show that authors' structural holes in both collaboration networks and knowledge networks have a positive impact on knowledge creation. Similarly, Fronzetti Colladon et al. (2020) find that the future success of scientific research can be predicted through social networks and semantic analysis. Moreover, the reputation of contributors has an inverted U-shaped influence on the popularity of answers (Osatuyi et al., 2022).

Second, the content features (e.g., level of detail, specificity, and clarity) can promote the appeal of questions. For example, Chua and Banerjee (2015) find the level of detail, specificity, clarity, and socio-emotional value of questions are important content features for predicting whether a question would be answered or not (Chua and Banerjee, 2015). Further, the content quality and source credibility are useful for finding good answers in online knowledge communities (Neshati, 2017). Osatuyi et al. (2022) find that the information quality of answers has a positive effect on their probability of being selected as the best answer. Sussman and Siegal (2003) posit that information cues' influence on information adoption is mediated by perceived information usefulness. Liu et al. (2020) propose a text analytic framework for predicting the usefulness of answers and show that their model performs better than the other models (Sussman and Siegal, 2003).

Third, knowledge networks also play an important role. Guan et al. (2017) find that a paper's centralities in both the collaboration network and knowledge network have an inverted U-shaped impact on citations. Zhang et al. (2019) quantify the degree of knowledge recombination in the knowledge graph, and they also find an inverted U-shaped relationship between recombination distance and a firm's innovation. Moreover, Liu et al. (2020) find that the intensity of science (i.e., the degree to which scientific knowledge is adopted in technology research and development) has an inverted U-shaped influence on knowledge convergence and the impact is moderated by relational embeddedness.

Knowledge spanning and knowledge hierarchy. Questions' positions in the knowledge space shape their popular success (Askin and Mauskopf, 2017; Guan et al., 2017, 2018; Zhang et al., 2019). The features of popular questions tend to collectively signal their quality and reflect the taste of the public (Chua and Banerjee, 2015). The ensemble of different features can be represented with the knowledge space. Knowledge space is a multi-dimensional representation of various forms of knowledge (e.g., questions, patents, research articles). It constitutes a broader ecosystem of knowledge production and consumption. Each question's features determine its position and interrelations with the other questions in the knowledge space. The interrelations of questions help people organize and evaluate questions. To summarize, people perceive and evaluate questions according to their positions in the knowledge space.

Category spanning is one of the most commonly adopted innovation strategies for recombination. Categorization plays an essential role in human behaviors. As we have illustrated, categorization is conducive to the retrieval, browsing, and creation of knowledge. As an indispensable part of the online knowledge market, social markers (i.e., folk taxonomy) are widely used to mark and classify questions. A category could be a specific field, thing, or concept. Questions are categorized and indexed by their categories of knowledge. The online knowledge market usually urges the user who asks questions to add categories that help questions reach those who are interested in these questions.

Categories help users organize and discover content in the knowledge market. Users can retrieve, filter, share, and organize questions through this flexible and unrestricted system. Just as Zuckerman (1999, p. 1398) has put it, "social objects are evaluated by legitimate categories". The categories of questions have a large impact on the evaluation of the question quality. Toba et al. (2014) find that training a type-based quality classifier is effective in discovering high-quality answers. Hu et al. (2020) use topic models to extract keywords from articles and construct five keyword popularity features and they show that these keyword popularity features can effectively improve the prediction models of highly cited papers.

The social object of category spanning is penalized for "perceived quality and audience attention" (Keuschnigg and Wimmer, 2017). The objects related to multiple categories are difficult to interpret. Thus, the public may underestimate the quality of these social objects and pay less attention to them. However, the findings are not entirely consistent with each other. For example, Askin and Mauskopf (2017) find that the songs of optimal differentiation are more likely to succeed than the typical songs. Similarly, Shi et al. (2021) find that the tag distance has an inverted U-shaped influence on the popularity of questions.

There are many other bottlenecks of knowledge spanning. For example, Kuhn (1977) argues that education plays a crucial role in maintaining tradition and restricting innovation. It is difficult to transfer knowledge across groups and generations. Since knowledge accumulates over time, successive generations of innovators face an increasingly massive amount of accumulated knowledge. Human beings are the carrier of knowledge. The difficulty in the transferring of knowledge (i.e., human capital) becomes the bottleneck of innovations. Therefore, Jones (2009) argues that there is an increasing *burden of knowledge* for innovators. Gruber et al. (2012) find that scientific education helps inventors carry out knowledge recombination across technological boundaries. Nevertheless, the breadth of inventors' knowledge recombination decays over time after their education.

Prior research on category spanning proposes two mechanisms to explain the negative consequence of category spanning (Keuschnigg and Wimmer, 2017). First, category-spanning tends to reduce the niche fitness of cultural products. Category

spanners are usually generalists who allocate their time and effort to multiple categories. Compared with the specialists who focus more on a particular field, they gain less experience, ability, and niche fitness. Second, category-spanning tends to increase the audience's confusion towards cultural products. It is more difficult for the audience to form stable expectations for category spanners than non-spanners. There are cognitive difficulties when evaluating questions raised by knowledge spanners. Since the questions of a high degree of knowledge recombination usually span multiple disciplines, people need to consume more cognitive resources to understand it. Keuschnigg and Wimmer (2017) find that the confusion mechanism is relatively more important than the fitness mechanism in explaining the spanning effect.

Going beyond the limit of recombination is as bad as falling short. From the perspective of Bourdieu's field theory of science, Foster et al. (2015) conceptualize the choice of research question as strategic action and habitus. They find that conservative strategies are common, while innovative strategies are rare. Although innovative publications can achieve a higher impact than conservative ones, they face a more considerable risk of failing to publish. Specifically, they distinguished five scientific strategies: repeat bridges, repeat consolidations, new bridges, new consolidations, and jumps. More interestingly, they find that the impact of research depends quadratically on the proportion of each scientific strategy, whether innovative or not.

Similarly, Uzzi and Spiro (2005) find an inverted U-shaped relation between the extent of recombination and musical success. Askin and Mauskopf (2017) argue that there is a tradeoff between similarity and differentiation. Uzzi and Spiro (2005) measure the genre-weighted typicality of popular songs with Cosine similarity and find an inverted U-shaped relation between typicality and popularity. Since the degree of spanning can be measured with 1-Cosine similarity and the inverted U-shape is symmetric, their findings suggest an inverted U-shaped relation between boundary spanning and popularity. Using the number of answers to represent the popularity of a question, Shi et al. (2021) construct the knowledge tag network and examine the impact of tag distance for the questions on Stack Overflow. The tag distance measures the degree of knowledge spanning (how close the tags are to each other in a question). Their findings suggest that the tag distance has an inverted U-shaped influence on the popularity of questions and this nonlinear influence is moderated by tag frequency (Shi et al., 2021). Taken together, knowledge spanning has a parabolic effect on the success of questions. Below a critical point, the increase of knowledge spanning would enhance the appeal of questions; otherwise, it would reduce the appeal of questions. Thus, we propose the following hypothesis:

H1: *There exists an inverted U-shaped relation between the distance of knowledge spanning and the appeal of questions.*

Knowledge is also featured by its hierarchical structure (Cole, 1983). Knowledge is not uniformly distributed. The knowledge continuum can be divided into two main parts: the core of knowledge and the research frontier (Cole, 1983). The core usually consists of a small number of well-recognized theories. It is the starting point based on which people can produce new knowledge. The social process of evaluation connects the core and the frontier. As time goes by, only a tiny part of the frontier knowledge is still evaluated as valuable and eventually becomes the core of knowledge. There is greater consensus toward the core of knowledge in the scientific fields at the top of the hierarchy compared with that at the bottom. But there is no such difference in the consensus on the research frontier between the scientific field at the top and that at the bottom of the hierarchy (Cole, 1983).

The present research examines the hierarchical structure of knowledge. The representation of knowledge or thought is

primarily organized in the form of language. Language is characterized by its hierarchical structure (Tang et al., 2019). The American linguists Hayakawa and Hayakawa explain it with the ladder of abstraction (Hayakawa and Hayakawa, 1939). Human beings are capable of reasoning on different levels of abstraction. For example, the cow named Bessie living on a farm could be described on eight levels of abstraction (from concrete to abstract): process level (e.g., an object composed of atoms), name, Bessie, cow, livestock, farm assets, assets, wealth. Hayakawa et al. (1939) argue that the ladder of abstraction widely exists in our thought and communication. The kids of the farmer consider Bessie a lovely cow. In comparison, the farmer describes Bessie as wealthy. It is reasonable to infer that choosing different levels of abstraction will produce different social consequences.

According to the level of abstraction, the categories used to label the types of questions on the Q&A website constitute a knowledge tree. Coarse-grained categories occupy higher positions in the hierarchical structure of knowledge. Accordingly, questions can be classified into two types: abstract questions and concrete questions. Since abstract questions are framed in a more general way, they cover a much larger group of people. In contrast, concrete questions can only cover a much smaller number of users. Yet, just as Hannan and Freeman (1977) argued, although the width of the niche market for concrete questions is narrower than that of abstract questions, they are more manageable, operational, and competitive. However, the superiority of the narrow niche is contingent upon the stability of the environment. If the environment is unstable, choosing a wide niche is a better strategy. Lerner and Lomi (2018) find that Wikipedia articles with coarse-grained categories can get more attention but lower evaluations (e.g., they are not featured articles). Overall, the position of the question in the hierarchical structure, or the hierarchy of knowledge, also plays an important role in knowledge sharing (Lerner and Lomi, 2018). Based on the analysis above, we propose the following hypothesis:

H2: Knowledge hierarchy (i.e., the level of abstraction) influences the appeal of questions.

The effect of boundary spanning is contingent on various social factors (Keuschnigg and Wimmer, 2017; Goldberg et al., 2016; Shi et al., 2021; Liu et al., 2020). Keuschnigg and Wimmer (2017) find that the impact of category spanning on the price of the film DVD is moderated by market context (e.g., the major films or not), product familiarity (e.g., past box-office success), and social-cultural distance (e.g., the regularity of co-occurrence). When these moderators are considered, the negative effects of category spanning are far from ubiquitous. People who prefer typicality dislike cultural products that span boundaries (Goldberg et al., 2016). Goldberg et al. (2016) construct a geometric model to measure users' taste for variety and typicality. Their findings suggest that the impact of cultural products' object atypicality (category spanning) on users' latent appeal is moderated by users' taste for variety and typicality.

The effect of category spanning is contingent on the fuzziness of categories (Kovács and Hannan, 2010). The strength of category boundaries reflects the partiality of membership in categorization research (Hannan et al., 2007). If the partial assignments of membership become common, the boundaries will become blurred or fuzzy. Kovács and Hannan (2010) propose that when the category boundaries are fuzzy, spanning them will not increase the audiences' confusion. If so, the penalties for boundary spanning would be small or even negligible. Else, the penalties would be more significant. As a result, it is easier to span the categories with blurry boundaries. This idea is captured by the concept of category contrast. Low contrast or high fuzziness reduces the appeal of the object (Negro et al., 2011). Since coarse-

grained categories are in the higher positions of the hierarchical structure of knowledge, they are more abstract and do not have clear-cut boundaries. As a result, the effect of category spanning for the question with abstract categories will be larger than that with concrete categories. Based on the logic above, we propose our hypothesis as follows:

H3: The impact of knowledge spanning on the appeal of questions is moderated by the question's position in the knowledge hierarchy. To be specific, the inverted U-shaped influence of knowledge spanning is more prominent for the questions with more abstract categories.

Methods

Data. We collect the data from Zhihu.com. Zhihu is the largest Q&A website in China as an alternative to Quora (Wikipedia, 2021a, 2021b). Quora, one of the most famous Q&A websites, has attracted more than 326 million registered users (Wikipedia, 2021a). Similarly, the number of registered users of Zhihu.com reached 220 million in 2018 (Wikipedia, 2021b). In our dataset, there are 312,053 questions asked by the users on Zhihu.com from December 2010 to May 2018. Each question record contains the question id, the number of answers, the number of question followers, the content of the question, and a list of categories. Most questions (66%) have two or more categories.

We choose to study knowledge spanning Zhihu.com, not only because it has a vast number of users in China and is playing an increasingly important role in the formation of public opinion, but also because Zhihu.com officially maintains a well-structured knowledge tree. The node of Zhihu's knowledge tree represents the category of questions, and the direct link from the father node to the child node suggests the hierarchical relationship between them. Based on the knowledge tree and the questions' category lists, we design our research as follows (see Fig. 1).

First, we construct the knowledge space using word embedding models. Viewing the list of categories appearing on the question webpage as a sentence, we employ the Word2Vec model to train a neural network model. After that, we can represent each category and question as an N -dimensional vector. Thus, we can measure the distance of knowledge spanning each question based on its categories' positions in the N -dimensional knowledge space. Second, Zhihu.com officially maintains a knowledge tree for all the categories of the questions ($N=108,432$). We can calculate the distance from each category to the root node in the knowledge tree and measure the knowledge hierarchy for each question. Third, we quantify the appeal of questions based on the number of followers of each question. Finally, we build up non-linear regression models to test our hypotheses.

Measures

Knowledge spanning. Natural language processing researchers have made significant progress in representing relationships between words. The words in a corpus can be embedded as vectors into a dense, continuous, high-dimensional space. These vector space models, known collectively as word embeddings models, have attracted widespread interest among computer scientists and computational linguists due to their ability to capture and represent complex semantic relations. Word embedding models have inspired a wide range of item-context embedding models beyond words, ranging from images (Xian et al., 2016) and audio clips (Xie and Virtanen, 2019) to graphs (Perozzi et al., 2014; Grover and Leskovec, 2016) and journals (Tshitoyan et al., 2019).

The ongoing development of neural networks and deep learning also provide potent tools (especially the word2vec algorithm) to explicitly model the latent knowledge space. Based

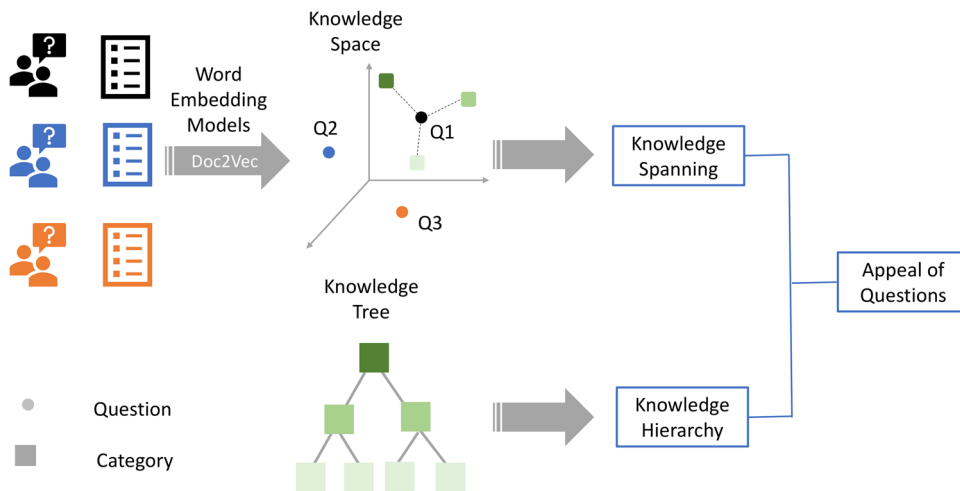
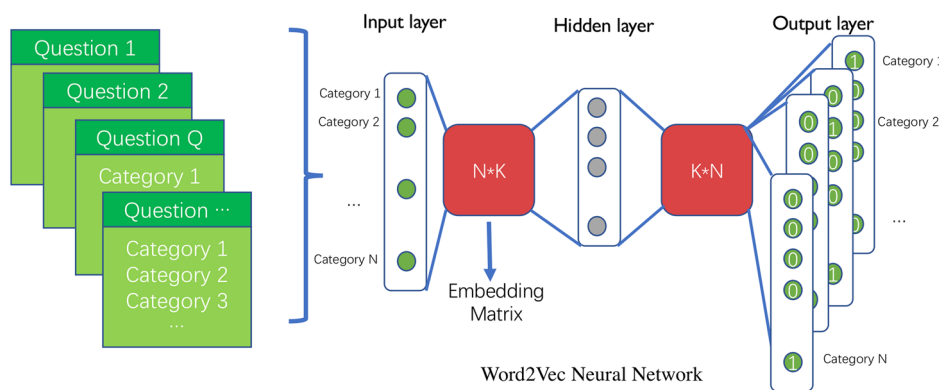


Fig. 1 The research design of knowledge spanning. The circle represents the question, and the square represents the category. When a user asks a question, they mark it with several categories. We construct the knowledge space with word embedding models and build the knowledge tree based on the hierarchical relationship of categories. Then, we can measure knowledge spanning with the knowledge space, quantify the knowledge hierarchy with the knowledge tree, and investigate their relationships with the appeal of questions.

Step 1: Training the Word2vec Neural Network



Step 2 : Getting the Vector of Category i from Embedding Matrix of trained Word2Vec Neural Network

$$T_i = (t_1, t_2, \dots, t_N), \text{ for } i = 1, 2, \dots, N$$

N: number of Categories
 K: dimension of the embedding vector
 S_Q : a set of categories of question Q

Step 3: Calculating Knowledge Spanning (KS) of the given Question Q

$$KS(Q) = \frac{2}{n(n-1)} * [\sum_{\{i,j\} \in S_Q \otimes S_Q, i \neq j} 1 - \text{CosineSimilarity}(T_i, T_j)]$$

Fig. 2 The neural network model of word embeddings. It shows the three steps employed in our study.

on the knowledge space, the concept of knowledge spanning can be reformulated. In our word embedding model, each category is represented as a vector in a vector space. The categories in similar contexts will be positioned nearby, whereas categories in distinct contexts will be positioned farther apart.

Word embedding models aim to represent all words from a corpus within the K -dimensional space that best preserves distances between n words across m semantic contexts. In this study, we view a question as a document and its categories as words. Given the 99th percentile of the length of the category sequence is five, we set the window size of our word2vec model to five. For a given category sequence of a question, we make each category in the sequence as the center words and the other words in the window as surrounding words. In this case, the specific

order of the category sequence does not matter. For example, given a sequence of five categories $[x_1, x_2, x_3, x_4, x_5]$, we will create five copies of the category sequence, in which x_i ($i = 1, 2, 3, 4, 5$) is the center word respectively, and the other four categories serve as the surrounding words. Thus, we can get 20 pairs of center and surrounding words from these five categories. Figure 2 schematically illustrates the neural network structure of the word embedding models with three steps.

Step 1, we train a word2vec neural network using the skip-gram model with negative sampling. Each category of a question is treated as a center category, and the other categories are used as surrounding categories (i.e., contexts). The skip-gram model proposes that we can use the center category (input) to predict its surrounding categories (output), and it calculates the conditional

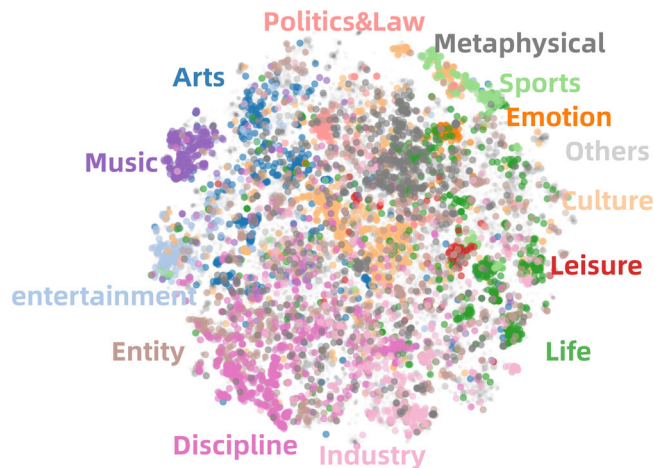


Fig. 3 The visualization of 2D knowledge space. Each point represents a category, and the points of the same category are visualized with the same color.

probability of generating the surrounding categories given a center category. The model parameters of our skip-gram model are the embedding matrix which represents N categories with K dimensions. We set $K = 50$ in this study. After model parameters are initialized, we alternatively perform forward propagation and backpropagation and update model parameters using gradients given by backpropagation. As shown in Fig. 2, the solution is a matrix with N rows and K columns. According to Mikolov et al. (2013), we can denote the maximum distance of the words as C , the vocabulary size as V , and the embedding size (i.e., the number of dimensions) as D . Then the training complexity of the skip-gram model is proportional to $Q = C \times (D + D \times \log_2(V))$ (Mikolov et al., 2013).

Step 2, we can get the category vector from the trained word2vec model.

Step 3, we calculate the knowledge spanning for each question. Cosine similarity is widely used to measure the cosine of the angle between two vectors in natural language processing (Caliskan et al., 2017; Garg et al., 2018; Uzzi and Spiro, 2005). Cosine similarity is the dot product of two vectors normalized by their vector length (Jurafsky and Martin, 2023). High cosine means high similarity. For a question with two categories a and b , we define the semantic similarity as the cosine similarity of their vectors (T_a, T_b) in the knowledge space,

$$\text{CosineSimilarity}(T_a, T_b) = \frac{T_a \cdot T_b}{|T_a| |T_b|}$$

Thus, the distance of knowledge spanning is $1 - \text{CosineSimilarity}(T_a, T_b)$. For those questions with three or more categories, the distance of knowledge spanning in the knowledge space $KS(Q)$ is represented as follows:

$$KS(Q) = \frac{2}{n(n-1)} * \left[\sum_{\{i,j\} \in |S_Q \otimes S_Q|, i \neq j} 1 - \text{CosineSimilarity}(T_i, T_j) \right]$$

In the equation, n is the number of categories for question Q . For questions with only one category, $KS(Q) = 0$.

To summarize, we can get the knowledge space with this method and measure the distance of knowledge spanning each question. As the dimension of knowledge space is much larger than two, we need to conduct dimension reduction before visualization using T-SNE. Figure 3 shows the 2d knowledge space. Each node represents a category, and the nodes of the same parent category are visualized with the same color. Because the distribution of knowledge spanning is highly skewed, we transform it with its logarithmic form ($M = -2.06, SD = 0.71$).

Table 1 Descriptive information of the measurements.

Variables	Mean	Std	Min	Median	Max
Appeal of question (log)	4.018	2.224	0.000	3.555	13.021
Knowledge spanning (log)	-1.450	0.918	-3.000	-1.096	0.000
Hierarchy	5.397	1.801	0.000	5.000	14.000
Title length	21.580	10.186	3.000	19.000	125.00
Lasting days (log)	5.439	1.697	2.197	5.929	7.901377

Knowledge hierarchy. Each category has a category page on Zhihu.com. The category page displays the number of followers, the number of discussions related to the category, a list of father categories, a list of child categories, a brief introduction of the category, a list of discussions, a list of highly rated answers, and a list of new questions. Zhihu.com adopts a hybrid classification system to facilitate the categorization of questions. It was initially designed by the website developers and further improved by active users. First, when a user raises a question on Zhihu.com, she has to choose up to five categories for the question. Second, Zhihu.com will automatically suggest three categories based on the content of the question.

The knowledge tree of Zhihu.com is constructed based on the hierarchical relationship between different categories. To make the category structure concise, non-repetitive, and non-overlapping, Zhihu.com adopts the scientific taxonomy method to construct the knowledge tree. Based on the parent-child relationship, the categories form a directed acyclic graph (DAG). The root node of the knowledge tree is the top-level parent category of all other categories. It has six child nodes or sub-categories: subject, entity, metaphysics, industry, life/art/culture/activities, and unclassified. Because the category of unclassified reveals little information about the essence of questions, it is not considered in this research.

We use the knowledge tree to measure the knowledge hierarchy of each question ($M = 8.60, SD = 1.80$). The hierarchical level of categories reflects the level of abstraction of questions. We set the lowest hierarchical level as zero. The category on the higher hierarchical level in the knowledge tree has a larger abstraction level. For the questions with more than one category, we measure the knowledge hierarchy by computing the average of the abstraction level for all categories of the question.

Appeal of questions. Consistent with prior research, we measure the appeal of questions by content popularity. The popularity of specific content can be quantified by users' particular behaviors stimulated by the content. There are different ways to measure content popularity. For example, the popularity of a post can be expressed by the number of views, likes, followers, reposts, and replies (Ravi et al., 2014; Toba et al., 2014; Yao et al., 2015). There is a strong correlation between these measurements (Wang and Wu, 2016; Wang et al., 2016). In this present research, we use the number of followers to measure the appeal of the question. However, like the other measurements of success, the distribution of the appeal of questions is also highly skewed. Thus, we transform it into the logarithmic form ($M = 4.018, SD = 2.224$). In addition, we have also included the title length ($M = 21.580, SD = 10.186$) and lasting days (log) ($SD = 5.439, SD = 1.697$) as control variables. Table 1 illustrates the main variables and their statistical information in the research.

Results

The first hypothesis of this study proposes an inverted U-shaped relation between the distance of knowledge spanning and the

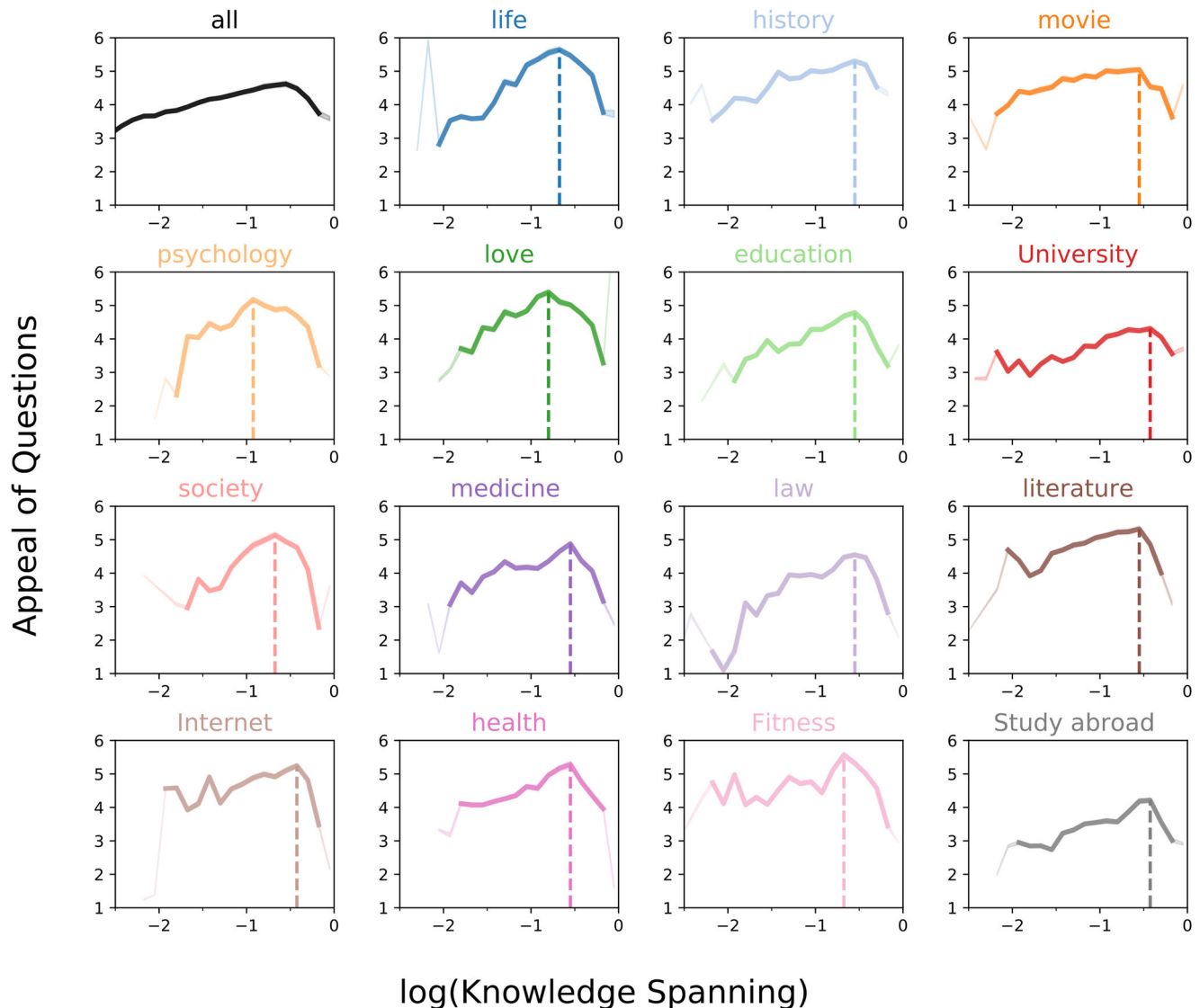


Fig. 4 The effect of knowledge spanning for the questions of different categories. It shows the inverted-U-shaped relationship between knowledge spanning and the appeal of questions for all questions and the questions in the categories of life, history, movie, psychology, love, education, university, society, medicine, law, literature, Internet, health, fitness, study abroad.

appeal of questions. As shown in the first subplot of Fig. 4, the appeal of questions has an inverted U-shaped relation with knowledge spanning. Using the bootstrap method, we get the average appeal of a question. Furthermore, the gray belt around the black line in Fig. 4 covers 95% of the data. The curve increases when knowledge spanning (\log) is less than -0.6 , and it decreases when knowledge spanning (\log) is larger than -0.6 (see Fig. 4), which supports our hypothesis H1. We have also examined H1 across categories. The other subplots of Fig. 4 demonstrate the relation between knowledge spanning and the appeal of questions for the fifteen most popular categories. The findings indicate that the inverted U-shaped relation exists across categories.

Further, we build multiple linear regression models to formally test our hypotheses (see Table 2). Model 1 is the main effect model in which we include the knowledge spanning (\log) and its square form into our model to test the parabolic relation, using knowledge hierarchy, the length of question title, lasting days (\log), and the days of the week as control variables. Model 2 is the main effect model (scaled) which includes the same variables as Model 1, but all the variables in Model 2 are scaled to facilitate model interpretation. Models 3 and 4 are full

models which include the nonlinear interaction terms between knowledge hierarchy and knowledge spanning. Compared to Model 3, all the variables in Model 4 are scaled for the sake of model interpretation. According to Table 2, the coefficients of the square of knowledge spanning (\log) are negative, indicating an inverted U-shaped relation. Thus, hypothesis H1 is well supported.

The second hypothesis argues that there is an association between knowledge hierarchy and the appeal of the questions. According to Model 1–4 in Table 2, the main effect of knowledge hierarchy on the appeal of questions is positive ($B = 0.01$, $p < 0.001$). In other words, the more abstract the question is, the more appealing it is. Thus, H2 is well supported.

The most important hypothesis of this study (H3) proposes that the nonlinear effect of knowledge spanning is moderated by the knowledge hierarchy of questions. As Models 3 and 4 in Table 2 show, there are significant interactions between knowledge spanning and question knowledge hierarchy. To better understand the moderation effect, we visually illustrate it in Fig. 5. Apparently, the existence of the inverted U-shape depends on the knowledge hierarchy of questions. If the knowledge hierarchy of

Table 2 Linear regression models of the appeal of questions.

	Main effect model Model 1	Main effect model (scaled) Model 2	Full model Model 3	Full model (scaled) Model 4
Spanning	280.625*** (1.831)	126.138*** (-0.823)	334.198*** (5.091)	124.979*** (0.845)
Spanning ²	-76.937*** (1.771)	-34.582*** (0.796)	-42.513*** (6.281)	-33.212*** (0.807)
Hierarchy	0.013*** (0.002)	0.008*** (0.001)	0.016*** (0.002)	0.013*** (0.001)
Spanning × hierarchy			10.404*** (0.938)	8.426*** (0.759)
Spanning ² × hierarchy			5.813*** (1.069)	4.708*** (0.866)
Title length	-0.011*** (0.0003)	-0.049*** (0.001)	-0.011*** (0.0003)	-0.049*** (0.001)
Lasting days (log)	0.724*** (0.002)	0.552*** (0.001)	0.724*** (0.002)	0.552*** (0.001)
Monday	0.007 (0.009)	0.003 (0.004)	0.007 (0.009)	0.003 (0.004)
Tuesday	0.008 (0.009)	0.004 (0.004)	0.009 (0.009)	0.004 (0.004)
Wednesday	0.019* (0.009)	0.009** (0.004)	0.020* (0.009)	0.009** (0.004)
Thursday	-0.0003 (0.009)	-0.0001 (0.004)	-0.0001 (0.009)	-0.00003 (0.004)
Friday	0.001 (0.009)	0.0003 (0.004)	0.001 (0.009)	0.0004 (0.004)
Constant	0.359*** (0.015)	-0.002 (0.002)	0.398*** (0.015)	-0.0005 (0.002)
Observations	404,577	404,577	404,577	404,577
R ²	0.389	0.390	0.390	0.390
Adjusted R ²	0.389	0.389	0.390	0.390

The variables of models 2 and 4 are scaled. The values in parentheses denote standard errors.
p* < 0.05, *p* < 0.01, and ****p* < 0.001.

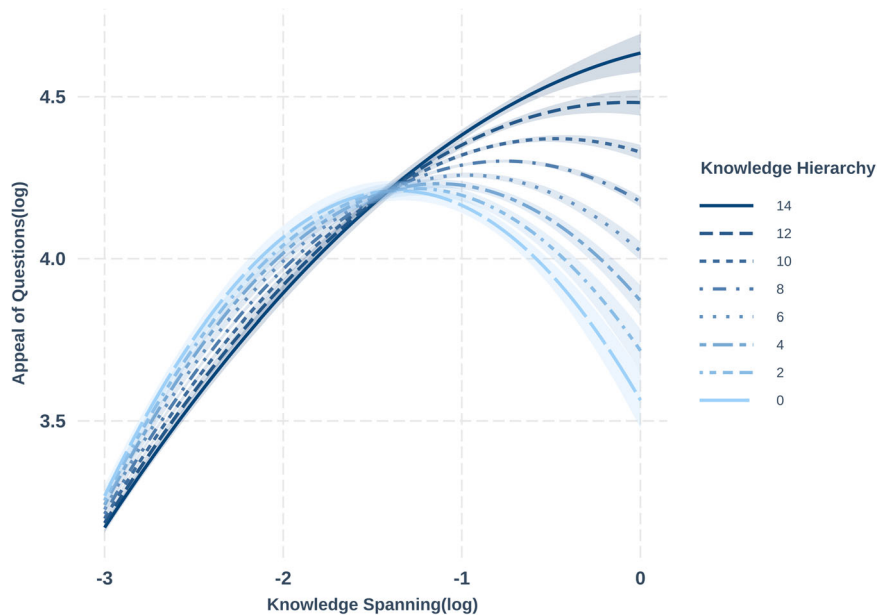


Fig. 5 The moderation effect of knowledge hierarchy. The parabolic impact of knowledge spanning is weakened and disappears when the questions are on the higher level of the knowledge hierarchy.

questions is smaller than six, the inverted U-shape gets more prominent; In contrast, when the knowledge hierarchy of questions is larger than six, the inverted U-shape disappears quickly. Thus, H3 is well supported.

Discussions and conclusions

Drawing on the theories of knowledge recombination, we reformulate the hypothesis of knowledge spanning in the context of the Q&A website. Using the data collected from a large Q&A

website, we examine how the variation of knowledge spanning affects questions' success in attracting public attention. Our findings indicate that good questions have the capability to achieve the tradeoff between similarity and differentiation. The optimal differentiation in the knowledge space shapes the success of questions on the Q&A website (Askin and Mauskopf, 2017; Shi et al., 2021). However, the inverted U-shaped influence of knowledge spanning is contingent upon the position of questions on the ladder of abstraction. To be specific, concrete questions are more prevalent when knowledge spanning is small, while abstract questions are much more popular when knowledge spanning is large.

This study makes an essential contribution to the theories of knowledge recombination by supplying a new perspective on the popular success of questions in the online knowledge market. Our findings largely support the inverted U-shaped knowledge-spanning effect (Askin and Mauskopf, 2017; Foster et al., 2015; Shi et al., 2021; Uzzi and Spiro, 2005). Admittedly, knowledge recombination plays a crucial role in innovation (Schumpeter, 1934; Nelson and Winter, 1982; Koestler, 1964; Zhang et al., 2019; Shi et al., 2021). However, there is no making without breaking. If a question spans a long distance in the knowledge space, it would be unvalued or even completely ignored. In contrast, if a question only spans a very short distance in the knowledge space, it may suffer from a lack of novelty. Therefore, there is a tradeoff or essential tension between tradition and innovation. The tradeoff is achieved by negative feedback. The marginal return of information spanning diminishes with the increase of knowledge spanning. Moreover, when information spanning is above a threshold, the marginal return even becomes negative. We demonstrate that excessive knowledge spanning hinders the appeal of questions. If the risk of recombination is a form of destruction of the tradition, negative feedback suppresses this destruction and improves the stability of the knowledge market.

Interestingly, in comparison with prior research (Askin and Mauskopf, 2017; Foster et al., 2015; Shi et al., 2021; Uzzi and Spiro, 2005), we further show that the existence of the inverted U-shaped effect is contingent on the knowledge hierarchy of questions. According to our findings, only when the knowledge hierarchy of questions is very small (i.e., concrete questions) can a noticeable inverted U-shaped effect be observed. As the knowledge hierarchy of questions decreases, the negative feedback is significantly weakened. And when the knowledge hierarchy of questions is larger than six, the negative feedback almost disappears. Further, because the mean value of knowledge hierarchy is relatively large ($M = 8.60$), most users of the knowledge market would like to ask relatively more abstract questions. Consequently, the negative feedback on knowledge recombination is minimal on the Q&A website. Thus, the Q&A website is more tolerant or supportive of knowledge spanning.

Compared with the situation of high knowledge spanning, the moderation effect of knowledge hierarchy is much smaller when knowledge spanning is below a threshold. When knowledge spanning is below a threshold, people prefer concrete questions to abstract ones and the questions of low knowledge hierarchy lost their advantage over those questions of high hierarchy. On the one hand, according to the *contrast hypothesis* (Kovács and Hannan, 2010), since a low knowledge hierarchy lacks contrast, the penalties associated with low knowledge hierarchy are smaller than that of high hierarchy. On the other hand, according to the *heterogeneity hypothesis* (Negro et al., 2011), the low knowledge hierarchy implies that the people attracted by the same label are heterogeneous. Thus, they share fewer common features and are more likely to

disagree with each other. When knowledge spanning is below a threshold, the promoting effect is smaller than the inhibiting effect, and the questions of high knowledge hierarchy (i.e., level of abstraction) lose their competitive advantage.

Practical implications. First, maintaining a hierarchical tree of categories is a good idea, especially for the knowledge market. The popular success of questions is socially constructed (Barabási, 2018). The feedback mechanism is strongly driven by the social process of evaluation (Cole, 1983). Social knowledge as public goods is like fictitious capital (e.g., stocks). Surprisingly, stock market prefers unfamiliar explorative patents to incremental exploitative patents (Fitzgerald et al., 2021). Although the firms focusing on the exploitation strategy generate much better operating performance in the short term, they are undervalued by the stock market. We show that the cognitive bias towards knowledge spanning has a similar effect on the Q&A website. Consequently, the Q&A website has much fewer constraints on knowledge recombination, especially when the knowledge hierarchy of questions is large. Consistent with Kovács et al. (2021), our findings also indicate that the knowledge hierarchy can flatten or even reverse the U-shaped impact of knowledge spanning (see Fig. 5). Thus, the knowledge market can better encourage knowledge spanning by controlling the knowledge hierarchy.

Second, the perspectives of geometry and network provide a powerful lens to investigate knowledge recombination (Uzzi and Spiro, 2005; Foster et al., 2015; Kozłowski et al., 2019; Min et al., 2021; Lee and Sohn, 2021). Knowledge recombination occurs in social fields (Bourdieu, 1975). The social field of the Q&A website can be described with the knowledge space which helps people navigate in it. But there is a challenge in capturing agents' positions in the social fields. Prior research on categorizations attempts to solve this question by constructing the feature space and label space (Pontikes and Hannan, 2014). When the number of variables is large, there exists a problem of representation. If we represent them with dummy coding or one-hot encoding, most values of these variables are zero. As a result, it is impossible to compute the distance between the two questions. Fortunately, through building a geometric model of knowledge spanning, we can embed questions into a high-dimensional geometric space, construct the knowledge space, and quantify the distance of knowledge spanning.

Limitations. We acknowledge the limitations of this research which pave the road for future research. First, we have only examined knowledge spanning in asking questions on the Q&A website. How to understand the knowledge spanning in *answering* questions remains a question. Second, we quantify similarity-based boundary spanning with word embeddings and capture the hierarchy-based boundary spanning with network analysis. It is necessary to consider both hierarchy and similarity within one model. One possible solution is to embed the categories of questions into a hyperbolic space (Papadopoulos et al., 2012). Third, we limit our research scope to free knowledge market because it can turn private knowledge into public knowledge, not the opposite. Yet, the fee-based knowledge market is also developing rapidly. For example, Zhihu.com has also begun to adopt the mechanism of paid Q&A. It is necessary to test the findings of this study in the other forms of question-asking or information systems.

It is also noteworthy to clarify the latent premises and restrictions of the study. First, the most important premise of our study is to conceptualize innovation or creativity as recombination. Admittedly, logical deductions, intuitions, and

new discoveries can also bring innovation or creativity. Second, we argue that the visible social knowledge in the online knowledge market can be summarized into given knowledge categories. However, these given categories may not cover the questions that are less frequently raised.

Conclusions. In all, this research supports that knowledge spanning has an inverted U-shaped effect on the success of questions. However, the parabolic impact of knowledge spanning is weakened and disappears when the questions are on a higher level of abstraction. In other words, moving up the ladder of abstraction helps break the boundaries of knowledge space. Social evaluation plays a crucial role in linking the knowledge core and the knowledge in communication (Cole, 1983). Human beings evaluate questions from two dimensions: one is the similarity between questions, and the other is the hierarchy of the questions on the ladder of abstraction. Our study demonstrates that the Q&A website and its users are more supportive of knowledge recombination because people can raise questions in an abstract way. The advantage of the knowledge market lies in constructing a tolerant knowledge space and building a hierarchical knowledge tree.

Data availability

The dataset and code employed in this study are available on Open Science Framework <https://doi.org/10.17605/OSF.IO/EPQTX>.

Received: 29 January 2023; Accepted: 18 May 2023;

Published online: 23 June 2023

References

- Askin N, Mauskopf M (2017) What makes popular culture popular? Product features and optimal differentiation in music. *Am Sociol Rev* 82(5):910–944
- Badilescu-Buga E (2013) Knowledge behaviour and social adoption of innovation. *Inf Process Manag* 49(4):902–911
- Barabási A-L (2018) *The formula: the universal laws of success*, 1st edn. Little, Brown and Company
- Bourdieu P (1975) The specificity of the scientific field and the social conditions of the progress of reason. *Soc Sci Inf* 14(6):19–47
- Bourdieu P, Nice R (2004) *Science of science and reflexivity*. University of Chicago Press
- Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186
- Chua AYK, Banerjee S (2015) Answers or no answers: studying question answerability in Stack Overflow. *J Inf Sci* 41(5):720–731
- Cole S (1983) The hierarchy of the sciences? *Am J Sociol* 89(1):111–139
- Ferguson J-P, Carnabuci G (2017) Risky recombinations: institutional gatekeeping in the innovation process. *Organ Sci* 28(1):133–151
- Fitzgerald T, Balsmeier B, Fleming L, Manso G (2021) Innovation Search Strategy and Predictable Returns. *Manage Sci* 67(2):1109–1137. <https://doi.org/10.1287/mnsc.2019.3480>
- Foster J. G., Rzhetsky A, Evans J. A (2015) Tradition and Innovation in Scientists' Research Strategies. *Am Sociol Rev* 80(5):875–908. <https://doi.org/10.1177/0003122415601618>
- Fronzetti Colladon A, D'Angelo CA, Gloor PA (2020) Predicting the future success of scientific publications through social network and semantic analysis. *Scientometrics* 124(1):357–377
- Fu H, Oh S (2019) Quality assessment of answers with user-identified criteria and data-driven features in social Q&A. *Inf Process Manag* 56(1):14–28
- Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci USA* 115(16):E3635–E3644
- Gazan R (2010) Microcollaborations in a social Q&A community. *Inf Process Manag* 46(6):693–702
- Goldberg A, Hannan MT, Kovács B (2016) What does it mean to span cultural boundaries? Variety and atypicality in cultural consumption. *Am Sociol Rev* 81(2):215–241
- Grover A, Leskovec J (2016) node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864. <https://doi.org/10.1145/2939672.2939754>
- Gruber M, Harhoff D, Hoisl K (2012) Knowledge Recombination Across Technological Boundaries: Scientists vs. Engineers. *Manage Sci* 59(4):837–851. <https://doi.org/10.1287/mnsc.1120.1572>
- Guan J, Pang L (2018) Bidirectional relationship between network position and knowledge creation in Scientometrics. *Scientometrics* 115(1):201–222
- Guan J, Yan Y, Zhang JJ (2017) The impact of collaboration and knowledge networks on citations. *J Informetr* 11(2):407–422
- Hannan MT (2010) Partiality of memberships in categories and audiences. *Annu Rev Sociol* 36(1):159–181
- Hannan MT, Freeman J (1977) The population ecology of organizations. *Am J Sociol* 82(5):929–964
- Hannan MT, Pólos L, Carroll G (2007) *Logics of organization theory: Audiences, codes, and ecologies*. Princeton University Press
- Hardin G (1968) The tragedy of the commons. *Science* 162(3859):1243–1248
- Hayakawa SI, Hayakawa AR (1939) *Language in thought and action*, 1st edn. Harcourt Brace Jovanovich
- Hsu G, Hannan MT, Koçak Ö (2009) Multiple category memberships in markets: an integrative theory and two empirical tests. *Am Sociol Rev* 74(1):150–169
- Hsu G, Negro G, Perretti F (2012) Hybrids in Hollywood: a study of the production and performance of genre-spanning films. *Ind Corp Change* 21(6):1427–1450
- Hu Y-H, Tai C-T, Liu KE, Cai C-F (2020) Identification of highly-cited papers using topic-model-based and bibliometric features: the consideration of keyword popularity. *J Informetr* 14(1):101004
- Jones BF (2009) The burden of knowledge and the “death of the renaissance man”: is innovation getting harder? *Rev Econ Stud* 76(1):283–317
- Jurafsky D, Martin HJ (2023) *Speech and language processing* (3rd edn draft), Chapter 6 <https://web.stanford.edu/~jurafsky/slp3/6.pdf>
- Keuschnigg M, Wimmer T (2017) Is category spanning truly disadvantageous? New evidence from primary and secondary movie markets. *Soc Forces* 96(1):449–479
- Koestler A (1964) *The act of creation*. Penguin Books
- Kovács B, Carnabuci G, Wezel FC (2021) Categories, attention, and the impact of inventions. *Strateg Manag J* 42(5):992–1023
- Kovács B, Hannan M (2010) The consequences of category spanning depend on contrast. In: *Categories in markets: origins and evolution*. Emerald Group Publishing Limited, pp. 175–201
- Kozlowski AC, Taddy M, Evans JA (2019) The geometry of culture: analyzing the meanings of class through word embeddings. *Am Sociol Rev* 84(5):905–949
- Kuhn TS (1977) *The essential tension: Selected studies in scientific tradition and change*. University of Chicago Press
- Lee J, Sohn SY (2021) Recommendation system for technology convergence opportunities based on self-supervised representation learning. *Scientometrics* 126(1):1–25
- Lerner J, Lomi A (2018) Knowledge categorization affects popularity and quality of Wikipedia articles. *PLoS ONE* 13(1):e0190674
- Liu N, Mao J, Guan J (2020) Knowledge convergence and organization innovation: the moderating role of relational embeddedness. *Scientometrics* 125(3):1899–1921
- Liu X, Wang GA, Fan W, Zhang Z (2020) Finding useful solutions in online knowledge communities: a theory-driven design and multilevel analysis. *Inf Syst Res* 31(3):731–752
- Liu Z, Jansen BJ (2017) Identifying and predicting the desire to help in social question and answering. *Inf Process Manag* 53(2):490–504
- Liu Z, Jansen BJ (2018) Questioner or question: predicting the response rate in social question and answering on Sina Weibo. *Inf Process Manag* 54(2):159–174
- Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*
- Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 3111–3119
- Min C, Bu Y, Sun J (2021) Predicting scientific breakthroughs based on knowledge structure variations. *Technol Forecast Soc Change* 164:120502
- Moaniba I. M, Su H.-N, Lee P.-C (2018) Knowledge recombination and technological innovation: the important role of cross-disciplinary knowledge. *Innovation* 20(4):326–352. <https://doi.org/10.1080/14479338.2018.1478735>
- Morita K, Atlam E-S, Fuketra M, Tsuda K, Oono M, Aoe J (2004) Word classification and hierarchy using co-occurrence word information. *Inf Process Manag* 40(6):957–972
- Negro G, Hannan MT, Rao H (2011) Category reinterpretation and defection: modernism and tradition in Italian winemaking. *Organ Sci* 22(6):1449–1463
- Nelson RR, Winter SG (1982) *An evolutionary theory of economic change*. The Belknap Press of Harvard University Press
- Neshati M (2017) On early detection of high voted Q&A on stack overflow. *Inf Process Manag* 53(4):780–798

- Ordanini A, Nunes JC, Nanni A (2018) The featuring phenomenon in music: how combining artists of different genres increases a song's popularity. *Mark Lett* 29(4):485–499
- Osatuyi B, Passerini K, Turel O (2022) Diminishing returns of information quality: untangling the determinants of best answer selection. *Comput Hum Behav* 126:107009
- Ostrom E (1990) *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press
- Papadopoulos F, Kitsak M, Serrano MÁ, Boguñá M, Krioukov D (2012) Popularity versus similarity in growing networks. *Nature* 489(7417):537–540
- Peirce CS, Bisanz E (2016) *Prolegomena to a science of reasoning: phaneroscopy, semeiotic, logic*. Peter Lang edn
- Perozzi B, Al-Rfou R, Skiena S (2014) DeepWalk: online learning of social representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710. <https://doi.org/10.1145/2623330.2623732>
- Pontikes E, Hannan M (2014) An ecology of social categories. *Sociol Sci* 1:311–343
- Qi T, Wang T, Chen N (2020) Analysis of sponsorship networks and cross-domain knowledge exchange: an empirical study on Zhihu. *Int J Crowd Sci* 4(3):255–271
- Ravi S, Pang B, Rastogi V, Kumar R (2014) Great question! Question quality in community Q&A. In: *Eighth International AAAI conference on weblogs and social media*. 8(1):426–435. <https://doi.org/10.1609/icwsm.v8i1.14529>
- Schumpeter AJ (1934) *The theory of economic development: an inquiry into profits, capital, credit, interest, and the business cycle*. Harvard University Press
- Shi Y, Chen S, Kang L (2021) Which questions are valuable in online Q&A communities? A question's position in a knowledge network matters. *Scientometrics* 126(10):8239–8258
- Shin M, Holden T, Schmidt RA (2001) From knowledge theory to management practice: towards an integrated approach. *Inf Process Manag* 37(2):335–355
- Stewart TA (1997) *Intellectual capital: the new wealth of organizations*, 1st edn. Doubleday/Currency
- Sussman SW, Siegal WS (2003) Informational influence in organizations: an integrated approach to knowledge adoption. *Inf Syst Res* 14(1):47–65
- Tang X, Chen L, Cui J, Wei B (2019) Knowledge representation learning with entity descriptions, hierarchical types, and textual relations. *Inf Process Manag* 56(3):809–822
- Tilly C (2006) *Why? What happens when people give reasons...and why*. Princeton University Press
- Toba H, Ming Z-Y, Adriani M, Chua T-S (2014) Discovering high quality answers in community question answering archives using a hierarchy of classifiers. *Inf Sci* 261:101–115
- Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K. A, Ceder G, Jain A (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 571(7763):95–98. <https://doi.org/10.1038/s41586-019-1335-8>
- Uzzi B, Spiro J (2005) Collaboration and creativity: the small world problem. *Am J Sociol* 111(2):447–504
- Wang C-J, Wu L (2016) The scaling of attention networks. *Physica A* 448:196–204
- Wang C-J, Wu L, Zhang J, Janssen MA (2016) The collective direction of attention diffusion. *Sci Rep* 6:34059
- Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):440–442
- Wikipedia (2021a) Quora. <https://en.wikipedia.org/w/index.php?title=Quora>. Accessed 9 June 2021
- Wikipedia (2021b) Zhihu. <https://en.wikipedia.org/w/index.php?title=Zhihu>. Accessed 9 June 2021
- Xian Y, Akata Z, Sharma G, Nguyen Q, Hein M, Schiele B (2016) Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 69–77)
- Xie H, Virtanen T (2019) Zero-Shot Audio Classification Based On Class Label Embeddings. 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 264–267. <https://doi.org/10.1109/WASPAA.2019.8937283>
- Yao Y, Tong H, Xie T, Akoglu L, Xu F, Lu J (2015) Detecting high-quality posts in community question answering sites. *Inf Sci* 302:70–82
- Zhang J, Yan Y, Guan J (2019) Recombinant distance, network governance and recombinant innovation. *Technol Forecast Soc Change* 143:260–272
- Zuckerman EW (1999) The categorical imperative: securities analysts and the illegitimacy discount. *Am J Sociol* 104(5):1398–1438

Acknowledgements

This research work is funded by the National Social Science Fund of China (Grant No. 22BXW032), the Social Science Foundation of Jiangsu Province of China (Grant No. 19JD001), and the Chinese Scholarship Council (Grant No. 202006040168).

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Tiewei Li or Cheng-Jun Wang.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023