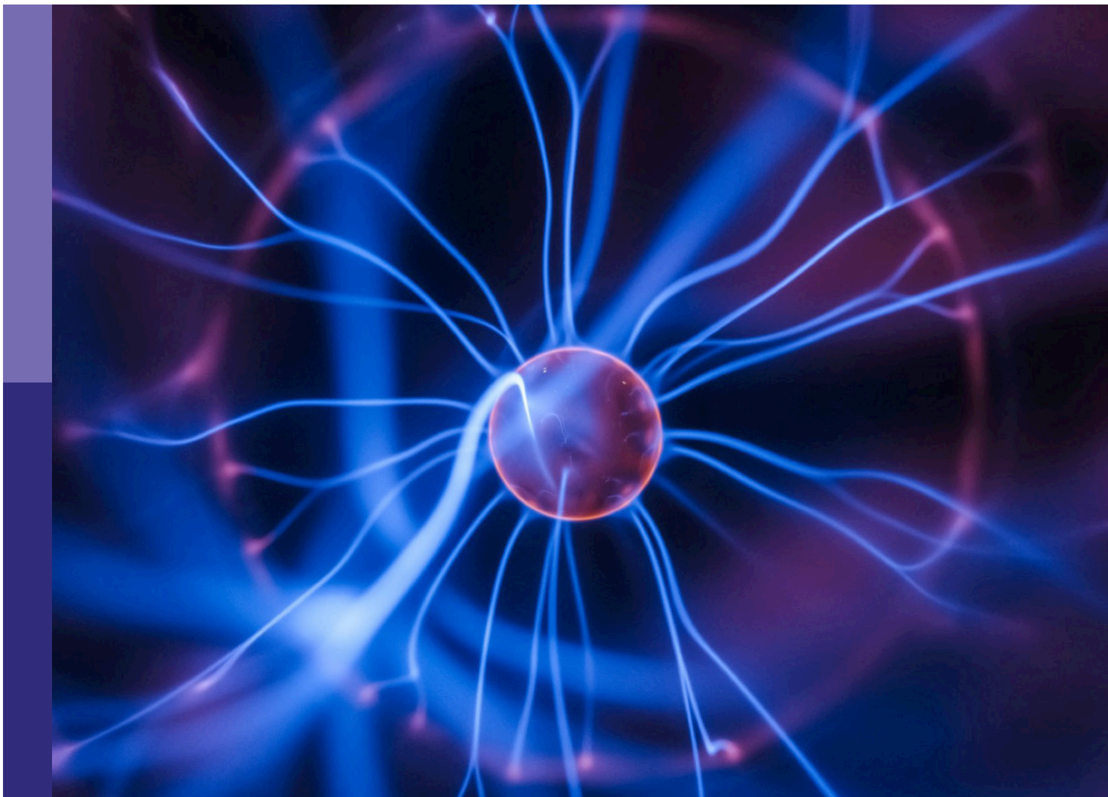
frontiers | Research Topics

Social economic networks

Edited by
Jianguo Liu and Claudio J. Tessone

Published in
Frontiers in Physics



ISSN 1664-8714
ISBN 978-2-8325-3238-6
DOI 10.3389/978-2-8325-3238-6

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Social economic networks

Topic editors

Jianguo Liu — Shanghai University of Finance and Economics, China
Claudio J. Tessone — University of Zurich, Switzerland

Citation

Liu, J., Tessone, C. J., eds. (2023). *Social economic networks*.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3238-6

Table of contents

04 Reconstruction of Unfolding Sub-Events From Social Media Posts
Ren-De Li, Qiang Guo, Xue-Kui Zhang and Jian-Guo Liu

14 Optimal Path Planning With Minimum Inspection Teams and Balanced Working Hours For Power Line Inspection
Zhao-Long Hu, Yuan-Zhang Deng, Hao Peng, Jian-Min Han, Xiang-Bin Zhu, Dan-Dan Zhao, Hui Wang and Jun Zhang

26 Structural evolution of international crop trade networks
Yin-Ting Zhang and Wei-Xing Zhou

38 Evaluating the connectedness of commodity future markets via the cross-correlation network
Lei Hou and Yueling Pan

47 Features of intercity bus passenger group mobility behaviors in the context of smart tourism
Shao-Yong Han, Jing-Chun Zhan, Cui-Hua Xie and Zhen Wang

58 Structural centrality of networks can improve the diffusion-based recommendation algorithm
Yixiu Kong, Yizhong Hu, Xinyu Zhang and Cheng Wang

67 Universal scaling behavior and Hawkes process of videos' views on Bilibili.com
Jiarui Dong, Yuping He, Jiayin Song, Hao Ding and Yixiu Kong

76 Core-middle-periphery network model for China banking system
Na Chen, Jianguo Liu, Yihui Chen, Ding Tu, Yang Ou and Mingzhu Jiang

87 The effect from elimination mechanism on information diffusion on entertainment programs in Weibo
Nannan Xu, Qiaoting Lin, Haibo Hu and Ying Li

102 Mechanism of supply chain coordination based on price discount with privacy protection in one-supplier-one-buyer system
Cui-Hua Xie, Jing-Chun Zhan, Le-Tian Zeng and Shao-Yong Han

frontiers | Frontiers in Physics

ORIGINAL RESEARCH
published: 24 June 2022
doi: 10.3389/fphy.2022.918663

Reconstruction of Unfolding Sub-Events From Social Media Posts

Ren-De Li[1], *Qiang Guo*[1], *Xue-Kui Zhang*[2] *and Jian-Guo Liu*[3]*

[1] Library and Business School, University of Shanghai for Science and Technology, Shanghai, China, [2] Institute of Journalism, Shanghai Academy of Social Science, Shanghai, China, [3] Institute of Accounting and Finance, Shanghai University of Finance and Economics, Shanghai, China

Event detection plays a crucial role in social media analysis, which usually concludes sub-event detection and correlation. In this article, we present a method for reconstructing the unfolding sub-event relations in terms of external expert knowledge. First, a Single Pass Clustering method is utilized to summarize massive social media posts. Second, a Label Propagation Algorithm is introduced to detect the sub-event according to the expert labeling. Third, a Word Mover's Distance method is used to measure the correlation between the relevant sub-events. Finally, the Markov Chain Monte Carlo simulation method is presented to regenerate the popularity of social media posts. The experimental results show that the popularity dynamic of the empirical social media sub-events is consistent with the data generated by the proposed method. The evaluation of the unfolding model is 50.52% ~ 88% higher than that of the random null model in the case of "Shanghai Tesla self-ignition incident." This work is helpful for understanding the popularity mechanism of the unfolding events for online social media.

Keywords: sub-event mining, sub-event detection, sub-event correlation, sub-event summary, sub-event evolution, expert knowledge, social media

OPEN ACCESS

Edited by:
Mahdi Jalili,
RMIT University, Australia

Reviewed by:
Zhao-Long Hu,
Zhejiang Normal University, China
Zhuoming Ren,
Hangzhou Normal University, China

****Correspondence:***
Jian-Guo Liu
liujg004@ustc.edu.cn

Specialty section:
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

Received: *12 April 2022*
Accepted: *12 May 2022*
Published: *24 June 2022*

Citation:
Li R-D, Guo Q, Zhang X-K and Liu J-G (2022) Reconstruction of Unfolding Sub-Events From Social Media Posts. Front. Phys. 10:918663. doi: 10.3389/fphy.2022.918663

1 INTRODUCTION

Unfolding sub-events of a social media event could tell a storyline of public opinions during the event development [1]. Every time when a large-scale incident occurs, around the theme, it will be accompanied by the generation of a lot of discussion and various opinions. A sub-event is a component of a complex event since the topic of public opinions evolves with the development of events. When individuals, celebrities, enterprises, or governments encounter a public relations (PR) crisis, it is difficult to grasp the direction of public opinion from the uncontrolled interpretation of thousands of people. It is vital for PR managers to clarify the trend of public opinion from sub-events of the incident.

For PR crisis events, it has similar characteristics of emergency or epidemic events, such as natural disasters [2, 3], epidemic spreading [4, 5], and sports competitions [6, 7]. The information related to disaster events can be uploaded and reported, which contributes to the disaster reporting [8]. On social media, events and their related sub-events can be discussed or explored through public online posts.

Sub-event identification faces two challenges of ambiguous distinguishability. First is whether similar expressions are effectively distinguished. Online posts contain a massive amount of re-posts or similar user expressions. Second is whether the related expressions can be effectively distinguished. The discussions and expressions will form different topics, reflecting the sub-events from the perspective of user-generated content. But a post belonging to which sub-event

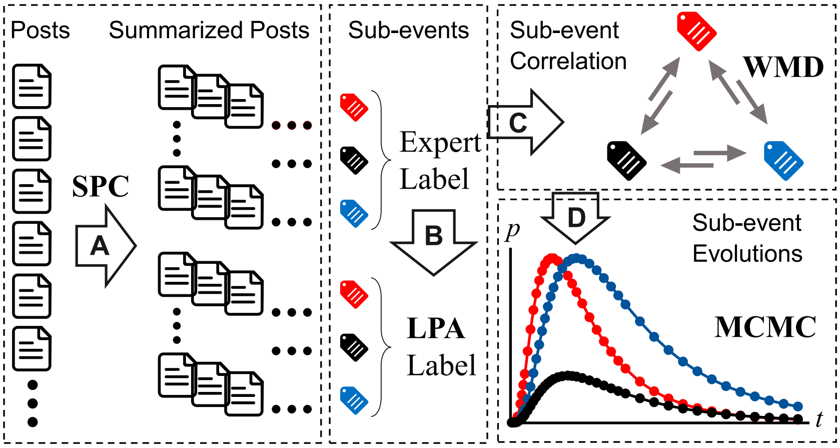


FIGURE 1 | (Color online) Schematic illustration of the proposed method. **(A)** Post summary is conducted by Single Pass Clustering (SPC). **(B)** Sub-event labeling is executed by the Label Propagation Algorithm (LPA). **(C)** Sub-event correlation is carried out by Word Mover's Distance (WMD). **(D)** Sub-event evolution is regenerated by Markov Chain Monte Carlo (MCMC).

needs to be classified. A clear division of sub-events can provide effective support for correlation and evolution analyses.

Inspired by the idea, we present a mode to detect and correlate the sub-events, which aims to unfold a complex event into correlated sub-events and predict the popularity dynamic of social media events. During the modeling process, it is about to solve the two issues which are the ambiguities of sub-event classification (the former two steps of **Figure 1**) and correlation between sub-events (the latter two steps of **Figure 1**). As shown in **Figure 1A**, after collecting the social media posts, a fast clustering method is used to cluster similar posts. The procedure is to reduce the redundancy among replicate posts and each classification stands for a summarized post. In order to unfold the sub-event to meet with the knowledge of PR managers, expert labeling is given and used to predict the unlabeled summarized posts (**Figure 1B**). Each label represents a topic concerned by PR managers, which is defined as a sub-event. The topic correlation is measured by the number of paired posts between sub-events (**Figure 1C**). Finally, by using the Markov Chain Monte Carlo simulation, each development trend of the sub-event can be depicted and compared to the real world topic evolution (**Figure 1D**). This procedure regenerates the results of sub-event popularity curves and will be verified by a null model with random labels.

2 RELATED WORK

2.1 Unfolding Events From Public Information

In order to correctly observe the filtering of the results from public information, a classic model considers the impact of sharing such information on the analytical foundations of reliable sensing [9]. The observations can be obtained by the text, image, video, and voice message provided by social media users. [10]. Based on these observations, several unfolding methods have been developed. CrisisTracker's clustering system [11] includes event detection, content ranking, and summarization while retaining the drill-down functionality to raw reports. The security information and event management systems could also connect events by pattern matching [12]. An ontology method systematizes the available solutions under a modular- and platform-independent conceptual framework [13]. An iterative expectation-maximization algorithm is proposed to find the truth of the events in social sensing with information flows. Among these studies, the verification of events or sub-events is based on the supervised learning with specific labels, whereas PR crisis usually has no label for identification.

Although some research has examined the use of social media for mitigating crises and emergencies [14–16], the use of specialized detection methods [17] for clarifying the ambiguity of classification is still lacking. The main challenge is to find the popularity mechanism of social media events. In this article, we use public observations to sort out the sub-events by combining the expert knowledge and correlate these sub-events to a topic tree and popularity trends for the event storyline.

2.2 Sub-Event Detection

An event usually contains the cause and result stages, where the sub-event refers to one of the stages of an event [18]. The sub-event detection can be achieved by many classic unsupervised methods as follows: 1) the burst-topic detection is used to identify important moments, which argues that the sharp increase in the number of status updates corresponds to the occurrence of important moments in the event [19]. 2) The event summarization usually contains machine learning techniques such as hidden Markov model [20], hierarchical Dirichlet processes [21], and graph optimization formulation [7]. 3) The clustering approaches include word co-occurrence [22], hierarchical clustering algorithm [23], K-nearest neighbor clustering approach [24], artificial neural networks [10], support vector machine [25]. 4) The spatial and temporal distribution methods are also widely used [3, 26, 27].

One major theoretical issue that has dominated the unsupervised detection field for many years concerns the ambiguity of classification for a sub-event. Semi-supervised approaches have also been explored for this task, especially concerning crisis events [28, 29]. However, due to a lack of expert knowledge, the effect of classification may derive from the common sense of PR management. In this article, we proposed a simple procedure to summarize the sub-events by combining the clustering-based single pass algorithm and graph-based label propagation algorithm by introducing the expert knowledge. The Single Pass Clustering (SPC) is a method to simply merge similar posts. The Label Propagation Algorithm (LPA) is to solve the ambiguity and gives a clear classification based on expert knowledge.

2.3 Sub-Event Correlation

The correlation approach contains a causality or correlation pattern of sub-events. Two kinds of methods can reveal the unfolding event to evolve. The first one is graph-based methods, which concerns the correlation pattern of sub-events. A maximum-weighted bipartite graph matching is created to correlate events [30]. The recurrent sequence model [31, 32] has experimented with a recurrent neural network of LSTM for script learning to predict the probability of the next event. An event-oriented similarity graph is designed to represent the relationship among sub-events [18]. A subgraph similarity is used to measure the event relationships and generate an evolution correlation [33]. The second one is causal inference methods, which concern the causality patterns of sub-events. The generalization of redefining mining aims to find the correlation between disjoint sets of related objects [1]. An event–level attention mechanism is utilized to represent the relations between subsequent events [34]. A logical correlation is proposed for common sense inference of the given event [35]. An event ontology knowledge model is built to construct the evolution patterns [36].

These methods are based on a network or sequential perspective. However, if sub-event correlation refers to topic-level correlation, there will be a multiple pair problem. One sub-event contains several posts about a topic and so does the other sub-events. The correlation of sub-events happens between the topic posts. PR managers are sensitive to the posts that change with the topic evolving [37], but few studies have supported the topic-level correlation. Although the LDA-based model could extract the topics [2, 38], the correlation between the posts inside of topics is still an open question. In this article, the Word Mover's Distance (WMD) method is applied to calculate the correlation of the posts in different topics (sub-events). Then, the Markov Chain Monte Carlo (MCMC) simulation method is introduced to predict topics' evolutionary trends.

3 METHODS

3.1 Single Pass Clustering

The SPC method is a classical method for streaming data clustering. For data streams arriving in sequence, the method processes the data once at a time in the order of input. It is an incremental algorithm, which has a high time efficiency. The shortcoming is that the method depends on the input order. If the data streams arrive in different orders, different clustering results will appear.

Given the Weibo post document set $d = \{d_1, d_2, \ldots, d_m\}$, each document d_i contains a variable length sequence of words $w_i^1, w_i^2, \ldots, w_i^{T_i}$. We use Doc2VecC to vectorize each post and the words in it. The Doc2VecC method defines the probability of observing a target word w^t:

$$P\left(w^t|\mathbf{c}^t, \hat{\mathbf{x}}\right) = \frac{exp\left(\mathbf{v}_{w^t}^T\left(\mathbf{U}\mathbf{c}^t + \frac{1}{T}\mathbf{U}\hat{\mathbf{x}}\right)\right)}{\sum_{w'\in V} exp\left(\mathbf{v}_{w'}^T\left(\mathbf{U}\mathbf{c}^t + \frac{1}{T}\mathbf{U}\hat{\mathbf{x}}\right)\right)}, \quad (1)$$

where w^t is the target word, $\mathbf{c}^t$ is the word's local context, $\hat{\mathbf{x}}$ is the global context, $\mathbf{v}^T$ is a trainable parameter, V is the vocabulary used in the training corpus, $\mathbf{U}$ is the learned matrix in which each row represents a vector for one word, and T is the length of document.

The loss function is:

$$l = -\sum_{i=1}^{n}\sum_{t=1}^{T_i} P\left(w^t|\mathbf{c}^t, \hat{\mathbf{x}}\right). \quad (2)$$

Using the training model, each document can be represented as an average of embeddings of the words:

$$\mathbf{d}_i = \frac{1}{T}\sum_{w\in d_i}\mathbf{w}, \quad (3)$$

where $\mathbf{d_i}$ is the vector for document d_i and $\mathbf{w}$ is a row in $\mathbf{U}$ and is the embedding for word w.

The similarity of the two post document vectors $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$ is measured by cosine metric:

$$S\left(\mathbf{d}_i, \mathbf{d}_j\right) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_i| \cdot |\mathbf{d}_j|}. \quad (4)$$

The SPC method is used to cluster the posts roughly since it only process the post documents once. The algorithm is as follows:

Algorithm 1. Single Pass Clustering (SPC)

```
Input: post documents set $d = \{d_1, d_2, \cdots, d_m\}$;
       class set $D = \{\}$ ;
       similarity threshold $S_T$;
Output: summarized documents set $D = \{D_1, D_2, \cdots, D_n\}$
$d = \{d_1, d_2, \cdots, d_m\}$ Doc2VecC to $\{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_m\}$;
for $\mathbf{d}_i \in d$ do
    let $max = 0$, $simD = \{\}$;
    for $D_j \in D$ do
        let $s = S(\mathbf{d}_i, center(D_j))$;
        if $s > max$ then
            $max = s$, $simD = D_j$
    if $s < S_T$ then
        add $\{\mathbf{d}_i\}$ to $D$ ;
    else
        add $\{\mathbf{d}_i\}$ to simD ;
        update centroid $simD$ with $mean(simD)$
for $D_j \in D$ do
    summray={};
    for $\mathbf{d}_i \in D_j$ do
        add $\{w \in d_{i,j} | \mathbf{d}_{i,j} \rightarrow d_{i,j}\}$ to summray
    $D_j$ = summray
```

Step 1: Assign the first document $\mathbf{d}_1$ as the representative for D_1.

Step 2: For $\mathbf{d}_i$, calculate the document similarity S with the representative for each existing cluster.

Step 3: If $S_{\max}$ is greater than a threshold value S_T, add the item to the corresponding cluster and recalculate the cluster representative; otherwise, use $\mathbf{d}_i$ to initiate a new cluster.

Step 4: If $\mathbf{d}_i$ remains to be clustered, return to step 2.

The representative is the mean vector of a cluster. After the SPC process, we denote the document vector $i \in [1, \mathrm{m}]$ from cluster $j \in [1, n]$ as $\mathbf{d}_{i,j}$, and the corresponding document as $d_{i,j}$. The clustering set is expressed as $D = \{D_1, D_2, \ldots, D_n\}$.

The number of cluster n is much smaller than the length of posts m. The micro-blog's posts have the attributes of redundancy since a large proportion of user's re-posts. The SPC method is to largely reduce the redundancy among posts.

In order to summarize the words of each clustering, we define

$$D_j = \cup \left\{ w | w \in d_{i,j} \right\}. \tag{5}$$

Then, the vector of the summarized document $\mathbf{D}_j$ can also be calculated by **Eq 3**. After we get the summarized posts, the next task is to label these data.

3.2 Label Propagation Algorithm

The expert knowledge is introduced to label the summarized posts. Experts need to label a small part of the summarized posts to feed the LPA. The LPA considers that the label of each node should be similar to most of its neighbors, and the label is "propagated" to form the same "label" within the same "community" based on the network perspective.

Given annotated data $(\mathbf{D}_1, y_1), \ldots (\mathbf{D}_l, y_l)$ and the labeled set $Y_l = \{y_1, \ldots, y_l\} \in \{1, \ldots, C\}$, where the category C is given by expert and present in the labeled data. Unlabeled data are $(\mathbf{D}_{l+1}, y_{l+1}), \ldots (\mathbf{D}_{l+u}, y_{l+u})$, and $Y_u = \{y_{l+1}, \ldots, y_{l+u}\}$ is the labeled set to predict, where $l + u = n$ and $L \ll u$. The Label Propagation Algorithm (LPA) is used to predict Y_u by Y_l and $X = X_l \cup X_u = \{\mathbf{D}_1, \ldots, \mathbf{D}_{l+u}\}$.

Algorithm 2. Label Propagation Algorithm (LPA)

```
Input: $X = \{\mathbf{D}_1, \cdots, \mathbf{D}_{l+u}\}$;
       labeled set $Y_l = \{y_1, \cdots, y_l\}$;
       Unlabeled set $Y_u = \{y_{l+1}, \cdots, y_{l+u}\}$;
Output: labeled data $\{(\mathbf{D}_1, y_1), \cdots, (\mathbf{D}_{l+u}, y_{l+u})\}$
$Y_u = \{\}$;
for $\mathbf{D}_i \in X$ do
    for $\mathbf{D}_j \in X$ do
        if $i \neq j$ then
            $\omega_{ij} = \exp(-\frac{S(\mathbf{D}_i, \mathbf{D}_j)}{\sigma^2})$;
            $T_{ij} = \frac{\omega_{ij}}{\sum_{k=1}^{l+u} \omega_{kj}}$
for $inter = 1 : t$ do
    for $\mathbf{D}_i \in X$ do
        for $\mathbf{D}_j \in X$ do
            if $i \neq j$ & $random > T_{ij}$ & $\mathbf{D}_j$ is labeled then
                label[i] ← label[j]
```

A fully connected graph is created so that each sample point (labeled and unlabeled) is treated as a node. The following weight calculation is used to set the weights of the edges between two points i,j:

$$\omega_{ij} = \exp\left(-\frac{S\left(\mathbf{D}_i, \mathbf{D}_j\right)}{\sigma^2} \right), \tag{6}$$

where the parameter σ is adjustable. Then, the probabilistic transition matrix $\mathbf{T} \in (l + u) \times (l + u)$ is defined as:

$$T_{ij} = \frac{\omega_{ij}}{\sum_{k=1}^{l+u} \omega_{kj}}. \tag{7}$$

The element T_{ij} is the probability of label j propagating to label i. By probability propagation, the probability distribution is concentrated in a given class, and then the node labels are passed through the weights of the edges. We can express the random walks as given below:

$$y_i[c] = \sum_{j \in X_l} T_{ij}^t \cdot y_j[c], \tag{8}$$

where $y_i[c]$ is the probability of node $\mathbf{D}_i \in X_u$ to have label c. The probability T_{ij}^t is to jump from node $\mathbf{D}_j$ and end up in node $\mathbf{D}_i$ in t steps. The number of steps is a large number (infinity). Since the probabilistic transition matrix $\mathbf{T}$ can be written as a block matrix:

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{ll} & \mathbf{T}_{lu} \\ \mathbf{T}_{ul} & \mathbf{T}_{uu} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ \mathbf{T}_{ul} & \mathbf{T}_{uu} \end{bmatrix}. \tag{9}$$

In the matrix form, **Eq 8** can be induced as flowing:

$$\begin{bmatrix} \hat{\mathbf{Y}}_l \\ \hat{\mathbf{Y}}_u \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ (\mathbf{I} - \mathbf{T}_{ul})^{-1} \cdot \mathbf{T}_{uu} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{Y}_l \\ 0 \end{bmatrix}, \tag{10}$$

where the label vectors of labeled nodes $\hat{\mathbf{Y}}_l = \mathbf{Y}_l$ and the label vectors of unlabeled nodes $\hat{\mathbf{Y}}_u = (\mathbf{I} - \mathbf{T}_{ul})^{-1} \cdot \mathbf{T}_{uu}$. Finally, one can get the label of

$$\mathbf{D}_i \in X_u = \underset{c}{\operatorname{argmax}}\, \hat{\mathbf{Y}}_u[i]. \tag{11}$$

3.3 Word Mover's Distance

In order to correlate the posts between the sub-events, the WMD method is introduced. According to the LPA results, each label represents a sub-event and includes several summarized posts. The WMD is used to calculate the pairs between summarized posts of sub-events. The WMD method measures the semantic distance of the two documents. Each document is a summarized post.

The post document with labeled c is added into the set $\mathbb{C}_c = \{d_i[c]\}, i \in \{1, \ldots, n\}, c \in \{1, \ldots, C\}$, representing a sub-event c of summarized documents.

In order to build the correlation between sub-events, Word Mover's Distance (WMD) is used to identify the similarity between classifications. WMD is a distance between two text documents x, y. Let $|x|, |y|$ be the number of distinct words in x, y. The normalized frequency vectors of each word in x and y are respectively expressed as $\mathbf{f}_x \in \mathbb{R}^{|x|}$ and $\mathbf{f}_y \in \mathbb{R}^{|y|}$ (so $\mathbf{f}_x^{\mathrm{T}}\mathbf{1} = \mathbf{f}_y^{\mathrm{T}}\mathbf{1} = \mathbf{1}$). Then, the WMD is defined as

$$\begin{aligned}\mathrm{WMD}(x,y) &= \min_{\mathbf{F}\in\mathbb{R}^{|x|\times|y|}} \langle \mathbf{S},\mathbf{F}\rangle,\\ \text{s.t.}\quad \mathbf{F}\mathbf{1} &= \mathbf{f}_x,\\ \mathbf{F}^{\mathrm{T}}\mathbf{1} &= \mathbf{f}_y,\end{aligned} \tag{12}$$

where $\mathbf{F}$ is the transportation flow matrix with F_{ij} denoting the amount of flow traveling from word i in x to word j in y and $\mathbf{S}$ is the transportation cost with $S_{ij} = S\,(\mathbf{w}_i, \mathbf{w}_j)$ being the distance between two words measured by the Doc2VecC.

Algorithm 3. Word Mover's Distance (WMD)

```
Input: sub-event set $\mathbb{C} = \{\mathbb{C}_1, \cdots, \mathbb{C}_C\}$;
       similarity threshold $\Theta$;
Output: the number of paired posts $E = \{e_{k,l}\}$
Initialize $E = \{e_{k,l}\}$ with $e_{k,l} = 0$;
for $\mathbb{C}_k \in \mathbb{C}$ do
    for $\mathbb{C}_l \in \mathbb{C}$ do
        for $d_i \in \mathbb{C}_k$ do
            for $d_j \in \mathbb{C}_l$ do
                if WMD$(d_i, d_j) \geq \Theta\}$ then
                    $e_{k,l} += 1$
```

According to the WMD method, one can establish relevant relationships of sub-events according to the similarity between the post d_i in sub-event classifications $\mathbb{C}_k$ and the post d_j in sub-event classifications $\mathbb{C}_l$. We denote the set of paired posts between classifications as

$$e_{k,l} = |\{(d_i, d_j)|d_i \in \mathbb{C}_k, d_j \in \mathbb{C}_l, \mathrm{WMD}(d_i, d_j) \geq \Theta\}|, \tag{13}$$

where Θ is a threshold value.

3.4 Markov Chain Monte Carlo

The WMD method gives the pairs between different sub-events. The core task of our method is to acquire the prior probability and evolution probability, so that the correlation and evolutionary trends can be built.

The prior probability of each sub-event is calculated by using the statistical probability:

$$\pi(k) = \frac{|\mathbb{C}_k|}{\sum_{k=1}^{C}|\mathbb{C}_k|}, \tag{14}$$

where $|\mathbb{C}_i|$ is the number of summarized documents for sub-event i.

The evolution probability between sub-event pairs is calculated using the conditional probability:

$$Q(k,l) = p(\mathbb{C}_l|\mathbb{C}_k) = \frac{p(\mathbb{C}_k|\mathbb{C}_l)}{p(\mathbb{C}_k)} = \frac{|e_{k,l}|}{|\mathbb{C}_k|}. \tag{15}$$

According to the Metropolis rejection defined by Hastings, the acceptance probability is:

$$\alpha(k,l) = \min\left\{\frac{\pi(l)Q(l,k)}{\pi(k)Q(k,l)}, 1\right\}. \tag{16}$$

The Metropolis–Hastings update makes one proposal l, which is the new state with probability $\alpha(k, l)$ but otherwise, the new state is the same as the old state k. By using the Metropolis–Hastings algorithm, one can get the sample collection, which the element is the type of sub-event. Given the length of sample collection T and the number of time slice, each time step t includes the Δn samples. The probability of a sub-event $\mathbb{C}_k$ in the time step t is defined as:

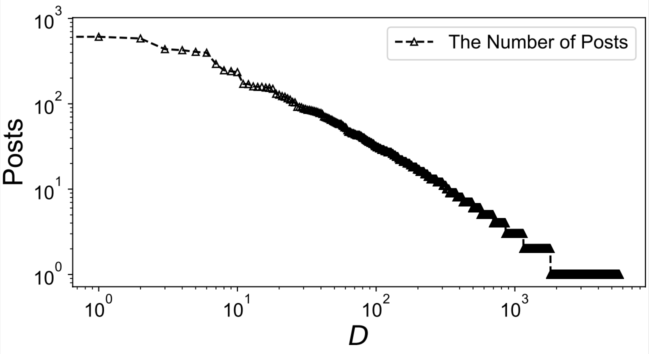


FIGURE 2 | Number of posts in each summarized post D.

$$p_t(\mathbb{C}_k) = |\mathbb{C}_k(t)|/\Delta n. \tag{17}$$

Algorithm 4. MCMC: Metropolis–Hastings algorithm

```
Input: prior probability $\pi(k)$;
       conditional probability $Q(k,l)$;
       time interval $\Delta n = T/slice$
Output: the evolutionary probability of each state $p_t(\mathbb{C}_k)$
Initialize the initial state value of Markov Chain $X_0 = x_0$;
for $inter = 1 : n$ do
    $X_n = x_n, y \sim Q(x, x_n)$;
    $u \sim Uniform[0, 1]$;
    $\alpha(x_n, y) = \min\{\frac{\pi(y)Q(y,x_n)}{\pi(x_n)Q(x_n,y)}, 1\}$;
    if $u < \alpha(x_n, y)$ then
        $X_{n+1} = y$
    else
        $X_{n+1} = x_n$
Let A be a zero C*T matrix;
for $c = 1 : C$ do
    for $t = 1 : T$ do
        $A[c,t] = |X(c)|/\Delta n$
```

In the end of the model process, the regenerated popularity curves of every sub-event can be obtained.

3.5 Model Evaluation

The regenerated popularities have to be evaluated by comparing the real dynamic model and a random model for reference.

3.5.1 The Real Popularity Dynamic

The real evolution of the "Shanghai Tesla self-ignition incident" is measured by

$$p_t(\widehat{\mathbb{C}_k}) = |\widehat{\mathbb{C}_k(t)}|/\Delta\hat{n}, \tag{18}$$

where each time step t includes the $\Delta\hat{n}$ overall documents in 2 days and $|\widehat{\mathbb{C}_k(t)}|$ is the number of real sub-events $\widehat{\mathbb{C}_k(t)}$ in each time step.

TABLE 1 | Example of summarized posts.

No.	Original post (part of the sample)	Similarly	Keywords	Expert Label
1	Suspicious Tesla sudden **self-burning** cause heavy **losses** in a Shanghai **parking** space. A part of surveillance **video** of an underground **parking** space popped up and spread on Weibo. In the **video**, a parked Tesla erupted 'like a **flamethrower**'. The **fire** at the scene has been put down. Except the Audi next to it, several **cars** were **burned** which cause heavy **losses**	0.92	self-burning; video; parking; flame; loss	Event Happen
2	A Shanghai Tesla caught on **fire** in underground **parking**, all surrounding **cars destroyed** in the **video**. A Tesla Model S was in **flames spontaneously** in an underground **parking** of Shanghai Xuhui district. The **fire** caused other **vehicles** parked around **loss**			
3	Tesla **responds** to the **self-burning** of Shanghai Tesla: **Verifying** the situation. In response to reports that a Tesla car suddenly **self-burning** in a Shanghai community **parking** space, Tesla's official Weibo **responded** that 'After learning of the **accident** in Shanghai, we sent a **team** to the **scene** at the first time. We are actively contacting relevant **departments** and **cooperating** to **verify** the situation. According to the current information, there were no **casualties**	0.95	responds; self-burning; accident; verify; casualty	Corporate Respond
4	In response to reports that a Tesla car suddenly **self-burning** in Shanghai community **parking** space, Tesla's official Weibo **responded** that 'After learning of the **accident** in Shanghai, we sent a **team** to the **scene** at the first time. We are actively contacting relevant **departments** and **cooperating** to **verify** the situation. According to the current information, there were no **casualties**			
5	'It would be me, if I left the car half hour later!' The **car** owner said, 'The car was, **burned** to the **frame**, it was **terrified**. The **owner** said that this Tesla was **bought** three and a half **years** ago, and it has never been **broken**. The time of the incident was about 30 min after he **parked** the car. 'what if I **parked** the car 30 min later? Or if I stay, in the **car** for another 30 min? I dare not think further . . . '	0.63	car owner; broken; terrified; charging; fire	Client Respond
6	The owner responded: It was not **charging** at the **time** of the **incident**, and it has just finished **supercharging** a few **hours** ago. The **car owner** said that he **parked** the car 1 h before the **fire** without **charging**. In fact, the car finished the **supercharging** only a few **hours** before the **fire**, which increased its cruising range to another 350 kms			

3.5.2 Jensen–Shannon Divergence

Jensen–Shannon Divergence (short for *JSD*) [39] is introduced to measure the similarity between real distribution p_1 and MCMC distribution p_2 and is defined as:

$$JSD(p_1, p_2) = H\left[\frac{1}{2}p_1 + \frac{1}{2}p_2\right] - \frac{1}{2}\left[H(p_1) + H(p_2)\right], \tag{19}$$

$$H(p) = -\sum_{r=1}^{R} p(r)\log p(r), \tag{20}$$

where p_1 and p_2 are the two distributions to be compared and $H(p)$ represents the Shannon entropy. The lower bound is $JSD = 0$ only when two distributions are identical. The smaller the *JSD* value is, the more similar the two distributions are.

3.6 Null Model

Then, a null model is built for the reference effect. Keeping the other steps of the proposed method, the null model replaces the LPA process with random labels. The evaluation still compares the simulated popularity curve and real evolutionary curve of each sub-event. The improvement rate is calculated by the difference of *JSD* between the null model and the proposed model divided by the *JSD* value of the null model.

4 EXPERIMENTAL RESULTS

The experiment dataset comes from the competition of WRD Big Data, which are about the "Shanghai Tesla self-ignition incident" Weibo data, with 61,688 blog posts from 21 April 2019 to 5 May 2019. The incident is about a Tesla car suddenly smoking and self-igniting, which caused heated public debates on safety and the enterprise's responsibility. Data pre-processing process is conducted to delete the data labeled as robots, the data of re-tweets without own comment, and microblogging texts less than 10 words. In the remaining 40,119 blog posts, after replacing the deleted stop-words, emojis, special characters, HTML tags, and URLs of various hyperlinks, the TextRank algorithm is used to extract the keywords from the set of blog posts after the word segmentation, and each blog post contains 10 keywords. The unfolding model is conducted as follows.

The first step is to cluster similar posts. By using the SPC method, the original 40,119 blog posts are summarized to 4,050 posts. Each summarized post contains a number of similar documents, in which users are talking about the same content. After sorting the number of documents in descending order, the number of original posts in each summarized post approximately follows the power-law distribution (**Figure 2**). The results indicate that a large number of post documents are concentrated in a small number of clusters.

TABLE 2 | Label information of sub-events.

Sub-event $\mathbb{C}$	Standard	Frequency	Probability (%)
Event Happen $\mathbb{C}_1$	Tesla sudden self-burning	439	10.84
Corporate Respond $\mathbb{C}_2$	Corporate releases statement Corporate responds to owners Corporate responds to media	901	22.25
Client Respond $\mathbb{C}_3$	Owners elaborate on events Owners respond to corporate	386	9.53
Media Report $\mathbb{C}_4$	Media coverage Media interviews	418	10.32
Fire Control $\mathbb{C}_5$	Site information Survey results	379	9.36
Weibo Discuss $\mathbb{C}_6$	About the event About similar events	1,057	26.10
Event Processing $\mathbb{C}_7$	Event inspection Announcement of survey	223	5.51
Expert Opinion $\mathbb{C}_8$	Media opinions Personal opinions	247	6.10

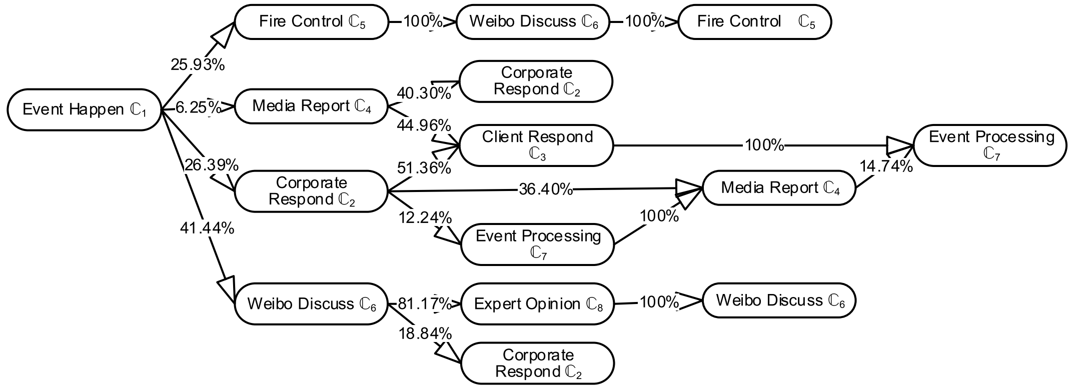

Fire Control $\mathbb{C}_5$
100%
Weibo Discuss $\mathbb{C}_6$
100%
Fire Control $\mathbb{C}_5$
25.93%
Corporate Respond $\mathbb{C}_2$
40.30%
Event Happen $\mathbb{C}_1$
6.25%
Media Report $\mathbb{C}_4$
44.96%
Client Respond $\mathbb{C}_3$
100%
Event Processing $\mathbb{C}_7$
26.39%
51.36%
14.74%
Corporate Respond $\mathbb{C}_2$
36.40%
Media Report $\mathbb{C}_4$
41.44%
12.24%
100%
Event Processing $\mathbb{C}_7$
Weibo Discuss $\mathbb{C}_6$
81.17%
Expert Opinion $\mathbb{C}_8$
100%
Weibo Discuss $\mathbb{C}_6$
18.84%
Corporate Respond $\mathbb{C}_2$

FIGURE 3 | Correlation of sub-events as a topic tree.

As is shown in **Table 1**, there are two typical posts that can be summarized according to the similarity threshold. Here, we set the similarity threshold as 0.75 in SPC. The first kind of similarity is the posts talking about the same content, such as the records 1 and 2 can be seen as one. The second is simply the same content's re-post, such as the records 3 and 4 are also summarized as one. When the similarity of the post is smaller than the threshold, the records would not be summarized. The records 5 and 6 still stand respectively for two posts. In the last two columns, experts label the summarized posts according to the keywords of the events. There are 8 labels concluded by three experts, i.e. Event Happen, Corporate Respond, Client Respond, Media Report, Fire Control, Weibo Discuss, Event Processing, and Expert Opinion, which are labeled in the first 600 summarized posts.

The second step is to extract the sub-events. The results are in the form of labeling, which can be seen in **Table 2**. It gives the standards of expert labeling and the number and prior probability of labeling after the process of the LPA method. The standards of labeling are defined by experts when the first 600 summarized posts are labeled. The frequency of each sub-event $\mathbb{C}$ is counted by expert labeling and LPA labeling. The prior probability of labeling is calculated by averaging the number of summarized posts.

The third step is to correlate the sub-events. Through the WMD method, the numbers of pairs between sub-events are used to calculate the evolution probability. The results are shown in **Figure 3** as a topic-changing tree. Based on prior probability and evolution probability, the MCMC simulation gives the probability distribution of each sub-event.

Finally, the fourth step is to verify the development of the sub-event. The regenerated sub-event curves are compared with the real popularity curves as shown in **Figure 4**. The *JSD* value equals 0.0950, 0.0841, 0.0635, 0.06804, 0.2304, 0.2135, 0.3727, and 0.1377 respectively for Event Happen $\mathbb{C}_1$, Corporate Respond $\mathbb{C}_2$, Client Respond $\mathbb{C}_3$, Media Report $\mathbb{C}_4$, Fire Control $\mathbb{C}_5$, Weibo Discuss $\mathbb{C}_6$, event processing $\mathbb{C}_7$, and expert opinions $\mathbb{C}_8$. The results are 87.03, 88, 86.87, 57.37, 75.48, 65.33, 50.52, and 80.54% higher than that of the null model (seen in **Table 3**).

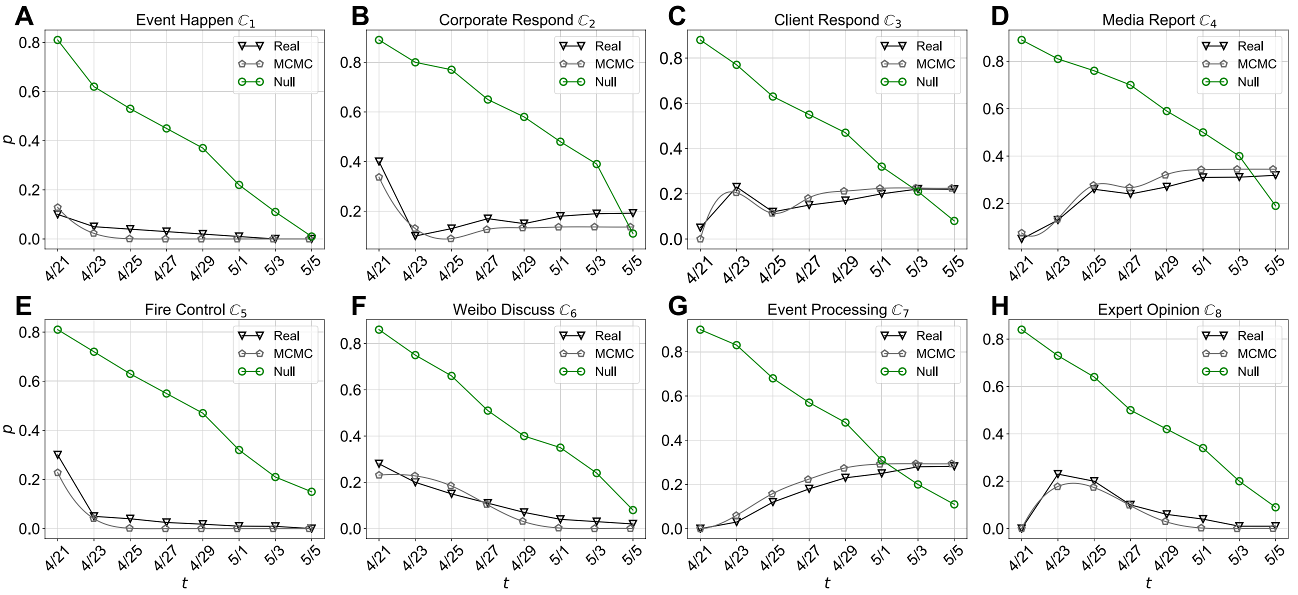


FIGURE 4 | Popularity curve of sub-event development. Three curves are the real popularity dynamic, the popularity of unfolding model regenerated by MCMC, and the reference popularity of null model. The evaluations are between the three curves by *JSD*. For example in **(A)**, the *JSD* value between real and MCMC popularity is 0.095, which shows the close trends between unfolding model and real dynamic. The *JSD* value between MCMC and null popularity is 0.7329, indicating the significant difference between the unfolding model and the random model. The rest of *JSD* values **(B–H)** can be seen in Table 3.

TABLE 3 | Model evaluation.

JSD	Unfolding model	Null model	Improvement (%)
Event Happen $\mathbb{C}_1$	0.0950	0.7329	87.03
Corporate Respond $\mathbb{C}_2$	0.0841	0.5299	88.00
Client Respond $\mathbb{C}_3$	0.0635	0.5183	86.87
Media Report $\mathbb{C}_4$	0.0680	0.5406	57.37
Fire Control $\mathbb{C}_5$	0.2304	0.8709	75.48
Weibo Discuss $\mathbb{C}_6$	0.2135	0.5095	65.33
Event Processing $\mathbb{C}_7$	0.3727	0.7533	50.52
Expert Opinion $\mathbb{C}_8$	0.1379	0.7077	80.54

5 CONCLUSION AND DISCUSSION

In this article, we use Single Pass Clustering (SPC) to summarize the massive posts. The step is to reduce the redundancy among similar posts and form summarized posts. Then, the Label Propagation Algorithm (LPA) is introduced so that the small-scale expert labels can spread to the whole datasets. Each label is a topic concerned by PR managers and represents a sub-event. The SPC and LPA processes complete the sub-event detection. Among the summarized posts between sub-events, we use Word Mover's Distance (WMD) to pair the correlated documents. Markov Chain Monte Carlo (MCMC) simulation is finally used to correlate the sub-events and predict each sub-event evolutionary. The WMD and MCMC complete the sub-event correlation. The results show that the procedure is 50.52% ~ 88% higher than the random null model in the case of "Shanghai Tesla self-ignition incident".

The reconstruction method can help to intuitively understand different sides of the events and the hotspot shift of public opinion. But there are several limitations of this article. First, external knowledge deserves further study to enhance the comprehensibility and accuracy of sub-events. Second, similarity measurements are essential for the results of classification [40], and which measurement is stable for Weibo post classification is an open question. Third, time-line correlation should be introduced into topic-level sub-event development trends [41]. Lastly, the approach of network reconstruction [42, 43, 44] can be integrated into content reconstruction.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

R-DL provided this topic and wrote the manuscript. QG, X-KZ, and J-GL guided, discussed, and modified the manuscript. All authors contributed to the manuscript and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China (NNSFC) (Grant Nos. 71901144, 71771152, 61773248, 72032003, and 72171150), the Major Program of the National Fund of Philosophy and Social Science of China (Nos. 20ZDA060 and 18ZDA088), and the Fund of University of Shanghai for Science and Technology (No. CFTD223049).

ACKNOWLEDGMENTS

We thank Chun-ping Cao, Yu Li, Kai-ge Yuan, and Yue Lu for preliminary experience in the second competition of WRD Big Data.

REFERENCES

1. Kumar D, Ramakrishnan N, Helm RF, Potts M. Algorithms for Storytelling. *IEEE Trans Knowl Data Eng* (2008) 20:736–51. doi:10.1109/tkde.2008.32
2. Wu Q, Ma S, Liu Y. Sub-Event Discovery and Retrieval During Natural Hazards on Social Media Data. *World Wide Web* (2016) 19:277–97. doi:10.1007/s11280-015-0359-8
3. Pohl D, Bouchachia A, Hellwagner H. Online Indexing and Clustering of Social media Data for Emergency Management. *Neurocomputing* (2016) 172: 168–79. doi:10.1016/j.neucom.2015.01.084
4. Rashid MT, Wang D. Covidsens: a Vision on Reliable Social Sensing for Covid-19. *Artif intelligence Rev* (2020) 1–25. doi:10.1007/s10462-020-09852-3
5. Nolasco D, Oliveira J. Mining Social Influence in Science and Vice-Versa: A Topic Correlation Approach. *Int J Inf Management* (2020) 51:102017. doi:10.1016/j.ijinfomgt.2019.10.002
6. Huang Y, Shen C, Li T. Event Summarization for Sports Games Using Twitter Streams. *World Wide Web* (2018) 21:609–27. doi:10.1007/s11280-017-0477-6
7. Meladianos P, Xypolopoulos C, Nikolentzos G, Vazirgiannis M. An Optimization Approach for Sub-event Detection and Summarization in Twitter. In: European Conference on Information Retrieval (Springer) (2018). p. 481–93. doi:10.1007/978-3-319-76941-7_36
8. Phengsuwan J, Shah T, Thekkummal NB, Wen Z, Sun R, Pullarkatt D, et al. Use of Social media Data in Disaster Management: A Survey. *Future Internet* (2021) 13:46. doi:10.3390/fi13020046
9. Wang D, Amin MT, Li S, Abdelzaher T, Kaplan L, Gu S, et al. Using Humans as Sensors: an Estimation-Theoretic Perspective. In: IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks (IEEE) (2014). p. 35–46. doi:10.1109/ipsn.2014.6846739
10. Pohl D, Bouchachia A, Hellwagner H. Automatic Sub-event Detection in Emergency Management Using Social media. In: Proceedings of the 21st international conference on world wide web (2012). p. 683–6. doi:10.1145/2187980.2188180
11. Rogstadius J, Vukovic M, Teixeira CA, Kostakos V, Karapanos E, Laredo JA. Crisistracker: Crowdsourced Social media Curation for Disaster Awareness. *IBM J Res Development* (2013) 57:4–1. doi:10.1147/jrd.2013.2260692
12. Vielberth M, Menges F, Pernul G. Human-as-a-security-sensor for Harvesting Threat Intelligence. *Cybersecurity* (2019) 2:1–15. doi:10.1186/s42400-019-0040-0
13. Avvenuti M, Cimino MG, Cresci S, Marchetti A, Tesconi M. A Framework for Detecting Unfolding Emergencies Using Humans as Sensors. *SpringerPlus* (2016) 5:43–23. doi:10.1186/s40064-016-1674-y
14. Jin Y, Liu BF, Austin LL. Examining the Role of Social Media in Effective Crisis Management: The Effects of Crisis Origin, Information Form, and Source on Publics' Crisis Responses. *Commun Res* (2014) 41:74–94. doi:10.1177/0093650211423918
15. Lachlan KA, Spence PR, Lin X. Expressions of Risk Awareness and Concern through Twitter: On the Utility of Using the Medium as an Indication of Audience Needs. *Comput Hum Behav* (2014) 35:554–9. doi:10.1016/j.chb.2014.02.029
16. Veil SR, Buehner T, Palenchar MJ. A Work-In-Process Literature Review: Incorporating Social media in Risk and Crisis Communication. *J contingencies crisis Manag* (2011) 19:110–22. doi:10.1111/j.1468-5973.2011.00639.x
17. Lachlan KA, Spence PR, Lin X, Najarian K, Del Greco M. Social media and Crisis Management: Cerc, Search Strategies, and Twitter Content. *Comput Hum Behav* (2016) 54:647–52. doi:10.1016/j.chb.2015.05.027
18. Lv S, Huang L, Zang L, Zhou W, Han J, Hu S. Yet Another Approach to Understanding News Event Evolution. *World Wide Web* (2020) 23:2449–70. doi:10.1007/s11280-020-00818-7
19. Nichols J, Mahmud J, Drews C. Summarizing Sporting Events Using Twitter. In: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (2012). p. 189–98. doi:10.1145/2166966.2166999
20. Shen C, Liu F, Weng F, Li T. A Participant-Based Approach for Event Summarization Using Twitter Streams. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2013). p. 1152–62.
21. Srijith PK, Hepple M, Bontcheva K, Preotiuc-Pietro D. Sub-story Detection in Twitter with Hierarchical Dirichlet Processes. *Inf Process Management* (2017) 53:989–1003. doi:10.1016/j.ipm.2016.10.004
22. Huang L. Optimized Event Storyline Generation Based on Mixture-Event-Aspect Model. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013). p. 726–35.
23. Jin Z, Cao J, Jiang YG, Zhang Y. News Credibility Evaluation on Microblog with a Hierarchical Propagation Model. In: 2014 IEEE International Conference on Data Mining (IEEE) (2014). p. 230–9. doi:10.1109/icdm.2014.91
24. Kumar NP, Rao MV, Krishna PR, Bapi RS. Using Sub-sequence Information with Knn for Classification of Sequential Data. In: International Conference on Distributed Computing and Internet Technology (Springer) (2005). p. 536–46. doi:10.1007/11604655_60
25. Sreenivasulu M, Sridevi M. Comparative Study of Statistical Features to Detect the Target Event during Disaster. *Big Data Min Anal* (2020) 3:121–30. doi:10.26599/bdma.2019.9020021
26. Khurdiya A, Dey L, Mahajan D, Verma I. Extraction and Compilation of Events and Sub-events from Twitter. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (IEEE), 1 (2012). p. 504–8. doi:10.1109/wi-iat.2012.192
27. Piergiovanni A, Ryoo MS. Learning Latent Super-Events to Detect Multiple Activities in Videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018). p. 5304–13. doi:10.1109/cvpr.2018.00556
28. Alam F, Joty SR, Imran M. Domain Adaptation with Adversarial Training and Graph Embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018). 1077–1087. doi:10.18653/v1/P18-1099
29. Alam F, Joty S, Imran M. Graph Based Semi-supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets. In: Proceedings of the International AAAI Conference on Web and Social Media (2018).
30. Long R, Wang H, Chen Y, Jin O, Yu Y. Towards Effective Event Detection, Tracking and Summarization on Microblog Data. In: International conference on web-age information management (Springer) (2011). p. 652–63. doi:10.1007/978-3-642-23535-1_55
31. Li Z, Ding X, Liu T. Constructing Narrative Event Evolutionary Graph for Script Event Prediction. In: International Joint Conference on Artificial Intelligence (2018). p. 4201–4207. doi:10.24963/ijcai.2018/584

32. Pichotta K, Mooney R. Learning Statistical Scripts with Lstm Recurrent Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence (2016).
33. Liu Y, Peng H, Guo J, He T, Li X, Song Y, et al. Event Detection and Evolution Based on Knowledge Base. In: Proceedings of the KBCOM 2018, WSDM (2018). p. 1–7. doi:10.475/123_4
34. Lv S, Qian W, Huang L, Han J, Hu S. Sam-net: Integrating Event-Level and Chain-Level Attentions to Predict what Happens Next. In: Proceedings of the AAAI Conference on Artificial Intelligence, 33 (2019). p. 6802–9. doi:10.1609/aaai.v33i01.33016802
35. Yuan C, Yuan C, Bai Y, Li Z. Logic Enhanced Commonsense Inference with Chain Transformer. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management (2020). p. 1763–72. doi:10.1145/3340531.3411895
36. Mao Q, Li X, Peng H, Li J, He D, Guo S, et al. Event Prediction Based on Evolutionary Event Ontology Knowledge. *Future Generation Computer Syst* (2021) 115:76–89. doi:10.1016/j.future.2020.07.041
37. Reuter C, Stieglitz S, Imran M. Social media in Conflicts and Crises. *Behav Inf Technology* (2020) 39:241–51. doi:10.1080/0144929x.2019.1629025
38. Chen X, Zhou X, Sellis T, Li X. Social Event Detection with Retweeting Behavior Correlation. *Expert Syst Appl* (2018) 114:516–23. doi:10.1016/j.eswa.2018.08.022
39. Li RD, Liu JG, Guo Q, Zhang YC. Social Signature Identification of Dynamical Social Networks. *Physica A: Stat Mech its Appl* (2018) 508:213–22. doi:10.1016/j.physa.2018.05.094
40. Liu JG, Hou L, Pan X, Guo Q, Zhou T. Stability of Similarity Measurements for Bipartite Networks. *Sci Rep* (2016) 6:18653–10. doi:10.1038/srep18653
41. Nolasco D, Oliveira J. Subevents Detection through Topic Modeling in Social media Posts. *Future Generation Computer Syst* (2019) 93:290–303. doi:10.1016/j.future.2018.09.008
42. Hu ZL, Han X, Ma L. Network Structure Transmission with Limited Data via Compressed Sensing. *IEEE Trans Netw Sci Eng* (2020) 7:3200–11. doi:10.1109/tnse.2020.3018134
43. Hu ZL, Wang L, Tang CB. Locating the Source Node of Diffusion Process in Cyber-Physical Networks via Minimum Observers. *Chaos* (2019) 29:063117. doi:10.1063/1.5092772
44. Hu ZL, Shen Z, Han J, Peng H, Lu J, Jia R, et al. Localization of Diffusion Sources in Complex Networks: A Maximum-Largest Method. In: Physica A: Statistical Mechanics and its Applications (2019). doi:10.1016/j.physa.2019.121262

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Physics

ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/fphy.2022.955499

Optimal Path Planning With Minimum Inspection Teams and Balanced Working Hours For Power Line Inspection

*Zhao-Long Hu *, Yuan-Zhang Deng, Hao Peng, Jian-Min Han, Xiang-Bin Zhu, Dan-Dan Zhao, Hui Wang and Jun Zhang*

College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua, China

Power line inspection plays a significant role in the normal operation of power systems. Although there is much research on power line inspection, the question of how to balance the working hours of each worker and minimize the total working hours, which is related to social fairness and maximization of social benefits, is still challenging. Experience-based assignment methods tend to lead to extremely uneven working hours among the working/inspection teams. Therefore, it is of great significance to establish a theoretical framework that minimizes the number of working teams and the total working hours as well as balances the working hours of inspection teams. Based on two real power lines in Jinhua city, we first provide the theoretical range of the minimum number of inspection teams and also present a fast method to obtain the optimal solution. Second, we propose a transfer-swap algorithm to balance working hours. Combined with an intelligent optimization algorithm, we put forward a theoretical framework to balance the working hours and minimize the total working hours. The results based on the two real power lines verify the effectiveness of the proposed framework. Compared with the algorithm without swap, the total working hours obtained by the transfer-swap algorithm are shorter. In addition, there is an interesting finding: for our transfer-swap algorithm, the trivial greedy algorithm has almost the same optimization results as the simulated annealing algorithm, but the greedy algorithm has an extremely short running time.

Keywords: path planning, optimization algorithms, complex networks, balancing working hours, power line inspection

OPEN ACCESS

Edited by:
Claudio J. Tessone,
University of Zurich, Switzerland

Reviewed by:
Kai Yang,
Yangzhou University, China
Rende LI,
the University of Shanghai for Science and Technology, China

****Correspondence:***
Zhao-Long Hu
huzhaolong@zjnu.edu.cn

Specialty section:
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

Received: *28 May 2022*
Accepted: *22 June 2022*
Published: *22 July 2022*

Citation:
Hu Z-L, Deng Y-Z, Peng H, Han J-M, Zhu X-B, Zhao D-D, Wang H and Zhang J (2022) Optimal Path Planning With Minimum Inspection Teams and Balanced Working Hours For Power Line Inspection. Front. Phys. 10:955499. doi: 10.3389/fphy.2022.955499

1 INTRODUCTION

With the development of the society and smart grid, the demand for electricity is increasing, and the range of power lines is also getting wider [1]. The safe operation and maintenance of power lines is related to our high quality of life, but subtle disturbances in the power system may cause great harm [2]. For example, on December 23, 2015, Ukraine reported a service outage [3, 4]. As a part of power systems, the power line inspection is a very important step to ensure the normal operation and maintenance of the power system. However, power lines are always exposed to the outside and are vulnerable to earthquakes, floods, storms, building collapses, etc., so it is necessary to regularly inspect power lines [5, 6]. The power line inspection includes tower inspection and wire inspection. The traditional method of overhead power line inspection is to manually walk along the line or by vehicles and use telescopes and infrared thermal imagers to conduct an inspection. The main problems of traditional inspection are as follows: on the one hand, the distance of the inspected power lines is long and the workload is heavy and the efficiency is very slow. In the event of natural disasters such as earthquakes and landslides, the inspection task will not be carried out. On the other hand, inspection in mountainous areas is of high risk, threatening the life safety of workers. Therefore, how to efficiently complete power line inspection is a very important issue. With the development of technology, unmanned aerial vehicles (UAVs) can be used instead of a human in some cases. Compared with traditional inspection, UAVs have the advantages of strong adaptability, high accuracy, and high work efficiency in power line inspection [7–10]. Workers can control UAVs to take photos and then send them to the terminal. From the terminal, we can find out where the problem is and then use robots to repair it. At present, the robot can only do some simple repair work. If it encounters complex problems, it still needs to be carried out manually [11–14]. Usually, UAVs face limited battery life and controllable distance. As we know, the target power lines are usually far away from the office (work unit). Therefore, the inspection task is implemented with the following two steps. The first step is to drive from the office to the drop-off points of the towers. The second step is to take photos by UAVs and repair them with robots or workers once a problem is detected. At present, the work office mainly relies on experience to assign several inspection teams to inspect the corresponding power lines, resulting in extremely uneven working hours among inspection teams. What is worse is some inspection teams work overtime for a long time. Thus, an efficient solution for balancing the working hours of each inspection team can solve this social fairness problem to a certain extent. With regard to the second step of the inspection work, the inspection time for each tower and the corresponding power line can be considered a constant. This assumption is reasonable because we usually do not know in advance whether these towers need to be repaired. Thus, in order to balance the working hours and minimize the total working hours, we only need to provide optimal power line inspection path planning before performing the inspection task.

Given a target power line, the optimal path planning problem can be transformed into a traveling salesman problem (TSP) or vehicle routing problem (VRP). That is, the inspection team starts from the office and finally returns to the original starting location. Here, all the target towers and the corresponding power lines should be inspected and can only be inspected once. Therefore, the problem belongs to the NP-hard problem encountered in combinatorial optimization [15, 16]. The TSP or VRP is a very old and classic problem in graph theory, and many research methods for optimal path planning were proposed [17], such as integer programming [18], dynamic programming [19], and branch and bound algorithm [20, 21]. However, the high computational complexity of those exact algorithms prevents them from being applied to large-scale networks. To improve computational efficiency, agent-based or multi-agent-based heuristic intelligent optimization algorithms were successively proposed. For example, genetic algorithms [22], ant colony algorithms [23], simulated annealing algorithms [24, 25], particle swarm algorithms [26], and some derivative algorithms, or hybrid algorithms [17, 27, 28]. Recently, some optimization methods by machine learning were presented [29], such as graph neural network [30] and reinforcement learning [31].

If there are too many power lines to be inspected, it is not practical to assign one inspection team to complete the inspection task. The problem can be transformed into classical multiple traveling salesman problems (MTSP), which is also NP-hard [32, 33]. The constraint conditions are 1) all the inspection teams should start from the same location (office) and return to the office. 2) Each inspection team must inspect at least one tower. 3) Each tower and the corresponding power line can only be inspected once. Despite this problem seeming difficult, it can still be solved by the abovementioned method [17, 34, 35]. For example, based on k-means clustering, the optimal path planning was carried out in each cluster [36–38]. However, those works did not take into account the balanced workload of each cluster. As workload can reflect social fairness, and the MTSP associated with balancing workload has attracted increasing attention [39–41]. For example, Alves et al. minimized the distance and balanced the routes for the MTSP by genetic algorithms [39]. Xu et al. proposed a two-phase heuristic algorithm to balance the number of destinations of travel agents [40]. Compared with balancing the routes or the number of traveling destinations, it is fairer to balance the working hours. Lee et al. studied the balance of the traveling time for MTSP, but the traveling time among each pair of destinations is linear with the distance [42]. Recently, Vandermeulen et al. investigated the balanced working hours for MTSP by translating the task assignment problem into the minimum Hamiltonian partition problem; however, the traveling/cost time was obtained by simulation [43]. Hu et al. proposed a transfer method to balance the working time with the minimum number of inspection teams with two real power lines [6]. But, there is no need to consider the walking time because the task of taking photos can be implemented by the UAVs.

In this article, considering both driving time and inspection time, we give the theoretical solution for the minimum number of inspection teams. In addition, we propose a transfer-swap algorithm to balance the working hours among inspection teams and minimize to total working hours. Combined with the minimum number of inspection teams and intelligent optimization algorithms, a framework for optimal path planning is presented. Concretely, based on the latitude and longitude of the power grid (line-5876 and line-5803) in Jinhua City and the latitude and longitude of the drop-off points of towers of the two power lines, we obtain both the driving time from the office to all the towers and the driving time among each pair of towers through the web crawler [6]. Compared with the real driving time from the office to each tower, it is found that the driving time obtained by the crawler is not much different from that of the real one. Using the crawled driving time, we can construct a fully connected network of driving time between the office and all towers. Simulations verify the provided theoretical solution for the minimum number of inspection teams, and results from four optimization algorithms prove that the proposed transfer-swap algorithm can well-balance the working hours.

The article is organized as follows. In **Section 2**, first, we describe our model. Second, we present the theoretical analysis

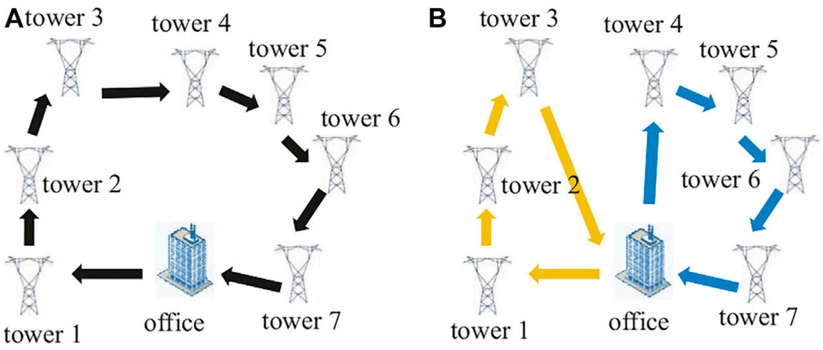


FIGURE 1 | (color online) A simple example about path planning with one office and seven towers. **(A)** Optimal path for one inspection team to complete a given inspection task. The total working hours (spending time) is the driving time of the path plus the inspection time by UAVs, robots, or workers, that is, $t_d + 7t_{ins}$. **(B)** Optimal path that balances the working hours between the two inspection teams. The working hours is t^1_{wh} and t^2_{wh} for the two inspection teams. The difference of working hours between the tow inspection teams is $t_{diff} = | t^1_{wh} - t^2_{wh} |$, where $| \cdot |$ is the absolute value of ·.

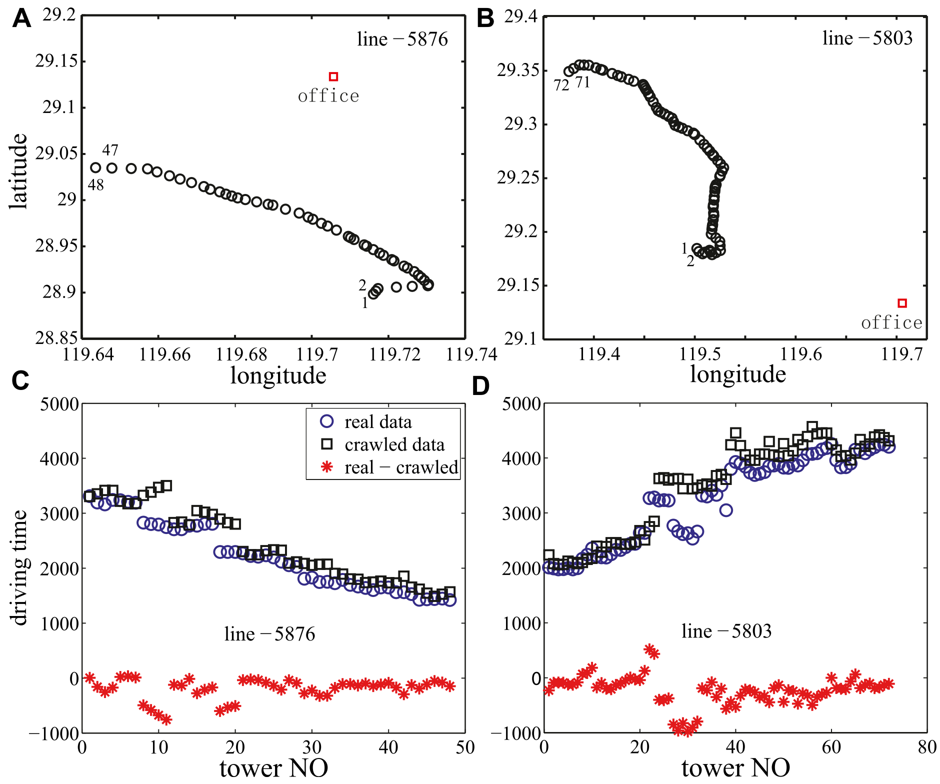


FIGURE 2 | (color online) **(A,B)** Longitudes and latitudes of towers for line-5876 and line-5803. Only the first two towers and the last two towers, as well as the location of the office, are marked. **(C,D)** Real driving time and the crawled driving time from the office to each tower for line-5876 and line-5803, and the red star curve stands for the difference. There are 48 towers in line-5876 and 72 towers in line-5803.

for the minimum number of inspection teams. Third, we provide the transfer-swap algorithm to balance the working hours among inspection teams, which is the key to the general framework for optimal path planning. In **Section 3**, we analyze two real power lines and verify the framework of optimal path planning with the minimum inspection teams and balancing the working hours. Finally, we conclude and discuss this article in **Section 4**.

2 MODEL AND METHODS

2.1 Optimal Inspection Path Planning Model

Because the location of the office is fixed, we can construct a fully connected network once the target towers are identified. The fully connected network can be described by G = (V, E, T). Here, V is the node-set $V = \{v_0, v_1, \ldots, v_N\}$. v_0 represents the office and the

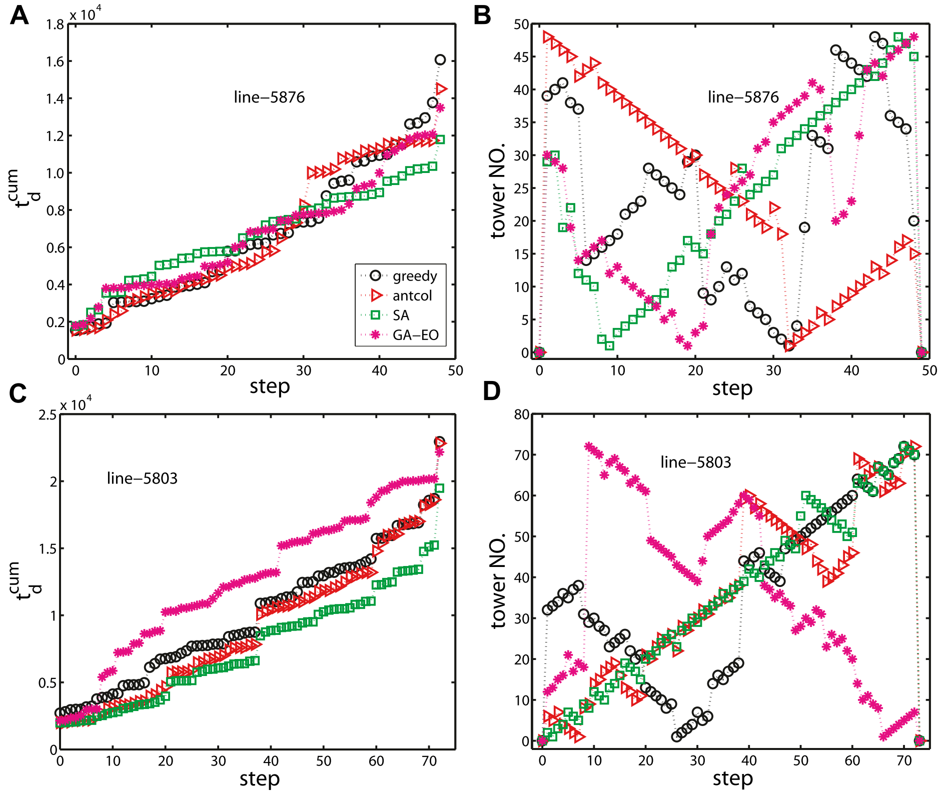


FIGURE 3 | (color online) Under the four algorithms (greedy, antcol, SA, and GA-EO); **(A,C)** the cumulative driving time for a single inspection team to complete the task; **(B,D)** the optimal inspection path. 0 in tower NO. represents the office, and tower NO means the tower number is to be inspected at that step. **(A,B)** line-5876. **(C,D)** Line-5803. The perturbation parameter $p = 0.05$ for the greedy algorithm.

others are the number of the N target towers, see **Figure 1**. The edge set $\mathrm{E} = \{e_{v_i v_j} = (v_i, v_j) \mid v_i, v_j \in \mathrm{V}, i \neq j\}$ stands for the edges among nodes. The term $\mathrm{T} \in \mathbb{R}^{(N+1)\times(N+1)}$ is the driving time matrix, and the entry $t_{v_i v_j}$ in T is the driving time from v_i to v_j. The other related variables are defined as follows:

t_d, the driving time for one single inspection team to visit all the target towers and return to the office.

t_ins, the inspection time for inspecting each tower and the corresponding power line, which can be completed by UAVs, robots, or workers. The term can be set to a constant, say 15 min. Thus, the total working hours (i.e., spending time) for one single inspection team to complete the task is $t_\mathrm{d} + t_\mathrm{ins}N$.

t_max, the maximum working hours on workday for every inspection team. Generally, let us define $t_\mathrm{max} = 8$ h or 28,800 s.

t_wh^r, the working hours of rth inspection team, which includes both driving time and inspection time over the assigned power lines. Here, $r = 1, 2, \ldots, k$ and k are the total number of inspection teams. In general, the working hours for each inspection team should satisfy $t_\mathrm{wh}^r \le t_\mathrm{max}$.

t_diff, the maximum difference of the working hours among inspection teams, quantified by $\max(t_\mathrm{wh}^{r_1}) - \min(t_\mathrm{wh}^{r_2})$ with r_1, $r_2 = 1, 2, \ldots, k$. Here, $\max(\cdot)$ and $\min(\cdot)$ stand for the maximum and minimum of $\cdot$, respectively. The smaller the indicator t_diff is, the fairer it is.

The objective is to balance the working hours of each inspection team and minimize the total working hours, see **Figure 1**. The system model can be written by

$$\min\ Z = \sum_{i=0}^{N} \sum_{j=0}^{N} t_{v_i v_j} x_{v_i v_j}. \tag{1}$$

Subject to

$$\sum_{j=1}^{N} x_{v_0 v_j} = k, \tag{2}$$

$$\sum_{i=1}^{N} x_{v_i v_0} = k, \tag{3}$$

$$\sum_{i=0, i\neq j}^{N} x_{v_i v_j} = 1, \forall j = 1, 2, \ldots, N, \tag{4}$$

$$\sum_{j=0, i\neq j}^{N} x_{v_i v_j} = 1, \forall i = 1, 2, \ldots, N, \tag{5}$$

$$u_{v_i} - u_{v_j} + p x_{v_i v_j} \le p - 1, \forall i,\ \ j = 1, 2, \ldots, N, i \neq j, \tag{6}$$

$$1 \le u_{v_i} \le p, \forall i = 1, 2, \ldots, N, \tag{7}$$

$$t_\mathrm{wh}^r \le t_\mathrm{max}, \forall r = 1, 2, \ldots, k, \tag{8}$$

$$t_\mathrm{diff} \le \epsilon, \tag{9}$$

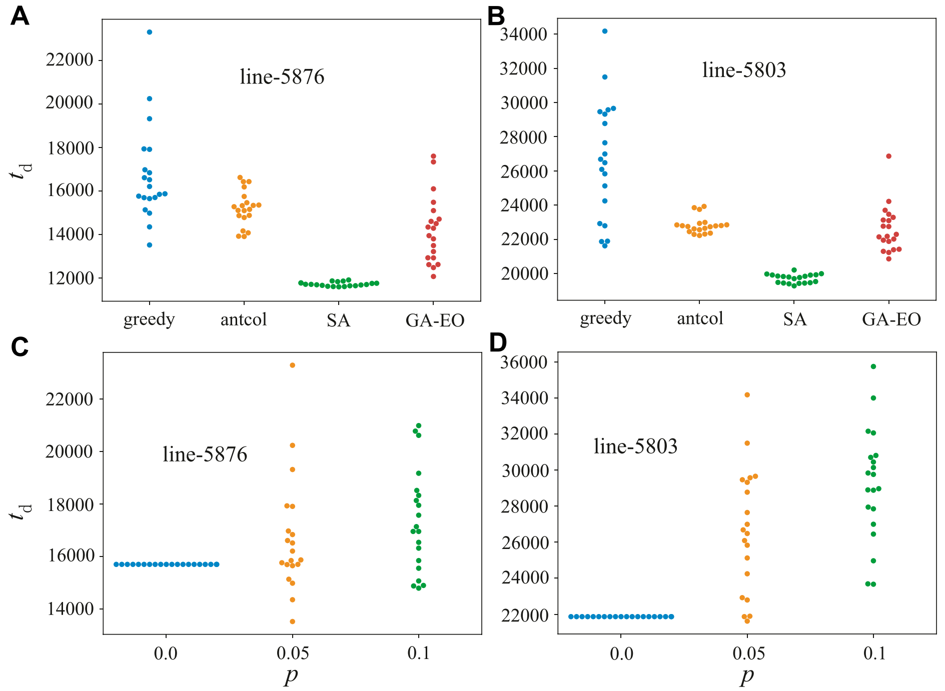


FIGURE 4 | (color online) **(A,B)** Under the four algorithms (greedy, antcol, SA, and GA-EO), the optimal driving time for a single inspection team to perform the inspection task with 20 times of independent running. **(C,D)** For different perturbation parameters *p* of the greedy algorithm, the optimal driving time for a single inspection team to perform the inspection task with 20 times running independently. **(A,C)** Line-5876. **(B,D)** Line-5803.

where

$$x_{v_i v_j} = \begin{cases} 1, & \text{if the tower } v_i \text{ precedes tower } v_j \text{ on a travel.} \\ 0, & \text{others.} \end{cases}, \quad (10)$$

Here, v_i, $v_j \in V$ and u_{v_i} are the visiting rank of tower v_i in order, $\forall i = 1, 2, \ldots, N$, and $2 \leq p \leq N + 1 - k$ represents the maximum number of towers that can be inspected by any inspection team r. In the interest of fairness, we introduce one threshold ϵ. If $t_{\text{diff}} > \epsilon$, the working hours among inspection teams are not balanced. For example, let us set $\epsilon = 1/4$ h, which means the maximum difference in working hours among all inspection teams should not be larger than 15 min.

Constraints 2) and 3) ensure all the k inspection teams starting from the office and returning to the office. Constraint sets 4) and 5) are the assignment constraints to make sure that each tower should be preceded by and precedes exactly one another tower. Constraint sets 6) and 7) are the Miller–Tucker–Zemlin sub-tour elimination constraints [18]. Constraints 8) and 9) are weak and are also our optimization objectives.

In this article, we introduce three artificial intelligence algorithms to optimize the inspection path for a given k, $t_{\max,}$ and ϵ. The three algorithms are the ant colony algorithm (antcol), simulated annealing algorithm (SA), and one hybrid algorithm made up of the genetic algorithm and extremal optimization (GA-EO). The genetic algorithm (GA) has poor local search ability, large computation, poor adaptability to large search space, and easy convergence to a local minimum. The GA-EO algorithm is adopted by combining the EO algorithm with the traditional genetic algorithm [44].

In order to compare with the abovementioned three optimal algorithms, we also present a greedy algorithm. The greedy algorithm is described as follows: For each step, with probability $1 - p$, the tower with the shortest driving time is selected as the next inspection target. Otherwise, one unselected tower is randomly selected as the next inspection target with the probability p. If $p = 0$, it is equivalent to the pure greedy algorithm, so p can be seen as a perturbation parameter. Concretely, it is assumed that the currently visited tower is v_i and the tower set containing all the towers that has been visited is defined as V_c. The next tower v_j that will be selected to visit with a probability $1 - p$ should satisfy the condition $\min_{v_j \epsilon V \backslash V_c} t_{v_i v_j}$. Here, $V \setminus V_c$ is the tower set in V but not in V_c.

2.2 Theoretical Analysis of Minimum Number of Inspection Teams

The minimum number of inspection teams is denoted as the capacitated vehicle routing problem (CVRP) [45]. Several general algorithms were proposed for a minimum number of vehicles, such as greedy algorithm [46], and integer programming [47]. Here, we show a theoretical solution to the minimum number of inspection teams. For one single inspection team to inspect the power line, the working hours are $t_d + Nt_{\text{ins}}$. Here, t_d is the total driving time on the path and t_{ins} is the inspection time to inspect

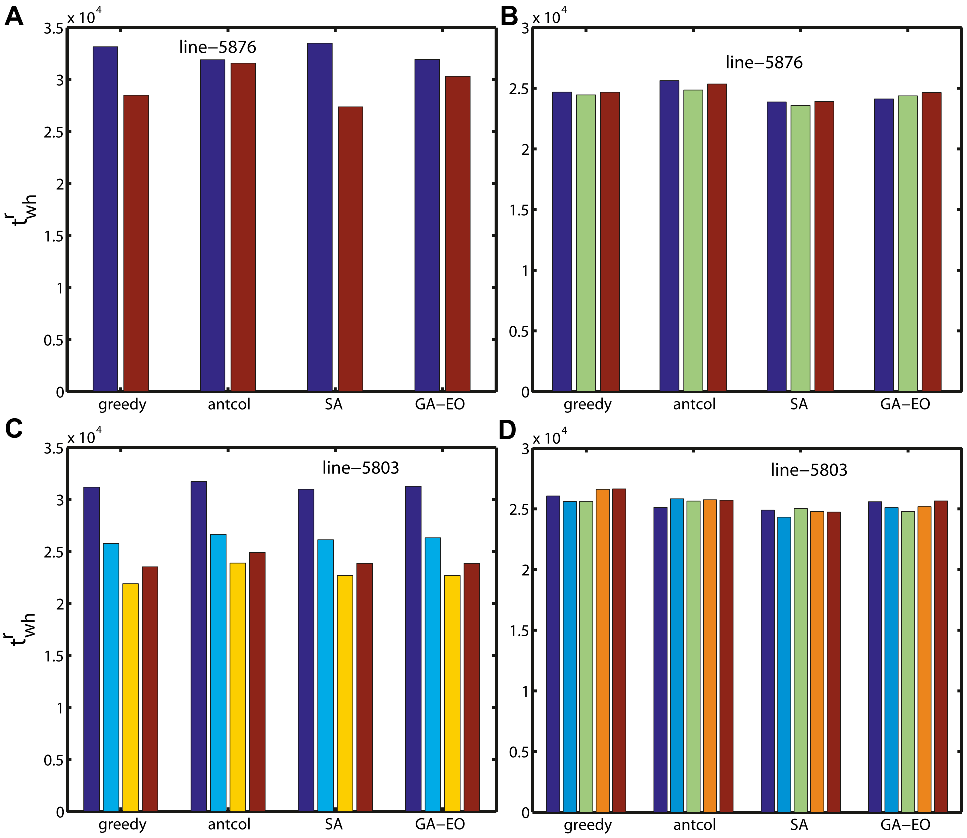


FIGURE 5 | (color online) Working hours t_{wh} for different numbers of inspection teams by using our framework without swap. **(A,B)** Line-5876 with $k = 2$ and $k = 3$ and **(C,D)** line-5803 with $k = 4$, $k = 5$. The perturbation parameter p is set to 0.05 for the greedy algorithm. Because the working hours are not well-balanced for line-5803 when $k = 4$, so the result of $k = 3$ for line-5803 is not shown in this figure.

one tower and the power line between the tower and the next tower, which can be set to a constant, such as 15 min. If $t_{\text{d}} + Nt_{\text{ins}} > t_{\max}$, more than one inspection team is required. The required number of inspection teams k satisfies

$$k \geq (t_{\text{d}} + Nt_{\text{ins}})/t_{\max}. \tag{11}$$

The equal sign holds if k is equal to 1. If $(t_{\text{d}} + Nt_{\text{ins}}) > t_{\max}$, then $k > 1$.

Because $t_{v_0 v_i}$ is the driving time from the office to tower v_i, then we have

$$\begin{cases} t_{\text{d}} + Nt_{\text{ins}} + (k-1)\max\left(t_{v_0 v_i}\right) \geq k\left[t_{\max} - \max\left(t_{v_0 v_i}\right)\right], v_i \in \text{V}\backslash \text{v}_0 \\ t_{\text{d}} + Nt_{\text{ins}} + (k-1)\min\left(t_{v_0 v_i}\right) \leq k\left[t_{\max} - \min\left(t_{v_0 v_i}\right)\right], v_i \in \text{V}\backslash \text{v}_0. \end{cases}. \tag{12}$$

From the previousequation, the minimum number of inspection teams satisfies

$$\frac{t_{\text{d}} + Nt_{\text{ins}} - \min\left(t_{v_0 v_i}\right)}{t_{\min} - \min\left(t_{v_0 v_i}\right)} \leq k \leq \frac{t_{\text{d}} + Nt_{\text{ins}} - \max\left(t_{v_0 v_i}\right)}{t_{\max} - \min\left(t_{v_0 v_i}\right)}. \tag{13}$$

To find the exact minimum number of inspection teams as quickly as possible, here, we present one alternative approach to estimate the value of k by using the average driving time from the office to all the towers. Specifically, the round-trip time of k inspection teams can be approximately computed by the average round-trip time from the office to all towers, which can be expressed as

$$t_{\text{ave}} = \frac{\sum_{v_i \in V}\left(t_{v_0 v_i} + t_{v_i v_0}\right)}{N}. \tag{14}$$

Thus, the total working hours for all teams is $t_{\text{d}} + N \times t_{\text{ins}} + (k - 1) \times t_{\text{ave}}$. Because there is one round-trip time in t_{d}, so we use $(k - 1) \times t_{\text{ave}}$ instead of $k \times t_{\text{ave}}$. Therefore, the minimum number of inspection teams should satisfy

$$\frac{t_{\text{d}} + t_{\text{ins}}N - t_{\text{ave}}}{t_{\max} - t_{\text{ave}}} \leq k. \tag{15}$$

Here, k is the smallest integer and is not less than $\frac{t_{\text{d}} + t_{\text{ins}}N - t_{\text{ave}}}{t_{\max} - t_{\text{ave}}}$. In general, because of the round-trip time, so the total working hours of completing the task for multiple inspection teams is larger than that for one single inspection team. As the round-trip time is obtained by the estimated value t_{ave}, so we relax the conditions to compute the minimum k, which leads to

$$k \in \left\{\left\lceil \frac{t_{\text{d}} + t_{\text{ins}}N - t_{\text{ave}}}{t_{\max} - t_{\text{ave}}} \right\rceil - 1, \left\lceil \frac{t_{\text{d}} + t_{\text{ins}}N - t_{\text{ave}}}{t_{\max} - t_{\text{ave}}} \right\rceil, \left\lceil \frac{t_{\text{d}} + t_{\text{ins}}N - t_{\text{ave}}}{t_{\max} - t_{\text{ave}}} \right\rceil + 1\right\},, \tag{16}$$

where $\lceil\cdot\rceil$ represents the smallest integer not less than $\cdot$.

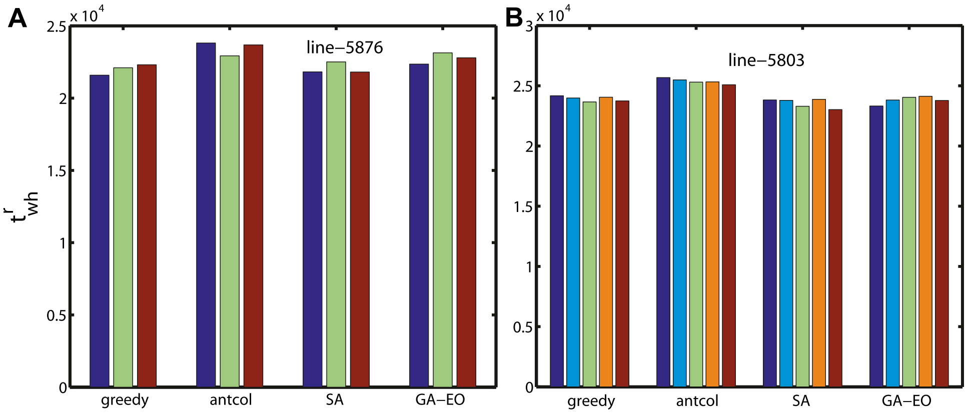


FIGURE 6 | (color online) Working hours t_{wh} for the minimum number of inspection teams with our framework and transfer-swap algorithm. **(A)** $k = 3$ for line-5876 and **(B)** $k = 5$ for line-5803. $p = 0.05$ for the greedy algorithm.

In conclusion, the minimum k can be obtained by **Eqn. 13**; however, we can use **Eqs. 11**, **16** to estimate it. Without loss of generality, in this article, we assume that the maximum working hours $t_{\max} = 8$ h of each inspection team on a workday and the $t_{\mathrm{ins}} = 15$ min.

2.3 One Framework For Optimal Inspection Path With Balancing Working Hours By the Transfer-Swap Algorithm

Based on the minimum number of inspection teams k in the last section, we propose a new algorithm to balance the working hours and minimum the total working hours by the transfer-swap algorithm and provide the framework for optimal inspection path with several intelligent optimization algorithms.

To be a concert, we first randomly select k towers as center nodes and obtain the k set by K-means based on the driving time matrix T. For example, let us set node v_i to be one center node. For non-central node $v_j \in$ V, add v_j to the set of v_i if the driving time from v_j to v_i is minimum among all the center nodes.

Second, based on the optimization algorithms, we can get the optimal path and compute the working hours for each set t^r_{wh} with $r = 1, 2, \ldots, k$ and add them to obtain the total working hours $\sum_{r=1}^{k} t^r_{\mathrm{wh}}$.

Third, if $\max(t^r_{\mathrm{wh}}) - \min(t^r_{\mathrm{wh}}) \le \epsilon$ and $t^r_{\mathrm{wh}} \le t_{\max}, \forall r \in \{1, 2, \ldots, k\}$ and the total working hours $\sum_{r=1}^{k} t^r_{\mathrm{wh}}$ is smaller than that of the previous value, then we get the final optimization result.

Fourth, if the condition in the third step is not satisfied, we transfer one tower in the tower set with $\max(t^r_{\mathrm{wh}})$ to the tower set with $\min(t^r_{\mathrm{wh}})$ by random. Implement the third step with given iterations. If the condition in the third step is still not satisfied, we select one tower in each tower set by random and swap them. Implement the last two steps with given iterations. The detail of this algorithm is shown in Algorithm 1.

Algorithm 1. Optimal inspection path planning with the transfer-swap algorithm.

```
Require: V, T, $t_{\mathrm{d}}$, $k$, ite = 0, $\mathrm{ite}_{\max} = 1000$, $t_{\max} = 28800\mathrm{s}$, $\epsilon = 900\mathrm{s}$,
    $t_{\mathrm{swap}} = 0$, $t^{\max}_{\mathrm{swap}} = 100$, $\mathrm{tower}_{\mathrm{NO}} = \{\mathrm{set}_1 = \emptyset, \mathrm{set}_2 = \emptyset, \cdots, \mathrm{set}_k = \emptyset\}$,
    $t^0_{\mathrm{total}} = \sum_{r=1}^{k} t^r_{\mathrm{wh}}$=inf, $\mathrm{std}(t^0_{\mathrm{wh}})$ =inf.
Ensure: The $\mathrm{tower}_{\mathrm{NO}}$ and $t^r_{\mathrm{wh}}$.
    ♯♯♯ Initialize $\mathrm{tower}_{\mathrm{NO}}$ by using K-Means algorithm ♯♯♯
    Select $k$ towers from V \ $v_0$ by random, and add them to each set in
    $\mathrm{tower}_{\mathrm{NO}}$, respectively. Namely, $\mathrm{set}_r = \mathrm{tower}_r$ with $r = 1, 2, \cdots, k$.
    for $v_i$ in V \ $\{v_0, \mathrm{tower}_{\mathrm{NO}}\}$ do
        Compute $t_{v_i\mathrm{tower}_r}$ for $r \in \{1, 2, \cdots, k\}$
        if $t_{v_i\mathrm{tower}_{r1}} = \min(t_{v_i\mathrm{tower}_r})$ with $r1, r \in \{1, 2, \cdots, k\}$ then
            Add $v_i$ to $\mathrm{set}_{r1}$
        end if
    end for
    ♯♯♯ Balance the working hours with transfer and swap ♯♯♯
    while $t_{\mathrm{swap}} \le t^{\max}_{\mathrm{swap}}$ do
        while ite $\le \mathrm{ite}_{\max}$ do
            Obtain the optimal paths of $\mathrm{set}_r$ for $r \in \{1, 2, \cdots, k\}$ by greedy,
    antcol, SA, and GA-EO.
            Compute the working hours $t^r_{\mathrm{wh}}$ for $r \in \{1, 2, \cdots, k\}$. Compute the
    total working hours to complete this task by $\sum_{r=1}^{k} t^r_{\mathrm{wh}}$.
            if $t^r_{\mathrm{wh}} \le t_{\max}, \forall r \in \{1, 2, \cdots, k\}$, and $\max(t^r_{\mathrm{wh}}) - \min(t^r_{\mathrm{wh}}) \le \epsilon$ and
    $\sum_{r=1}^{k} t^r_{\mathrm{wh}} \le t^0_{\mathrm{total}}$ then
                The new $\mathrm{tower}_{\mathrm{NO}}$ is the optimal path, and terminate the
    program.
            else if $\sum_{r=1}^{k} t^r_{\mathrm{wh}} \le t^0_{\mathrm{total}}$ and the standard deviation $\mathrm{std}(t^r_{\mathrm{wh}}) \le$
    $\mathrm{std}(t^0_{\mathrm{wh}})$ then
                The new $\mathrm{tower}_{\mathrm{NO}}$ replaces the old one, set $t^0_{\mathrm{total}} = \sum_{r=1}^{k} t^r_{\mathrm{wh}}$,
    $\mathrm{std}(t^0_{\mathrm{wh}}) = \mathrm{std}(t^r_{\mathrm{wh}})$
            end if
            Remove one tower in the tower set with $\max(t^r_{\mathrm{wh}})$ to the tower set
    with $\min(t^r_{\mathrm{wh}})$ for $r \in \{1, 2, \cdots, k\}$
            ite = ite + 1
        end while
        Select one tower in each tower set in $\mathrm{tower}_{\mathrm{NO}}$ by random, and swap
    them
        $t_{\mathrm{swap}} = t_{\mathrm{swap}} + 1$
    end while
```

3 RESULTS

3.1 The Analysis of Two Real Power Lines

In this section, we analyze the driving time of two real power lines, line-5876 and line-5803, in Jinhua City, Zhejiang Province.

The real data contain the following details [4]:

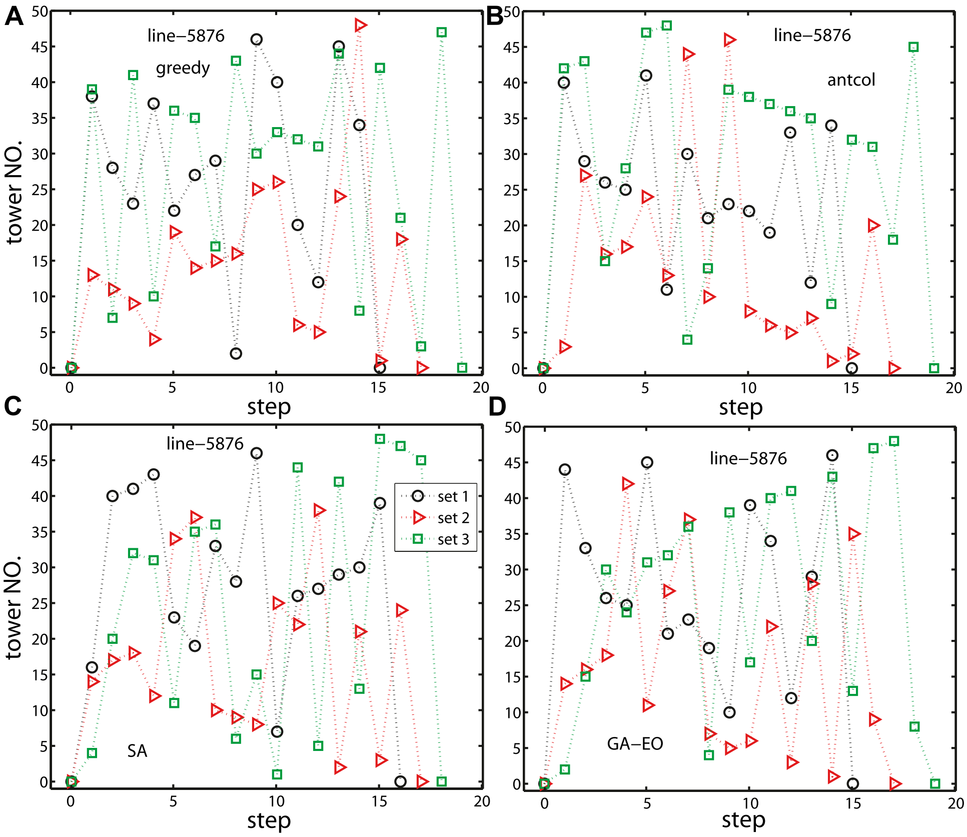


FIGURE 7 | (color online) Optimal inspection paths with our framework with **(A)** greedy, **(B)** antcol, **(C)** SA and **(D)** GA-EO. $k = 3$ for line-5876. $p = 0.05$ for the greedy algorithm. Here, the coordinate (0, 0) is the starting point, namely, the office.

i) Longitudes and latitudes of all towers for the two power lines and the office. As shown in **Figures 2A,B**.
ii) Longitudes and latitudes of drop-off points of each tower.
iii) The driving times $t_{v_0 v_i}$ from the office to drop off points of each tower.

For our question, the optimal path should start from the office, after inspecting all the target towers and power lines and finally returns to the office. Although data contain the driving time from the office to the drop-off point of each tower, it does not contain the driving time $t_{v_i v_j}$ between the drop-off point of each pair of towers. The incomplete data prevent us from using the framework immediately. Here, we provide an alternative solution. By using the API interface of Baidu Map, we can get the driving times between the drop-off of each pair of towers with a web crawler. In order to test the validity of the data obtained by the web crawler, we also crawl the driving time of the office to the drop-off point of each tower, as shown in **Figures 2C,D**. We find that the crawled data of the two power lines are not much different from the real data. The crawled data are slightly larger, but the difference is generally located near 0, as shown by the red star curve in **Figures 2C,D**. Therefore, we can optimize the path by our proposed framework with the crawled data.

3.2 Optimize Inspection Path With One Single Inspection Team

In this section, we test our framework with two real power lines. The configuration of our computer is Intel Core (TM) i7-7700 CPU, 16 GB RAM, and 3.6 GHz processing speed. First, we study the inspection path planning for one single inspection team by antcol, SA, GA-EO, and the greedy algorithm. Because the inspection time of each tower and its corresponding power line, t_{ins} is considered to be 900s, and it is only necessary to optimize the driving time when assigning a single inspection team. As can be seen from **Figures 3A,C**, when all targets are inspected, the SA algorithm takes the shortest driving time, followed by the GA-EO algorithm. The greedy algorithm in line 5876 has the longest driving time, while the total driving time of antcol and the greedy algorithm in line 5803 is nearly the same. **Figures 3B,D** shows the inspection sequence of the two power lines under the four algorithms.

The result in **Figure 3** is from a single simulation. In order to reduce the randomness of the four algorithms, we analyze the results from running independently on 20 times, see **Figures 4A,B**. It can be seen that the optimal driving time calculated by the SA algorithm is the most stable, and the other three algorithms have a large fluctuation, among which the

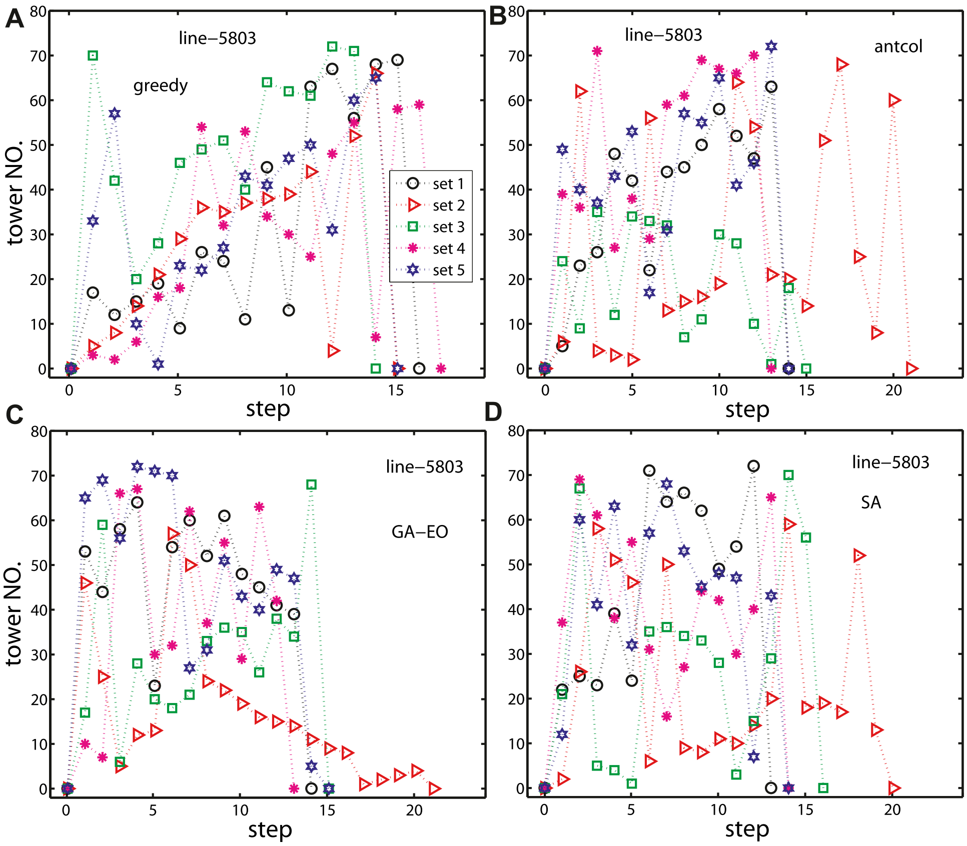


FIGURE 8 | (color online) Optimal inspection paths with our framework with **(A)** greedy, **(B)** antcol, **(C)** SA and **(D)** GA-EO. $k = 5$ for line-5803. $p = 0.05$ for the greedy algorithm. Here, the coordinate (0, 0) is the starting point, namely, the office.

fluctuation of the greedy algorithm is the largest. In addition, we also find that the driving time by using the SA algorithm is always the shortest, which suggests that SA performs the best path planning. Considering the perturbation parameter p in our greedy algorithm, we further study the driving time with different p for the greedy algorithm in **Figures 4C,D**. When $p = 0$, the probability of jumping out of the local optimal value is 0, so the driving time is a constant. When $p = 0.1$, the result is slightly worse than of $p = 0.05$. As we can give the optimal path before executing the task, we can simulate it several independent times to find out the inspection path associated with the shortest driving time.

In general, the working hours for everyone is not more than 8 h on a workday, namely, $t_{\max} = 28800$s. Let us assume that the inspection time for each tower is 15 min, that is, $t_{\text{ins}} = 900$s. From **Figure 3** and **Figure 4**, we can find that when assigning one single inspection team to perform the inspection task, the working hours of the four algorithms for line-5876 are about 59914s, 58403s, 54918s, and 57409s. For line-5803, the corresponding working hours are about 91433s, 87623s, 84507s, and 87396s. All the working hours exceed the maximum working hours $t_{\max}$; therefore, more inspection teams are needed. For line-5876, there are 48 towers and the average driving time from the office to each tower is 2221s. For line-5803, there are 72 towers, and the average driving time from the office to each tower is 3252s. From **Eqn. 16**, the minimum number of inspection teams is one of the elements in the set {2, 3, 4} for line-5876, and the minimum number of inspection teams is one of the elements in {3, 4, 5} for line-5803.

3.3 Optimize Path With the Minimum Number of Inspection Teams

In this section, we analyze and verify the theory of a minimum number of inspection teams and further study and verify our framework for the optimal path planning with the two power lines.

First, we study the optimal path planning with our framework without the swap strategy. From **Figure 5**, it can be found that when $k = 2$ and $k = 4$ for line-5876 and line-5803, respectively, the working hours are not well-balanced, and the working hours of some inspection teams exceed the given $t_{\max} = 28800$s. For line-5876 with $k = 3$ and line-5803 with $k = 5$, the working hours of all inspection teams are within 28800s, and at the same time, the working hours are well-balanced, which is coincided with the theoretical results. Furthermore, we can find that the optimal inspection path of SA has the shortest working hours, while the results of the other three algorithms are almost the same.

To compare with the method without a swap strategy, here we embed the swap strategy to optimize the inspection path. From

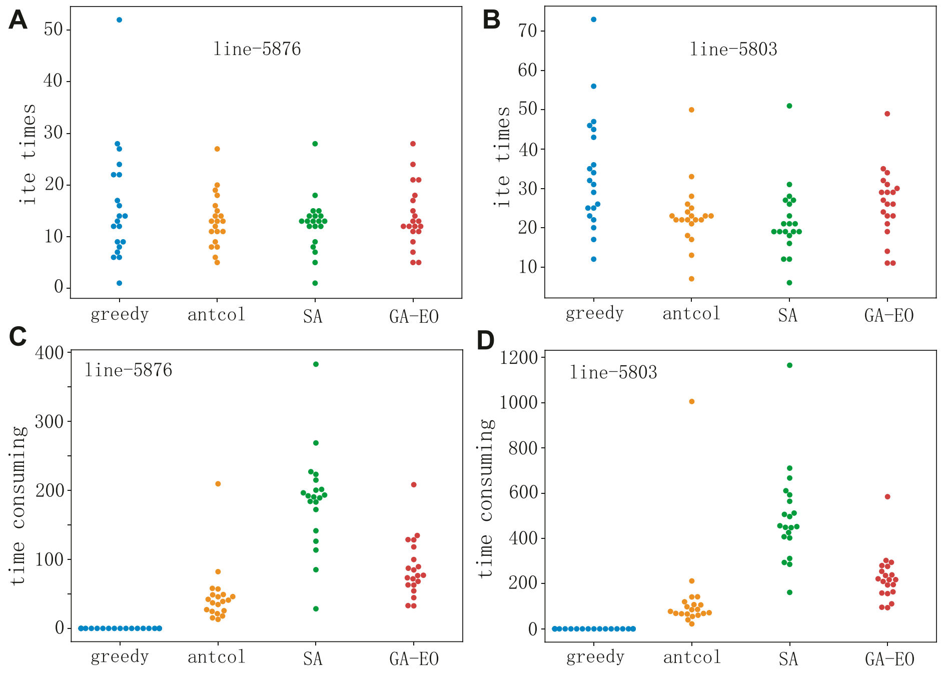


FIGURE 9 | (color online) **(A,B)** Number of iterations converging to the optimal solution with our framework under the four algorithms. The results are obtained by 20 independent simulations. **(C,D)** Time-consuming (seconds) of converging to the optimal solution with our framework under the four algorithms.

Figures 5B,D) and **Figure 6**, we can find that the working hours are shorter when combining the transfer and swap strategy for line-5876 and line-5803, which verifies the validity of our proposed transfer-swap algorithm for balancing working hours and minimizing the total working hours. A very interesting result is that with the transfer-swap algorithm, the working hours for the greedy algorithm and SA are very close and work the best, while the working hours for antcol are relatively long. The optimal inspection paths of the k inspection teams with our framework for line-5876 and line-5803 are shown in **Figure 7** and **Figure 8**.

Furthermore, we analyze the number of iterations and time-consuming converging to the optimal solution to quantify the performance of our framework under the four algorithms, see **Figure 9**. It can be found that for line-5876, the number of iterations to converge to the optimal solution is about 15 times. For line-5803, the number of iterations is about 25 times. **Figures 9C,D** show that the greedy algorithm is very fast and only takes a few seconds, but SA is the slowest and requires about 200s for line-5876 and 500s for line-5803. Therefore, combined with the time-consuming and the working hours, the greedy algorithm performs better than SA, which is a counter-intuitive result. The reason may be that the swap strategy and the perturbation parameter are helpful in avoiding locally optimal solutions.

4 CONCLUSION AND DISCUSSION

Taking both driving time and inspection time into consideration, in this study, we study the optimal path planning with the balanced working hours of each inspection team. In order to study the working hours, we have analyzed two real power lines in Jinhua city. In addition, we have provided a range of theoretical solutions for the minimum number of working teams and further have presented a fast method to estimate the theoretical solution, which is the first contribution. In addition, we have proposed a path optimization framework for balancing working hours and minimizing the total working hours based on the minimum number of inspection teams. The key to the proposed framework lies in the transfer-swap algorithm, and it is the second contribution. The simulation results showed that the minimum number of working teams was consistent with our theoretical solution and also verified our framework could balance the working hours and minimize the total working hours. Compared with the optimal results without a swap strategy, the total working hours are shorter when using our proposed transfer-swap algorithm. An interesting finding is that the simulated annealing algorithm (SA) had the shortest total working hours among the four algorithms when the swap strategy is absent. However, when using the transfer-swap algorithm, the working hours obtained by the greedy algorithm were close to those obtained by SA, but the greedy algorithm has the shortest computation time. Thus, with integrated optimization results and running time, it is more efficient to use the greedy algorithm.

In this article, we studied the perturbation parameter $p = 0.05$ for the greedy algorithm with our framework and did not analyze the pure greedy algorithm with $p = 0$. This is because the optimization result remains the same for $p = 0$, so it is easy to fall into a locally optimal solution. In our work, we assumed that the inspection time of each tower was the same. Therefore, there may be some fluctuations in the optimal results when using our framework. A better way to deal with this question is to predict the inspection time of each tower based on historical data. In addition, in our framework, we can use some other optimal algorithms to replace the four methods (greedy, SA, antcol, and EO-GA), such as reinforcement learning or deep learning algorithms.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Z-LH designed and wrote this manuscript. Y-ZD made the numerical simulations. HP analyzed the results, and all authors discussed and wrote the manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China (Nos. 62103375, 62072412, and 62171413), Key Project of Philosophy and Social Science of Zhejiang Province (22NDJC009Z), Social Development Project of Zhejiang Provincial Public Technology Research (Nos. LQ19F030010), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 19YJCZH056).

REFERENCES

1. Paul G, Rowena G, Ioana M, Miguel M. Does Electricity Drive Structural Transformation? Evidence from the united states. *Labour Econ* (2021) 68: 101944. doi:10.3386/w26477
2. Sergey VB, Roni P. Catastrophic cascade of Failures in Interdependent Networks. *Nature* (2010) 464:1025–8.
3. Sullivan JE, Kamensky D. How Cyber-Attacks in ukraine Show the Vulnerability of the u.S. Power Grid. *Electricity J* (2017) 30(3):30–5. doi:10.1016/j.tej.2017.02.006
4. Hu ZL, Wang L, Tang CB. Locating the Source Node of Diffusion Process in Cyber-Physical Networks via Minimum Observers. *Chaos* (2019) 29(6): 063117. doi:10.1063/1.5092772
5. Van NN, Robert J, Davide R. Automatic Autonomous Vision-Based Power Line Inspection: A Review of Current Status and the Potential Role of Deep Learning. *Int J Electr Power Energ Syst* (2018) 99:107–20.
6. Hu Z-L, Li J-H, Chen A, Xu F, Jia R, Lin F-L, et al. Optimize Grouping and Path of Pylon Inspection in Power System. *IEEE Access* (2020) 8:108885–95. doi:10.1109/access.2020.3001435
7. Li Z, Liu Y, Hayward R, Zhang J, Cai J. Knowledge-based Power Line Detection for Uav Surveillance and Inspection Systems. In: 2008 23rd International Conference Image and Vision Computing New Zealand. Christchurch, New Zealand: IEEE (2008). p. 1–6. doi:10.1109/ivcnz.2008.4762118
8. Wang Z, Gao Q, Xu J, Li D. A Review of Uav Power Line Inspection. In: L Yan, H Duan, X Yu, editors. *Advances in Guidance, Navigation and Control*. Singapore: Springer (2022). p. 3147–59. doi:10.1007/978-981-15-8155-7_263
9. Pan J-S, Lv J-X, Yan L-J, Weng S-W, Chu S-C, Xue J-K. Golden eagle Optimizer with Double Learning Strategies for 3d Path Planning of Uav in Power Inspection. *Mathematics Comput Simulation* (2022) 193:509–32. doi:10.1016/j.matcom.2021.10.032
10. Guan H, Sun X, Su Y, Hu T, Wang H, Wang H, et al. Uav-lidar Aids Automatic Intelligent Powerline Inspection. *Int J Electr Power Energ Syst* (2021) 130: 106987. doi:10.1016/j.ijepes.2021.106987
11. Pouliot N, Richard P-L, Montambault S. Linescout Technology Opens the Way to Robotic Inspection and Maintenance of High-Voltage Power Lines. *IEEE Power Energ Technol. Syst. J.* (2015) 2(1):1–11. doi:10.1109/jpets.2015.2395388
12. Katrasnik J, Pernus F, Likar B. A Survey of mobile Robots for Distribution Power Line Inspection. *IEEE Trans Power Deliv* (2010) 25(1):485–93. doi:10.1109/tpwrd.2009.2035427
13. Silano G, Bednar J, Nascimento T, Capitan J, Saska M, Ollero A. A Multi-Layer Software Architecture for Aerial Cognitive Multi-Robot Systems in Power Line Inspection Tasks. In: 2021 International Conference on Unmanned Aircraft Systems (ICUAS). Athens, Greece: IEEE (2021). p. 1624–9. doi:10.1109/icuas51884.2021.9476813
14. Carlos HFd.S, Mohamed HA, Daniel M, Campos BAA. Geometrical Motion Planning for cable-climbing Robots Applied to Distribution Power Lines Inspection. *Int J Syst Sci* (2021) 52(8):1646–63.
15. Croes GA. A Method for Solving Traveling-Salesman Problems. *Operations Res* (1958) 6(6):791–812. doi:10.1287/opre.6.6.791
16. Dantzig GB, Ramser JH. The Truck Dispatching Problem. *Manag Sci* (1959) 6(1):80–91. doi:10.1287/mnsc.6.1.80
17. Jiang C, Wan Z, Peng Z. A New Efficient Hybrid Algorithm for Large Scale Multiple Traveling Salesman Problems. *Expert Syst Appl* (2020) 139:112867. doi:10.1016/j.eswa.2019.112867
18. Miller CE, Tucker AW, Zemlin RA. Integer Programming Formulation of Traveling Salesman Problems. *J Acm* (1960) 7(4):326–9. doi:10.1145/321043.321046
19. Bellman R. Dynamic Programming Treatment of the Travelling Salesman Problem. *J Acm* (1962) 9(1):61–3. doi:10.1145/321105.321111
20. Volgenant T, Jonker R. A branch and Bound Algorithm for the Symmetric Traveling Salesman Problem Based on the 1-tree Relaxation. *Eur J Oper Res* (1982) 9(1):83–9. doi:10.1016/0377-2217(82)90015-7
21. Dell'Amico M, Montemanni R, Novellani S. Algorithms Based on branch and Bound for the Flying Sidekick Traveling Salesman Problem. *Omega* (2021) 104: 102493.
22. Maity S, Roy A, Maiti M. An Imprecise Multi-Objective Genetic Algorithm for Uncertain Constrained Multi-Objective Solid Travelling Salesman Problem. *Expert Syst Appl* (2016) 46:196–223. doi:10.1016/j.eswa.2015.10.019
23. Escario JB, Jimenez JF, Giron-Sierra JM. Ant colony Extended: Experiments on the Travelling Salesman Problem. *Expert Syst Appl* (2015) 42(1):390–410. doi:10.1016/j.eswa.2014.07.054
24. Ezugwu AE-S, Adewumi AO, Frîncu ME. Simulated Annealing Based Symbiotic Organisms Search Optimization Algorithm for Traveling Salesman Problem. *Expert Syst Appl* (2017) 77:189–210. doi:10.1016/j.eswa.2017.01.053
25. Zhou Y, Xu W, Fu Z-H, Zhou M. Multi-neighborhood Simulated Annealing-Based Iterated Local Search for Colored Traveling Salesman Problems. In: *IEEE Trans. Intell. Transport. Syst.* IEEE (2022). p. 1–11. doi:10.1109/tits.2022.3147924
26. Wang K-P, Huang L, Zhou C-G, Pang W. Particle Swarm Optimization for Traveling Salesman Problem. *Proc 2003 Int Conf Machine Learn Cybernetics* (2003) 3:1583–5.
27. Baraglia R, Hidalgo JI, Perego R. A Hybrid Heuristic for the Traveling Salesman Problem. *IEEE Trans Evol Computat* (2001) 5(6):613–22. doi:10.1109/4235.974843
28. Osaba E, Villar-Rodriguez E, Oregi I, Moreno-Fernandez-de-Leceta A. Hybrid Quantum Computing - Tabu Search Algorithm for Partitioning Problems: Preliminary Study on the Traveling Salesman Problem. In: *2021 IEEE Congress on Evolutionary Computation (CEC)*. Kraków, Poland: IEEE (2021). p. 351–8. doi:10.1109/cec45853.2021.9504923

29. Mele UJ, Gambardella LM, Montemanni R. A New Constructive Heuristic Driven by Machine Learning for the Traveling Salesman Problem. *Algorithms* (2021) 14(9):267. doi:10.3390/a14090267
30. Yujiao H, Zhen Z, Yuan Y, Xingpeng H, Xingshe Z, Wee SL. A Bidirectional Graph Neural Network for Traveling Salesman Problems on Arbitrary Symmetric Graphs. *Eng Appl Artif Intelligence* (2021) 97:104061.
31. Zhang Z, Liu H, Zhou M, Wang J. Solving Dynamic Traveling Salesman Problems with Deep Reinforcement Learning. In: *IEEE Trans. Neural Netw. Learning Syst*. IEEE (2021). p. 1–14. doi:10.1109/tnnls.2021.3105905
32. Tolga B. The Multiple Traveling Salesman Problem: An Overview of Formulations and Solution Procedures. *Omega* (2006) 34(3):209–19.
33. Jain S. Solving the Traveling Salesman Problem on the D-Wave Quantum Computer. *Front Phys* (2021) 9:760783. doi:10.3389/fphy.2021.760783
34. Changdar C, Pal RK, Mahapatra GS. A Genetic Ant colony Optimization Based Algorithm for Solid Multiple Travelling Salesmen Problem in Fuzzy Rough Environment. *Soft Comput* (2017) 21:4661–75. doi:10.1007/s00500-016-2075-4
35. Yan Z, Xiaoxia H, Yingchao D, Jun X, Gang X, Xinying X. A Novel State Transition Simulated Annealing Algorithm for the Multiple Traveling Salesmen Problem. *The J Supercomputing* (2021) 77:11827–52.
36. Chandran N, Narendran TT, Ganesh K. A Clustering Approach to Solve the Multiple Travelling Salesmen Problem. *Int J Ind Syst Eng* (2006) 1(3):372–87. doi:10.1504/ijise.2006.009794
37. Yang C, Szeto KY. Solving the Traveling Salesman Problem with a Multi-Agent System. In: *2019 IEEE Congress on Evolutionary Computation (CEC)*. Wellington, New Zealand: IEEE (2019). p. 158–65. doi:10.1109/cec.2019.8789895
38. Nallusamy R, Duraiswamy K, Dhanalaksmi R, Parthiban P. Optimization of Non-linear Multiple Traveling Salesman Problem Using K-Means Clustering, Shrink Wrap Algorithm and Meta-Heuristics. *Int J Nonlinear Sci* (2010) 9(2):171–7.
39. Alves RMF, Lopes CR. Using Genetic Algorithms to Minimize the Distance and Balance the Routes for the Multiple Traveling Salesman Problem. In: *2015 IEEE Congress on Evolutionary Computation (CEC)*. Sendai, Japan: IEEE (2015). p. 3171–8. doi:10.1109/cec.2015.7257285
40. Xiaolong X, Hao Y, Mark L, Marcello T. Two Phase Heuristic Algorithm for the Multiple-Travelling Salesman Problem. *Soft Comput* (2018) 22:6567–81.
41. Yongzhen W, Yan C, Yan L. Memetic Algorithm Based on Sequential Variable Neighborhood Descent for the Minmax Multiple Traveling Salesman Problem. *Comput Ind Eng* (2017) 106:105–22.
42. Lee TR, Ueng JH. A Study of Vehicle Routing Problems with Load-balancing. *Int Jnl Phys Dist Log Manage* (1999) 29(10):646–57. doi:10.1108/09600039910300019
43. Vandermeulen I, Gross R, Kolling A. Balanced Task Allocation by Partitioning the MultipleTraveling Salesperson Problem. In: 2019 International Conference on Autonomous Agents and Multiagent Systems. Montreal, Canada: ACM (2019). p. 1479–87.
44. Yu-Wang C, Yong-Zai L, Gen-Ke Y. Hybrid Evolutionary Algorithm with Marriage of Genetic Algorithm and Extremal Optimization for Production Scheduling. *Int J Adv Manufacturing Tech* (2008) 36:959–68.
45. Hoff A, Andersson H, Christiansen M, Hasle G, Løkketangen A. Industrial Aspects and Literature Survey: Fleet Composition and Routing. *Comput Operations Res* (2010) 37(12):2041–61. doi:10.1016/j.cor.2010.03.015
46. Ball MO, Golden BL, Assad AA, Bodin LD. Planning for Truck Fleet Size in the Presence of a Common-Carrier Option. *Decis Sci* (1983) 14(1):103–20. doi:10.1111/j.1540-5915.1983.tb00172.x
47. Dantzig GB, Fulkerson DR. Minimizing the Number of Tankers to Meet a Fixed Schedule. *Naval Res Logistics Q* (1954) 1(3):217–22.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 24 August 2022
DOI 10.3389/fphy.2022.926764

Structural evolution of international crop trade networks

Yin-Ting Zhang[1] and Wei-Xing Zhou[1,2,3]*

[1]School of Business, East China University of Science and Technology, Shanghai, China, [2]Research Center for Econophysics, East China University of Science and Technology, Shanghai, China, [3]Department of Mathematics, East China University of Science and Technology, Shanghai, China

Food security is a critical issue closely linked to human being. With the increasing demand for food, international trade has become the main access to supplementing domestic food shortages, which not only alleviates local food shocks, but also exposes economies to global food crises. In this paper, we construct four temporal international crop trade networks (iCTNs) based on trade values of maize, rice, soybean and wheat, and describe the structural evolution of different iCTNs from 1993 to 2018. We find that the size of all the four iCTNs expanded from 1993 to 2018 with more participants and larger trade values. Our results show that the iCTNs not only become tighter according to the increasing in network density and clustering coefficient, but also get more similar. We also find that the iCTNs are not always disassortative, unlike the world cereal trade networks and other international commodity trade networks. The degree assortative coefficients depend on degree directions and crop types. The analysis about assortativity also indicates that economies with high out-degree tend to connect with economies with low in-degree and low out-degree. Additionally, we compare the structure of the four iCTNs to enhance our understanding of the international food trade system. Although the overall evolutionary patterns of different iCTNs are similar, some crops exhibit idiosyncratic trade patterns. It highlights the need to consider different crop networks' idiosyncratic features while making food policies. Our findings about the dynamics of the iCTNs play an important role in understanding vulnerabilities in the global food system.

KEYWORDS

international crop trade, temporal network, directed network, structural evolution, food security

EDITED BY
Jianguo Liu,
Shanghai University of Finance and Economics, China

REVIEWED BY
Tian Qiu,
Nanchang Hangkong University, China
Sunil Kumar,
University of Delhi, India

*CORRESPONDENCE
Wei-Xing Zhou,
wxzhou@ecust.edu.cn

SPECIALTY SECTION
This article was submitted to Social Physics,
a section of the journal
Frontiers in Physics

RECEIVED 23 April 2022
ACCEPTED 29 July 2022
PUBLISHED 24 August 2022

CITATION
Zhang Y-T and Zhou W-X (2022), Structural evolution of international crop trade networks.
Front. Phys. 10:926764.
doi: 10.3389/fphy.2022.926764

1 Introduction

Food security is a mainstay in national security and has become one of the global hot spots [1]. Due to the impact of the COVID-19 pandemic, the number of global hungry people continued to rise in 2020, from 2.05 billion to 2.37 billion, and about 30 million in 2030 more than if the pandemic had not happened[1]. Food supplies face unknown potential risks, with factors such as global pandemics, climate extremes, conflicts and so on. Globalization confers pros and cons with regard to food security [2], providing access to international food trade [3]. On the one hand, international trade meets food demand of some economies with food shortages by supplying food produced elsewhere beyond self-consumption and strategic reserves [4]. On the other hand, trade might multiply disruption to food supply chains [5] and exacerbate economies' vulnerability to sudden shock in global food system [6]. Therefore, international food trade has a crucial impact on food security [7, 8].

Before evaluating underlying benefits and risks for food security, it is necessary to explore characteristics of international food trade. Recently, complex networks have become an important method to study trade relationships existing between pairs of economies in the world [9]. Therefore, many studies have contributed insights into the structure and dynamics of the global food trade system based on network science [10–12]. Some literature focused on one major crop feeding a large number of population, such as maize [13] and wheat [14]. These studies described the trade patterns of international crop trade system [13], and explored the factors that impact the food supply [15]. Investigating the international virtual water network (iVWN) is another common approach to understand global food security [16]. By quantifying water embodied in several food commodities, researchers link the properties of the iVTW to the resilience of the global food system to shocks [17, 18]. However, there are many different definition of resilience [4, 19], or indicators measuring network vulnerability [15, 20].

The topological properties of the international food trade networks (iFTNs) are closely related to the assessment of global food security and should be investigated carefully. Previous literature has focused on the complexity of iFTNs [11] and studied the impact of shocks to the iFTN [21, 22]. However, the microstructure of the iFTN is worth discussing and studying [23]. The evolution of the international food trade system and comparison of trade patterns between different crops still remain a spectrum of investigation. Here, we consider four dietary staples (maize, rice, soybean and wheat), which make up more than 75% of the calories consumed by populations and animals [2, 11]. We construct four international crop trade networks (iCTNs) and quantify the evolution of these iCTNs from 1993 to 2018. Although our work does not specifically evaluate shocks or food security, the dynamics of four different iCTNs provide basic understanding of the global food trade system. The evolution of network features shows the change of iCTNs and the necessity of new methods measuring food security.

In this paper, we attempt to explore and compare the main stylized characteristics pertaining to crop trade relationships and their evolution over time. We focus on structural characteristics such as node degrees, node strengths, link weights, density, the clustering coefficient, reciprocity and assortativity. Our study answers two questions: 1) How has the structure of iCTNs changed over time? 2) What are the differences in trade patterns between different crops? The remainder of this paper is organized as follows. Section 2 describes the data sets used in our work and the construction of the iCTNs. Section 3 presents the empirical results about the dynamics of the four iCTNs. We summarize our results in Section 4.

2 Data and method

2.1 Data description

We obtained the FAOSTAT data sets on international trade flows from the Food and Agriculture Organization (FAO, http://www.fao.org), which contain the annual bilateral export-import data during the period 1993–2018. The Soviet Union collapsed in 1991 and the world pattern changed dramatically, and Yugoslavia and Czechoslovakia also disintegrated one after another in 1992. Therefore, our data began in 1992 [17]. Since the data sets contain some inconsistencies between the declaration of importers and exporters, we first complied the crop trade matrix by using the import data, and then used the export data to fill data gaps. We got four crop trade matrices $W^{crop}(t)$, and denoted them with superscripts $crop \in \{\mathrm{M, R, S, W}\}$ for maize, rice, soybean and wheat. The number of economies changes as the evolution of political boundaries over time. However, this fact does not affect our analysis of iCTNs [24]. We excluded economies from the annual network analysis when their aggregated values of any kind of crop trade was zero. The final data sets for the network analysis covered 246 economies over the period from 1993 to 2018.

2.2 Network construction

We constructed the temporal iCTNs with respect to different crops. The annual iCTN in each year is a multi-layer network, where the nodes represent economies connected by multiple directed links (or links). The link weight $w_{ij}^{crop}(t)$ for a crop is the exports from the economy i to the economy j in a network $G^{crop}(t) = (\mathcal{V}^{crop}(t), W^{crop}(t))$, where $\mathcal{V}^{crop}(t)$ is the set of

[1] The State of Food Security and Nutrition in the World (2020), available at https://www.fao.org.

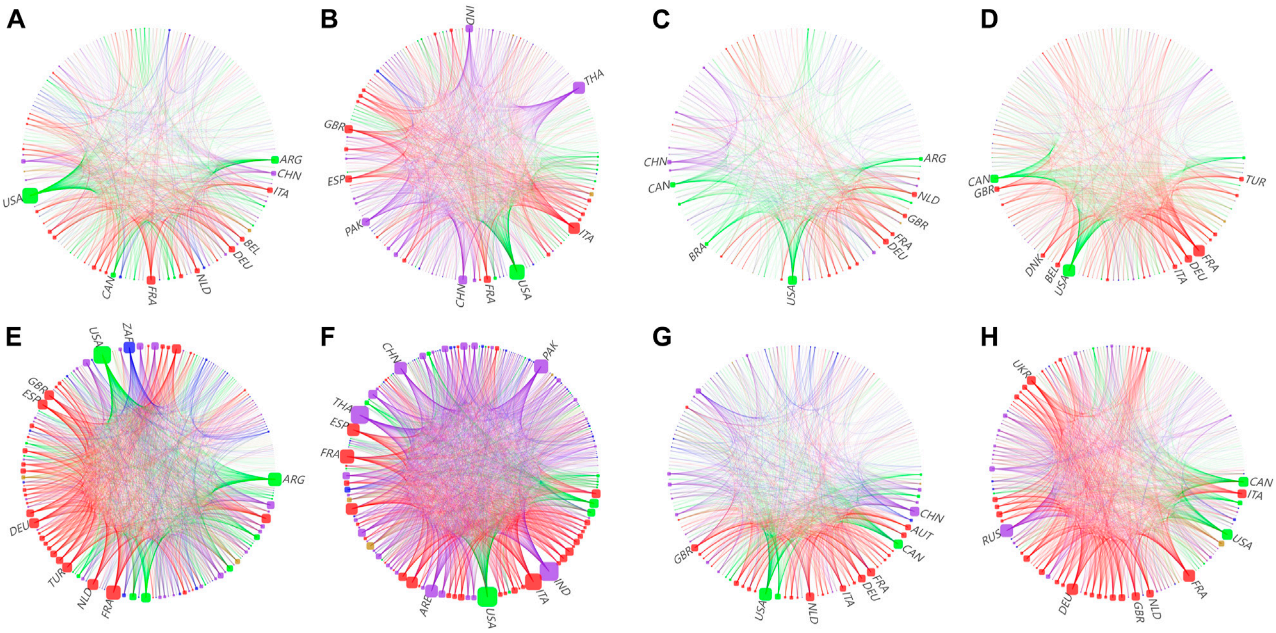


FIGURE 1
Four international crop trade networks (iCTNs) in 1993 **(A–D)** and 2018 **(E–H)**. The columns from left to right respectively describe maize, rice, soybean and wheat. For each chordal graph, nodes stand for economies participating in crop trade. The color of nodes is corresponding to different region. The nodes in blue stand for economies in Africa region; the nodes in green stand for economies in America region; the nodes in purple stand for economies in Asia region; the nodes in red stand for economies in Europe region and the nodes in brown stand for economies in Pacific region. Outgoing links from an economy are shown with the same color as the origin region.

nodes (that is, the set of economies involved in the trade of *crop* in year *t*). The total networks over 26 years include 246 economies. Not all economies engaged in crops trade in each year, so usually $N_{\mathcal{V}}^{crop}(t) < 246$. We obtained 26 × 4 yearly international crop trade networks and explored the evolution of structural properties for each iCTN in this work.

Figure 1 shows the four iCTNs in 1993 and 2018. For each economy (node), the symbol size represents its total export value. The thickness of a link represents the trade flow between two economies. It is evident that compared with the iCTNs of 1993, there were more links in the iCTNs of 2018, which indicates that new trade relationships were formed. The nodes became larger and the links became broader, corresponding to larger trade volumes. We note that economies in Asia and Europe are major exporters, especially for maize and wheat. What's more, the United States and Germany are the most important crop exporters that had very large export values in 2018.

3 Empirical results

3.1 Summary statistics

For each year, we computed network statistics and described the evolution of the four iCTNs. For simplicity, we omitted the superscript *crop* in the following description. The number of nodes $N_{\mathcal{V}}$ measures how many economies engaged in trade, and the number of links $N_{\mathcal{E}}$ measures the trade relationships between economies, where $\mathcal{E} = \{e_{ij}\}$ is the set of links e_{ij}. Figure 2 illustrates the size evolution of the four iCTNs from 1993–2018.

The number $N_{\mathcal{V}}$ of nodes involved in Figure 2A is the number of nodes, where $\mathcal{V}$ is the set of nodes. Compared with the iCTNs in 1993, the number of nodes of the maize and soybean networks increased markedly. It is consistent with previous literature [21]. However, the number of nodes of the rice and wheat networks kept stable with some slight fluctuations. Figure 2B shows the evolution of the number of links $N_{\mathcal{E}}$, which show excellent linear growth with respect to time *t*:

$$N_{\mathcal{E}}^{crop} = a^{crop} + b^{crop}t, \tag{1}$$

A linear regression gives that $b^{\mathrm{M}} = 50.09$ for maize, $b^{\mathrm{R}} = 65.85$ for rice, $b^{\mathrm{S}} = 27.90$ for soybean, and $b^{\mathrm{W}} = 34.64$ for wheat. This highlights the fact that the number of links approximately increased linearly year by year. In general, the size of four iCTNs has expanded from 1993 to 2018. The network of rice had the largest size, indicating that more rice trade relations have been established between economies. Similarly, for soybean trade, less trade links have been established between

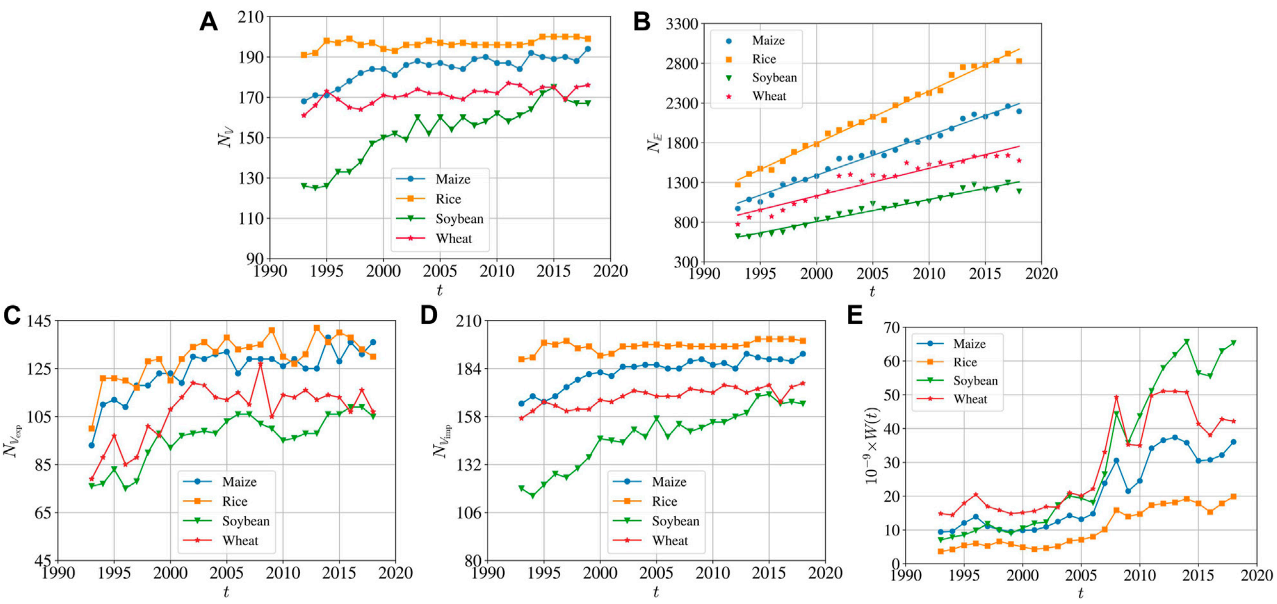


FIGURE 2
Evolution of the summary statistics of the four iCTNs from 1993 to 2018. **(A)** Number of nodes $N_{\mathcal{V}}$. **(B)** Number of links $N_{\mathcal{E}}$. **(C)** Number of exporting economies $N_{\mathcal{V}_{\rm exp}}$. **(D)** Number of importing economies $N_{\mathcal{V}_{\rm imp}}$. **(E)** Total link weight W in units of US dollars. Curves with different colored markers correspond to different crops.

economies. The increase in trade and network complexity differ for different iCTNs.

Figures 2C,D show the numbers of exporting and importing economies ($N_{\mathcal{V}_{\rm exp}}$ and $N_{\mathcal{V}_{\rm imp}}$) of the four iCTNs from 1993 to 2018. Compared with $N_{\mathcal{V}}$ in Figure 2A, we recognize that:

$$N_{\mathcal{V}_{\rm exp}} < N_{\mathcal{V}_{\rm imp}} < N_{\mathcal{V}} < N_{\mathcal{V}_{\rm exp}} + N_{\mathcal{V}_{\rm imp}}. \tag{2}$$

There are much more importing economies than exporting economies, and many economies both export and import the same crops, which is also observed for the international pesticide trade networks [25].

Link weight presents the trade value between two economies. We calculated the sum of link weights to show the total trade value of a given crop. Figure 2E describes the evolution of the international trade values $W(t)$ of the four crops from 1993 to 2018. We find that the trade values of the four crops have an overall increasing trend, but decreased locally. Remarkably, $W(t)$ increased sharply in 2007/2008 due to the 2008 food crisis, which results in a significant rise in food prices and food insecurity [26]. Contributing factors are various, and macro-level underlying causes include higher oil prices, which affect the costs for food production and processing. Indeed, the oil market crashed in the middle of 2008 [27], followed by the prices of agricultural goods. In particular, a general rise in agricultural prices could create a global food price bubble [28]. Agricultural commodities exhibited unexpected price spikes again in 2011, prompting an increase in crop trade values [29]. Hence, we observe that $W(t)$ experienced a marked increase in 2011. In addition, rice had the lowest trade values in each year, and $W(t)$ of soybean overtook wheat to the highest in 2009.

3.2 Degree and strength

The node degrees show how many trade partners each economy has. In a directed network, we consider both in-degree and out-degree of a node to measure import and export respectively. The in-degree of node is defined as follows

$$k_i^{\rm in} = \sum_{j\in\mathcal{V}-\{i\}} I_{\mathcal{E}}\left(e_{ji}\right) = \sum_{j=1}^{N_{\mathcal{V}}} I_{\mathcal{E}}\left(e_{ji}\right), \tag{3}$$

where $I_{\mathcal{E}}(e_{ji})$ is the indicator function:

$$I_{\mathcal{E}}\left(e_{ji}\right) = \begin{cases} 1, & \text{if} \quad e_{ji} \in \mathcal{E} \\ 0, & \text{if} \quad e_{ji} \notin \mathcal{E} \end{cases} \tag{4}$$

The out-degree of node is defined as follows

$$k_i^{\rm out} = \sum_{j\in\mathcal{V}-\{i\}} I_{\mathcal{E}}\left(e_{ij}\right) = \sum_{j=1}^{N_{\mathcal{V}}} I_{\mathcal{E}}\left(e_{ij}\right). \tag{5}$$

Since the networks are weighted, we quantity node strengths, including in-strength $s_i^{\rm in}$ and out-strength $s_i^{\rm out}$, which are defined as follows

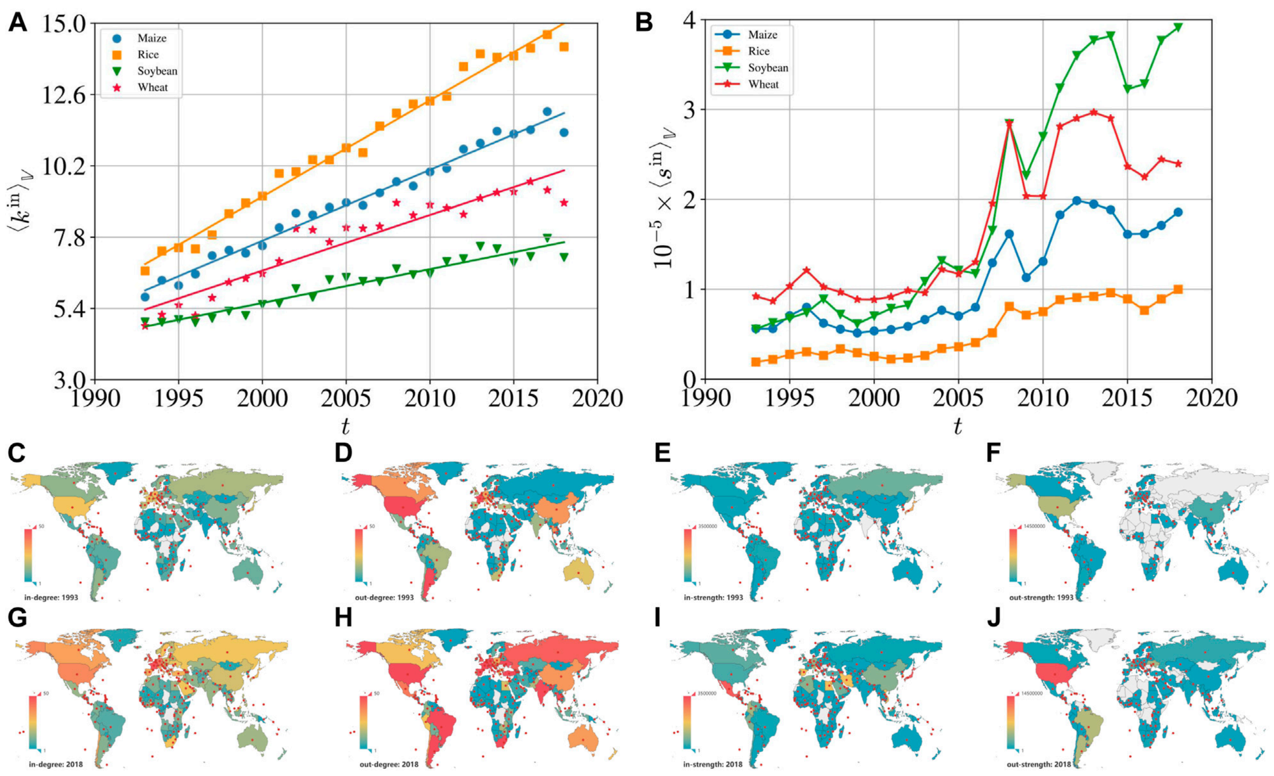


FIGURE 3
Yearly evolution of the degrees and the strengths of the four iCTNs from 1993 to 2018. The graphs in the first row **(A–B)** respectively show the evolution of the average in-degree and the average in-strength from 1993 to 2018. Curves in different colors correspond to different crops. The rows from middle to bottom show the distributions of the degrees and the strengths of the international maize trade network in 1993 and in 2018. The middle row shows the global map in 1993: **(C)** The distribution of the in-degree; **(D)** the distribution of the out-degree; **(E)** the distribution of the in-strength; **(F)** the distribution of the out-strength. The bottom row shows the global map in 2018: **(G)** The distribution of the in-degree; **(H)** the distribution of the out-degree; **(I)** the distribution of the in-strength; **(J)** the distribution of the out-strength.

$$s_i^{\text{in}} = \sum_{j\in\mathcal{V}-\{i\}} w_{ji} = \sum_{j=1}^{N_{\mathcal{V}}} w_{ji}, \tag{6}$$

$$s_i^{\text{out}} = \sum_{j\in\mathcal{V}-\{i\}} w_{ij} = \sum_{i=1}^{N_{\mathcal{V}}} w_{ij}, \tag{7}$$

where $w_{jj} = 0$ by definition.

The degrees and the strengths measure the importance of a node, and we used the average degrees and the average strengths to evaluate the overall structure of the networks. It is easy to get that the average in-degree of nodes $\langle k^{\text{in}}\rangle_{\mathcal{V}}$ is equal to the average out-degree of nodes $\langle k^{\text{out}}\rangle_{\mathcal{V}}$ [25]. The average in-strength $\langle s^{\text{in}}\rangle_{\mathcal{V}}$ and the average out-strength $\langle s^{\text{out}}\rangle_{\mathcal{V}}$ respectively measure the average values of imports and exports, which are also equal to each other.

Figures 3A,B show the yearly evolution of the average in-degree and in-strength from 1993 to 2018. The average node in-degree represents the average number of exporting partners owned to an economy [30], which is equal to the average node our-degree [25]. Figure 3A shows that the evolution of the average in-degree has an excellent linear growth with respect to time t:

$$k_{\mathcal{V}}^{\text{in, crop}} = a^{crop} + b^{crop}t. \tag{8}$$

Simple linear regressions give that $b^{\text{M}} = 0.24$ for maize, $b^{\text{R}} = 0.32$ for rice, $b^{\text{S}} = 0.11$ for soybean, and $b^{\text{W}} = 0.19$ for wheat. For all the iCTNs, the average in-degree increases with time, which indicates increasing active trade relationships among economies. The increasing trend of $k_{\mathcal{V}}^{\text{in}}$ is a result of the growth rate of $N_{\mathcal{E}}$ being greater than that of $N_{\mathcal{V}}$. We plotted the distributions of the in- and out-degree of the iCTNs in 1993 and 2018 to describe the explicit change. Since the results are similar in different iCTNs, here we only show the global maps of the international maize trade network. As shown in Figures 3C,D,G,H, both for in-degree and out-degree, the color of maps in 2018 was deeper than that in 1993, indicating an increase in the number of crop trade connections. However, $k_{\mathcal{V}}^{\text{in}}$ decreased markedly in 2018. It may be due to the fact that the growth of networks is not caused by simply adding new links to existing nodes which would disappear while new nodes are created [31].

As shown in Figure 3B, similar to the dynamics of link weights, the average in-strength increased across the sample and showed a potential upward trend. Before 2007, $s_{\mathcal{V}}^{\text{in}}$ kept a slight increase for each crop and occurred small fluctuations in some years. The average in-strength showed significant fluctuations in 1996–1997 and 2004–2005. The main factor that caused the average trade values to change dramatically is the food price. Prices for most crops started to climb slowly in 1990 and peaked in 1996 (maize, rice and wheat) and 1997 (soybean) before declining sharply. But financial crisis of 1997–99 quickly ended the crop price surge [32]. In 2004, due to bad harvests and high oil prices, the food prices increased, which caused an increase in global food trade values. The food prices slowed down as the global commodity prices were under control in 2005. The "world food crisis" of 2007–2008 inflated food prices significantly. This crisis originated from the long-term cycle of fossil-fuel dependence on industrial capitalism, coupled with the inflationary effect of current biofuel offset and financial speculation [33]. Under the influence of the food price crisis [34], the average trade values showed a significant upward trend from 2007 to 2008. After 2008, $s_{\mathcal{V}}^{\text{in}}$ reverted to increase with fluctuations. Overall, both for in-strength and out-strength, the color of maps in 2018 was deeper than that in 1993 as shown in Figures 3E,F,I,J, suggesting an increase in crop trade volumes.

3.3 Competitiveness

The density of a directed network refers to the ratio of the number of links that actually exist in the network to the number of all possible links:

$$\rho = \frac{N_{\mathcal{E}}}{N_{\mathcal{V}}\left(N_{\mathcal{V}}-1\right)}, \tag{9}$$

To capture the potential relations between an economy's trading partners, we used the clustering coefficient to measure the connectivity of the economy's trading partners [35]. For a weighted network, the clustering coefficient of a node i is the ratio of all directed triangles to all possible triangles [36],

$$c_i = \frac{2T_i}{k_i\left(k_i-1\right)-2k_i^R}, \tag{10}$$

where T_i is the number of directed triangles containing node i, k_i is the total degree of node i, and

$$k_i^R = \sharp\left(\left\{j: e_{ij} \in \mathcal{E} \ \& \ e_{ji} \in \mathcal{E}\right\}\right) = \sum_{j \neq i}\left(w_{ij}w_{ji}\right)^0, \text{ for } w_{ij}w_{ji} \neq 0 \tag{11}$$

is the reciprocal degree of node i. We use the average clustering coefficient to measure the overall concentration of the network:

$$\langle c \rangle_{\mathcal{V}} = \frac{1}{N_{\mathcal{V}}}\sum_{i=1}^{N_{\mathcal{V}}} c_i. \tag{12}$$

Figure 4 illustrates the density and average clustering coefficient of the four crop trade networks, introduced to describe the competitiveness of the entire network. The value of density represents the tightness of a network [37]. In a dense network, the number of connections approaches to the maximum number of potential ties. According to Figure 4A, the density of each crop network was small. Over the last 26 years, the density rose with some fluctuations, and it indicates that the global food trade is becoming more and more frequent and close. The increasing densities of the iCTNs are consistent with some other international trade networks [38], but are much smaller [25, 39] than the total world trade networks. The density curves for rice, maize and wheat showed a linear upward trend, where the density of the rice network had the largest slope of increase and has become the largest since 1998. Although the number of links for the rice trade network increased significantly, its density did not change dramatically after 2012. From 2009 to 2012, the network density of the soybean trade network changed slightly without a dramatic trend, and fluctuated significantly after 2012.

The average clustering coefficient measures the overall concentration of connections in the network. Figure 4B shows that the economies were inclined to cluster together in the four iCTNs. The clustering coefficients have an upward trend, especially for the rice and soybean trade networks, which is consistent with conclusions from previous literature [40, 41]. Likewise, the values of clustering coefficients for rice were the largest after 2001 and displayed relatively least fluctuations, since the export and import of rice concentrated in some economies [42]. Compared with the evolution of the network density, a particularly dense network was inclined to have high clustering because its modes are more likely to share partners [38].

3.4 Persistence

There are two types of complex networks: multi-layer networks, in which nodes are connected in different ways; and temporal networks, in which nodes and links may appear or disappear and their attributes to the networks might change over time [43]. Node similarity has been widely studied for simple networks [44, 45] and multi-layer networks [43, 46]. This paper adopts a simple indicator to measure the similarity coefficient between two successive networks [25], since the links in iCTNs might look similar or change significantly. Considering two successive networks $\mathcal{G}(t-1)$ and $\mathcal{G}(t)$, let $\mathcal{E}_{(t-1)\cup t} = \mathcal{E}(t-1) \cup \mathcal{E}(t)$ be the union set of directed links and $\mathcal{E}_{(t-1)\cap t} = \mathcal{E}(t-1) \cap \mathcal{E}(t)$ be the intersection of directed links. Based on previous studies [25, 47, 48], we define the temporal similarity between two successive networks $\mathcal{G}(t-1)$ and $\mathcal{G}(t)$

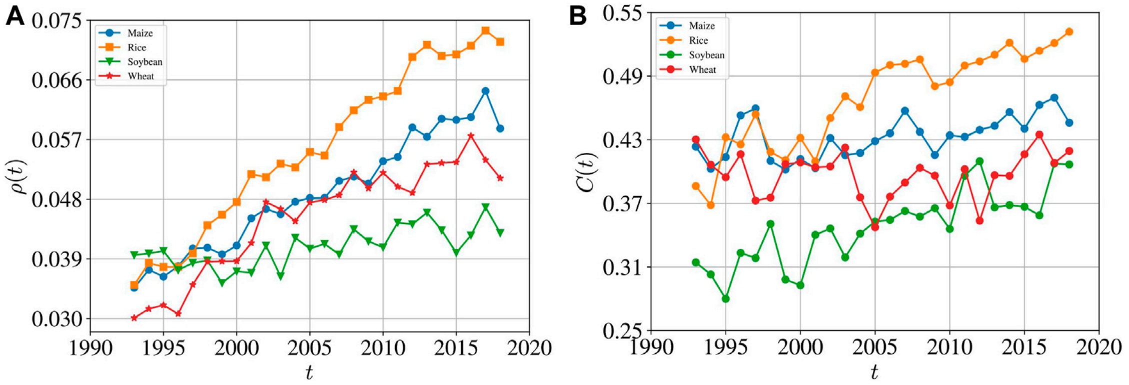


FIGURE 4
Yearly evolution of the network density **(A)** and average clustering coefficient **(B)** of the four international crop trade networks from 1993 to 2018.

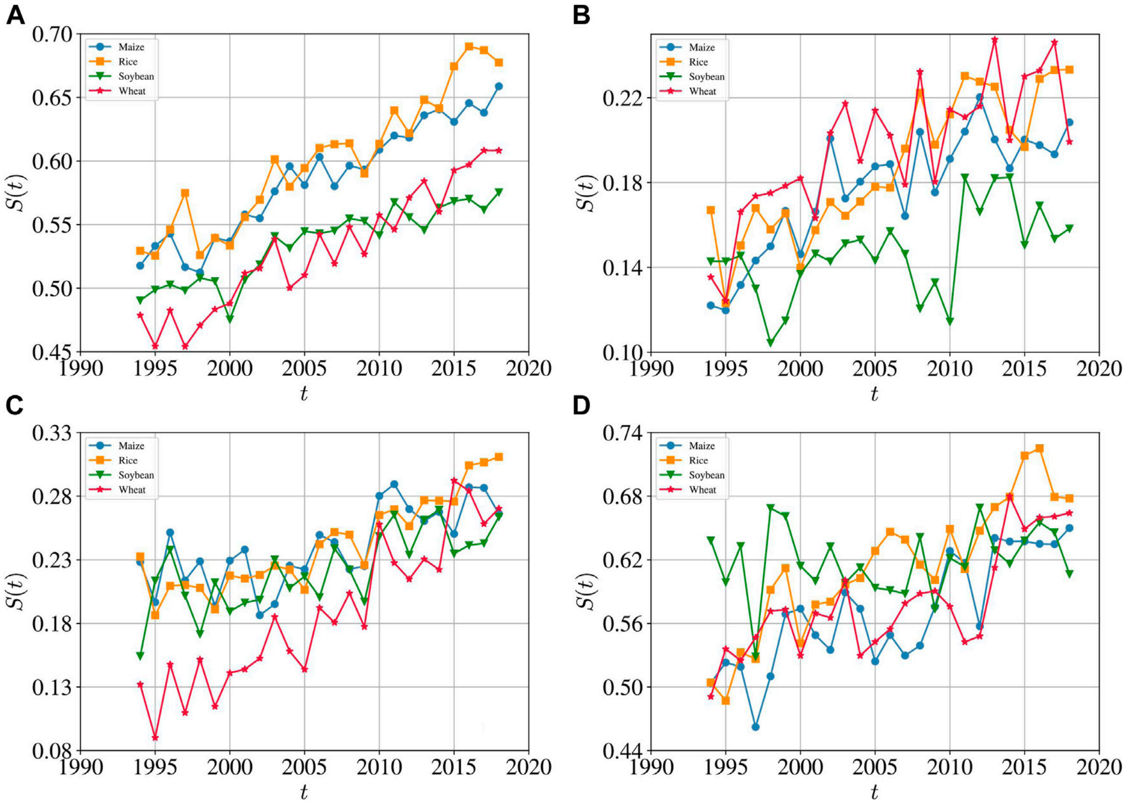


FIGURE 5
Evolution of the temporal similarity coefficient *S*(*t*) between two successive networks of the four crops from 1993 to 2018. **(A)** All links at each time. **(B)** Light links with the weights at each time less than the 20% percentile. **(C)** Medium links with the weights at each time between the 40 and 60% percentiles. **(D)** Heavy links with the weights at each time greater than the 80% percentile. The temporal similarity increased over time with slight fluctuations.

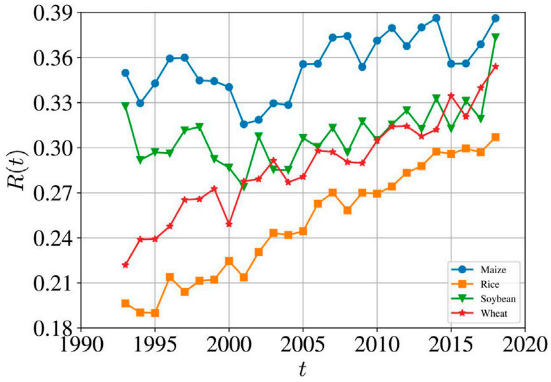


FIGURE 6
Evolution of overall reciprocity of the four iCTNs from 1993 to 2018. The overall reciprocity coefficients were between 0.1 and 0.4.

as the ratio of the number of overlapping directed links in the two networks over the number of all directed links in the two networks:

$$S(t) = \frac{\sharp\left(\mathcal{E}_{(t-1)\cap t}\right)}{\sharp\mathcal{E}_{(t-1)\cup t}}. \tag{13}$$

where $\sharp(\mathbf{X})$ denotes the cardinal number of set **X**. The value of the similarity coefficient $S(t)$ ranges between 0 and 1: $S(t) = 0$ indicates that the two networks are completely different in means of links, while $S(t) = 1$ means that the two networks are completely the same.

The analysis of the node similarity is a significant basis for understanding the evolution of features of the international crop trade system. The small value of the similarity coefficient S shows a high discrepancy in the structure of two successive networks [25, 49]. From Figure 5A, the temporal similarity increased over time with slight fluctuations, which indicates that the structure of successive iCTNs gets more similar. And the rice trade network had the largest temporal similarity recently. By comparing the similarity coefficient $S(t)$ of sub-networks containing links with different values of weight (light links in Figure 5B where the weights are less than the 20% percentile, medium links in Figure 5C where the weights are between the 40 and 60% percentiles, and heavy links in Figure 5D where the weights are greater than the 80% percentile), it can be found that the $S(t)$ curves have similar patterns qualitatively and the heavier links with greater trade flows have more stable.

3.5 Reciprocity

The reciprocity is critical to dynamical processes and network growth [50]. The reciprocity of a directed network is defined as the ratio of the number of bilateral links (i.e., links pointing in both directions) to the total number of links in the network [51, 52]:

$$R = \frac{\sharp\left(\left\{(i,j): e_{ij} \in \mathcal{E} \;\&\; e_{ji} \in \mathcal{E}\right\}\right)}{\sharp\left(\left\{(i,j): e_{ij} \in \mathcal{E}\right\}\right)} = \frac{1}{N_{\mathcal{E}}}\sum_{i\in V} k_i^R, \tag{14}$$

where

$$\sharp\left(\left\{(i,j): e_{ij} \in \mathcal{E} \;\&\; e_{ji} \in \mathcal{E}\right\}\right) = \sum_{i\in V} k_i^R \tag{15}$$

and

$$\sharp\left(\left\{(i,j): e_{ij} \in \mathcal{E}\right\}\right) = N_{\mathcal{E}}. \tag{16}$$

Reciprocity R is an indicator of the degree of bilateral trade relationships between economies in a network and plays an important role in the transmission mechanism of international trade information.

Figure 6 shows the evolution of overall reciprocity of the four iCTNs from 1993 to 2018. The overall reciprocity coefficients were between 0.1 and 0.4. It is found that the overall reciprocity was relatively stable with slight fluctuations for maize and soybean. In terms of wheat and rice, the reciprocity coefficients were always smaller than those of the maize and soybean trade networks, but showed an increasing trend. Especially for rice, the reciprocal coefficient $R(t)$ showed a nice linear relationship with time t. We note that the reciprocity coefficients of the iCTNs are much smaller than those of the international trade networks (larger than 0.5) [39, 51], which contain remarkably more commodities and thus more reciprocal links.

3.6 Assortativity

Assortativity quantifies the mixing pattern of complex networks, which measures whether the node is preferentially connected to a node with a similar scale [53]. In a directed network, we consider the correlation of four degree directions. The degree assortative coefficient $r_{\rm in,in}(t)$ between the in-degree of exporting economies and the in-degree of importing economies:

$$r_{\rm in,in}(t) = \frac{1}{N_{\mathcal{E}}}\sum_{e_{ij}\in\mathcal{E}} \frac{\left[\left(k_i^{\rm in} - \langle k_i^{\rm in}\rangle_{\mathcal{E}}\right)\left(k_j^{\rm in} - \langle k_j^{\rm in}\rangle_{\mathcal{E}}\right)\right]}{\sigma_{i,\mathcal{E}}^{\rm in}\sigma_{j,\mathcal{E}}^{\rm in}}, \tag{17}$$

where $\langle k_i^{\rm in}\rangle_{\mathcal{E}}$ and $\langle k_j^{\rm in}\rangle_{\mathcal{E}}$ are respectively the mean in-degrees of exporting economies and importing economies, and the variance of in-degrees of exporting economies is

$$\left(\sigma_{i,\mathcal{E}}^{\rm in}\right)^2 = \frac{1}{N_{\mathcal{E}}}\sum_{e_{ij}\in\mathcal{E}} \left(k_i^{\rm in} - \langle k_i^{\rm in}\rangle_{\mathcal{E}}\right)^2, \tag{18}$$

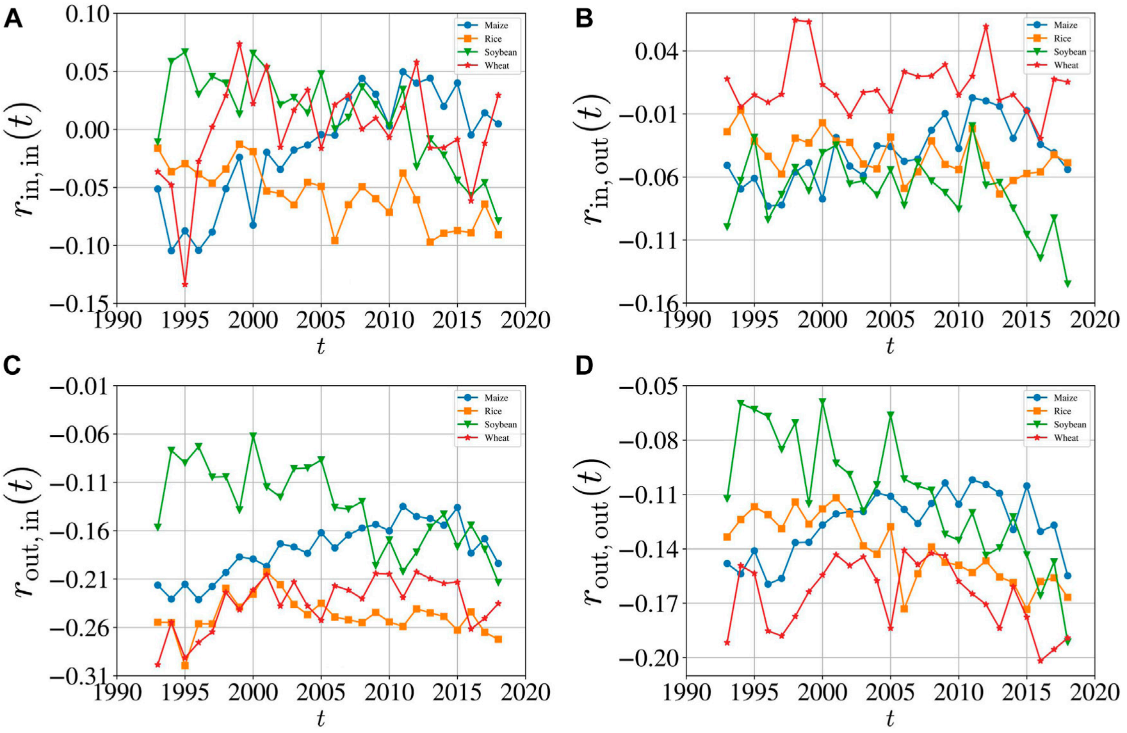


FIGURE 7
Evolution of degree assortative coefficients of the four iCTNs from 1993 to 2018. **(A)** The degree assortative coefficient $r_{\mathrm{in,in}}(t)$ between the in-degree of exporting economies and the in-degree of importing economies. **(B)** The degree assortative coefficient $r_{\mathrm{in,out}}(t)$ between the in-degree of exporting economies and the out-degree of importing economies. **(C)** The degree assortative coefficient $r_{\mathrm{out,in}}(t)$ between the out-degree of exporting economies and the in-degree of importing economies. **(D)** The degree assortative coefficient $r_{\mathrm{out,out}}(t)$ between the out-degree of exporting economies and the out-degree of importing economies.

and the variance $(\sigma_{j,\mathcal{E}}^{\mathrm{in}})^2$ of in-degrees of importing economies is defined in the same way.

Similarly, the degree assortative coefficient $r_{\mathrm{in,out}}(t)$ between the in-degree of exporting economies and the out-degree of importing economies is

$$r_{\mathrm{in,out}}(t) = \frac{1}{N_{\mathcal{E}}}\sum_{e_{ij}\in\mathcal{E}}\frac{\left[\left(k_i^{\mathrm{in}} - \langle k_i^{\mathrm{in}}\rangle_{\mathcal{E}}\right)\left(k_j^{\mathrm{out}} - \langle k_j^{\mathrm{out}}\rangle_{\mathcal{E}}\right)\right]}{\sigma_{i,\mathcal{E}}^{\mathrm{in}}\sigma_{j,\mathcal{E}}^{\mathrm{out}}}, \tag{19}$$

the degree assortative coefficient $r_{\mathrm{out,in}}(t)$ between the out-degree of exporting economies and the in-degree of importing economies is

$$r_{\mathrm{out,in}}(t) = \frac{1}{N_{\mathcal{E}}}\sum_{e_{ij}\in\mathcal{E}}\frac{\left[\left(k_i^{\mathrm{out}} - \langle k_i^{\mathrm{out}}\rangle_{\mathcal{E}}\right)\left(k_j^{\mathrm{in}} - \langle k_j^{\mathrm{in}}\rangle_{\mathcal{E}}\right)\right]}{\sigma_{i,\mathcal{E}}^{\mathrm{out}}\sigma_{j,\mathcal{E}}^{\mathrm{in}}}, \tag{20}$$

and the degree assortative coefficient $r_{\mathrm{out,out}}(t)$ between the out-degree of exporting economies and the out-degree of importing economies is

$$r_{\mathrm{out,out}}(t) = \frac{1}{N_{\mathcal{E}}}\sum_{e_{ij}\in\mathcal{E}}\frac{\left[\left(k_i^{\mathrm{out}} - \langle k_i^{\mathrm{out}}\rangle_{\mathcal{E}}\right)\left(k_j^{\mathrm{out}} - \langle k_j^{\mathrm{out}}\rangle_{\mathcal{E}}\right)\right]}{\sigma_{i,\mathcal{E}}^{\mathrm{out}}\sigma_{j,\mathcal{E}}^{\mathrm{out}}}. \tag{21}$$

The four degree assortative coefficients of different directions can be used to describe the relevance of two nodes connected by a directed link through their in- and out-degrees to accurately explore the mixing patterns of the international crop trade networks.

Figure 7 shows the evolution of the degree assortative coefficients of the four iCTNs from 1993 to 2018. The r values fluctuated sharply before 2000, followed by relatively mild fluctuations. Previous research that did not consider the direction of the degree has shown that world cereal trade networks [54] or other international trade networks [39, 55] are disassortative. In this paper we find that the degree assortative coefficients of different directions for different crop networks have different assortative patterns. As shown in Figure 7A, the degree assortative coefficients $r_{\mathrm{in,in}}(t)$ ranged from −0.2 to 0.1. For maize, the coefficients were almost negative before 2006, and showed an upward trend until 2015. For rice, the coefficients were always negative. For soybean, the coefficients showed significant fluctuations before 1994, ranged from 0 to 0.1 with fluctuations during 1995–2011, and finally dropped to less than zero. From Figure 7B, the coefficients $r_{\mathrm{in,out}}(t)$ for maize, rice and soybean were almost negative, while the coefficients for wheat were mainly positive. According to Figures 7C,D, for all the iCTNs, the degree assortative coefficients $r_{\text{out,in}}(t)$ and $r_{\text{out,out}}(t)$ were generally negative. In summary, except that almost all the assortative coefficients for the international rice trade network were negative, the iCTNs exhibit complex mixing patterns.

4 Conclusion

Achieving global food security is one of the major challenges of the coming decades [56], and network analysis has been a popular approach to understanding the international food trade system. In this paper, we focused on four important crops (maize, rice, soybean and wheat), and provided a time series analysis of the four international crop trade networks from 1993 to 2018. Rather than investigating one multiplex trade network *via* combining several goods, we analyzed the international trade networks of individual crops and carried out comparisons. We revealed the evolution of topological properties, including degrees, strengths, link weights, density, clustering coefficient, reciprocity, and assortativity.

We found that the sizes of all the four iCTNs expanded from 1993 to 2018 with more involved international trading participants and larger trade values. The number of links also significantly increased, indicating that many new trade relationships were formed in the global food trade system over the past decades. The link weights decreased sometimes, but showed an increasing trend in general for the four crops. As the networks are directed, we calculated the in-degree, out-degree, in-strength and out-strength to explicitly understand the trade flow in the global food system. The average in- and out-degree increased, representing a larger number of active trade relationships among economies. The increasing trade partnerships, network density, clustering coefficients and similarity coefficients consistently witness the globalization of the international crop trade.

We found that the density of each crop network was low. Over the last 26 years, the density rose with local fluctuations. Our findings are consistent with some other international trade networks [38], but are much smaller [25, 39] than the total world trade networks. The clustering coefficients also showed an upward trend, especially for the rice and soybean trade networks. The structure of the iCTNs become not only tighter but also more similar. In addition, the networks with greater trade flows have more stable relationships. In each iCTN, the reciprocity coefficients were between 0.1 and 0.4, and much smaller than those of the international trade networks. We also obtained some interesting results. For example, although most iCTNs were disassortatively mixed, there were iCTNs exhibiting assortative mixing patterns in certain years, which unveils more complicated mixing behavior than an overall assortative coefficient for the world cereal trade networks [54] or other international trade networks [39, 55].

We compared the structure of four iCTNs to enhance our understanding of the global food system. Although the overall evolution of different iCTNs is similar, some crops have unique trade patterns. For example, the average in-degree of the international wheat trade network decreased in 2011, contrary to other crops. It might be affected by the Russian wheat export ban in 2010–2011, which caused a decrease in the trade flow [57]. The density of the international rice trade network has the largest increase and has become the largest since 1998. The evolution of the clustering coefficients shows that the international rice trade network became more clustered, since the rice exporting and importing concentrated in some economies [42].

Our findings about the topology of the iCTNs play an important role in understanding vulnerabilities in the global food system [11]. These results also highlight the need to consider unique features of different crop networks while making food policies [11]. Since each iCTN has its own structural properties, they are expected to have different reactions to external disturbances and shocks. The global food system is sensitive and easily affected by climate change, water scarcity, and land reclamation [58]. For example, we could assume that an extreme climate decreases the production of crops in some areas which are main global crop suppliers. These economies would cut down crop exports and even implement export bans if their domestic food reserves are insufficient. However, we found that the density of the international rice trade network showed an upward trend during the recent food crisis (e.g., in 2007–2009). As the international rice trade network is increasingly connected, the rice trade tends to concentrate on some regions. A few large producers account for the bulk of net exports and absorb more shocks because of their centrality in the network [59]. These economies are not sensitive to global changes since they have proportionately higher reserves. Therefore the international rice trade network is relatively stable and its structure would not shift dramatically.

In addition to environmental factors, global price shocks also exert a significant influence on the global food system [60], especially for rice, the main staple crop. Many economies rely on rice imports to feed domestic consumption and the rice price hike would put more pressure on importing economies [61], limiting the poor to buying rice [2]. Demand for substitute staple foods increases to soften the impact of rice price shocks [61]. The iCTNs are characterized by substantial heterogeneity across different crops, but crops are traded as complements which indicates that different iCTNs might have a correlation [11]. This paper discussed the global food system as a collection of independent food-staple trade players, and ignored the substitution across crops. However, our findings are still relevant from a policy perspective. As noted above, the similarities and differences between different iCTNs provide more details of the global food trade linkages and address the need to adjust trade policies for different crop importers or exporters. Future research should consider the nonlinear interactions between different iCTNs from the framework of multi-layer networks.

Data availability statement

Publicly available datasets were analyzed in this study, which can be found here: https://www.fao.org.

Author contributions

Funding acquisition, W-XZ; Investigation, Y-TZ; Methodology, Y-TZ and W-XZ; Supervision, W-XZ; Writing—original draft, Y-TZ and W-XZ; Writing—review and editing, Y-TZ, and W-XZ.

Funding

The National Natural Science Foundation of China (72171083), the Shanghai Outstanding Academic Leaders Plan, and the Fundamental Research Funds for the Central Universities.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Farsund AA, Daugbjerg C, Langhelle O. Food security and trade: Reconciling discourses in the food and agriculture organization and the world trade organization. *Food Secur* (2015) 7:383–91. doi:10.1007/s12571-015-0428-y

2. D'Odorico P, Carr JA, Laio F, Ridolfi L, Vandoni S. Feeding humanity through global food trade. *Earth's Future* (2014) 2:458–69. doi:10.1002/2014EF000250

3. Gephart JA, Pace ML. Structure and evolution of the global seafood trade network. *Environ Res Lett* (2015) 10:125014. doi:10.1088/1748-9326/10/12/125014

4. Suweis S, Carr JA, Maritan A, Rinaldo A, D'Odorico P. Resilience and reactivity of global food security. *Proc Natl Acad Sci U S A* (2015) 112:6902–7. doi:10.1073/pnas.1507366112

5. Centeno MA, Nag M, Patterson TS, Shaver A, Windawi AJ. The emergence of global systemic risk. *Annu Rev Sociol* (2015) 41:65–85. doi:10.1146/annurev-soc-073014-112317

6. Wellesley L, Preston F, Lehne J, Bailey R. Chokepoints in global food trade: Assessing the risk. *Res Transportation Business Manage* (2017) 25:15–28. doi:10.1016/j.rtbm.2017.07.007

7. Baldos ULC, Hertel TW. The role of international trade in managing food security risks from climate change. *Food Secur* (2015) 7:275–90. doi:10.1007/s12571-015-0435-z

8. Wood SA, Smith MR, Fanzo J, Remans R, DeFries RS. Trade and the equitability of global food nutrient distribution. *Nat Sustain* (2018) 1:34–7. doi:10.1038/s41893-017-0008-6

9. Almog A, Bird R, Garlaschelli D. Enhanced gravity model of trade: Reconciling macroeconomic and network models. *Front Phys* (2019) 7:55. doi:10.3389/fphy.2019.00055

10. Puma MJ, Bose S, Chon SY, Cook BI. Assessing the evolving fragility of the global food system. *Environ Res Lett* (2015) 10:024007. doi:10.1088/1748-9326/10/2/024007

11. Torreggiani S, Mangioni G, Puma MJ, Fagiolo G. Identifying the community structure of the food-trade international multi-network. *Environ Res Lett* (2018) 13:054026. doi:10.1088/1748-9326/aabf23

12. Dolfing AG, Leuven JRFW, Dermody BJ. The effects of network topology, climate variability and shocks on the evolution and resilience of a food trade network. *PLoS One* (2019) 14:e0213378. doi:10.1371/journal.pone.0213378

13. Wu F, Guclu H. Global maize trade and food security: Implications from a social network model. *Risk Anal* (2013) 33:2168–78. doi:10.1111/risa.12064

14. Fair KR, Bauch CT, Anand M. Dynamics of the global wheat trade network and resilience to shocks. *Sci Rep* (2017) 7:7177. doi:10.1038/s41598-017-07202-y

15. Gutierrez-Moya E, Adenso-Diaz B, Lozano S. Analysis and vulnerability of the international wheat trade network. *Food Secur* (2021) 13:113–28. doi:10.1007/s12571-020-01117-9

16. Suweis S, Konar M, Dalin C, Hanasaki N, Rinaldo A, Rodriguez-Iturbe I. Structure and controls of the global virtual water trade network. *Geophys Res Lett* (2011) 38:L10403. doi:10.1029/2011GL046837

17. Carr JA, D'Odorico P, Laio F, Ridolfi L. On the temporal variability of the virtual water network. *Geophys Res Lett* (2012) 39:L06404. doi:10.1029/2012GL051247

18. Sartori M, Schiavo S. Connected we stand: A network perspective on trade and global food security. *Food Policy* (2015) 57:114–27. doi:10.1016/j.foodpol.2015.10.004

19. Kummu M, Kinnunen P, Lehikoinen E, Porkka M, Queiroz C, Roos E, et al. Interplay of trade and food system resilience: Gains on supply diversity over time at the cost of trade independency. *Glob Food Sec* (2020) 24:100360. doi:10.1016/j.gfs.2020.100360

20. Larochez-Dupraz C, Huchet-Bourdon M. Agricultural support and vulnerability of food security to trade in developing countries. *Food Secur* (2016) 8:1191–206. doi:10.1007/s12571-016-0623-5

21. Burkholz R, Schweitzer F. International crop trade networks: The impact of shocks and cascades. *Environ Res Lett* (2019) 14:114013. doi:10.1088/1748-9326/ab4864

22. Distefano T, Laio F, Ridolfi L, Schiavo S. Correction: Shock transmission in the international food trade network. *PLoS One* (2021) 16:e0254327. doi:10.1371/journal.pone.0254327

23. Zhang YT, Zhou WX. Microstructural characteristics of the weighted and directed international crop trade networks. *Entropy* (2021) 23:1250. doi:10.3390/e23101250

24. Distefano T, Laio F, Ridolfi L, Schiavo S. Shock transmission in the international food trade network. *PLoS One* (2018) 13:e0200639. doi:10.1371/journal.pone.0200639

25. Li JA, Xie WJ, Zhou WX. Structure and evolution of the international pesticide trade networks. *Front Phys* (2021) 9:681788. doi:10.3389/fphy.2021.681788

26. Hadley C, Linzer DA, Belachew T, Mariam AG, Tessema F, Lindstrom D. Household capacities, vulnerabilities and food insecurity: Shifts in food insecurity in urban and rural Ethiopia during the 2008 food crisis. *Soc Sci Med* (2011) 73: 1534–42. doi:10.1016/j.socscimed.2011.09.004

27. Sornette D, Woodard R, Zhou WX. The 2006-2008 oil bubble: Evidence of speculation, and prediction. *Physica A: Stat Mech its Appl* (2009) 388:1571–6. doi:10.1016/j.physa.2009.01.011

28. Nawrotzki RJ, Robson K, Gutilla MJ, Hunter LM, Twine W, Norlund P. Exploring the impact of the 2008 global food crisis on food security among vulnerable households in rural South Africa. *Food Secur* (2014) 6:283–97. doi:10.1007/s12571-014-0336-6

29. Antunes de Araujo FH, Bejan L, Stosic B, Stosic T. An analysis of Brazilian agricultural commodities using permutation - information theory quantifiers: The influence of food crisis. *Chaos Solitons Fractals* (2020) 139:110081. doi:10.1016/j.chaos.2020.110081

30. Zhang S, Wang L, Liu Z, Wang X. Evolution of international trade and investment networks. *Physica A: Stat Mech its Appl* (2016) 462:752–63. doi:10.1016/j.physa.2016.06.117

31. Carr JA, D'Odorico P, Laio F, Ridolfi L. On the temporal variability of the virtual water network. *Geophys Res Lett* (2012) 39:L06404. doi:10.1029/2012GL051247

32. Trostle R. *Global agricultural supply and demand: Factors contributing to the recent increase in food commodity prices*. Washington, D.C.: United States Department of Agriculture (2008).

33. McMichael P. A food regime analysis of the 'world food crisis. *Agric Hum Values* (2009) 26:281–95. doi:10.1007/s10460-009-9218-5

34. Goetz L, Glauben T, Bruemmer B. Wheat export restrictions and domestic market effects in Russia and Ukraine during the food crisis. *Food Policy* (2013) 38: 214–26. doi:10.1016/j.foodpol.2012.12.001

35. Zhao Y, Gao X, An H, Xi X, Sun Q, Jiang M. The effect of the mined cobalt trade dependence network's structure on trade price. *Resour Pol* (2020) 65:101589. doi:10.1016/j.resourpol.2020.101589

36. Fagiolo G. Clustering in complex directed networks. *Phys Rev E* (2007) 76: 026107. doi:10.1103/PhysRevE.76.026107

37. Hou W, Liu H, Wang H, Wu F. Structure and patterns of the international rare earths trade: A complex network analysis. *Resour Pol* (2018) 55:133–42. doi:10.1016/j.resourpol.2017.11.008

38. Cepeda-Lopez F, Gamboa-Estrada F, Leon C, Rincon-Castro H. The evolution of world trade from 1995 to 2014: A network approach. *J Int Trade Econ Dev* (2019) 28:452–85. doi:10.1080/09638199.2018.1549588

39. Fagiolo G, Reyes J, Schiavo S. The evolution of the world trade web: A weighted-network analysis. *J Evol Econ* (2010) 20:479–514. doi:10.1007/s00191-009-0160-x

40. Kou Y, Xian G, Dong C, Ye S, Zhao R. Dynamic evolution research and system implementation of international soybean trade network based on complex network. *Proc 2nd Int Conf Comput Sci Appl Eng* (2018) 2018:3278055. doi:10.1145/3207677.3278055

41. Duenas M, Fagiolo G. Global trade imbalances: A network approach. *Adv Complex Syst* (2014) 17:1450014. doi:10.1142/S0219525914500143

42. Muthayya S, Sugimoto JD, Montgomery S, Maberly GF. An overview of global rice production, supply, trade, and consumption. *Ann N Y Acad Sci* (2014) 1324: 7–14. doi:10.1111/nyas.12540

43. Lv L, Zhang K, Zhang T, Li X, Zhang J, Xue W. Eigenvector centrality measure based on node similarity for multilayer and temporal networks. *IEEE Access* (2019) 7:115725–33. doi:10.1109/ACCESS.2019.2936217

44. Hou L, Liu K. Common neighbour structure and similarity intensity in complex networks. *Phys Lett A* (2017) 381:3377–83. doi:10.1016/j.physleta.2017.08.050

45. Jiang W, Wang Y. Node similarity measure in directed weighted complex network based on node nearest neighbor local network relative weighted entropy. *IEEE Access* (2020) 8:32432–41. doi:10.1109/ACCESS.2020.2971968

46. Zhang RJ, Ye FY. Measuring similarity for clarifying layer difference in multiplex ad hoc duplex information networks. *J Informetr* (2020) 14:100987. doi:10.1016/j.joi.2019.100987

47. Tang J, Scellato S, Musolesi M, Mascolo C, Latora V. Small-world behavior in time-varying graphs. *Phys Rev E* (2010) 81:055101. doi:10.1103/PhysRevE.81.055101

48. Gunes I, Gunduz-Oguducu S, Cataltepe Z. Link prediction using time series of neighborhood-based node similarity scores. *Data Min Knowl Discov* (2016) 30: 147–80. doi:10.1007/s10618-015-0407-0

49. Fan X, Li X, Yin J, Tian L, Liang J. Similarity and heterogeneity of price dynamics across China's regional carbon markets: A visibility graph network approach. *Appl Energ* (2019) 235:739–46. doi:10.1016/j.apenergy.2018.11.007

50. Squartini T, Picciolo F, Ruzzenenti F, Garlaschelli D. Reciprocity of weighted networks. *Sci Rep* (2013) 3:2729. doi:10.1038/srep02729

51. Serrano MA, Boguñá M. Topology of the world trade web. *Phys Rev E* (2003) 68:015101(R). doi:10.1103/PhysRevE.68.015101

52. Garlaschelli D, Loffredo M. Patterns of link reciprocity in directed networks. *Phys Rev Lett* (2004) 93:268701. doi:10.1103/PhysRevLett.93.268701

53. Mou N, Fang Y, Yang T, Zhang L. Assortative analysis of bulk trade complex network on maritime silk road. *IEEE Access* (2020) 8:131928–38. doi:10.1109/ACCESS.2020.3009970

54. Dupas MC, Halloy J, Chatzimpiros P. Time dynamics and invariant subnetwork structures in the world cereals trade network. *PLoS One* (2019) 14: e0216318. doi:10.1371/journal.pone.0216318

55. Fagiolo G, Reyes J, Schiavo S. On the topological properties of the world trade web: A weighted network analysis. *Physica A: Stat Mech its Appl* 387 (2008) 3868–73. doi:10.1016/j.physa.2008.01.050

56. Porkka M, Kummu M, Siebert S, Varis O. From food insufficiency towards trade dependency: A historical analysis of global food availability. *PLoS One* (2013) 8:e82714. doi:10.1371/journal.pone.0082714

57. Svanidze M, Gotz L, Serebrennikov D. The influence of Russia's 2010/2011 wheat export ban on spatial market integration and transaction costs of grain markets. *Appl Econ Perspect Pol* (2021) 44:1083–99. doi:10.1002/aepp.13168

58. Premanandh J. Factors affecting food security and contribution of modern technologies in food sustainability. *J Sci Food Agric* (2011) 91:2707–14. doi:10.1002/jsfa.4666

59. Marchand P, Carr JA, Dell'Angelo J, Fader M, Gephart JA, Kummu M, et al. Reserves and trade jointly determine exposure to food supply shocks. *Environ Res Lett* (2016) 11:095009. doi:10.1088/1748-9326/11/9/095009

60. Baffes J, Kshirsagar V. Shocks to food market systems: A network approach. *Agric Econ* (2020) 51:111–29. doi:10.1111/agec.12544

61. Haggblade S, Me-Nsope NM, Staatz JM. Food security implications of staple food substitution in Sahelian West Africa. *Food Policy* (2017) 71:27–38. doi:10.1016/j.foodpol.2017.06.003

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fphy.2022.1017009

Check for updates

OPEN ACCESS

EDITED BY
Jianguo Liu,
Shanghai University of Finance and Economics, China

REVIEWED BY
Ou Yang,
University of Shanghai for Science and Technology, China
Yong Luo,
University of Shanghai for Science and Technology, China

*CORRESPONDENCE
Lei Hou,
l.hou@nuist.edu.cn

SPECIALTY SECTION
This article was submitted to Social Physics,
a section of the journal
Frontiers in Physics

RECEIVED 11 August 2022
ACCEPTED 24 August 2022
PUBLISHED 15 September 2022

CITATION
Hou L and Pan Y (2022), Evaluating the connectedness of commodity future markets via the cross-correlation network.
Front. Phys. 10:1017009.
doi: 10.3389/fphy.2022.1017009

Evaluating the connectedness of commodity future markets via the cross-correlation network

Lei Hou* and Yueling Pan

School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing, China

Financial markets are widely believed to be complex systems where interdependencies exist among individual entities in the system enabling the risk spillover effect. The detrended cross-correlation analysis (DCCA) has found wide applications in examining the comovement of fluctuations among financial time series. However, to what extent can such cross-correlation represent the spillover effect is still unknown. This article constructs the DCCA network of commodity future markets and explores its proximity to the volatility spillover network. Results show a moderate agreement between the two networks. Centrality measures applied to the DCCA networks are able to identify key commodity futures that are transmitting or receiving risk spillovers. The evolution of the DCCA network reveals a significant change in the network structure during the COVID-19 pandemic in comparison to that of the pre- and post-pandemic periods. The pandemic made the commodity future markets more interconnected leading to a shorter diameter for the network. The intensified connections happen mostly between commodities from different categories. Accordingly, cross-category risk spillovers are more likely to happen during the pandemic. The analysis enriches the applications of the DCCA approach and provides useful insights into understanding the risk dynamics in commodity future markets.

KEYWORDS

cross-correlation, commodity network, volatility spillover, connectedness, centrality

Introduction

Financial markets play a critical role in economic development but have been severely threatened by a wide range of socio-economic events in recent decades, such as the subprime mortgage crisis in 2008, the US–China trade war, and the COVID-19 pandemic [1–3]. Rich and in-depth investigations into these events have demonstrated that the risks not only influence each individual entity in the system but also spread among the entities and evolve into system-wide crises. In other words, the entities in a financial market are interdependent on each other, forming a complex networked system that enables the contagion of risks through the interdependencies [4–6]. Financial systems are thus normally modeled as networks, such as the networks of financial institutions [7, 8], the network of stock indices [9], and the network of commodity futures [10]. The key technique for the construction of a financial network is the quantification of the interdependencies among individual entities. However, for many financial systems (e.g., stock market and future market), such interdependencies between entities (e.g., stock indices and commodity futures) cannot be directly observed. To quantify the pair-wise and system-wise connectedness is accordingly vital to the understanding of the dynamics of risk contagion in financial systems.

In the literature on the volatility spillover effect, connectedness is normally explored as the extent to which a shock in one entity's time series (e.g., stock price and return) could lead to changes in other entities [11, 12]. Techniques based on vector auto-regression (VAR) are widely applied to study such a problem, and various measures have been accordingly developed. One of the most acknowledged metric frameworks is proposed by Diebold and Yilmaz [13–15]. Instead of studying the spillover effect from one time series to another, they apply variance decomposition to an N-variable VAR. Accordingly, the share of the forecast error variation for a target time series from each of the other time series in the system can be quantified simultaneously. The pair-wise spillover effect is thus directly measured by the results of variance decomposition. Since such a spillover effect is regarded as directional, the ability of an entity to transmit risks can be quantified by totaling the spillover effect from it to all others (out-degree), while the extent of an entity being influenced by others can be quantified by totaling the spillover effect received by the entity (in-degree). Such a method and its variations have been applied to construct and analyze a wide range of financial networks. For example, Yang and Zhou constructed a time-varying volatility spillover network of countries according to the VIX of several major national stock market indices and uncovered the central role of the US market [16]. The spillover effect from the US market to others has intensified since the 2008 global financial crisis. Balcilar et al. investigated the spillover effect among the prices of agricultural futures and crude oil futures and identified two sets of commodity futures to be risk transmitters and risk receivers, respectively [17]. Shen et al. explored the connectedness of different economic sectors in China and found that the sectors such as mechanical equipment act as risk transmitters, while sectors such as banking are the main risk takers [18]. Overall, the variance decomposition framework based on the VAR model has shown effectiveness in representing the volatility spillover effect in financial systems.

Given the nature of financial markets as complex systems, the interdependencies among financial entities have also caught widespread attention in the field of econophysics and complexity science. The detrended cross-correlation analysis (DCCA) [19, 20] has been the most acknowledged and applied technique in the analysis of cross-correlations between financial time series, such as commodity future prices [21, 22] and stock trading volumes or prices [23, 24]. Since there could potentially be cross-correlations between any two financial time series, financial markets can thus be linked into networks [25–27]. The analysis of DCCA networks also has the potential to measure the importance of each individual entity in the whole system. For example, Pereira et al. applied centrality measures of weighted degree and PageRank to the DCCA network of 20 regional stock markets and concluded that European markets play a central role in the world's financial markets [28]. Mbatha and Alovokpinhou constructed the network of 134 companies from the South African stock market and found that the financial industry plays the most prominent role [29].

When the VAR-based methods characterize the directional relationship that a shock in one time series leads to the volatility change in another time series within a given lag time, the DCCA approach describes the bilateral relationship of co-fluctuation of two time series. In spite of the widespread applications of the DCCA approach in investigating the dynamics of financial networks [8, 27–32], whether, or to what extent, can such an approach represent the volatility spillover effect as indicated by the VAR-based measures is still unclear. The exploration of such a research question is crucial to deepen the understanding of the dynamics of complex financial systems, as well as enrich the application of the DCCA approach.

Focusing on the commodity future market, this article applies both the VAR-based volatility spillover measures and the DCCA coefficient to construct networks of the 19 commodities. Two research questions are thereby explored: 1) to what extent can the DCCA network depict the volatility spillover effect among commodity futures; and 2) how is the DCCA network of commodity futures evolving over time. Centrality measures are applied to the DCCA network, which are found with high effectiveness to identify the key risk takers, while moderate effectiveness to uncover key risk transmitters. Further dynamical analysis of the DCCA network reveals the dramatic impact of COVID-19 on the topology of the DCCA network with intensified cross-category risk spillovers.

Materials and methods

Detrended cross-correlation analysis

The fluctuation of a wide range of real-world time series is found with strong scaling behavior, and the detrended fluctuation analysis (DFA) is proposed to analyze such a phenomenon [33, 34]. Given a time series x_t, $t = 1, \cdots, N$, its profile time series is thus $X(t) = \sum_{k=1}^{t} (x_k - \bar{x})$, where $\bar{x}$ is the mean value of the original time series. To assess the local trends, the profile time series is further divided into small intervals with an equal size of s. Accordingly, this results in $N_s = int(N/s)$ intervals. The local trend of each interval can be quantified by applying an ordinary least square regression, resulting in a fitted time series $X_f(t)$. The detrended fluctuations can thus be represented as a new time series by subtracting the local trend from the profile time series, i.e., $X(t) - X_f(t)$. The detrended fluctuation of the original time series x_t can be written as a function of the window size s, which reads

$$F_{DFA}(x,s) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[X(t) - X_f(t)\right]^2}. \tag{1}$$

Normally, the relationship between the detrended fluctuation and the window size follows a power-law, that is, $F_{DFA}(x,s) \propto s^{\alpha}$. The scaling exponent α is normally used to describe the long-range auto-correlation of the time series.

While DFA deals with the fluctuations of a single time series, the DCCA approach is proposed to investigate the co-fluctuation of two time series [19, 20]. For the scenario of two time series, say, x_t and y_t, the same process in DFA can be applied to each of the time series, to obtain the profile time series $X(t)$ and $Y(t)$, and the fitted local trends $X_f(t)$ and $Y_f(t)$. Instead of the fluctuation of a single time series, the co-fluctuation of the two time series can be accordingly calculated as

$$F^2_{DCCA}(xy,s) = \frac{1}{N}\sum_{i=1}^{N}\left[X(t) - X_f(t)\right]\left[Y(t) - Y_f(t)\right]. \tag{2}$$

Similar to DFA, the co-fluctuation of the two time series is also expected to follow a power-law relationship with the window size, i.e., $F_{DCCA}(xy,s) \propto s^{\alpha}$. If the scaling exponent α takes a nonzero value, a long-range cross-correlation can be concluded between the time series. To obtain a more generalized value to capture the cross-correlation between the time series, the DCCA coefficient can be defined as

$$\rho_{DCCA}(xy,s) = \frac{F^2_{DCCA}(xy,s)}{F_{DFA}(x,s)\cdot F_{DFA}(y,s)}. \tag{3}$$

The DCCA coefficient ρ_{DCCA} takes values ranging from -1 to 1, with -1 indicating the perfect anti-cross-correlation, 1 indicating the perfect cross-correlation, and 0 indicating no cross-correlation. When the window size s is a free parameter, we set $s = 16$ throughout the following analysis.

Measures for the volatility spillover

While a number of measures have been proposed to characterize the connectedness and volatility spillover effect in financial systems, this article adopts a widely used approach developed by Diebold and Yilmaz [13–15].

Considering a set of K variables (time series) with $V_t = (v_{1,t}, v_{2,t}, \cdots, v_{K,t})'$ as the vector of variables at a time, each time series is thus $V(k) = \{v_{k,t}\}$, $t = 1, 2, \cdots, N$. The system can be described by a K-variable VAR model as $V_t = \Theta_1 V_{t-1} + \Theta_2 V_{t-2} + \cdots + \Theta_l V_{t-l} + \epsilon_t$, where $\Theta_1, \cdots \Theta_l$ are parameter matrices, ϵ_t is the vector of white noise, and l is the time lag. In other words, each variable is modeled as a function of the l lags of its own as well as all the other variables in the system. A moving average representation for the model can be given by

$$V_t = \sum_{i=0}^{\infty} A_i \epsilon_{t-i}, \tag{4}$$

where $A_i = \Theta_1 A_{i-1} + \Theta_2 A_{i-2} + \cdots + \Theta_l A_{i-l}$ is a $K \times K$ coefficient matrix with A_0 as an identity matrix and $A_i = 0$ for $i < 0$. According to the moving average representation, H-step-ahead forecast error variance decomposition can be calculated and denoted as $\Pi^H = [\pi^H_{ij}]$, where $H = 1, 2, \cdots$ is the predictive horizon. The element π^H_{ij} depicts the fraction of time series $V(i)$'s forecast error variance caused by the shock in time series $V(j)$, which can be written as

$$\pi^H_{ij} = \psi^{-1}_{jj}\frac{\sum_{h=0}^{H-1}\left(e'_i A_h \Psi e_j\right)}{\sum_{h=0}^{H-1}\left(e'_i A_h \Psi A'_h e_j\right)}, \tag{5}$$

where Ψ is the covariance matrix for the vector of errors ϵ, ψ^{-1}_{jj} is the jth diagonal element in matrix Ψ, and e_i is a vector with only the ith value being 1 while others being 0. The value π^H_{ij} can thus be used to quantify the spillover effect from a shock in time series $V(j)$ to time series $V(i)$, i.e., the directional connectedness from j to i. Such a value can be further normalized as

$$\tilde{\pi}^H_{ij} = \frac{\pi^H_{ij}}{\sum_{k=1}^{K}\pi^H_{ik}}. \tag{6}$$

Accordingly, for each time series $V(i)$, the summation of the connectedness from other time series equals 1, i.e., $\sum_j \tilde{\pi}^H_{ij} = 1, \forall i$.With the directional connectedness defined, the K time series can be linked as a directed volatility spillover network, where each node is a time series (a financial entity) and each weighted and directed link describes the relative intensity of the spillover effect. Diebold and Yilmaz further defined several measures for node-level connectedness, including to-connectedness and from-connectedness [13–15]. The to-connectedness is defined as

$$C^{to}_i = \sum_{j=1, j\neq i}^{K} \tilde{\pi}^H_{ij}, \tag{7}$$

which corresponds to the out-degree of i in the spillover network describing the total spillover effect transmitted by i to others. Similarly, the from-connectedness is defined as

$$C^{from}_i = \sum_{j=1, j\neq i}^{K} \tilde{\pi}^H_{ji}, \tag{8}$$

which is basically the in-degree of i in the spillover network describing the total spillovers received by i.

Throughout the analysis, we set the predictive horizon to $H = 5$, i.e., the volatility spillover effects are calculated based on the 5-step-ahead forecast error.

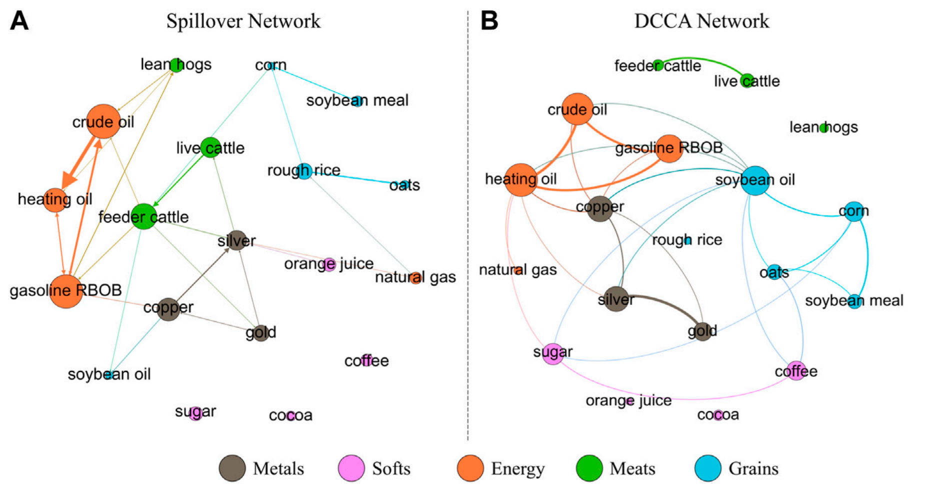


FIGURE 1
Volatility spillover network **(A)** and DCCA network **(B)** of 19 commodity futures. The links in the spillover network are directed, and only those with a weight of $\tilde{\pi}_{ij} > 0.15$ are displayed. The links in the DCCA network are undirected, and all the links with $w_{ij} > 0.2$ are displayed. The node size in both networks is proportional to the degree (out-degree for the spillover network).

TABLE 1 Pearson correlation coefficients between the centrality measures in the DCCA network and spillover effects, as measured by to-connectedness C_i^{to} and from-connectedness C_i^{from}, respectively.

Centrality measure	To-connectedness		From-connectedness	
	Correlation	*p*-value	Correlation	*p*-value
Degree	0.489	0.034	0.701	0.0008
Eigenvector	0.549	0.015	0.725	0.0004
Closeness	0.225	0.354	0.533	0.0189
PageRank	0.479	0.038	0.695	0.0009

Data collection

The future market has been one of the major financial systems that attracted widespread attention in the literature, where strong spillover effects have been frequently uncovered [34, 35]. Meanwhile, the DCCA approach has also found applications in characterizing the cross-correlation among different future markets [21, 22, 36]. The present study thereby adopts the future market as the detailed context to explore the proximity of the DCCA network to the volatility spillover effect and the dynamics of the commodity future network.

Given the purpose of the present analysis, we mainly focus on the commodity contracts in the US market. The various commodities can be divided into five major categories, namely, metals, softs, energy, meats, and grain. While there are normally many commodity futures in each category, here we only consider the commodity futures that are most traded for each category. To be more specific, gold, copper, and silver are selected for metal future contracts; coffee, sugar, orange juice, and cocoa are selected for soft crop future contracts; crude oil, natural gas, heating oil, and gasoline are selected for energy future contracts; live cattle, lean hogs, and feeder cattle are selected for meat future contracts; and rough rice, soybean oil, soybean meal, corn, and oats are selected for grain future contracts. The detailed data were downloaded from Thomson Reuters Datastream, which is a live database for various financial systems. Our data span 9 years, from 1 January 2013 to 31 December 2021. For each trading day, we collect the open, high, low, and close indexes. In other words, the time series to be analyzed are the 9-year-long daily prices of 19 commodity futures.

TABLE 2 Top five commodity future markets with the highest values for to-connectedness C_i^{to}, from-connectedness C_i^{from}, degree centrality DC_i, eigenvector centrality EC_i, closeness centrality CC_i, and PageRank centrality RC_i.

Measure	1	2	3	4	5
To-connectedness	Heating oil	Crude oil	Gasoline	Feeder cattle	Silver
From-connectedness	Crude oil	Gasoline	Feeder cattle	Heating oil	Copper
Degree	Heating oil	Crude oil	Soybean oil	Gasoline	Copper
Eigenvector	Heating oil	Crude oil	Gasoline	Soybean oil	Copper
Closeness	Soybean oil	Heating oil	Copper	Crude oil	Corn
PageRank	Heating oil	Crude oil	Soybean oil	Gasoline	Silver

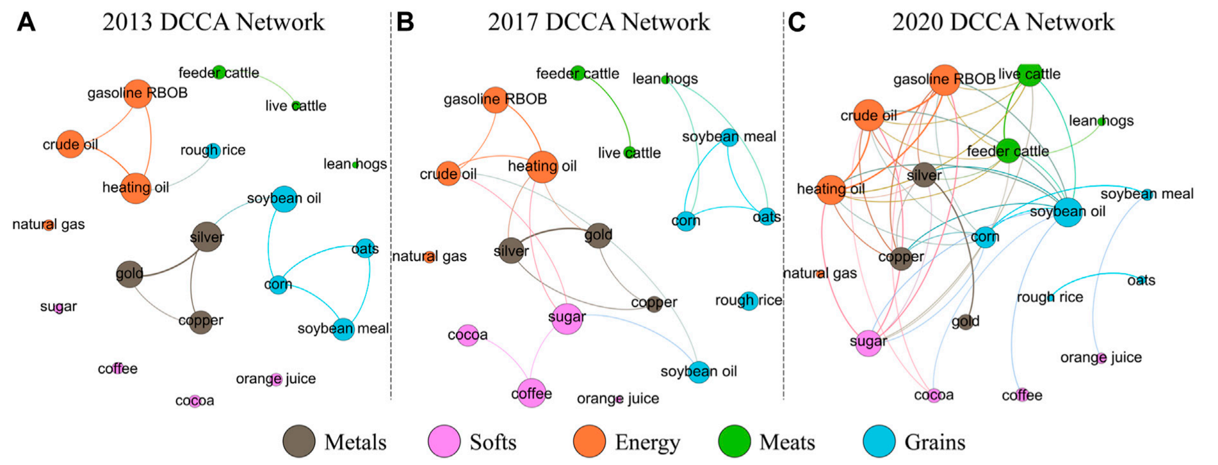


FIGURE 2
DCCA networks of the commodity futures markets in 2013 **(A)**, 2017 **(B)**, and 2020 **(C)** respectively. Only links with a weight larger than 0.35 are displayed, and the node size is proportional to the degree

Results

Static analysis

We first construct and analyze the volatility spillover network and DCCA network using the full 9-year data. Both networks consist of 19 nodes, with each being one commodity future market. The networks are fully connected with different weights on links. The weight of a link in the DCCA network is the absolute value of the cross-correlation coefficient $w_{ij} = |\rho_{DCCA}(ij)|$ between two commodity futures' time series of daily close price, c_t. In other words, we consider the intensity of the cross-correlation, regardless of its direction. For the volatility spillover network, we first calculate the close-to-close volatility of commodity future price in each week t as $\sigma_t = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(r_i - \bar{r})^2}$, where T is the trading days in the week, and $r_i = \log\ (c_i/c_{i-1})$ is the return of the ith day in the week. The variance decomposition is applied to the close-to-close volatility of commodity future prices. The pair-wise connectedness value $\tilde{\pi}_{ij}$, as calculated by Eq. 6, is thus regarded as the weight for the link from commodity i to commodity j.

As shown in Figure 1, the spillover network and DCCA network of the 19 commodity futures show similar structures, in spite of the fact that the former is directed while the latter is undirected. Energy futures of crude oil, heating oil, and gasoline form a strongly connected triad in both networks. The metal futures of copper, silver, and gold are also closely interconnected. On the other hand, the soft futures, including orange juice, sugar, cocoa, and coffee, are loosely connected to others in either the spillover network or the DCCA network. To get a more generalized quantification of the similarity between the cross-correlation and spillover effect, we calculate the Pearson correlation coefficient between the values of w_{ij} and $\tilde{\pi}_{ij}$. The analysis shows that the weights on the matched links from two networks have a correlation of 0.511 ($p = 4.107 \times 10^{-24}$), indicating a moderate positive correlation. Thus, the DCCA coefficient between the future prices of two commodities can, to a moderate degree, depict the directed volatility spillover effect. Despite the different underlying logics,

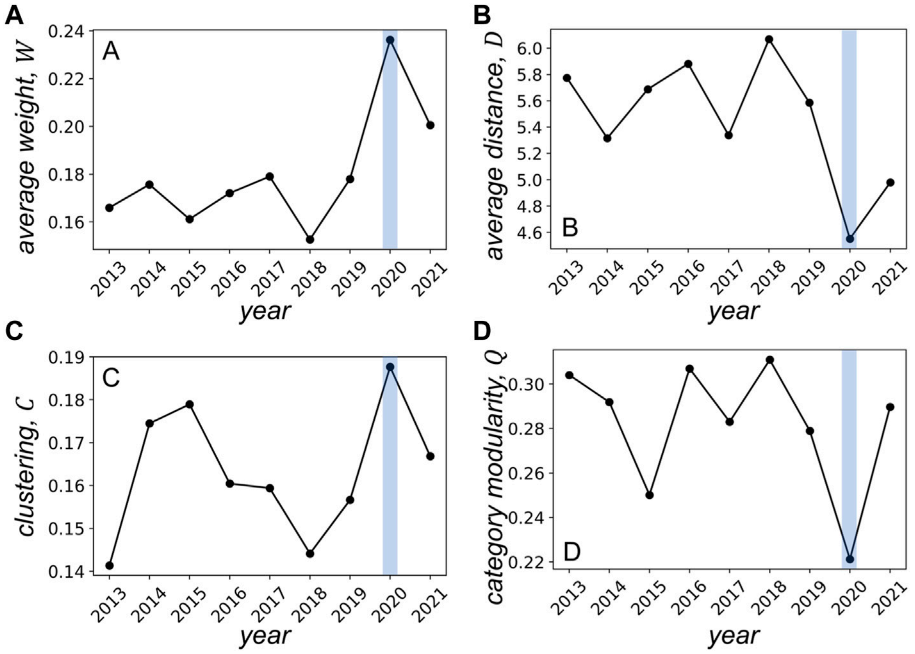


FIGURE 3
Average weight **(A)**, average distance **(B)**, clustering coefficient **(C)**, and category modularity **(D)** of the yearly DCCA network of commodity future markets.

the two measures, namely, the DCCA coefficient and volatility spillover, depict the relationship between two time series' fluctuations. Accordingly, if the volatility of one time series largely influences that of another time series (large value for the spillover), the two time series would tend to co-fluctuate regardless of the time lag, resulting in a strong cross-correlation. However, the volatility spillover is directed and considers not only the two time series but also all the time series in the system. As a consequence, the correlation between the DCCA coefficient and volatility spillover is only moderate but significant.

The volatility spillovers are actually directed, and thus the risk transmitters and risk receivers can be identified by the spillover network via the measures of to-connectedness (out-degree) and from-connectedness (in-degree), respectively. However, the DCCA coefficient is bilateral with no direction. An apparent question is whether the DCCA network of commodities can help to identify the key risk transmitters and risk receivers. Here, we further apply four basic centrality measures to the DCCA network to examine the accuracy of predicting the risk transmitters and risk receivers.

Since the links in the DCCA network are weighted, the degree centrality of a commodity i is thus $DC_i = \sum_{j\neq i} w_{ij}$. The eigenvector centrality not only considers the number of neighbors of a node but also evaluates the importance of the neighbors. Thus, the eigenvector centrality of a commodity i can be calculated as the weighted average of the centrality values of its neighbors, i.e., $EC_i = \frac{1}{\lambda} \sum_{j\neq i} w_{ij} \cdot EC_j$, where λ is the largest eigenvalue of the adjacency matrix $P = \{w_{ij}\}$. The closeness centrality of a commodity is the average value of its shortest distance to each other commodity. While the DCCA network is weighted, the length of a link is assumed to be the reciprocal of the weight, i.e., $1/w_{ij}$. The distance between two commodities i and j, denoted with d_{ij}, is thus the summation of length for the shortest path connecting i and j, which has the minimal value. Note that, although the network is fully connected, the shortest path is not necessarily the direct link connecting the two commodities. Accordingly, the closeness centrality for i can be calculated as $CC_i = (K - 1)/ \sum_{j\neq i} d_{ij}$, where K is the number of commodities in the DCCA network. The PageRank centrality also assumes a node's importance to be determined by its neighbors. The centrality value can be achieved via an iterative process. At the initial step, each node has a centrality value $PC_i(t = 0) = 1$. For each following step, the centrality value updates as $PC_i(t) = \sum_{j\neq i} w_{ij} \cdot \frac{PC_j(t-1)}{DC_j}$. The eventually stabilized values are then regarded as the PageRank centrality of a commodity node.

We apply the four centrality measures to the constructed DCCA network of commodity future markets to calculate the centralities for each commodity. To test the ability of these centrality measures in identifying the risk transmitters (to-

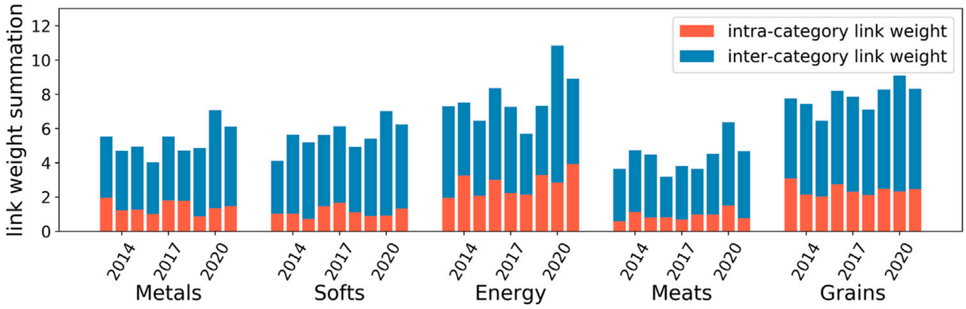


FIGURE 4
Stacked bar plot for yearly weight summation of intra- and inter-category links.

connectedness, as defined in Eq. 7) and risk receivers (from-connectedness, as defined in Eq. 8), we calculate the Pearson correlation coefficients which are reported in Table 1. For the to-connectedness, i.e., the spillovers transmitted by a commodity to others, the centrality measures of degree, eigenvector, and PageRank show moderate accuracies with correlations ranging from 0.479 to 0.549. However, the closeness centrality has a very low correlation of 0.225 to the to-connectedness of commodities. For the from-connectedness, i.e., the spillovers received by a commodity, these centrality measures show higher accuracies. In other words, the centrality measures, including degree, eigenvector, and PageRank, applied to the DCCA network are strongly correlated to the from-connectedness while moderately correlated to the to-connectedness.

We also compare the most important top five commodities as identified by different measures, as shown in Table 2. According to the volatility spillover effect, the energy futures, including heating oil, crude oil, and gasoline, are the key risk transmitters and at the same time risk receivers. These commodities are also identified by degree centrality, eigenvector centrality, and PageRank centrality as the most influential node in the DCCA network. However, differences between the spillover network and the DCCA network can also be observed. While feeder cattle are also an important risk transmitter and risk receiver, centrality measures in DCCA failed to uncover such an important role. In contrast, soybean oil is evaluated to be an imported commodity in the DCCA network, but it does not transmit nor receive much spillover effect. Despite the different focuses on the two approaches, the DCCA network can be used to identify the key risk transmitters and risk receivers with moderate accuracy.

Dynamics of the DCCA network

We further analyze how the DCCA network of commodity future markets has evolved over the past 9 years by constructing a DCCA network for each year. Figure 2 visualizes the DCCA network for 2013, 2017, and 2020, respectively. Intuitively, the cross-correlations among commodities are becoming stronger, and thus the DCCA network gets more connected over the years.

To quantitatively explore the dynamics of the network, we focus on four structural features, namely, the average weight, average distance, clustering coefficient, and category modularity. The average weight of the DCCA network is calculated as $W = \langle w_{ij} \rangle$, which describes the connectedness of the network. The average distance of the DCCA network is calculated as $D = \langle d_{ij} \rangle$, measuring how easy it is for the nodes to reach each other. The clustering coefficient of a node describes how strongly its neighbors are connected to each other. Following the definition proposed by Saramäki et al. [37], the clustering coefficient is calculated as $clustering_i = \frac{1}{K(K-1)} \sum_{j,k,j\neq k} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3}$, where $\hat{w}_{ij} = w_{ij}/\max\ (w)$. The clustering coefficient of the DCCA network is averaged over that of every node, i.e., $C = \langle clustering_i \rangle$. Since the 19 commodities considered in the present study come from five different categories, we measure the extent to which the links connect commodities within the same category. Following the modularity measure proposed for the community structure in networks [38], we define the category modularity in the commodity network as $Q = \frac{\sum_{ij} a_{ij} w_{ij}}{\sum_{ij} w_{ij}}$, where $a_{ij} = 1$ if the two commodities i and j subject to the same category, and $a_{ij} = 0$ otherwise.

As shown in Figure 3, the connectedness of the DCCA network, i.e., the average weight, has remained at a relatively stable level ranging from 0.15 to 0.18 during the period from 2013 to 2019. However, the connectedness dramatically increased to 0.236 in 2020. Such a result indicates that the COVID pandemic that broke out at the end of 2019 significantly affected the commodity future markets, making them more strongly interconnected. Due to the intensified connections among the commodities, the average distance of the DCCA network decreased, meaning that it becomes easier for risks to spread from one commodity future market to another. Meanwhile, the clustering coefficient largely increased in 2020, indicating that strong triadic cross-correlations are formed under the impact of the pandemic. The overall category modularity saw a dramatic decrease in 2020, that is, the ratio of intra-category links over all links has decreased.

To have a closer examination of the dynamics of connection patterns in the DCCA network, we investigate how the intra-category links and inter-category links for each category of commodities are evolving. For each category of commodities c, we compare the intra-category link weight summation $S_c^{intra} = \sum_{i\in\Gamma_c, j\in\Gamma_c, i\neq j} w_{ij}$, where Γ_c is the set of commodities of category c and the inter-category link weight summation $S_c^{inter} = \sum_{i\in\Gamma_c, j\notin\Gamma_c} w_{ij}$, which are reported in Figure 4. In addition the differences among different categories, the evolutions of intra- and inter-category links also show different patterns. Despite the fluctuations, the intra-category link weight summation S_c^{intra} has remained at a stable level for each category. Even in 2020, there is no significant change in the value of S_c^{intra}. On the other hand, the inter-category link weight summation S_c^{inter} increased in 2020, especially for the category of energy and meat. As such, the increase in average cross-correlations, reported in Figure 3A, majorly comes from the inter-category links. This is also the reason for the decrease in category modularity.

Despite the dramatic impact the pandemic has made on the connectedness of the DCCA network of commodity future in 2020, such impact does not maintain. As reported in Figure 3, all the network features recovered, to some extent, from the pandemic's impact in 2021, especially for the clustering (Figure 3C) and category modularity (Figure 3D), the 2021 network shows very similar values as compared to the pre-pandemic networks. The average weight (Figure 3A) and average distance (Figure 3B) of the 2021 network are also not as dramatic as that of 2020. Such recovery of the network structure is partially because of the ease of the pandemic situation in 2021 and also indicates that the extreme external events normally would only make a temporary impact on financial markets.

Conclusion and discussion

Risk spreading in complex financial systems has been widely acknowledged to be central to the understanding of the system dynamics. Different streams of research have developed various approaches to construct networks of financial systems, including the VAR-based approach which measures the extent to which the shock in one financial market influences another with a given time lag, and the DCCA-based approach which measures the comovement of fluctuations between two financial time series. The present article offers a comparison between the networks of commodity future markets constructed by such two streams of approach. The cross-correlation is found with moderate proximity to the spillover network. The centrality measures applied to the DCCA network, including degree, eigenvector, and PageRank, are able to identify risk transmitters and risk receivers. The results indicate the effectiveness of the DCCA network in characterizing the structure of the volatility spillover effect. The cross-correlations among financial time series can thus also serve as an important approach for investors to monitor the risks in financial systems and develop appropriate investment strategies accordingly. However, the DCCA network is not always accurate. For example, soybean oil is identified by the DCCA network as one of the most important commodity future markets, but it is not a key risk transmitter nor a risk receiver. Thus, the difference between the cross-correlation and volatility spillover effect should be considered in the application of DCCA when investigating risk dynamics in financial systems.

The COVID-19 pandemic is revealed to be influential on the connectedness of the commodity future markets. The DCCA network of 2020 is found with stronger average cross-correlations, shorter average distance, and stronger clustering features. In particular, it is found that the cross-correlations between commodities from the same category did not change much, while that between commodities from different categories have become stronger in 2020. Such a result suggests a higher risk of cross-category spillover during the pandemic. This observation is in line with previous findings that financial systems tend to have stronger connectedness during a wide range of extreme external events such as financial crises and pandemics [16, 17]. Thus, investors should be cautious about the intensified risk contagions among commodity future markets during extreme events, especially the cross-category risk spillovers. An interesting observation in this article is that the average degree, average distance, clustering, and category modularity in the 2021 network began to recover to almost the level of pre-pandemic. However, due to the limited time range of the applied data, the present article is unable to track the recovery dynamics of the network of commodity future markets. Future research shall further explore the mechanism and timeliness of the recovery process of financial networks after dramatic structural changes caused by external events.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

Both authors contributed to the study design, data collection, data analysis, and writing.

Funding

The study was partially supported by the Social Science Fund of Jiangsu Province (Grant No. 21TQC005) and the Social

Science Foundation of the Jiangsu Higher Education Institutions (Grant No. 2021SJA0164).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, Vespignani A, White DR. Economic networks: The new challenges. *Science* (2009) 325(5939):422–5. doi:10.1126/science.1173644

2. Zhang D, Hu M, Ji Q. Financial markets under the global pandemic of COVID-19. *Financ Res Lett* (2020) 36:101528. doi:10.1016/j.frl.2020.101528

3. An J, Mikhaylov A, Richter UH. Trade war effects: Evidence from sectors of energy and resources in Africa. *Heliyon* (2020) 6(12):e05693. doi:10.1016/j.heliyon.2020.e05693

4. Allen F, Gale D. Financial contagion. *J Polit Econ* (2000) 108(1):1–33. doi:10.1086/262109

5. Gai P, Kapadia S. Contagion in financial networks. *Proc R Soc A* (2010) 466(2120):2401–23. doi:10.1098/rspa.2009.0410

6. Jackson MO, Pernoud A. Systemic risk in financial networks: A survey. *Annu Rev Econ* (2021) 13:171–202. doi:10.1146/annurev-economics-083120-111540

7. Elyasiani E, Kalotychou E, Staikouras SK, Zhao G. Return and volatility spillover among banks and insurers: Evidence from pre-crisis and crisis periods. *J Financ Serv Res* (2015) 48(1):21–52. doi:10.1007/s10693-014-0200-z

8. Wang GJ, Yi S, Xie C, Stanley HE. Multilayer information spillover networks: measuring interconnectedness of financial institutions. *Quant Finance* (2021) 21(7):1163–85. doi:10.1080/14697688.2020.1831047

9. Heiberger RH. Stock network stability in times of crisis. *Physica A: Stat Mech its Appl* (2014) 393:376–81. doi:10.1016/j.physa.2013.08.053

10. Xiao B, Yu H, Fang L, Ding S. Estimating the connectedness of commodity futures using a network approach. *J Futures Markets* (2020) 40(4):598–616. doi:10.1002/fut.22086

11. Akkoc U, Civcir I. Dynamic linkages between strategic commodities and stock market in Turkey: Evidence from SVAR-DCC-GARCH model. *Resour Pol* (2019) 62:231–9. doi:10.1016/j.resourpol.2019.03.017

12. Jung RC, Maderitsch R. Structural breaks in volatility spillovers between international financial markets: Contagion or mere interdependence? *J Banking Finance* (2014) 47:331–42. doi:10.1016/j.jbankfin.2013.12.023

13. Diebold FX, Yilmaz K. Measuring financial asset return and volatility spillovers, with application to global equity markets. *Econ J* (2009) 119(534):158–71. doi:10.1111/j.1468-0297.2008.02208.x

14. Diebold FX, Yilmaz K. Better to give than to receive: predictive directional measurement of volatility spillovers. *Int J Forecast* (2012) 28(1):57–66. doi:10.1016/j.ijforecast.2011.02.006

15. Diebold FX, Yılmaz K. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *J Econom* (2014) 182(1):119–34. doi:10.1016/j.jeconom.2014.04.012

16. Yang Z, Zhou Y. Quantitative easing and volatility spillovers across countries and asset classes. *Manage Sci* (2017) 63(2):333–54. doi:10.1287/mnsc.2015.2305

17. Balcilar M, Gabauer D, Umar Z. Crude Oil futures contracts and commodity markets: New evidence from a TVP-VAR extended joint connectedness approach. *Resour Pol* (2021) 73:102219. doi:10.1016/j.resourpol.2021.102219

18. Shen YY, Jiang ZQ, Ma JC, Wang GJ, Zhou WX. Sector connectedness in the Chinese stock markets. *Empir Econ* (2022) 62(2):825–52. doi:10.1007/s00181-021-02036-0

19. Podobnik B, Stanley HE. Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. *Phys Rev Lett* (2008) 100(8):084102. doi:10.1103/physrevlett.100.084102

20. Zebende GF. DCCA cross-correlation coefficient: Quantifying level of cross-correlation. *Physica A: Stat Mech its Appl* (2011) 390(4):614–8. doi:10.1016/j.physa.2010.10.022

21. Liu L. Cross-correlations between crude oil and agricultural commodity markets. *Physica A: Stat Mech its Appl* (2014) 395:293–302. doi:10.1016/j.physa.2013.10.021

22. Wang J, Shao W, Kim J. Analysis of the impact of COVID-19 on the correlations between crude oil and agricultural futures. *Chaos Solitons Fractals* (2020) 136:109896. doi:10.1016/j.chaos.2020.109896

23. Podobnik B, Horvatic D, Petersen AM, Stanley HE. Cross-correlations between volume change and price change. *Proc Natl Acad Sci U S A* (2009) 106(52):22079–84. doi:10.1073/pnas.0911983106

24. Pan Y, Hou L, Pan X. Interplay between stock trading volume, policy, and investor sentiment: A multifractal approach. *Physica A: Stat Mech its Appl* (2022) 603:127706. doi:10.1016/j.physa.2022.127706

25. Adam AM, Kyei K, Moyo S, Gill R, Gyamfi EN. Multifrequency network for SADC exchange rate markets using EEMD-based DCCA. *J Econ Finan* (2022) 46(1):145–66. doi:10.1007/s12197-021-09560-w

26. Li J, Shi Y, Cao G. Topology structure based on detrended cross-correlation coefficient of exchange rate network of the belt and road countries. *Physica A: Stat Mech its Appl* (2018) 509:1140–51. doi:10.1016/j.physa.2018.06.059

27. Ferreira P, Tilfani O, Pereira E, Tavares C, Pereira H, El Boukfaoui MY. Dynamic connectivity in a financial network using time-varying DCCA correlation coefficients. *Econometric Res Finance* (2021) 6(1):57–75. doi:10.2478/erfin-2021-0004

28. Pereira E, Ferreira P, da Silva M, Miranda J, Pereira H. Multiscale network for 20 stock markets using DCCA. *Physica A: Stat Mech its Appl* (2019) 529:121542. doi:10.1016/j.physa.2019.121542

29. Mbatha VM, Alovokpinhou SA. The structure of the South African stock market network during COVID-19 hard lockdown. *Physica A: Stat Mech its Appl* (2022) 590:126770. doi:10.1016/j.physa.2021.126770

30. Wang GJ, Xie C, ChenYJChen S. Statistical properties of the foreign exchange network at different time scales: evidence from detrended cross-correlation coefficient and minimum spanning tree. *Entropy* (2013) 15(5):1643–62. doi:10.3390/e15051643

31. Shin KH, Lim G, Min S. Dynamics of the global stock market networks generated by DCCA methodology. *Appl Sci* (2020) 10(6):2171. doi:10.3390/app10062171

32. Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. Mosaic organization of DNA nucleotides. *Phys Rev E* (1994) 49(2):1685–9. doi:10.1103/physreve.49.1685

33. Hu K, Ivanov PC, Chen Z, Carpena P, Stanley HE. Effect of trends on detrended fluctuation analysis. *Phys Rev E* (2001) 64(1):011114. doi:10.1103/physreve.64.011114

34. Kang SH, McIver R, Yoon SM. Dynamic spillover effects among crude oil, precious metal, and agricultural commodity futures markets. *Energy Econ* (2017) 62:19–32. doi:10.1016/j.eneco.2016.12.011

35. Gong X, Liu Y, Wang X. Dynamic volatility spillovers across oil and natural gas futures markets based on a time-varying spillover method. *Int Rev Financial Anal* (2021) 76:101790. doi:10.1016/j.irfa.2021.101790

36. Zhang S, Guo Y, Cheng H, Zhang H. Cross-correlations between price and volume in China's crude oil futures market: A study based on multifractal approaches. *Chaos Solitons Fractals* (2021) 144:110642. doi:10.1016/j.chaos.2020.110642

37. Saramäki J, Kivelä M, Onnela JP, Kaski K, Kertesz J. Generalizations of the clustering coefficient to weighted complex networks. *Phys Rev E* (2007) 75(2):027105. doi:10.1103/physreve.75.027105

38. Newman ME. Modularity and community structure in networks. *Proc Natl Acad Sci U S A* (2006) 103(23):8577–82. doi:10.1073/pnas.0601602103

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fphy.2022.1017309

OPEN ACCESS

EDITED BY
Jianguo Liu,
Shanghai University of Finance and Economics, China

REVIEWED BY
Jianfeng Qiu,
Anhui University, China
Hongbo Zhu,
Huaiyin Institute of Technology, China

*CORRESPONDENCE
Zhen Wang,
63710710@qq.com

SPECIALTY SECTION
This article was submitted to Social Physics,
a section of the journal
Frontiers in Physics

RECEIVED 11 August 2022
ACCEPTED 30 August 2022
PUBLISHED 23 September 2022

CITATION
Han S-Y, Zhan J-C, Xie C-H and Wang Z (2022), Features of intercity bus passenger group mobility behaviors in the context of smart tourism.
Front. Phys. 10:1017309.
doi: 10.3389/fphy.2022.1017309

Features of intercity bus passenger group mobility behaviors in the context of smart tourism

Shao-Yong Han[1,2,3], Jing-Chun Zhan[4], Cui-Hua Xie[3] and Zhen Wang[3,4]*

[1]School of Information Engineering and Technology, Changzhou Institute of Industry Technology, Changzhou, China, [2]Postdoctoral Scientific Research Workstation, Bank of Zhengzhou, Zhengzhou, China, [3]School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou, China, [4]School of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

The features of intercity bus passenger group mobility behaviors have important guiding significance for the transportation department. Based on passengers' intercity bus ticket reservation records (roundtrips from Shanghai or Chongqing city) from a smart tourism app, the travel behaviors of these two groups of bus passengers are analyzed and compared. In each group, the passengers' travelling interval time presents a power-law with a cutoff index, and the passengers' travelling behaviors have negative memory and low burstiness. Also, travel distance displays a scale-free property, and it is more likely to have an exponential distribution. Furthermore, the difference in cyclotron radius between these two groups' travelling distances is quite significant; roundtrips from Shanghai are frequent. Last, holidays have a significant influence on passengers' travel behaviors, which leads to more trips. The research conclusions are helpful to deeply understand the features of human mobility behaviors in theory, and can assist the transportation department in traffic planning in the application.

KEYWORDS

smart tourism, intercity bus passenger, mobility features, statistical analysis, decision support

1 Introduction

In recent years, green, low-carbon travel modes have been vigorously promoted in China. More and more medium and short-distance passengers choose to travel by intercity bus instead of self-driving. In the "Internet +" era, smart travel software has been widely used. An increasing number of passengers use smart travel apps to book bus tickets. The booking records of numerous passengers on the website provide feasibility for analyzing the mobility behaviors of passengers who travel by intercity bus. Mining and analyzing the travel behavior of bus passengers, demonstrating the travel behavior of bus passengers, and finding the evolution of bus passenger flow can provide decision support for the transportation bureau, bus operation companies, and other departments to plan bus operation routes and improve operation efficiency, which has substantial practical value.

The earliest analysis of human behavior can be traced back to 2005. Barabási [1] published a paper in Nature and proposed that human behavior's inter-event time distribution presents the power-law feature, which made many scholars start to study human behavior [2, 3].

The time features of human behavior refer to the statistical law of time shown by people engaged in a specific event many times [4]. Through many empirical statistics, such as email communication [1], web browsing [5], and online movie on demand [6], in these people's behaviors, scientists found that the inter-event time presents a power-law curve. Goh [7] proposed to analyze the "burstiness" feature and the "memory" feature of human behavior time series.

Burstiness [8] refers to the intermittent increase and decrease of activity or event frequency, calculated by the variation coefficient to get the result. For an individual user, the similarity between his two consecutive activities is called memory [7]. First, the time series of these activities is calculated. Then, the memory value can be obtained by calculating their first-order autocorrelation function.

The spatial features of human mobility behavior refer to the features of the movement activities in space, and the most intuitive is the travel distance distribution. Brockmann et al. [9] indirectly reflected human travel trajectory by studying the movement trajectory of banknotes; Gonzalez et al. [10] described the mobility features of individuals through the cyclotron radius and the mean square displacement of the individual movement; Song et al. [11] explored the predictability of human travel patterns by using mobile phone communication records.

Using the data records of human travel by taking public transport to explore the behavior features of passengers has become a research hotspot [12], and many scholars have conducted research and made discoveries to varying degrees. Huang et al. [13] used the data from one airline company, carried out a series of statistical analyses, and then obtained the spatiotemporal features of air passengers' group travel mobility behavior. With the help of Chengdu bus rapid transit smart card data, Yang Guang [14] defined and quantitatively studied the travel laws of passengers. Sui et al. [15] conducted a spatiotemporal analysis of passengers' travel behaviors based on bus smart card data in Brisbane, Australia. According to the booking records on the website, Han et al. [16] analyzed the travel inter-event time of passengers who take different transportation, then carried out modeling and simulation.

With the help of the booking records of intercity bus tickets in the smart tourism app, regarding the passengers taking the intercity bus as the research object, we analyzed their travelling behaviors and explored their mobility features. We chose passengers located in two municipalities directly under the central government in China, which are Shanghai and Chongqing. Shanghai is located at the mouth of the Yangtze River and on the edge of the East China Sea. There are almost no mountains in Shanghai, a typical plain city. In comparison, Chongqing is a specific mountain city near the Sichuan Basin, located upriver of the Yangtze River [17]. Although the two cities are municipalities directly under the central government, they have different topography, different geographical locations, and considerable differences in economic development levels [18]. The travelling behavior features of bus passengers in these two cities are analyzed and compared, may leading to some discoveries.

2 Experiments and results

We used the intercity bus ticket reservation records (roundtrips from Shanghai or Chongqing) from the website background database of a smart tourism app, and then analyzed and compared the temporal and spatial features of passengers' group mobility behaviors. Temporal features include the travel time interval, the burstiness, the memory, and the inter-event time between ticket purchase time and travel time. Spatial features include the travel distance distribution, the cyclotron radius, and the mean square displacement. In addition, the proportion of bus passengers in the two cities per month in 1 year is also counted and analyzed from an overall perspective.

2.1 Data sources and preprocessing

The data was taken from the intercity bus ticket reservation records from the database of a smart tourism app, and the intercity buses here refer to the long-distance buses, which undertake passenger transportation between cities.

During the data acquisition process, we deleted the passenger's name, gender, and other personal privacy data, and only retained the passenger's user id (UID). UID is a string, including a total of 16 English letters or numbers, which is unique in the database. In the database, through the passenger's UID, the booking date, departure date, departure city, arrival city and other fields are obtained by association, and then the passenger's travel distance is calculated according to the latitude and longitude information of the departure city and the arrival city.

We obtained the records of passengers travelling by the intercity buses in 2019, and limited the departure city to Shanghai (Chongqing) and the arrival city to other cities in China, and obtained some data; Then, the arrival city is limited to Shanghai (Chongqing) and the departure city is other domestic cities, and another part of data is obtained; Finally, these two parts of data are combined to form the dataset. In the dataset, the fields include the passenger's UID, the passenger's booking time (unit: day), the travel time (unit: day), the departure city, the longitude and latitude of the departure city, the arrival city, the

TABLE 1 Basic statistics of datasets.

Dataset	Number of passengers	Number of trips	Number of bus stops
Shanghai	1,263,454	3,620,287	741
Chongqing	774,010	1,932,633	632

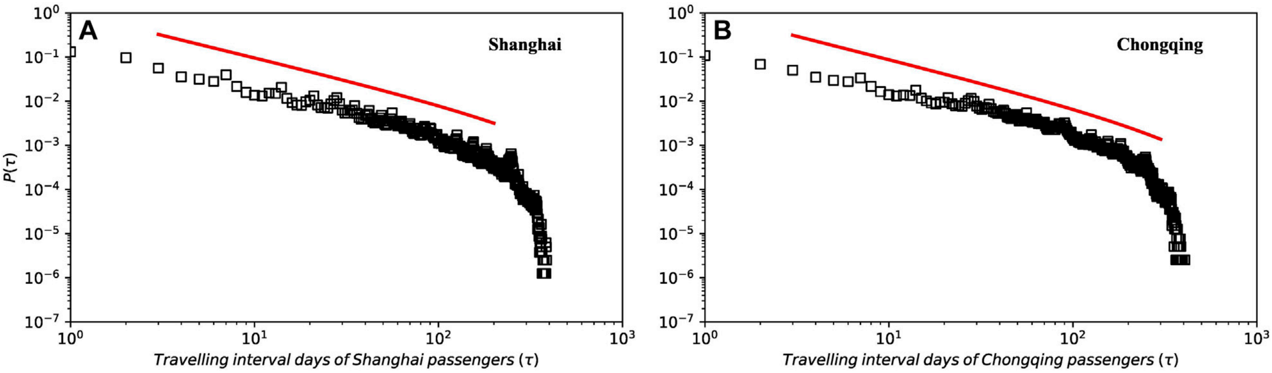


FIGURE 1
Distributions of travelling interval time.

longitude and latitude of the arrival city, and the passenger's travel distance.

The data was preprocessed on the Spark big data platform as follows:

1. 155 duplicate records were removed.
2. The departure (arrival) city value of some records is empty, so 17662 such records were removed.
3. The wrong longitude (latitude) data of the departure (arrival) city in the records were verified, and 162575 such records were repaired.

After the above data preprocessing, two experimental datasets were obtained, containing the basic statistical information of the two datasets (short for Shanghai and Chongqing), as listed in Table 1.

2.2 Temporal features analysis

2.2.1 Analysis of travel time interval

Passengers' travelling interval days reflect the frequency of passengers' travel and are an essential measure of passengers' travel by intercity bus. The distribution of the travel interval for all passengers in the two data sets is shown in Figure 1. This is different from the previous study that found that the mobility interval time follows a power-law form [6], and we can see that the passengers' travelling inter-event time presents a curve of power-law with a cutoff index. Similar results were observed in human mobility models [9, 10] and Huang's study [13].

According to the discovery of Huang [13] and Yan [19], the function of the curve of power-law with a cutoff index is shown as the following:

$$P(\tau) = \tau^{-\alpha} e^{-\beta\tau}, \tag{1}$$

Where τ represents the inter-event time of an individual's travelling event, and α represents a coefficient, and 'e' represents the natural constant, and β represents a coefficient.

The parameters of the distribution function of the two datasets are calculated by maximum likelihood estimation [20]. Two functions are obtained by computer fitting, which are $P(\tau) \propto \tau^{-1.01} e^{-0.002\tau}$ (Dataset Shanghai), and $P(\tau) \propto \tau^{-1.05} e^{-0.002\tau}$ (Dataset Chongqing). The distribution diagram is shown in Figure 1, and the truncated power-law tail indicates that the travelling time intervals of passengers are not uniformly distributed.

2.2.2 Burstiness and memory

To analyze an individual's behavior, the first step is to obtain the time of the behavior events, and then sort them into time

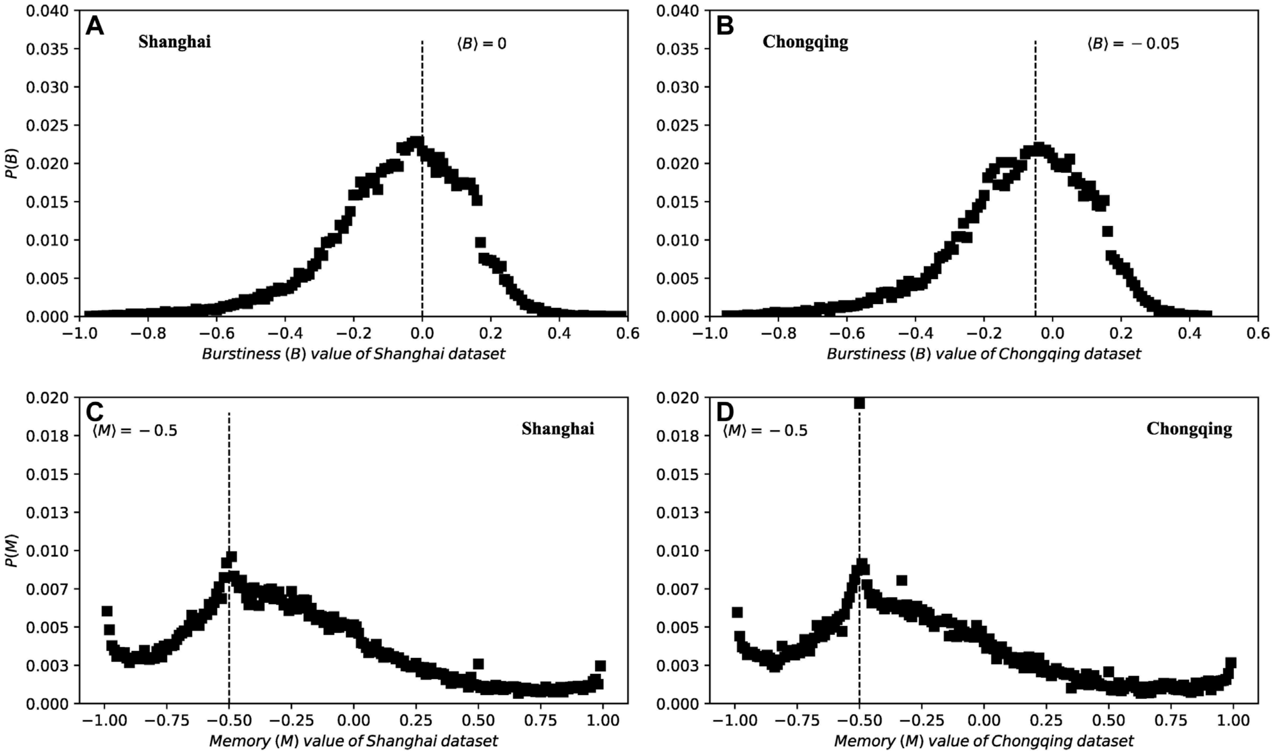


FIGURE 2
Distribution graphs (burstiness B and memory M of the Shanghai and Chongqing datasets).

series in order. Finally, these two indicators, the burstiness and the memory [7], are used to analyze. We use the same process and indicators to analyze the collective behavior of passengers who take intercity buses.

For a passenger, we get his boarding time for each travelling, and calculate the time series of travelling inter-event time. That is $\{\tau_1, \tau_2, ..., \tau_{jn_j}\}$, then we could figure out the mean value m_{τ_j} and the standard deviation value σ_{τ_j}. According to the research results of other scholars [7], the burstiness B_j can be calculated by the following equation:

$$B_j \equiv \frac{\left(\sigma_{\tau_j}/m_{\tau_j} - 1\right)}{\left(\sigma_{\tau_j}/m_{\tau_j} + 1\right)} = \frac{\sigma_{\tau_j} - m_{\tau_j}}{\sigma_{\tau_j} + m_{\tau_j}}, \tag{2}$$

where m_{τ_j} and σ_{τ_j} are the mean and standard deviation value. It can be seen from Eq. 2 that the value of B_j ranges from -1 to 1.

For passenger, calculating the time difference between the two continuous travelling events (the $(i+1)th$, the th), can get the inter-event time, and define it as $\tau_{j,i}$. We can obtain the $\tau_{j,i}$ value of each passenger, then get a dataset, and calculate the mean m_1 and standard deviation σ_1 of this data set. By the same token, for passenger, calculate the time difference between the two continuous travelling events (the $(i+2)th$, the $(i+1)th$), and define it as $\tau_{j,i+1}$, then calculate and get the mean m_2 and standard deviation σ_2. The memory M_j of a passenger j can be calculated by the following equation:

$$M_j \equiv \frac{1}{n_j - 1}\sum_{i=1}^{n_j-1}\frac{\left(\tau_{j,i} - m_1\right)\left(\tau_{j,i+1} - m_2\right)}{\sigma_1\sigma_2}, \tag{3}$$

where n_j represents the boarding times of a passenger j, $\tau_{j,i}$ is the inter-event time between the two continuous travelling events (the $(i+1)th$, the ith), and m_1, σ_1 are the mean, standard deviation value; $\tau_{j,i+1}$ is the inter-event time between the two continuous travelling events (the $(i+2)th$, the $(i+1)th$), and m_2, σ_2 are the mean, standard deviation value.

It can be seen from Eq. 3 that the value of M_j ranges from -1 to 1. When M_j equals 0, it represents that there is no memory effect; When M_j is positive, it shows that for the passenger j, in his travelling event time series, long inter-event time or short inter-event time continuously appears, and it's called memory effect. When M_j is negative, which indicates that in the passenger's travelling event time series, a long inter-event time alternates with a short inter-event time, and it's called negative memory effect.

According to the above calculation methods, from the collective behavior level, Figure 2 plots the distribution curve

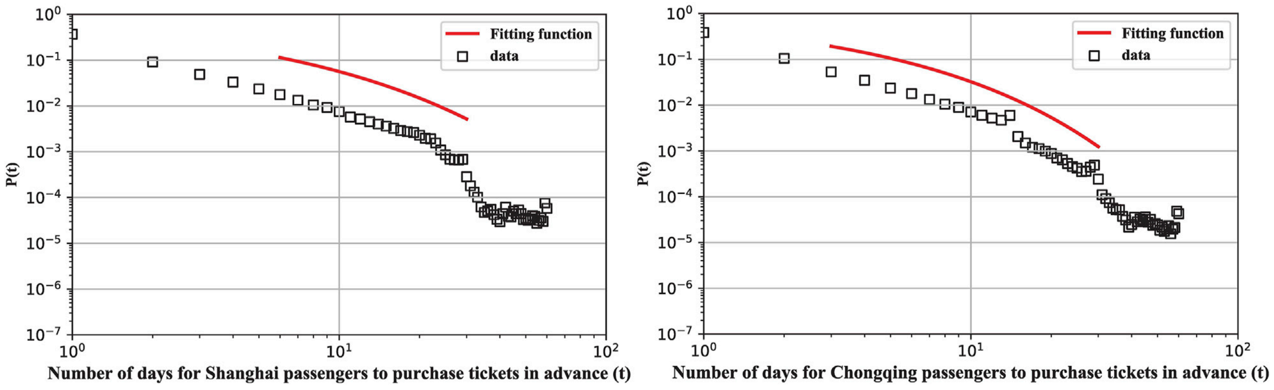


FIGURE 3
The distribution of advance ticket days to passengers in Shanghai (Chongqing) dataset.

for these two datasets. In Figure 2, we can find the burstiness mean $B = 0$ in the Shanghai dataset, the burstiness mean $B = -0.05$ in the Chongqing dataset, and the memory mean $M = -0.5$ in these two datasets. The distribution in Chongqing presents a weak burstiness. Negative memory indicates that the travelling time intervals of passengers show the situation of separation between a long time interval and a short time interval. This is very consistent with the current situation. Passengers go out by intercity bus and return by bus in a few days. Generally speaking, passengers take intercity buses for medium and long-distance travel, so the frequency of passengers taking the intercity bus cannot be as high as that of passengers taking the bus in the city at the same time. Usually, passengers will travel by intercity bus again at a long interval.

2.2.3 Analysis of inter-event time between ticket purchase time and travel time

Whether to purchase tickets in advance is closely related to the passenger’s situation, such as whether the trip is planned, or whether the trip is decided suddenly and not planned. So the number of days, that is, the inter-event time between ticket purchase time and travel time, is an essential temporal feature of our analysis and research.

The distribution of advance ticket days for bus passengers in Shanghai and Chongqing is demonstrated in Figure 3.

In Figure 3, the abscissa t represents the inter-event time (unit: day) between the time when a passenger books tickets and the passenger’s boarding time. The ordinate $p(t)$ represents the probability.

Based on the preliminary studies, we use the stretched exponential distribution [13] to fit the function, which is $(t) = e^{\lambda t^{\beta}}$. The fitting function of the Shanghai dataset is $(t) = e^{-0.81t^{0.55}}$, and the fitting function of the Chongqing dataset is $p(t) = e^{-0.84t^{0.61}}$. It can be found that the parameters of the two fitting functions are very close.

It can be seen from Figure 3 that the advance ticket days of passengers are concentrated in 1–60 days, which may be related to the reservation time limit of bus tickets. Whether in Shanghai or Chongqing, the probability of passengers purchasing tickets in advance reaches its highest value on 1 day and generally decreases with the increase of ticket purchasing days in advance. There are similar dynamics for advance ticket purchase time of bus passengers in these two cities.

Traditionally, the bus ticket price does not change as frequently as the price of airline tickets. Passengers who book bus tickets do not have an awareness of saving money by buying tickets in advance. However, the number of days to purchase tickets in advance can reflect the planning and urgency of passengers’ travel purposes to a certain extent. It can reflect the types of passengers on this basis.

Passengers who buy tickets many days in advance may have travel plans, while those who book tickets in a hurry to travel may have sudden events related to work to deal with. Therefore, passengers who buy tickets in advance for a long time mostly travel on holidays and rest days. On the contrary, passengers who purchase tickets in advance for a short time mostly travel on working days.

To prove our guess, we divide the types of travel dates of passengers into three categories, which are holiday, weekend, and workday, and then explore whether there is any difference in the time of ticket purchase in advance in different types of travel dates. Limited by the time tickets can be booked, passengers can book tickets 30 days in advance at the earliest. We plot the percentage stacked bar chart of these two factors, consisting of the types of travel dates and the days of advance ticket purchase for bus passengers in the two cities, illustrated in Figure 4.

The abscissa t represents the number of days that passengers have purchased tickets in advance, and the ordinate represents the percentage. The Blank columns represent that the passengers travel in holiday; the columns filled with horizontal lines

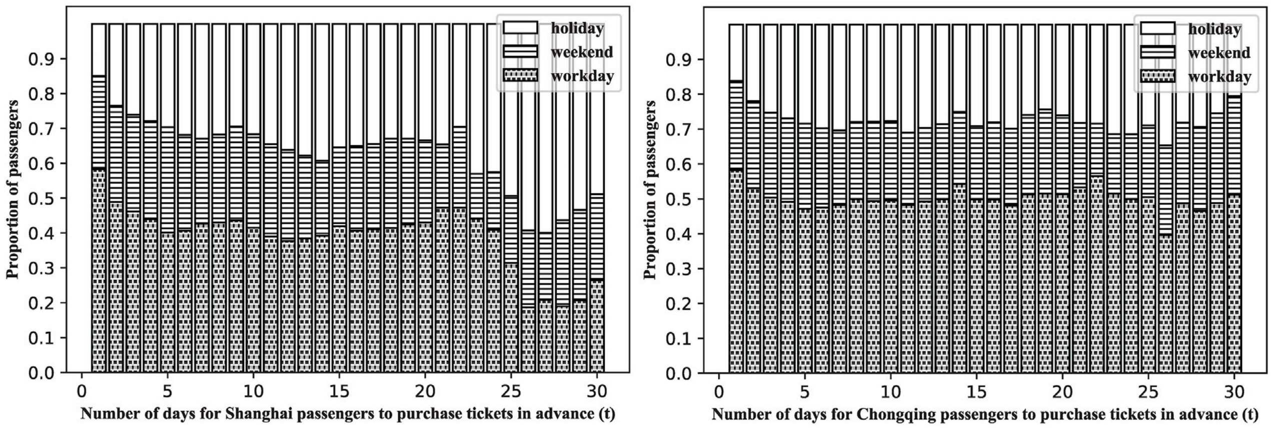


FIGURE 4
The percentage stacked bar chart of advance ticket purchase time and type of travel date in Shanghai (Chongqing) dataset.

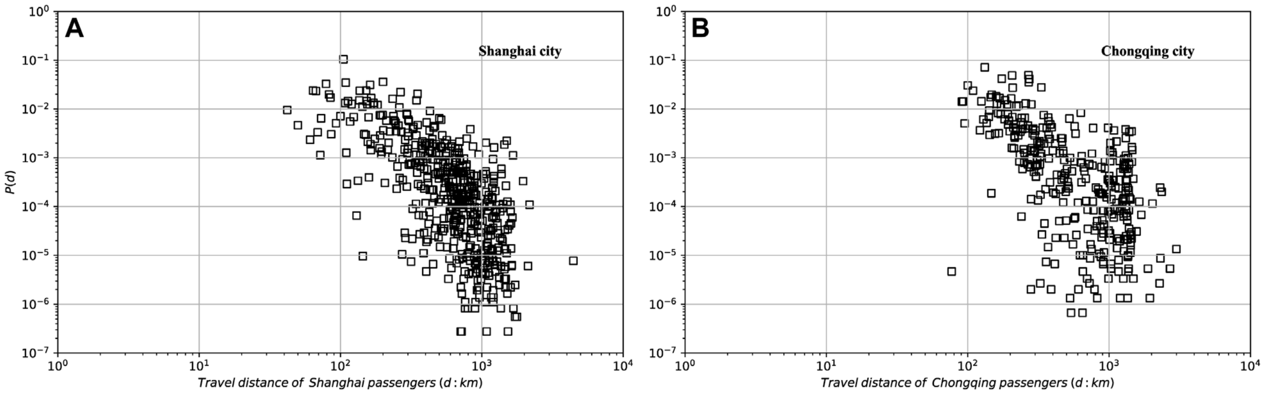


FIGURE 5
Distribution of passengers' travelling distance.

represent that the passengers travel at weekends; the column filled with crosses represents that the passengers travel on working days.

2.3 Spatial features analysis

2.3.1 Analysis of travel distance distribution

Travel distance has been widely studied in infrastructure networks such as public transportation [21]. Based on bus passenger survey data in Shijiazhuang city, Wang et al. [22] counted the passengers' travelling distance by bus, and found that it obeyed the exponential distribution.

We compare and analyze the travel distance distribution of bus passengers and the route distance distribution of intercity bus stops in these two cities. The travel distance distribution of bus passengers in the two cities is shown in Figure 5. Here, the travel distance refers to the distance between the starting stop and the terminal stop of each passenger's bus route.

In Figure 5, the abscissa d represents in these two cities, passengers' travelling distance, and the ordinate $p(d)$ represents probability.

According to Figure 5, most passengers in the two cities travel 100–1000 km, and some passengers in Shanghai travel less than 100 km, which may be because the distance between Shanghai and the surrounding cities of Shanghai is less than 100 km.

Since the travel distance of bus passengers is heavily dependent on the route distance operated by the bus company, we break up the length of distance into intervals

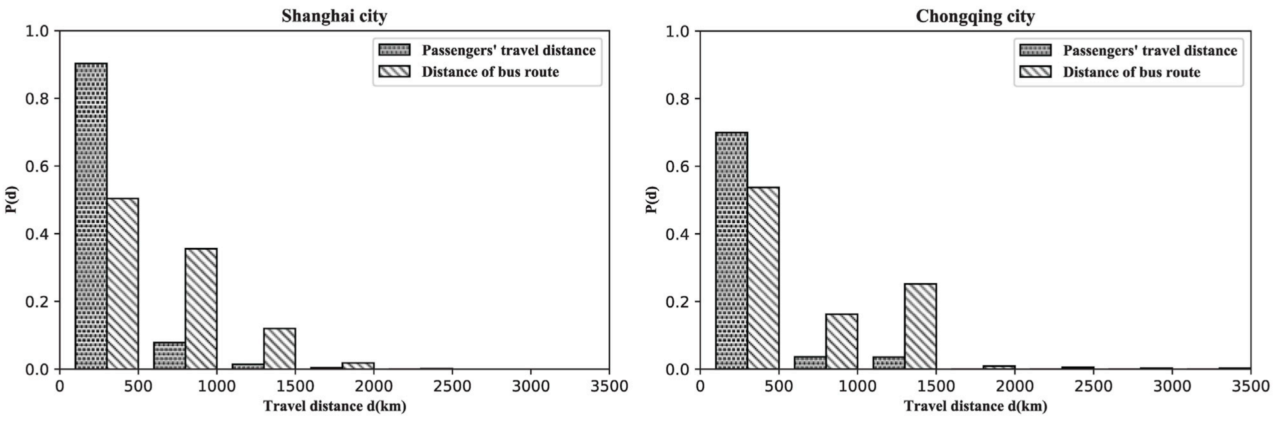


FIGURE 6
The compound bar chart of travel distance and bus route distance in Shanghai (Chongqing) dataset.

and then plot the compound bar chart of travel distance and bus route distance, which is demonstrated in Figure 6.

In Figure 6, the abscissa d represents the distance, and the ordinate represents probability. The columns filled with crosses represent the distribution of travel distance by passengers, and the columns filled with slashes represent the distribution of travel distance of bus routes operated by bus companies.

In Figure 6, we can find that for the highest column, the corresponding abscissa axis is between 0 and 500 km. Moreover, it can be seen that the height of the two types of columns is quite different. In the range of 0–500 km, the travel distance distribution probability of passengers is even much higher than that of bus routes. In the Shanghai city sub-picture, for the distance of bus routes, the proportion of 0–500 km is not much different from that of 500–1000 km, but the probability of passenger travel distance distribution in the range of 0–500 km is close to 0.9, and the probability of passenger travel distance distribution in the range of 500–1000 km is close to 0.1. The probability of passenger travel distance distribution in the field of 0–500 km is also close to 0.9 in Chongqing.

We guess that the intercity bus is the preferred means of transportation for passengers for short and medium-distance travel. For long-distance travel, passengers may choose high-speed rail or an airplane with a high probability. However, it can be seen that the number of long-distance buses and medium and short distance buses is unreasonable. The number of short-distance intercity buses should be increased to meet the passengers' demand.

According to Figure 5 and Figure 6, as the travel of bus passengers depends on the bus route transportation network, the travel distance distribution of bus passengers exhibits scale-free features. Brockmann et al. [9] studied the flow trajectory of banknotes as the banknotes were traded in different hands, and the flow distance of banknotes was not limited to the transportation network, so distance distribution obeyed the power law.

2.3.2 Cyclotron radius and mean square displacement

In the study of human mobility features, the diffusion features of humans can be represented by the growth law of the cyclotron radius and the mean square displacement (MSD) with time. The cyclotron radius [10] of all bus passengers in each city after n trips is given as:

$$r_g(n) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - x_t)^2}, \tag{4}$$

Where x_t represents the bus stop visited by all bus passengers most times on n trips, and x_k represents the bus stop that the passengers arrive at on each trip, and $x_k - x_t$ represents the distance between the two bus stops. In particular, for a passenger who only travels once, x_t and x_k denote the departure stop and the arrival stop of the bus route.Where, the abscissa n represents the travelling times of passengers, and the ordinate $r_g(n)$ represents the cyclotron radius.

In Figure 7, the overall cyclotron radius of bus passengers in the two cities reaches its maximum when the number of trips is 1. In the Shanghai data set, with the increase in the number of trips, the cyclotron radius decreases sharply. Until the number of trips n equals 28, the cyclotron radius begins to stabilize, at close to 100 km. In the Chongqing data set, when n ranges from 2 to 7, the cyclotron radius decreases slowly; when n ranges from 7 to 20, the cyclotron radius is relatively stable; and when n is greater than 20, the cyclotron radius is more divergent.

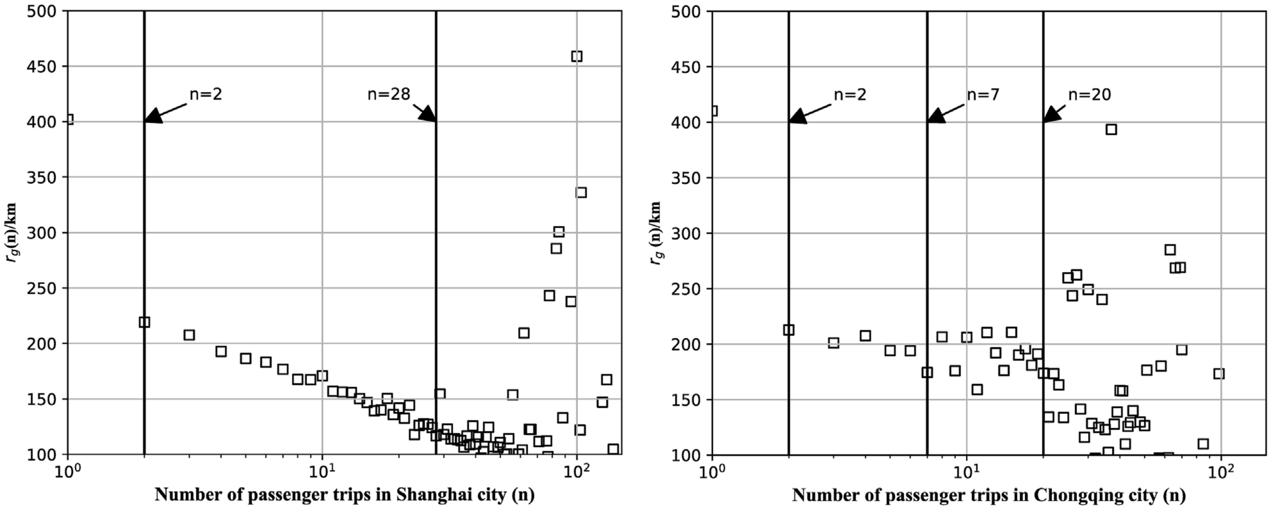


FIGURE 7
Distribution of cyclotron radius in Shanghai (Chongqing) dataset.

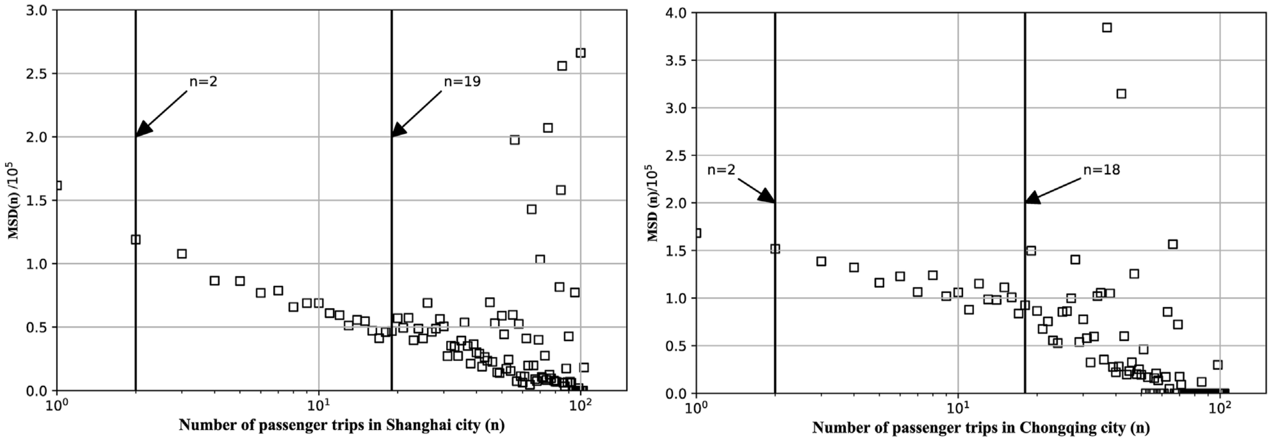


FIGURE 8
Distribution of MSD in Shanghai (Chongqing) dataset.

When the number of trips of bus passengers equals 1, the cyclotron radius is the largest, because when n equals 1, many passengers only travel once. Their cyclotron radius is the actual travel distance. When n is greater than 1, their round-trip rate is also very high. That is to say, when a passenger arrives at the destination bus stop, there is a high probability that he will return to the previous bus stop where he departed last, which results in the passengers' cyclotron radius decreasing rapidly.

In the Shanghai dataset, for passengers who travel 2–28 times, their destination cities are relatively close. Their travel frequency is very high, so the cyclotron radius generally shows a trend of decreasing with the increase in the number of trips. In the Chongqing dataset, passengers who travel 7–20 times often take intercity buses, and they have relatively stable and balanced trips in the bus stop network. The cyclotron radius decreases when the travelling times increase. Therefore, the features are different from those of other types of passengers.

The MSD distribution of bus passengers is shown in Figure 8. The MSD is given as:

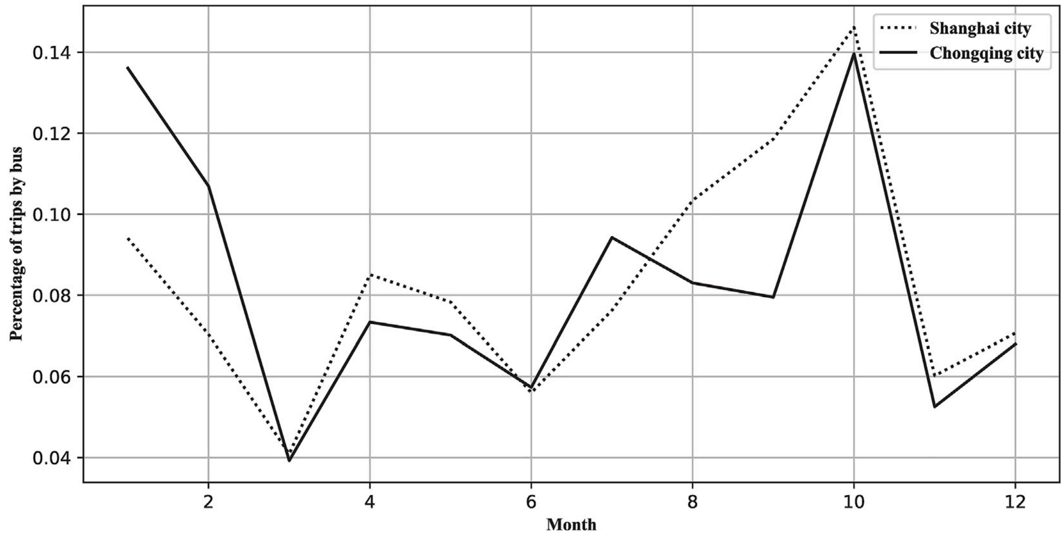


FIGURE 9
The proportion of passenger trips per month.

$$MSD(n) = \langle (x_n - x_0)^2 \rangle, \quad (5)$$

Where x_0 is the location of the first departure bus stop in the passenger's travel record, and the ⟨•⟩ operator refers to the mean distance of n trips.Where, the abscissa n represents passengers' travelling times in each city, and the ordinate represents the MSD of the passengers' trips. In Figure 8, the MSD has a peak value at the beginning. When n ranges from 1 to 18 (19), the MSD shows a downward trend. After that, the number of trips continues to increase, and the MSD presents a discrete distribution.

According to the previous analysis of the cyclotron radius and the MSD, we can see that the cyclotron radius and MSD of passengers' trips are closely related to the number of trips. The travel distance of most passengers does not increase with the increase in the number of trips but changes within a specific range, indicating the high boundedness of bus passenger trips. The distribution of the cyclotron radius in the Shanghai and Chongqing datasets is quite different. It is speculated that there are more passengers travelling for short distances in the surrounding areas of Shanghai, and they travel frequently. So the cyclotron radius decreases steadily with the increase in the number of trips.

2.4 Proportion of passenger trips per month

We counted the number of trips of all passengers in Shanghai (Chongqing) dataset in each month of the year, and the proportion of passenger trips per month in these two cities in year 2019 is shown in Figure 9.

The abscissa in Figure 9 represents the month, and the ordinate represents the proportion of passenger trips per month.

The spring Festival holiday in 2019 was from February 4 to February 10. Before the spring Festival, many passengers needed to take the intercity bus to go home, and in October, many passengers would travel on the National day. Therefore, January and October were the two periods with the highest passengers. April 5 to April 7 in 2019 was the Qingming Festival, and some passengers travelled during this period. However, during the period from July to October, the proportion of passengers in the two cities showed different trends. According to the data set from Chongqing, the travel volume of passengers was higher in July and continued to decrease in August and September. In the data set for Shanghai, the travel volume of passengers continued to rise from July to September. This period coincided with the summer vacation of students. It was speculated that during this period, many passengers with their children travelled to Shanghai and took intercity buses to travel to the cities around Shanghai. It also indicates that during festivals and summer holidays, bus companies should adjust their operation plans and increase the number of buses to meet the strong travel demand of passengers.

As shown in Figure 9, in January and October, the proportion of passenger trips was the highest in these two cities. We further analyzed the travel distance distribution of passengers in the two cities in these 2 months, as shown in Table 2. In Table 2, we can find that in Shanghai, the proportion of trips below 500 km is 88% in October and 78% in January; In Chongqing, the proportion of trips below 500 km is 93% in October and 88% in January, and this indicates that in October, passengers in Shanghai and Chongqing are more inclined to travel for short distances.

TABLE 2 Travel distance ratio (d, unit: km) in January and October in Shanghai and Chongqing.

Month	Dataset	d ≤ 500 (%)	500 < d ≤ 1000 (%)	1000 < d ≤ 1500 (%)	1500 < d ≤ 2000 (%)	2000 < d ≤ 2500 (%)
January	Shanghai	78	18	4	1	0
	Chongqing	88	4	7	0	0
October	Shanghai	88	10	1	0	0
	Chongqing	93	4	3	0	0

TABLE 3 Travel distance ratio (d ≤ 500, unit: km) in January and October in Shanghai and Chongqing.

Month	Dataset	d ≤ 100 (%)	100 < d ≤ 200 (%)	200 < d ≤ 300 (%)	300 < d ≤ 400 (%)	400 < d ≤ 500 (%)
January	Shanghai	12	29	13	13	10
	Chongqing	7	41	32	6	1
October	Shanghai	16	43	13	10	7
	Chongqing	6	36	42	8	1

In Table 2, at least 80% of the travel distance is less than 500 km, so we listed the proportion of travel distance below 500 km, which is shown in Table 3. According Table 3, we can find that in Chongqing, at least 70% of the travel distance is between 100 km and 300 km, and this may be due to that there is a strong demand for short-distance travel around the city, and Chongqing has a larger area than Shanghai.

3 Discussions

From the results of the experimental analysis, the passengers' travelling interval time presents a power-law curve with a cutoff index, and both exhibit weak burstiness and strong negative memory effects. Passengers take the intercity bus to go out and return in a few days. After a long period, the passengers will travel by intercity bus again, so the travelling time intervals of passengers present the situation that a long inter-event time alternates with a short inter-event time, which shows a negative memory effect.

There are similar dynamics for advance ticket purchase time of bus passengers in these two cities. When travelling on holiday, they prefer to buy tickets in advance in Shanghai. According to the analysis of the date of purchase of tickets in advance and the type of travel date, in Shanghai, among those who purchase tickets 25–30 days in advance, more than 50% of passengers plan to travel on holidays. Maybe these passengers are worried that they can't buy tickets near holidays. Therefore, on holidays, the transportation department and bus operation company should increase the amount of transportation.

The bus route network limits the travel distance of passengers. The distribution of travel distance of passengers by intercity bus in the two cities does not have scale-free features, which is more in line with the stretching index distribution.

According to the distribution of travel distance and bus route distance, in Shanghai and Chongqing, there is a strong demand for medium and short-distance travel within 500 km. In comparison, there is less demand for long-distance travel above 500 km. So the transportation departments and bus operation companies in the two places should adjust the bus routes, and concentrate the traffic on short-distance routes within 500 km.

The cyclotron radius and the MSD of passengers' travelling distance, which indicate that the mobility distance of most passengers does not increase infinitely with the increase of the number of trips and changes within a specific range. The two groups of bus passengers tend to move in a limited range, and their travel distance is highly bounded. There is a big difference in the distribution of the cyclotron radius of travel between the two groups. There are more short-distance passengers to and from Shanghai, so the cyclotron radius decreases steadily with the increase in the number of trips in the Shanghai dataset.

Holiday factors have a significant impact on the travel of the two groups, mainly traditional festivals such as spring Festival, Qingming Festival, and National Day. During these periods, the number of passengers travelling is large. The number of passengers travelling to and from Shanghai in the summer gradually increases. According to the proportion of passengers travelling in 1 year, in the two cities, passengers travel mainly in January and October, so the transportation department can consider increasing the amount of transportation in these two time periods.

4 Conclusion

The passengers' travelling inter-event time by taking the intercity bus presents a power-law with a cutoff index, and both exhibit weak burstiness and strong negative memory effects. The distribution of travel distance of passengers by intercity bus in the two cities does not have scale-free features, which is more in line with the stretching index distribution. The difference in cyclotron radius between these two groups' travelling distances is quite significant; roundtrips from Shanghai are frequent. Holidays have a significant influence on passengers' travel behaviors, which leads to more trips.

In terms of application, the transportation department and bus operation company can optimize the intercity bus operation and adjust their operation strategy based on the suggestions in this paper.

The travelling time intervals of passengers in the two cities show low burstiness and negative memory effects, and the underlying principles need to be further explored. In terms of theory, the results in this paper can provide help for more scholars' research to study the features of human mobility behaviors.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

S-YH provided this topic and wrote the manuscript. J-CZ, C-HX and ZW guided, discussed, and modified the manuscript. All authors contributed to the manuscript and approved the submitted version.

Funding

Supported by a project grant from Zhejiang Philosophy and Social Sciences Foundation (Grand No.19NDJC145 YB).

Conflict of interest

Author S-YH was employed by Bank of Zhengzhou.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Barabasi AL. The origin of bursts and heavy tails in human dynamics. *Nature* (2005) 435(7039):207–11. doi:10.1038/nature03459

2. Guo JL. A model of human behavior dynamics and exact results. *Acta Phys Sin* (2010) 59(6):3851–5. doi:10.7498/aps.59.3851

3. Yan XY, Zhao C, Fan Y, Di ZR, Wang WX. Universal predictability of mobility patterns in cities. *J R Soc Interf* (2014) 11(100):20140834. doi:10.1098/rsif.2014.0834

4. Zhou T, Han XP, Yan XY, Yang ZM, Zhao ZD, Wang BH. Statistical mechanics on temporal and spatial activities of human. *J Univ Electron Sci Tech China* (2013) 42(4):481–540. doi:10.3969/j.issn.1001-0548-2013.04.001

5. Dezsö Z, Almaas E, Lukács A, Rácz B, Szakadát I, Barabási AL. Dynamics of information access on the web. *Phys Rev E* (2006) 73(6):066132. doi:10.1103/PhysRevE.73.066132

6. Zhou T, Kiet HAT, Kim BJ, Wang BH, Holme P. Role of activity in human dynamics. *Europhys Lett* (2008) 82(2):28002. doi:10.1209/0295-5075/82/28002

7. Goh KI, Barabási AL. Burstiness and memory in complex systems. *Europhys Lett* (2008) 81(4):48002. doi:10.1209/0295-5075/81/48002

8. Lambiotte R, Tabourier L, Delvenne JC. Burstiness and spreading on temporal networks. *Eur Phys J B* (2013) 86(7):320–4. doi:10.1140/epjb/e2013-40456-9

9. Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. *Nature* (2006) 439(7075):462–5. doi:10.1038/nature04292

10. Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding individual human mobility patterns. *nature* (2008) 453(7196):779–82. doi:10.1038/nature06958

11. Song C, Qu Z, Blumm N, Barabási AL. Limits of predictability in human mobility. *Science* (2010) 327(5968):1018–21. doi:10.1126/science.1177170

12. Sun YF, Gan HC. Car commuters' travel behaviors with presence of multi-modal travel information. *J Univ Shanghai Sci Tech* (2018) 40(6):595–600. doi:10.13255/j.cnki.jusst.2018.06.013

13. Huang FH, Peng J, You MY. Analyses of characetristics of air passenger group mobility behaviors. *Acta Phys Sin* (2016) 65(22):228901–328. doi:10.7498/aps.65.228901

14. Yang G, Battie MC, Boyd SK, Videman T, Wang Y. Cranio-caudal asymmetries in trabecular architecture reflect vertebral fracture patterns. *Bone* (2017) 3:102–7. doi:10.1016/j.bone.2016.11.018

15. Tao S. Spatial-temporal analysis of travel behaviour using transit smart card data and its planning implications: A case study of Brisbane, Australia. *Shanghai Urban Plann Rev* (2017) 5:94–9. doi:10.11982/j.supr.20170594

16. Han SY, Guo Q, Yu K, Li RD, He B, Liu JG. Statistical mechanism of passenger mobility behaviors for different transportations. *Int J Mod Phys C* (2020) 31(6):2050082–13. doi:10.1142/S0129183120500825

17. Cai SL, Li SC. An analysis on location of chongqing city. *J Sichuan Normal Univ (Natural Science)* (2001) 24(4):423–5. doi:10.3969/j.issn.1001-8395.2001.04.030

18. Chen J, Chen Q, Li HP. Psychological influences on bus travel mode choice: A comparative analysis between two Chinese cities. *J Adv Transportation* (2020) 2:1–9. doi:10.1155/2020/8848741

19. Yan XY, Han XP, Wang BH, Zhou T. Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Sci Rep* (2015) 34(4):2678. doi:10.1038/srep02678

20. Zhao H, Zhang C. Maximum likelihood estimation for reflected Ornstein-Uhlenbeck processes with jumps. *Commun Stat - Theor Methods* (2019) 48(5):1221–33. doi:10.1080/03610926.2018.1425451

21. Peng CB, Jin XG, Wong KC, Shi MX, Liò P. Collective human mobility pattern from taxi trips in urban area. *PloS one* (2012) 7(4):e34487. doi:10.1371/journal.pone.0034487

22. Wang MS, Huang L, Yan XY. Exploring the mobility patterns of public transport passengers. *J Univ Electron Sci Tech China* (2012) 41(1):2–7. doi:10.3969/j.issn.1001-0548.2012.01.001

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 17 October 2022
DOI 10.3389/fphy.2022.1018781

OPEN ACCESS

EDITED BY
Jianguo Liu,
Shanghai University of Finance and Economics, China

REVIEWED BY
Xiaolong Ren,
University of Electronic Science and Technology of China, China
Zhuoming Ren,
Hangzhou Normal University, China

*CORRESPONDENCE
Yixiu Kong,
yixiu.kong@bupt.edu.cn
Yizhong Hu,
kuai@bupt.edu.cn

SPECIALTY SECTION
This article was submitted to Social Physics,
a section of the journal
Frontiers in Physics

RECEIVED 13 August 2022
ACCEPTED 20 September 2022
PUBLISHED 17 October 2022

CITATION
Kong Y, Hu Y, Zhang X and Wang C (2022), Structural centrality of networks can improve the diffusion-based recommendation algorithm.
Front. Phys. 10:1018781.
doi: 10.3389/fphy.2022.1018781

Structural centrality of networks can improve the diffusion-based recommendation algorithm

Yixiu Kong[1]*, Yizhong Hu[1]*, Xinyu Zhang[2] and Cheng Wang[1]

[1]School of Science, Beijing University of Posts and Telecommunications, Beijing, China, [2]Tsinghua Education Foundation, Tsinghua University, Beijing, China

The recommendation system has become an indispensable information technology in the real world. The recommendation system based on the diffusion model has been widely used because of its simplicity, scalability, interpretability, and many other advantages. However, the traditional diffusion-based recommendation model only uses the nearest neighbor information, which limits its efficiency and performance. Therefore, in this article, we introduce the centralities of complex networks into the diffusion-based recommendation system and test its performance. The results show that the overall performance of heat conduction algorithm can be improved by 184%–280%, using the centrality of complex networks, reaching almost the same accuracy level as the mass diffusion algorithm. Therefore, the recommendation system combining the high-order network structure information is a potentially promising research direction in the future.

KEYWORDS

recommendation system, centrality, complex network, diffusion model, collaborative filtering

1 Introduction

With the development of information technology, people are overwhelmed by an increasing amount of information. Although the development of technological innovations has made our lives easier, meanwhile, overloaded information consumes our time and efforts when we are searching online. Therefore, in response to this demand, searching systems and recommendation systems are evolving accordingly, and they both are the technologies that have been developed to deal with information overload. The searching systems solve the problem of directed search, while the recommendation systems can predict the possible preferences and interests of users based on the previous data. Recommendation systems have been developed for decades, and each part has been gradually improved and developed toward more multi-level and applicable models.

Collaborative filtering [1–3] is one of the most widely used, least computationally complex, and most effective information-filtering algorithms. The CF algorithm provides personalized recommendations for each user based on the user's past purchase history database and product search records. Breese et al. [2] classified CF into two broad categories, namely, memory-based and model-based approaches. Memory-based methods predict information and recommend products based on a measure of similarity between the user and the product [3–5]. Model-based algorithms use a collection of user and object information to generate information-filtering models [10, 11] through clustering [6], Bayesian approach [7], matrix factorization [8, 9], and other machine learning methods.

Different from computer science, the application of physics in the field of interdisciplinary science has also obtained some successful complex network theories, and various classical physical processes have provided some new insights and solutions for the active field of information filtering in recent years [1, 12, 13]. For example, the diffusion process like the heat conduction process on a dichotomous complex network [14], the principles of dynamic resource allocation in dichotomous complex networks [15], opinion diffusion [16], and gravity [17] have been applied in information filtering.

Overall, CF and other diffusion-based recommendation algorithms have been successfully applied to many well-known online e-commerce platforms. Meanwhile, in recent years, a lot of research studies, such as the heat conduction, mass diffusion, or hybrid method [20, 21], biased heat conduction [22, 23], multi-channel diffusion [24], preferential diffusion [25, 26] based on the CF direct random walk method [27], hypergraph models with social labels [28, 29], and multilinear interactive matrix factorization [30],are devoted to studying the two variations of the algorithm. These algorithms will further improve the efficiency of information-filtering systems. In addition, multiple explorations [31, 32] on information filtering considering external constraints have also been made.

Meanwhile, we should note that the compressed network structural information, including the information core and information backbone, provides some enlightenment for the in-depth understanding of information filtering [18, 19]. Some network structural centralities, such as PageRank and eigenvector centrality, can indicate the structural properties of networks in one-dimensional metrics, while traditional network structure statistics, such as degree, can only contain first-order structure information (nearest neighbor information). We can conclude that it is highly probable to obtain richer structural information about the network by replacing the influence of first-level nearest neighbors with one-dimensional complex network structural centralities, such as coreness and PageRank. This structural information considers the long-range correlations in the network structure, which is expected to help improve the accuracy of recommendation systems and improve the application prospect in real scenarios. In the following part of this article, we will show that the adoption of network structural information can greatly improve the performance of some recommendation algorithms.

2 Materials and methods

2.1 Dataset description

We obtained seven commonly used datasets in the research of recommendation systems. The datasets are listed in Table. 1.

2.2 Evaluation metrics

In the dataset, we know some of the items that users collect, and we need to recommend other items for users. Therefore, to compare the recommendation performance of recommendation algorithms [33], we divide the dataset into two categories: training set and test set. The training set is used for each user-recommended item, and then the test set is used to evaluate recommendation algorithm. When calculating the evaluation metrics, we will regard the selection of the user from the user's recommendation lists as a positive example, and otherwise, a negative example. For each user, in accordance with the actual results and predicted results of each item, 1 is expressed as a positive case, and 0 is expressed as a negative case. We can define the following three indicators: precision, recall, and F1-score.

Precision and recall are commonly used evaluation indices. Precision is defined as the proportion of data in the dataset that the label predicts correctly, and recall is the proportion of data that we predict to be correct in a certain class. They are calculated as

$$Recall = \frac{TP}{TP + FP},$$

But for each user, what we obtain is a list of users, namely, the TOP-K recommendations. So, we assume that the number of items recommended to the user is K, the precision is the number of recommended items in the test set, and the recall rate is the correct number of recommended items in the test set. For a target user, the precision and recall are defined as

$$P_i = \frac{d_i(K)}{K},\ R_i = \frac{d_i(K)}{D_i},$$

where $d_i(K)$ is the degree of the test set in the user i's TOP-K recommended list and $D_i$is the length of the test set for the *i*th user. After calculating the precision and recall of the TOP-K recommendation list of all users, we can obtain the precision and recall of the algorithm by calculating the average value of all users.

When recommending TOP-K items for users, what need to be considered are the precision and recall rate of the reality (whether the user likes it or not) and the predicted label (recommendation list) of each item. Therefore, F1-score, also called balanced F-score, is an index commonly used to express the precision rate of binary classification problems in statistical data analysis. F1-score can be regarded as the harmonic average result of two indices, and the value range is [0, 1]. F1-score is defined as

$$f_1 - \text{score} = \frac{2 \cdot precision \cdot recall}{precision + recall}.$$

2.3 The general Markovian form of diffusion-based recommendations

In this section, we briefly introduce the three algorithms adopted in this article. When they are applied to calculate the

TABLE 1 Introduction of datasets.

Dataset	Number of users	Number of items	Number of records	Sparsity
Movielens_latest small	610	9,784	100,836	1.70%
Movielens_100K	943	1,682	100,000	6.30%
Movielens_90K	2,113	10,109	855,598	4.00%
Movielens_1M	6,040	3,883	1,000,209	4.26%
Movielens_10M	69,878	10,681	10,000,054	1.33%
Last.fm	1,892	12,523	186,479	0.79%
Netflixdataset	5,967	16,977	1,261,097	1.24%

recommendation list of users on the dataset, it is necessary to calculate the probability transition matrix on each training set, but the time complexity is higher if the calculation is carried out directly through the formula. Therefore, through observation, it is found that the calculation form of the probability transition matrix is to multiply the corresponding elements after row normalization and column normalization of some two columns in the adjacent matrix. Therefore, the calculation formula can be simplified into a matrix form [34, 35] to reduce and improve the calculation efficiency of the algorithm. For example, the calculation formula of the probability transition matrix of the heat conduction algorithm is as follows:

$$\begin{aligned} W_{\alpha\beta}^{H} &= \frac{1}{k_\alpha}\sum_{i=1}^{u}\frac{a_{i\alpha}\cdot a_{i\beta}}{k_i} \\ &= \sum_{i=1}^{u}\frac{a_{i\alpha}\cdot a_{i\beta}}{k_\alpha\cdot k_i} \end{aligned}.$$

Here, $a_{i\alpha}$ and $a_{i\beta}$ are the elements of the ith row in two columns of the matrix a and β, respectively. The meaning of the aforementioned expression is to first normalize the elements in some two columns of the bipartite graph adjacency matrix, then matrix multiplication is performed, and the vector elements obtained are added after the multiplication step. Therefore, the formula of the probability transition matrix can be simplified to the following matrix form:

$$W_{\alpha\beta}^{H} = D_o^{-1}\cdot A^T\cdot D_u^{-1}\cdot A.$$

where

$$D_u^{-1} = \begin{pmatrix} \frac{1}{d_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{d_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{d_u} \end{pmatrix}, D_o^{-1} = \begin{pmatrix} \frac{1}{d_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{d_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{d_o} \end{pmatrix}.$$

Here, D_u^{-1} and D_o^{-1} represent a diagonal matrix formed by calculating the reciprocal of the node degree values of the two categories of users and items, respectively. If the node degree is 0, the derivative value is not necessarily to be 0. $\mathrm{d_i}$ is the degree of the ith user or item node. Similarly, the formulas of the other two algorithms are analyzed and can be simplified into the following matrix expressions:

$$W_{\alpha\beta}^{M} = \mathrm{A}^T\cdot D_u^{-1}\cdot A\cdot D_o^{-1},$$
$$W_{\alpha\beta}^{H+M} = D_o^{-1}\lambda\cdot A^T\cdot D_u^{-1}\cdot A\cdot D_o^{-1}1-\lambda.$$

If a network graph corresponding to a dataset is defined as $G = (V, E)$ and the adjacency matrix of the corresponding bipartite graph is defined as $A_{u\times o}$, then we can calculate the probability transition matrix of the users and regard the user's favorite vector in the adjacency matrix as the initial heat or resource amount for the item, so that we can calculate the initial heat or resource amount for the item according to the following expression:

$$result = A\cdot W^T,$$

where result is a matrix of size $u\times o$ and the ith row vector of the matrix represents a score list of all items for the user i. It was assumed that the set of all items in the dataset is V_{Item}, the set of items collected by the user is V_{collect}, and the uncollected set is

$$V_{notCollect} = V_{Item} - V_{collect}.$$

According to the final rating result, some items that are not collected by the user yet are sorted according to the score. Since the TOP-K recommendation with the highest score is selected in the dataset, it is the recommendation list of the user i that is recorded as a matrix of result_K. The ith row of the matrix is the K items recommended for the user i. According to the recommendation list and the test set, the evaluation metrics for the user i can be calculated.

2.4 A brief introduction to the centralities of networks

2.4.1 Closeness centrality

The closeness [36] indicates the degree of difficulty in arriving at other nodes from a certain node, and the larger the value is, the farther the distance from other nodes is; a lower value indicates a closer distance to other nodes. It is calculated using the following equation:

$$Closeness(V) = \frac{1}{\sum_{q-1}^{n-1} d(p,q)},$$

where p represents the node that is to be calculated, q represents other nodes in the network, and $d(p,q)$ is the shortest path length from node p to node q. The idea of this closeness is that the closer the nodes are to the network center, the more quickly they can reach other nodes. Therefore, the importance of each node obtained through closeness not only is influenced by the number of adjacent nodes but also reflects the minimum average shortest path to other nodes by utilizing the characteristics of the whole network.

2.4.2 Eigenvector centrality

Eigenvector centrality [37] is a metric to measure the impact of nodes on the network. This metric is used when nodes with the same number of links are present. A high score for eigenvector centrality means that the node is connected to a few nodes that have high scores themselves. For a given network with an adjacent matrix A, if two nodes i and j are not directly connected, then $A_{i,j} = 0$, and otherwise $A_{i,j} = 1$. The eigenvector of A must satisfy the following expression:

$$A\boldsymbol{x} = \lambda \boldsymbol{x},$$

and eigenvector centrality is given by the eigenvector corresponding to the largest eigenvalue of A.

2.4.3 Katz centrality

Like eigenvector centrality, Katz centrality [38] also measures the importance of nodes. The difference is that it considers the nodes that have an in-degree 0 by adding a decay coefficient α and a bias term β. Also, with the help of the adjacency matrix A, for a node i, Katz centrality x_i is calculated as follows:

$$x_i = \alpha \sum_{j \subseteq G} a_{i,j} x_j + \beta.$$

In practice, the attenuation coefficient $\alpha < 1/\lambda$ is usually chosen to ensure that the matrix is invertible and that centrality can be obtained.

2.4.4 PageRank

PageRank [39] algorithm was originally a calculation method for calculating weights to solve the ranking between web pages, which was developed based on eigenvector centrality. Although the method is proposed to solve the problem of a directed graph, this method can be used in any graph, and now it is often applied to the analysis of the importance of various networks.

For a directed graph $G = (V, E)$, we can define a Markovian process that has the probability transition matrix $M = (m_{i,j})$ in the graph, and the normalized initial centrality of all nodes is R_0. Thus, we can obtain the value after one-step transition as MR_0 and then proceed in turn until t-step probability transition:

$$R_0, MR_0, M^2R_0, \ldots, M^tR_0, \ldots$$

If this series converges, the final vector R represents the stationary distribution of the Markov chain and satisfies MR =R. Therefore, the value of the vector R is the PageRank value of the nodes in the network. In practice, a random distribution term E is added to ensure that the node with zero in-degree can also receive incoming links, and the weight of this random term is usually set to 0.15.

Finally, we can obtain the following expression:

$$\text{PageRank} = 0.85 \times \text{R} + 0.15 \times \frac{\text{E}}{n},$$

where n is the total number of nodes and E is a matrix whose elements are all equal to 1.

3 Results

The degree is the most commonly used metric in the study of a network structure model, which is defined as the number of neighboring nodes of a given node, thus reflecting the importance of a node in the network. In the diffusion-based recommendation algorithm, the process of heat conduction or mass diffusion of each node is completed governed by the degree of each node. But this idea is, in fact, too simple to be applied in the real world. In a social network, if a member B knows only one influential member A, although the node has a only degree that equals to 1, B's influence will increase due to the higher importance of node A. From this point of view, there is a room for improvement in the use of the network structure as an indicator for user recommendation.

Therefore, the method proposed in this article is to improve the recommendation algorithm based on the traditional diffusion process by replacing the degree with the network structural centralities like closeness, eigenvector centrality, PageRank, and Katz centrality. These four types of network structural centralities are used in this article.

3.1 The selection of test sets

We first randomly divide the dataset into 90% of the training set and 10% of the test set according to the conventional practice and use the algorithm to recommend items for users. However, since the size of the dataset of each user is not average, in real life, the length of the recommendation list required by each user is not the same. Therefore, to compare the difference of the recommendation performance among different algorithms, we

TABLE 2 Recommendation evaluations on seven datasets based on random selection of test sets.

Dataset	Algorithm	Precision (%)	Recall (%)	F1-score
ml_latest small	**Heats**	0.92	7.12	1.63%
	Probs	7.57	33.23	1.23%
	Hybrid	8.03	35.73	1.31%
ml_90K	**Heats**	0.12	0.57	0.20%
	Probs	15.84	23.16	18.8%
	Hybrid	16.16	23.93	19.3%
ml_100K	**Heats**	11.37	15.53	13.1%
	Probs	26.76	42.03	32.7%
	Hybrid	28.84	45.70	35.4%
ml_1M	**Heats**	5.10	18.88	8.04%
	Probs	7.94	28.12	12.4%
	Hybrid	8.99	33.27	14.14
Netflix	**Heats**	0.03	0.16	0.04%
	Probs	8.06	24.70	12.1%
	Hybrid	8.71	27.82	13.3%
ml_10M	**Heats**	5.58	29.84	9.40%
	Probs	7.51	36.68	12.5%
	Hybrid	7.94	39.40	13.2%
Last.fm	**Heats**	0.08	0.63	0.14%
	Probs	0.75	7.10	0.14%
	Hybrid	0.72	5.86	0.13%

only consider the TOP-K recommendation list of each user, if the length of the recommendation list is $K = 50$. The hybrid algorithm needs to set a weighting coefficient to determine the weight of the two algorithms, and the weighting coefficient is set to $\lambda = 0.5$. We obtain the performance of the data recommendation system under random segmentation, as shown in Table 1.

However, in the application of the recommendation algorithm, the random selection of the test set is inconsistent with the application of the recommendation system because it neglects the temporal information and can violate causality. Therefore, in this article, the test set can be selected in a temporal way. As all datasets are sorted in accordance with the timestamp in this article, we select the time of the latest 10% of the data as the test set of the data and the rest of the time before as the training set.

3.2 Results of datasets with temporal test sets

On the datasets, after selecting the test set with a temporal sequence, the results are as follows:

For the test set selected in a temporal sequence, several indicators of the performance are a bit lower than those of the results of random selection (shown in Table 2 and Table 3). This is mainly because of the following reason: if the dataset is small and the timestamps of the users are not likely to be evenly distributed, then it is likely that the user's activity will not be selected into the test set. This will reduce the number of samples in the training set; these users will be less connected to the whole system. This will also affect the algorithm for these users of the prediction results, and the calculation of the precision of the algorithm only considers this subset of the user's TOP-K recommended list. So, the accuracy of the four indicators will decline. However, in this way, we obtain the training set and the test set which are closer to the reality, and the result of the algorithm is more meaningful and applicable than that of random selection of test sets.

3.3 Results incorporating centralities

In the aforementioned Materials and methods section, it was shown that for the original algorithm, calculation is directly carried out through a probability transition matrix calculation

TABLE 3 Recommendation evaluations on seven datasets with temporal selection of test sets.

Dataset	Algorithm	Precision (%)	Recall (%)	F1-score (%)
ml_latest small	**Heats**	0.70	0.14	0.23
	Mass	4.78	3.61	4.12
	Hybrid	5.22	3.95	4.48
ml_90K	**Heats**	0.78	0.48	0.59
	Mass	6.31	6.53	6.42
	Hybrid	6.41	6.69	6.59
ml_100K	**Heats**	4.66	5.56	5.07
	Mass	8.81	14.80	11.0
	Hybrid	8.99	14.65	11.1
ml_1M	**Heats**	5.22	5.16	5.19
	Mass	18.95	15.72	17.2
	Hybrid	19.23	16.52	17.8
Netflix	**Heats**	0.15	0.22	0.18
	Mass	4.68	7.56	5.78
	Hybrid	4.55	8.15	5.84
ml_10M	**Heats**	2.40	2.02	2.20
	Mass	4.17	3.29	3.68
	Hybrid	4.13	3.28	3.66
Last.fm	**Heats)**	0.20	0.61	0.30
	Mass	1.74	5.10	2.60
	Hybrid	1.43	4.25	2.14

TABLE 4 Recommendation results adopting centralities.

	Precision (%)	Recall (%)	F1-score (%)
Degree (heats)	0.70	0.14	**0.23**
Degree (mass)	4.78	3.61	4.12
Degree (hybrid)	5.22	3.95	4.50
Closeness (heats)	4.51	3.32	**3.83**
Closeness (mass)	4.62	3.38	3.90
Closeness (hybrid)	4.54	3.31	3.82
Eigenvector (heats)	0.54	0.08	**1.32**
Eigenvector (mass)	4.95	3.88	4.35
Eigenvector (hybrid)	5.08	4.14	4.56
PageRank (heats)	4.24	3.28	**3.70**
PageRank (mass)	4.68	3.65	4.10
PageRank (hybrid)	4.70	3.70	4.14
Katz (heats)	4.35	3.27	**3.74**
Katz (mass)	4.38	3.29	3.76
Katz (hybrid)	4.41	3.30	3.78

formula, so that the time complexity is high; however, the formula can be converted into a matrix form through simplification, namely, the weight of heat or resource allocation can be changed, so that the performance of the algorithm can be improved. We calculate the five network structural metrics, and the results of the movielens_latest small dataset are as follows:

From the aforementioned Table 4, the following conclusions can be obtained between different network structural centralities and different algorithms:

1) The results of various network structural centralities basically meet the common knowledge that the precision of the hybrid algorithm is greater than that of the mass diffusion algorithm and is greater than that of the heat conduction algorithm. However, for the results of PageRank or eigenvector centrality, it can be seen that the results of mass diffusion algorithm are better than those of hybrid algorithm, mainly due to the influence of the fixed weighting coefficient of the hybrid algorithm.
2) More interestingly, by changing the network structural centrality, the precision of the algorithm is improved, especially the precision of the heat conduction algorithm is greatly improved by 5.73–16.7 times, with the F1-score measure, and almost reaches the level of mass diffusion. The results of HC are marked in bold font.

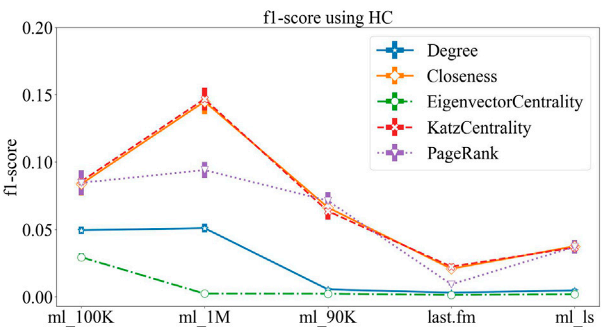


FIGURE 1
Results of different datasets using heat conduction.

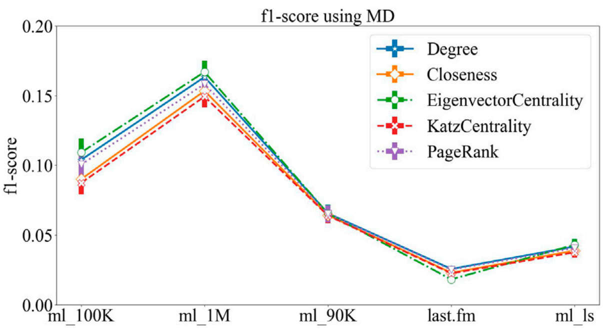


FIGURE 2
Recommendation results of five datasets using mass diffusion.

For further analysis, we choose the more comprehensive index F1-score to compare the differences of recommendation results obtained by different centralities on five datasets. As in the aforementioned dataset, the F1-score results obtained by the heat conduction algorithm using five network structural indicators are shown in Figure 1.

From the aforementioned figure, we can see that the method of introducing the network structural metrics improves the performance when using the HC algorithm. Among these metrics, Katz centrality and closeness centrality are the most effective, but the eigenvector centrality greatly reduces the precision of the original method. Moreover, we can see that the F1-score of the algorithm is improved by about 280% by using Katz centrality and closeness centrality on the movielens_1M dataset, and the F1-score of the algorithm is improved by 184% by using PageRank.

The results of MD and hybrid algorithms are shown in Figure 2 and Figure 3, respectively. The figures show that the results of the two algorithms are relatively similar, and the recommendation results when different structural centralities are applied are basically similar; also, there are a few small differences in different datasets. But one can see that for each dataset, when using the three metrics, the results of degree, eigenvector centrality, and PageRank are better, while the results of the other two metrics are slightly worse, but the differences between each metric are generally within 5%. On some datasets, such as movielens_100K and movielens_1M, we can see that the recommendation results obtained by using eigenvector centrality are marginally better than those obtained by using the degree.

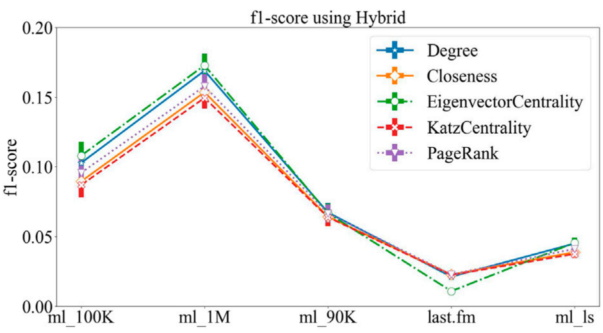


FIGURE 3
Recommendation results of five datasets using hybrid algorithm.

4 Discussion

In this study, we mainly focus on the traditional collaborative filtering (CF)-based recommendation systems to improve the information-filtering technique by utilizing the network structural centrality. This article focuses on the recommendation systems that are based on the diffusion processes: heat conduction algorithm and mass diffusion algorithm, which are based on physics theories. However, these two algorithms have their own focus in terms of accuracy and diversity. Combining the two algorithms by a weighting coefficient λ can lead to a hybrid algorithm that can obtain better performance in both the metrics [20].

The recommendation results of the aforementioned three algorithms that adopt degree information as the input of the recommendation system are obtained on the movielens and other datasets commonly used. Through the analysis, we find that the degree contains network structural information of nearest neighbors, which only reflects the number of neighbors of each node. Therefore, this article proposes an improved method that contains the information of higher-dimensional network structural information to improve the recommendation algorithm, by using metrics such as closeness centrality, eigenvector centrality, Katz centrality, and PageRank These metrics not only take the number of neighboring nodes into account but also contain the importance of neighboring nodes as well as their structural information.

The method proposed in this article shows that applying different network structural centralities improves the recommendation performance. The results of the HC algorithm obtained with all the tested network structural centralities show an improvement in performance in the range of 184%–280% (the only exception is that using eigenvector centrality causes a decrease in accuracy). For the MD algorithm, the differences in the results obtained after applying different metrics are small. Among them, the optimal results were obtained for three metrics: degree, eigenvector centrality, and PageRank. Meanwhile, the hybrid algorithm has overall better prediction results, but the results of some metrics are likely to be influenced by the weighting factor λ, and even better results could be obtained after adjusting the parameters.

Overall, in this article, we show that the centrality of networks contains higher-order structural information of the network topology than the traditional adoption of degree. Surprisingly, the recommendation algorithms incorporating such centralities, especially the heat conduction algorithm will have a significantly improved performance, almost comparable to that of the mass diffusion algorithm.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found at Stanford Large Network Dataset Collection: https://snap.stanford.edu/data/.

Author contributions

YK designed the research; YH and XZ collected the data; Cheng Wang, YK, and YH analyzed the data; and XZ performed the visualization. All authors wrote the manuscript.

Funding

This work was supported by the Research Funds for the Central Universities from Beijing University of Posts and Telecommunications, under grant No. 505022019, and NO. 500422415, and Research Innovation Fund for College Students of Beijing University of Posts and Telecommunications.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Lü L, Medo M, Yeung CH, Zhang YC, Zhang ZK, Zhou T. Recommender systems. *Phys Rep* (2012) 5191:1–49. doi:10.1016/j.physrep.2012.02.006

2. Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl JG. An open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM conference on Computer supported cooperative work; Chapel Hill North Carolina USA (1994). p. 175–86.

3. Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web; Hong Kong Hong Kong (2001). p. 285–95.

4. Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering (2013). Available from: https://arxiv.org/ftp/arxiv/papers/1301/1301.7363.pdf.1301.7363

5. Goldberg D, Nichols D, Oki BM, Terry D. Using collaborative filtering to weave an information tapestry. *Commun ACM* (1992) 3512:61–70. doi:10.1145/138859.138867

6. Ungar LH, Foster DP. Clustering methods for collaborative filtering. *AAAI Workshop recommendation Syst* (1998) 1:114–29.

7. Ungar L, Foster DP. A formal statistical approach to collaborative filtering. CONALD'98 (1998). Available from: https://www.cis.upenn.edu/~ungar/Datamining/Publications/CONALD.pdf.

8. Azar Y, Fiat A, Karlin A, McSherry F, Saia J Spectral analysis of data. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing: San Jose California USA (2001). 619–26.

9. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. *Computer* (2009) 42(8):30–7. doi:10.1109/mc.2009.263

10. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J machine Learn Res* (2003) 3(Jan):993–1022.

11. Keshavan R, Montanari A, Oh S. Matrix completion from a few entries. *IEEE Trans Inf Theor* (2010) 56(6):2980–98. doi:10.1109/tit.2010.2046205

12. Albert R, Barabási AL. Statistical mechanics of complex networks. *Rev Mod Phys* (2002) 74(1):47–97. doi:10.1103/revmodphys.74.47

13. Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. *Rev Mod Phys* (2009) 81(2):591–646. doi:10.1103/revmodphys.81.591

14. Zhang YC, Blattner M, Yu YK. Heat conduction process on community networks as a recommendation model. *Phys Rev Lett* (2007) 99(15):154301. doi:10.1103/physrevlett.99.154301

15. Zhou T, Ren J, Medo M, Zhang YC. Bipartite network projection and personal recommendation. *Phys Rev E* (2007) 76(4):046115. doi:10.1103/physreve.76.046115

16. Zhang YC, Medo M, Ren J, Zhou T, Li T, Yang F. Recommendation model based on opinion diffusion. *Europhys Lett* (2007) 80(6):68003. doi:10.1209/0295-5075/80/68003

17. Liu JH, Zhang ZK, Chen L, Liu C, Yang C, Wang X. Gravity effects on information filtering and network evolving. *PloS one* (2014) 9(3):e91070. doi:10.1371/journal.pone.0091070

18. Zeng W, Zeng A, Liu H, Shang MS, Zhou T. Uncovering the information core in recommender systems. *Sci Rep* (2014) 4(1):6140–8. doi:10.1038/srep06140

19. Zhang QM, Zeng A, Shang MS. Extracting the information backbone in online system. *PloS one* (2013) 8(5):e62624. doi:10.1371/journal.pone.0062624

20. Zhou T, Kuscsik Z, Liu JG, Medo M, Wakeling JR, Zhang YC. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc Natl Acad Sci U S A* (2010) 107(10):4511–5. doi:10.1073/pnas.1000488107

21. Fiasconaro A, Tumminello M, Nicosia V, Latora V, Mantegna RN. Hybrid recommendation methods in complex networks. *Phys Rev E* (2015) 92(1):012811. doi:10.1103/physreve.92.012811

22. Stojmirović A, Yu YK. Information flow in interaction networks. *J Comput Biol* (2007) 14(8):1115–43. doi:10.1089/cmb.2007.0069

23. Liu JG, Zhou T, Guo Q. Information filtering via biased heat conduction. *Phys Rev E* (2011) 84(3):037101. doi:10.1103/physreve.84.037101

24. Shang MS, Jin CH, Zhou T, Zhang YC. Collaborative filtering based on multi-channel diffusion. *Physica A: Stat Mech its Appl* (2009) 388(23):4867–71. doi:10.1016/j.physa.2009.08.011

25. Zhou T, Jiang LL, Su RQ, Zhang YC. Effect of initial configuration on network-based recommendation. *Europhys Lett* (2008) 81(5):58004. doi:10.1209/0295-5075/81/58004

26. Lü L, Liu W. Information filtering via preferential diffusion. *Phys Rev E* (2011) 83(6):066119. doi:10.1103/physreve.83.066119

27. Liu JG, Shi K, Guo Q. Solving the accuracy-diversity dilemma via directed random walks. *Phys Rev E* (2012) 85(1):016118. doi:10.1103/physreve.85.016118

28. Zhang ZK, Liu C. A hypergraph model of social tagging networks. *J Stat Mech* (2010) 10:P10005. doi:10.1088/1742-5468/2010/10/p10005

29. Zhang ZK, Zhou T, Zhang YC. Tag-aware recommender systems: A state-of-the-art survey. *J Comput Sci Technol* (2011) 26(5):767–77. doi:10.1007/s11390-011-0176-1

30. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. Available from: https://arxiv.org/abs/1511.07122.

31. Ren X, Lü L, Liu R, Zhang J. Avoiding congestion in recommender systems. *New J Phys* (2014) 16(6):063057. doi:10.1088/1367-2630/16/6/063057

32. Deng X, Wu L, Ren X, Jia C, Zhong Y, Lü L. Inferring users' preferences through leveraging their social relationships. In: Proceedings of the IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society; Beijing, China (October 2017). p. 5830–6.

33. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning; Pittsburgh Pennsylvania USA (June 2006). p. 233–40.

34. Zhang Y-C, Marcel B, Yi-Kuo Y. Heat conduction process on community networks as a recommendation model. *Phys Rev Lett* (2007) 99(15):154301. doi:10.1103/physrevlett.99.154301

35. Ren ZM, Kong Y, Shang MS, Zhang YC. A generalized model via random walks for information filtering. *Phys Lett A* (2016) 380(34):2608–14. doi:10.1016/j.physleta.2016.06.009

36. Chea E, Livesay DR. How accurate and statistically robust are catalytic site predictions based on closeness centrality? *Bmc Bioinformatics* (2007) 8(1):153–14. doi:10.1186/1471-2105-8-153

37. Gabriel P, Francis Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Inf Process Management* (1976) 12(5):297–312. doi:10.1016/0306-4573(76)90048-0

38. Katz L. A new status index derived from sociometric analysis. *Psychometrika* (1953) 18:39–43. doi:10.1007/bf02289026

39. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. *Computer networks ISDN Syst* (1998) 30(1-7):107–17. doi:10.1016/s0169-7552(98)00110-x

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 01 November 2022
DOI 10.3389/fphy.2022.1018704

Check for updates

OPEN ACCESS

EDITED BY
Jianguo Liu,
Shanghai University of Finance and Economics, China

REVIEWED BY
Long Guo,
China University of Geosciences, China
Can-Zhong Yao,
South China University of Technology, China

*CORRESPONDENCE
Yixiu Kong,
yixiu.kong@bupt.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

RECEIVED 13 August 2022
ACCEPTED 17 October 2022
PUBLISHED 01 November 2022

CITATION
Dong J, He Y, Song J, Ding H and Kong Y (2022), Universal scaling behavior and Hawkes process of videos' views on Bilibili.com.
Front. Phys. 10:1018704.
doi: 10.3389/fphy.2022.1018704

Universal scaling behavior and Hawkes process of videos' views on Bilibili.com

Jiarui Dong[1†], Yuping He[2†], Jiayin Song[3†], Hao Ding[2†] and Yixiu Kong[1]*

[1]School of science, Beijing University of Posts and Telecommunications, Beijing, China, [2]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, [3]School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China

Online videos have become the most popular method to obtain information for the public in recent years, such as TikTok and YouTube, and regional sites like Bilibili and Douyin. Compared with its growing influence, the analysis of user behavior on video sites is still less investigated. Herein, we fetch the video data from Bilibili.com and analyze the video views, comments, and other behaviors on the website. We found that the description model based on the Hawkes process can accurately predict the video views, which suggests that on the Bilibili website, the self-incentive mechanism of information cascade diffusion plays a decisive role in online views. Meanwhile, we also found that the view increment of the videos during the same period of time conforms to the general power-law distribution.

KEYWORDS

temporal social networks, Hawkes process, scaling behavior, Danmaku, Bilibili.com

1 Introduction

The online video platform, as an emerging tool for accessing information, has drawn significant attention from the commercial industry to scientific communities. However, it is still challenging to predict which videos or information will become popular in the near future. The basis for determining future popular videos is the future popularity of the video, which is usually measured by the number of times the video is played within a limited period of time. Therefore, video popularity prediction has become the key to discovering future popular videos. The goal of video popularity prediction is to predict the number of times a video will be played in the future based on the data available before or in the early stages of release [1–4]. Predicting the number of videos played in the future can not only help discover future popular videos but also directly help optimize strategies for online video services. [5] Shows a typical case of an optimizing service strategy using video popularity prediction. The researchers updated the caching strategy based on the predicted future views of the video. Compared with the LFU-/LRU-based caching strategy used in the previous online video service system, this strategy has significantly improved the caching efficiency.

TABLE 1 User and video metadata.

Field	Meaning
Send date	Uploading date
Tag	Tags
Duration	Duration
ID	ID number
Rank score	Official ranking score
Pub date	Upload date
Author	Author ID
Review	Comments
Mid	Message ID
Play	View counts
Pic	With picture or not
Description	Subjective description by uploader
Video review	Danmaku
Favorites	Number of likes
Arcurl	URL address
Bvid	Bilibili.com ID
Title	Title

In addition to the continuous pursuit of higher performance, video popularity prediction research also needs to consider the practicability of the model. Herein, our goal is to solve a series of problems in video popularity prediction research and contribute to the better application of this technology in online video services.

The related research on video popularity prediction started more than 10 years ago, and the first video popularity prediction model was formally proposed in 2010 [6]. Existing video popularity prediction models can be roughly divided into three main categories according to the types of tasks they target, data usage, and modeling methods for video popularity: prediction models based on mapping of popularity values; prediction models based on popularity time series; and the cold start prediction model. The prediction model based on the heat value mapping is the earliest proposed video heat prediction model. This type of model is generally implemented by modeling the distribution of the cumulative number of future video views with respect to the cumulative number of early video views using a mathematical function [6–8]. Subsequent proposed models of this type also use features partially extracted from the meta-information of the video [9–11]. With the introduction of multimodal features, the modeling methods of models based on heat value mapping have gradually become diversified, and both regression models and neural networks have been used to build such models. At the beginning of the research on video popularity prediction, the prediction model based on the popularity value mapping was effective in the popularity prediction task in the early stage of video service development. However, social information has a very limited impact on the number of videos played at that time. The early view number of a video can effectively reflect the viewing tendency of the user group toward the video. However, compared with the later proposed prediction model based on the popularity time series, the model based on the popularity value mapping lacks the sufficient ability to identify the trend of the number of videos played because it regards the number of videos played as a single cumulative value. However, this type of model has lower complexity and data requirements than the other two models, so it is easier to be used and deployed in online video service systems.

The second type of model regards the number of videos played as a sequence related to the video survival time and achieves prediction by modeling the correlation between the sequence of early video views and future views [12–14]. Time-series-based forecasting models usually assume that the number of videos played in different time periods early in their release is of different importance for inferring their future video views [15]. By modeling the sequence of video views, more information about the dynamics of early video views is utilized by the prediction model, so that the second type of model has better prediction performance than the other two types of models. Furthermore, since the second type of model treats the number of videos played as a sequence about the video survival time, it is possible to introduce more features sensitive to the video survival time. These features have been proven by existing research to be effective in helping models predict the possible burst of video views in future video viewing data [16–20]. In recent years, with the integration of social networks and online video services, time-series-based prediction models have also begun to extract multimodal features from diversified information, including social information, to cope with the impact of social information on video views [21–25]. The introduction of multimodal features, including social features, not only improves the performance of prediction models based on popularity time series but also makes such models increasingly complex, causing them to gradually lose their practicability in the video service environment. In today's online video service systems, in most cases, only the most basic prediction models based on popularity time series such as MLR and ARMA are still used.

On the other hand, Hawkes proposed a self/mutual-exciting process in his study [26]. The main characteristic of non-Markovian (with memory) self-exciting point processes is that the occurrence of any historical event affects the probability of future events for a long time and may lead to critical emergencies. The Hawkes process is an example of this. The standard linear self-excited Hawkes point process is a first-order non-Markovian stochastic model of intermittent explosive dynamics. The nonlinear Hawkes process is introduced to better describe the excitation (positive feedback) and inhibition (negative feedback) effects between events. The theoretical findings of [27] implied that the nonlinear Hawkes process has an asymptotically universal Zipf's law in the case of a mark distribution with zero mean.

In this study, we collect a new dataset containing various temporal data from openly accessible data from Bilibili. com. We propose that combined with the Danmaku data, the model based on the Hawkes process [28] can be used to evaluate the popularity of a video on Bilibili.com. In addition, the normalized popularity follows a power-law distribution, suggesting that it has a close

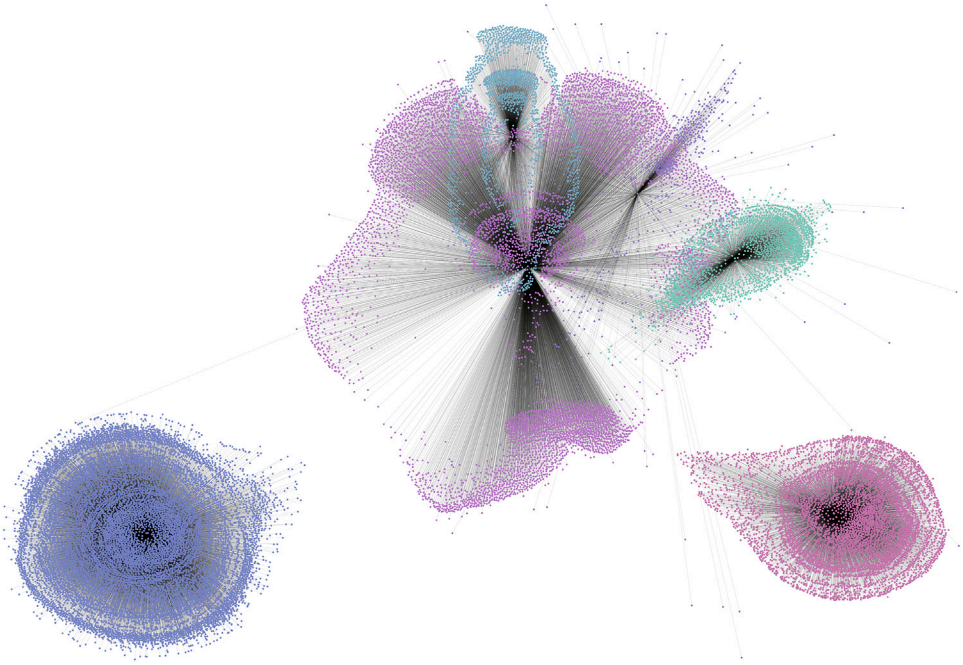


FIGURE 1
Clustering of a video–tag network. Each dot represents a video, and each color represents a unique tag cluster labeled by the uploader. The length represents the distance from the cluster center in the SVD space. The figures show there exist topics that are very distant from each other, and the topics have a wide diversity through the video–tag network. The random selection of videos by uploaders in general does not produce biases of the video content.

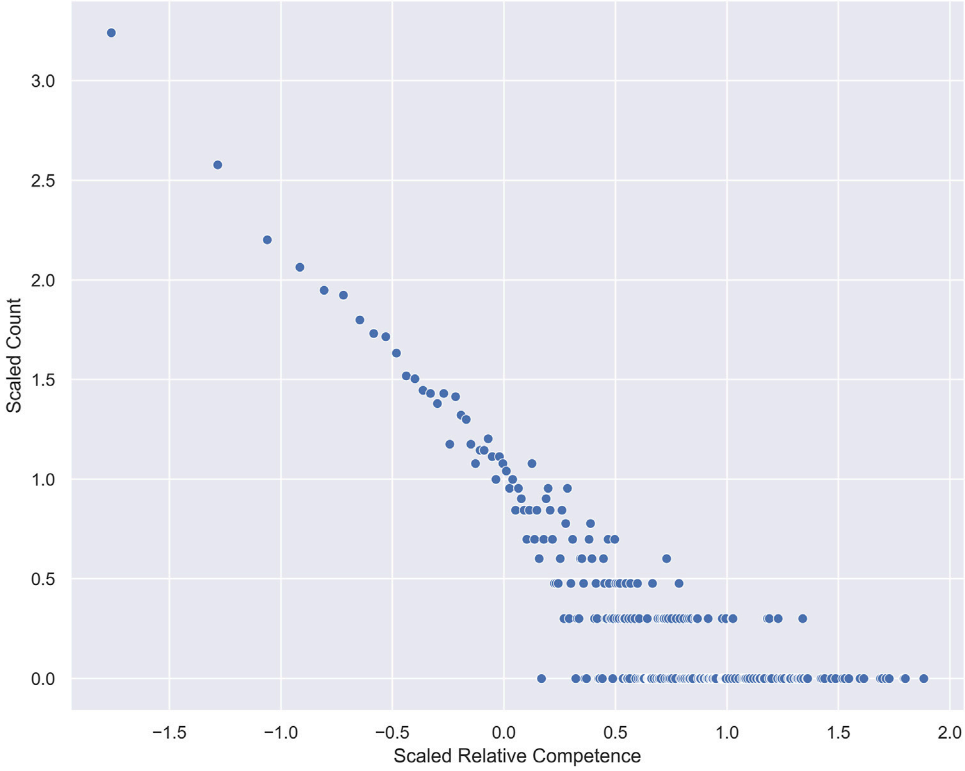


FIGURE 2
Power-law distribution of relative competence. The horizontal and vertical axes are scaled by logarithms.

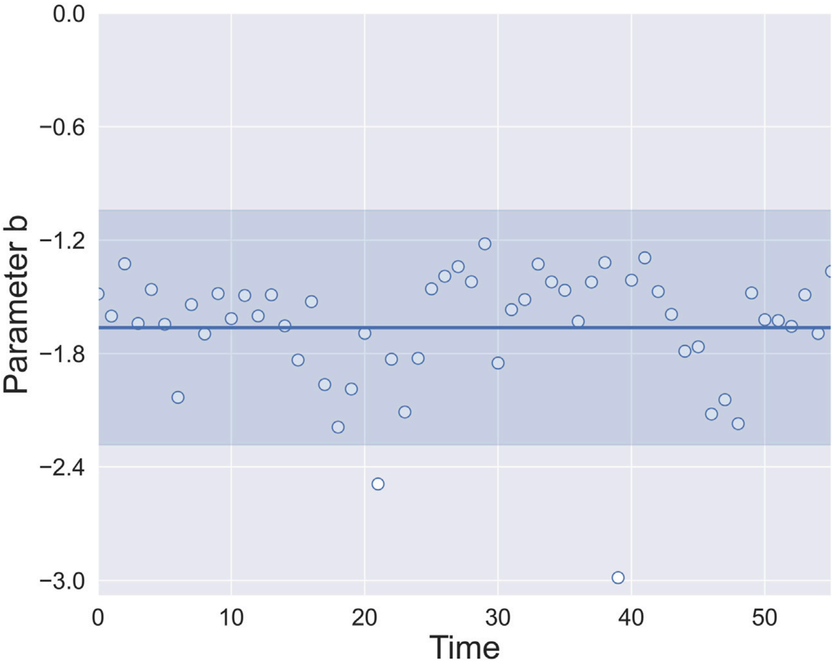


FIGURE 3
Power-law exponents of users' competence at different times. Parameter b is the exponents obtained by the previous power-law fitting. The solid line indicates the average of the power-law exponents, and the shaded area indicates the standard deviation of the power-law exponents. The data show parameter b is consistently around 1.5, disregarding the timestamps. This shows that relative competence exhibits a universal scaling behavior.

relationship with the self-exciting Hawkes process on the online video platform.

2 Materials and methods

2.1 Dataset description

We obtained the following datasets from the official website of Bilibili.Com.

2.1.1 User and video metadata

The dataset includes 224,672 uploaders, 1,425,882 videos, and 308,385 tags. The time span is from 2021.12 to 2022.2. There are a total of 1,425,882 records from the knowledge section of the Bilibili.com video website. There are 1,209,370 videos with tags, and the rest are videos without tags. Each record includes the following 17 fields, shown in Table 1.

2.1.2 Video views and Danmaku temporal data

This dataset is obtained from randomly selected 10,000 uploaders with more than 10,000 followers on the Bilibili website and the data of their latest released 50 videos. Videos that are too old will be replaced because uploaders will constantly publish new videos. The dataset starts from 2022.5.1 and ends at 2022.5.8 collects the number of videos played and the number of Danmaku of these videos every 3 hours, with a total of 56 timesteps. The data in each time step contain data from roughly 500,000 videos. The dataset is publicly available at https://github.com/luciidream/Universal-Scaling-behaviour-and-Hawkes-process-of- videos-views-on-Bilibili.com.git.

2.2 Hawkes process

The Hawkes process is a self-exciting point process [26, 29]. The process has arrivals at times $0 < t\_\{1\} < t\_\{2\} < t\_\{3\} < \cdots$ where the probability function of an arrival within $[t, t + dt)$:

$$\lambda_t dt = \left(\mu(t) + \sum_{t_i:\ t_i < t} \phi(t - t_1) \right) dt.$$

The function μ is the intensity of the underlying Poisson process. At time t_1, the first arrival occurs and then the intensity becomes $\mu(t) + \phi(t - t_1)$, and at the time t_2 of the second arrival, the intensity becomes $\mu(t) + \phi(t - t_1) + \phi(t - t_2)\}\ldots$ and so on [26].

During the time interval (t_j, t_{j+1}), the process represents the set of $j + 1$ independent processes with intensities $\mu(t), \phi(t - t_1), \ldots, \phi(t - t_j)$. The arrivals in the process whose

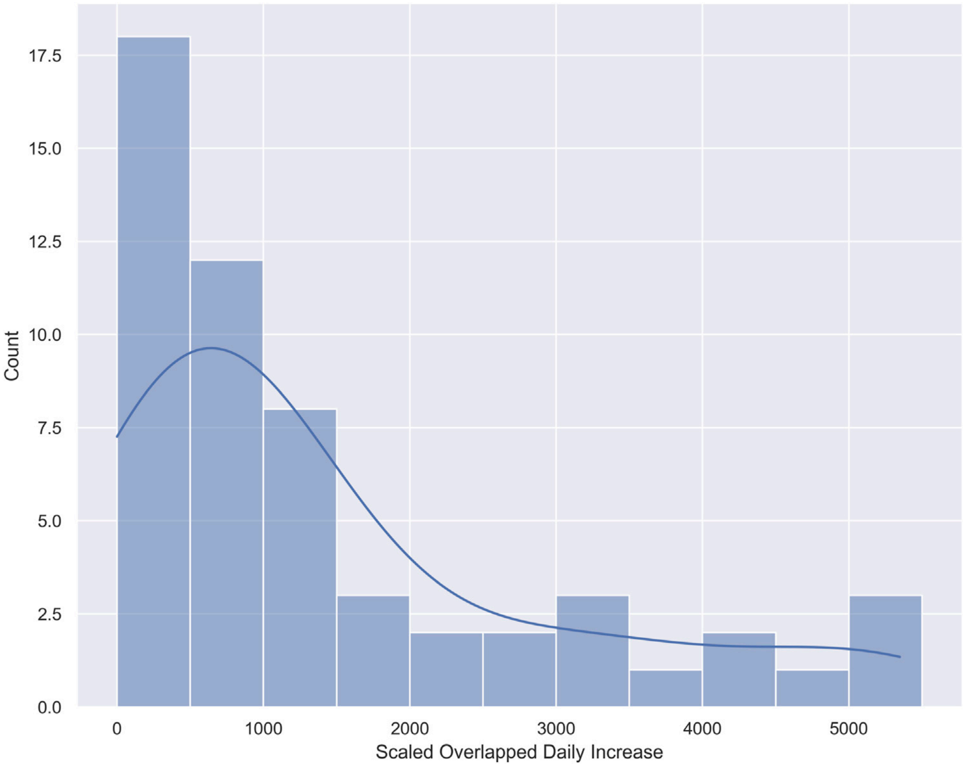


FIGURE 4
View counts of videos daily increase. The blue line is the indication line from a log-normal fitting. This is in accordance with the work of [31].

intensity is $\phi(t - t_j)$ are the descendants of the arrival at time t_k. The integral $\int_0^\infty \phi(t)\,dt$ is the average number of descendants of each arrival and is called the branching ratio. We consider the similar model used in [26], which regards the intensity function ϕ as a power-law decay over time.

$$\phi_m(\tau) = \kappa m^\beta (r + c)^{-(1+\theta)}, \tau \in \mathrm{R}^+,$$

where κ, β, c, and θ are the parameters that can be obtained from numerical fitting. In [28], parameter κ is the scaling factor that describes video quality, and m is the relative influence of the uploader. β measures the nonlinearity between the number of followers and popularity. $c > 0$ is the cutoff parameter that keeps ϕ bounded if τ is small. Finally, the exponents of the power-law distribution are given by $1 + \theta$.

3 Results

3.1 Tag clustering

In addition to the video label of the first section of the data, the corresponding one-hot vector of each video is obtained according to the label of each video. Then, all the vectors are combined to obtain a high-order matrix. We then perform singular value decomposition (SVD) on the matrix to reduce dimensionality and use the low-dimensional matrix to cluster videos. Connecting videos of the same class to the center point of the class constitutes a video network, as shown in Figure 1. As our videos are collected randomly from uploaders who have more than 10,000 followers, the figure is used to show the diversity of the topics and human tags and to validate that our data collection is of little bias referring to the topics or tags. The results indicate that there exist several clusters of topics of the videos we collected, and the dataset covers a broad range of topics and tags.

3.2 Uploader’s relative competence

We select uploaders who have released new videos in the same period and calculate the ratio of the growth of each person’s latest video at each moment to the total growth at that moment so as to express the influence of the uploader’s work relative to the works of other uploaders. This parameter shows the relative degree of attraction of uploaders toward the whole audience. The relative competitiveness of all creators is fitted according to the power-law distribution, and it is found that the relative competitiveness of the creators approximately obeys the power-law distribution [30], as shown in Figure 2.

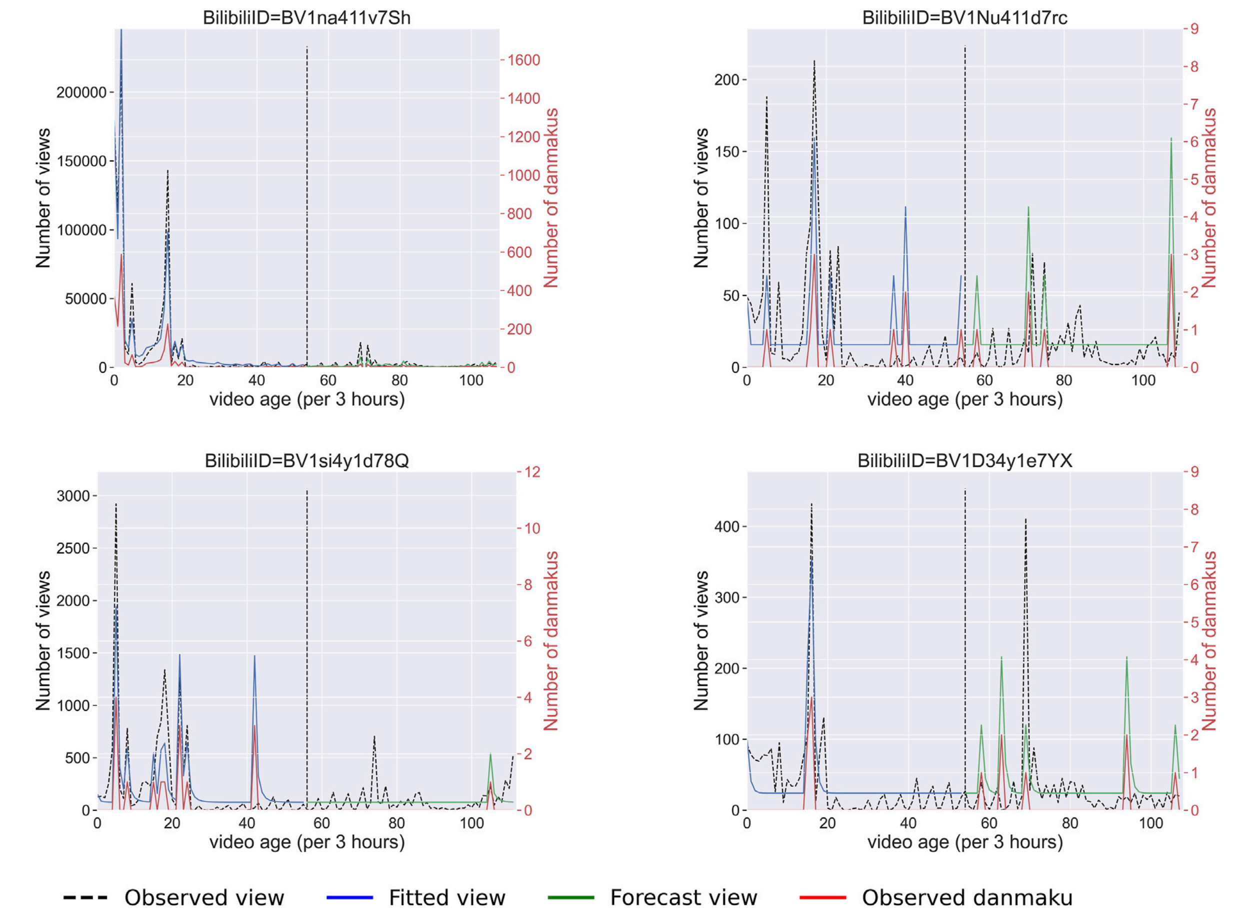


FIGURE 5
Hawkes model fitting on four Bilibili videos. The dashed line indicates where the prediction starts. The blue line is the fitted view curve, and the dashed line shows the original data. The green line is the prediction based on our Hawkes process model. The red line shows the Danmaku counts from the data.

We also estimate the power-law exponents, and surprisingly, the exponents through different times are all around 1.5, showing a universal scaling behavior, as illustrated in Figure 3. This result has been validated by statistical stability analysis.

Furthermore, using the video used in the previous section, we normalize the view counts recorded at each moment of each video, summing over the data of all videos and shifting the data to eliminate negative values and obtaining the average variation of the video views. This reveals the average daily views of each video and the mean field view pattern of the videos during their life span, as shown in Figure 4. For simplicity, the curve is well-fitted by a log-normal function, which is in accordance with the suggestion from [31].

3.3 Hawkes process model

From the previous introduction of the model, a point process model based on the Hawkes process is used to predict video views. Compared with feature-based prediction methods, the counting process does not require complex feature engineering and additional training, so it can be easily applied to real-time systems. The results show that, with the Danmaku data from the temporal dataset, our model can predict the popularity precisely even in the far future. The results are shown in Figures 5, 6.

Figure 5 shows the Hawkes model results of four randomly picked videos that are representative of the wellness of fitting. The videos have very different fundamental statistics and thus explain our method is general. The dashed line is the separation line of data used to train the model and obtain parameters and the test data that are used for evaluation of the results. The results show that our model can accurately predict the location of peaks in the video views in the far future, but the accuracy of the intensity of the peak varies from video to video.

Figure 6 shows the results of the Hawkes model fitting on the same video by varying the size of the training set. The four subplots represent the results using 20%, 30%, 40%, and 50% of

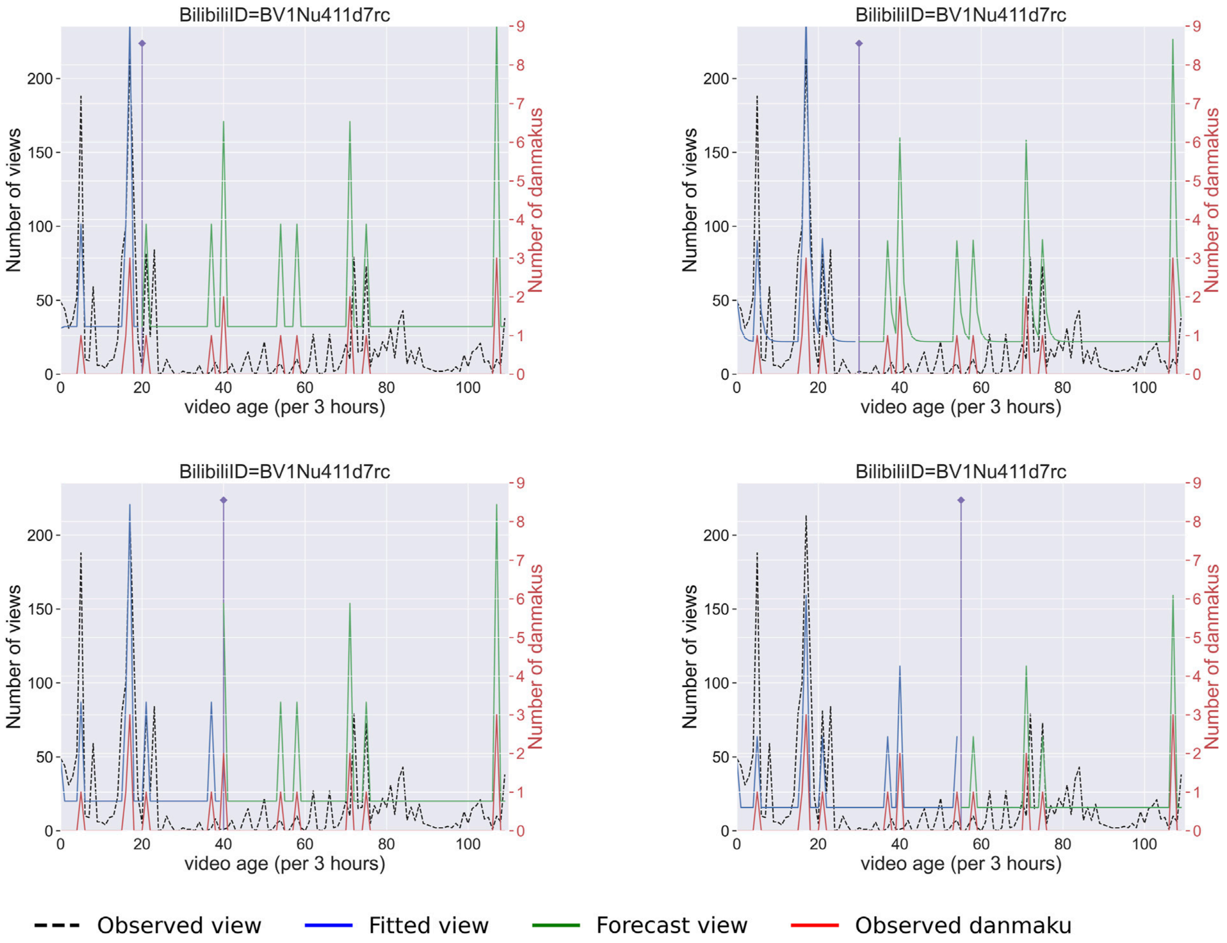


FIGURE 6
Predictions on the same video Bilibili ID BV1Nu411d7rc by varying the size of the training set. The four subplots represent different results corresponding to 20%, 30%, 40%, and 50% as training sets. The purple line indicates where the prediction starts. The blue line is the fitted view curve, and the dashed line shows the original data. The green line is the prediction based on our Hawkes process model. The red line shows the Danmaku counts from the data.

the time span as training sets, respectively. The results indicate that the predicted position of peaks by our method is resilient with different sizes of training sets. Also, the predicted intensity can improve with a larger training set.

It is worthwhile to mention that by adopting the model from [28], the parameters obtained from the fitting are not stable, as shown in Figure 7, suggesting the correlation of the six parameters. Further investigation can be conducted in this direction to clarify how these parameters influence each other.

4 Discussion

In this study, our results show that the view counts of online videos on Bilibili.com are significantly determined by the point processes, such as the Hawkes process. The relative view counts during a given time span among all videos (namely, the uploaders' relative competence) follow a universal scaling behavior. The video view model based on the Hawkes process effectively realizes the simultaneous prediction of video views. On one hand, it is conducive to the accurate placement of advertisements and improves commercial service quality, and on the other hand, by analyzing the results, it can provide decision support for content management or provide a basis for network storage optimization or expansion, low network storage utilization, and avoid problems such as large capacity redundancy caused by expansion, thereby reducing operating costs and producing practical application value.

Sornette et al. [27] studied the relationship between the universal power-law distribution and the nonlinear Hawkes process and proved that the Hawkes process can form the power-law distribution of events through theoretical analysis. This effect is

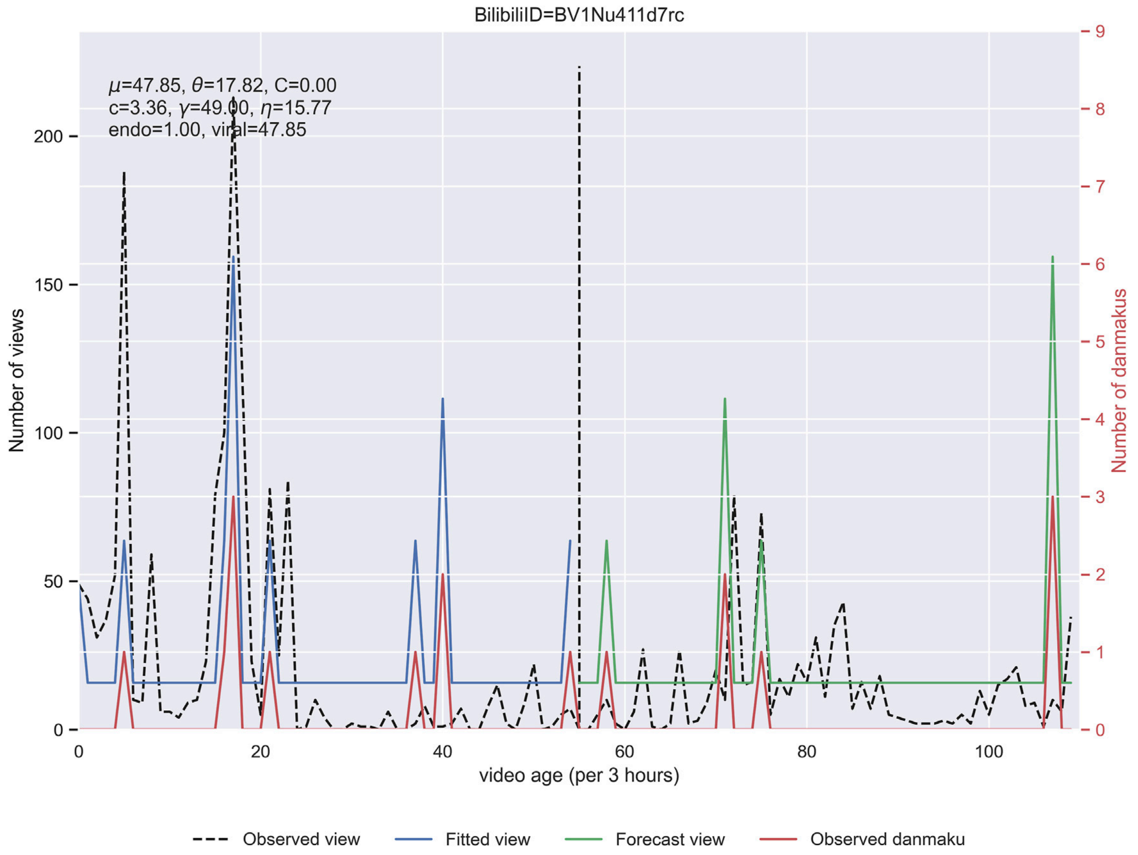


FIGURE 7
Parameters in the model fitting on video Bilibili ID BV1D34y1e7YX. The parameters in the fitting are listed at the top left corner. Parameters γ and η indicate the estimation of unobserved influence out of the data. Parameter endo describes the intrinsic popularity of videos. Parameter viral is a measurement of whether the video will become immensely popular and is a combined effect of both intrinsic and external impact.

also observed in our data, where the video views at different times and the relative competence indicators of uploaders also form a power-law distribution with a relatively stable power-law behavior. Sornette’s theory states that the Hawkes process resembles a nonlinear self-excited process, in which the properties arise from a complex interaction between a multiplicative process, memory, and endogeneity or reflexivity. The Hawkes model fitting results show that in our context, a video’s current views and its historical view counts, as well as the number of Danmaku comments, have a nonlinear mutualistic interacting process. The specific dynamics of the process is complex and awaits further study.

In summary, we collected a new set of data, which recorded the video view data on Bilibili.com, including the temporal records of video views, the number of Danmaku comments, and the metadata of uploaders and videos. The social networks formed by these data will help further explore the complex interaction between videos, users, and Danmaku comments. It also helps develop new models and verify existing theories, which contribute to the knowledge of understanding social interactions and networks. The Hawkes model analysis of the time-series data shows that there exists a Hawkes process with memory characteristics with the video views on Bilibili.com, and at the same time, the consistent power-law distribution characteristics are observed in many data statistics. This connection can be enlightened by the related theory by Sornette et al. Despite the analytical analysis that is specifically based on three assumptions, the general explanation of how the Hawkes process can produce power-law distribution suggests the possibility of a direct causal relationship, irrespective of the theoretical assumptions made in [27]. In the near future, a more general model can be developed to further clarify the phenomena with this newly collected dataset.

Data availability statement

The dataset is publicly available at https://github.com/luciidream/Universal-Scaling-behaviour-and-Hawkes-process-of- videos-views-on-Bilibili.com.git.

Author contributions

YK designed the research; HD and JD collected the data; JD, YH, and JS analyzed the data; HD and JD performed the visualization; all authors wrote the manuscript.

Funding

This work was supported by Research Funds for the Central Universities from the Beijing University of Posts and Telecommunications under grant nos. 505022019 and 500422415 and Research Innovation Fund for College Students of the Beijing University of Posts and Telecommunications.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Yu H, Zheng D, Zhao BY, Zheng W. Understanding user behavior in large-scale video-on-demand systems. *SIGOPS Oper Syst Rev* (2006) 40(4):333–44. doi:10.1145/1218063.1217968

2. Avramova Z, Wittevrongel S, Bruneel H, De Vleeschauwer D. Analysis and modeling of video popularity evolution in various online video content systems: Power-law versus exponential decay. In: Proceedings of the 2009 First International Conference on Evolving Internet; 23-29 August 2009; Cannes/La Bocca, France (2009). doi:10.1109/INTERNET.2009.22

3. Niu D, Liu Z, Li B, Zhao S. Demand forecast and performance prediction in peer-assisted on-demand streaming systems. In: Proceedings of the 2011 Proceedings IEEE INFOCOM; 10-15 April 2011; Shanghai, China (2011). doi:10.1109/INFCOM.2011.5935196

4. Mahajan V, Muller E, Bass FM. New product diffusion models in marketing: A review and directions for research. *J Marketing* (1990) 54(1):1–26. doi:10.2307/1252170

5. Hu W, Wang Z, Ma M, Sun LF. Edge video cdn: A wi-fi content hotspot solution. *J Comput Sci Technol* (2016) 31(6):1072–86. doi:10.1007/s11390-016-1683-x

6. Szabo G, Huberman BA. Predicting the popularity of online content. *Commun ACM* (2010) 53(8):80–8. doi:10.1145/1787234.1787254

7. Lerman K, Hogg T. Using a model of social dynamics to predict popularity of news. '10. In: Proceedings of the 19th International Conference on World Wide Web; April 2010; Raleigh, North Carolina, USA (2010). doi:10.1145/1772690.1772754

8. Figueiredo F, Benevenuto F, Almeida JM. The tube over time: Characterizing popularity growth of youtube videos. In: '11, Proceedings of the Fourth ACM International Conference on Web Search and Data Mining; February 2011; Hong Kong, China (2011). doi:10.1145/1935826.1935925

9. Lee JG, Moon S, Salamatian K. Modeling and predicting the popularity of online contents with Cox proportional hazard regression model. *Neurocomputing* (2012) 76(1):134–45. doi:10.1016/j.neucom.2011.04.040

10. Bandari R, Asur S, Huberman BA. The pulse of news in social media: Forecasting popularity. *Proc Int AAAI Conf Web Soc Media* (2012) 6(1): 26–33. doi:10.48550/arXiv.1202.0332

11. Ahmed M, Spagna S, Huici F, Niccolini S. A peek into the future: Predicting the evolution of popularity in user generated content. *Proc sixth ACM Int Conf Web search Data mining* (2013) 607–16. doi:10.1145/2433396.2433473

12. Asur S, Huberman BA, Szabo G, Wang C. Trends in social media: Persistence and decay. *Proc Int AAAI Conf Web Soc Media* (2011) 5(1):434–7. doi:10.48550/arXiv.1102.1402

13. Wu J, Zhou Y, Chiu DM, Zhu Z. Modeling dynamics of online video popularity. *IEEE Trans Multimedia* (2016) 18(9):1882–95. doi:10.1109/tmm.2016.2579600

14. Ratkiewicz J, Fortunato S, Flammini A, Menczer F, Vespignani A. Characterizing and modeling the dynamics of online popularity. *Phys Rev Lett* (2010) 105(15):158701. doi:10.1103/physrevlett.105.158701

15. Pinto H, Almeida JM, Gonçalves MA. Using early view patterns to predict the popularity of youtube videos. In: '13, Proceedings of the Sixth ACM International Conference on Web Search and Data Mining; February 2013; Rome, Italy (2013).

16. Hennig-Thurau T, Wiertz C, Feldhaus F. Does twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *J Acad Mark Sci* (2015) 43(3):375–94. doi:10.1007/s11747-014-0388-3

17. Matsubara Y, Sakurai Y, Prakash BA, Li L, Faloutsos C. Rise and fall patterns of information di-usion: Model and implications. KDD. In: 12 Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2012; Beijing, China (2012).

18. Li C, Liu J, Ouyang S. Characterizing and predicting the popularity of online videos. *IEEE Access* (2016) 4:1630–41. doi:10.1109/access.2016.2552218

19. Tatar A, De Amorim MD, Fdida S, Antoniadis P, A survey on predicting the popularity of web content[J]. *J Internet Serv Appl*, 2014, 5(1): 8, doi:10.1186/s13174-014-0008-y

20. Deng Z, Yan M, Sang J, Xu C. Twitter is faster: Personalized time-aware video recommendation from twitter to YouTube. *ACM Trans Multimedia Comput Commun Appl* (2015) 11(2):31–3123. doi:10.1145/2637285

21. Xu J, Van Der Schaar M, Liu J, Li H. Forecasting popularity of videos using social media. *IEEE J Sel Top Signal Process* (2015) 9(2):330–43. doi:10.1109/jstsp.2014.2370942

22. Fontanini G, Bertini M, Del Bimbo A. Web video popularity prediction using sentiment and content visual features. '16. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval; June 2016; New York, USA (2016).

23. Hoiles W, Aprem A, Krishnamurthy V. Engagement and popularity dynamics of YouTube videos and sensitivity to meta-data. *IEEE Trans Knowl Data Eng* (2017) 29(7):1426–37. doi:10.1109/tkde.2017.2682858

24. Trzciński T, Rokita P. Predicting popularity of online videos using support vector regression. *IEEE Trans Multimedia* (2017) 19(11):2561–70. doi:10.1109/tmm.2017.2695439

25. Chen X, Chen J, Ma L, Yao J, Liu W, Luo J, Zhang T. Fine-grained video attractiveness prediction using multimodal deep learning on a large real-world dataset. In: 18 Companion Proceedings of the Web Conference 2018; April 2018; Lyon, France (2018).

26. Hawkes AG. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* (1971) 58(1):83–90. doi:10.1093/biomet/58.1.83

27. Kanazawa K, Sornette D. Ubiquitous power law scaling in nonlinear self-excited Hawkes processes. *Phys Rev Lett* (2021) 127(18):188301. doi:10.1103/physrevlett.127.188301

28. Rizoiu MA, Xie L, Sanner S, Cebrian M, Yu H, Van Hentenryck P. Expecting to be HIP: Hawkes intensity processes for social media popularity. In: Proceedings of the 26th International Conference on World Wide Web; April 2017; Australia (2017).

29. Laub PJ, Lee Y, Taimre T. *The elements of Hawkes processes.* (2022). Springer Nature.

30. Albert R, Barabási AL. Statistical mechanics of complex networks. *Rev Mod Phys* (2002) 74(1):47–97. doi:10.1103/revmodphys.74.47

31. Wang D, Song C, Barabási AL. Quantifying long-term scientific impact. *Science* (2013) 342(6154):127–32. doi:10.1126/science.1237825

frontiers | Frontiers in Physics

TYPE Brief Research Report
PUBLISHED 11 November 2022
DOI 10.3389/fphy.2022.981650

Check for updates

OPEN ACCESS

EDITED BY
Gaogao Dong,
Jiangsu University, China

REVIEWED BY
Zhao-Long Hu,
Zhejiang Normal University, China
Jianhong Lin,
ETH Zürich, Switzerland

*CORRESPONDENCE
Na Chen,
332904003@qq.com

SPECIALTY SECTION
This article was submitted to Social Physics,
a section of the journal
Frontiers in Physics

RECEIVED 29 June 2022
ACCEPTED 18 August 2022
PUBLISHED 11 November 2022

CITATION
Chen N, Liu J, Chen Y, Tu D, Ou Y and Jiang M (2022), Core-middle-periphery network model for China banking system.
Front. Phys. 10:981650.
doi: 10.3389/fphy.2022.981650

Core-middle-periphery network model for China banking system

Na Chen[1]*, Jianguo Liu[2], Yihui Chen[3], Ding Tu[4], Yang Ou[5] and Mingzhu Jiang[6]

[1]School of Economics, Fudan University, Shanghai, China, [2]Institute of Accounting and Finance, Shanghai University of Finance and Economics, Shanghai, China, [3]Independent Researcher, Shanghai, China, [4]China Foreign Exchange Trade System, Shanghai, China, [5]Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai, China, [6]School of Management, Shanghai University, Shanghai, China

The banking system could be mapped by the network model to generate the structural properties of evolution dynamics. In this study, we empirically investigate the evolution properties of the China bank network from 2008 to 2019 where the banks and lending relationships are set as the nodes and links. By introducing the middle layer into the core–periphery (CP) model, we present the core–middle–periphery (CMP) model where the nodes belonging to the core layer are fully connected and the ones belonging to the middle layer connect the core and periphery layer. Compared with the traditional CP model, the reconstruction error of the CMP model is decreased by 64% compared with the one obtained by the CP model, and the transition stability probability is enhanced greatly. This work is helpful for deeply understanding the evolution properties of the banking system.

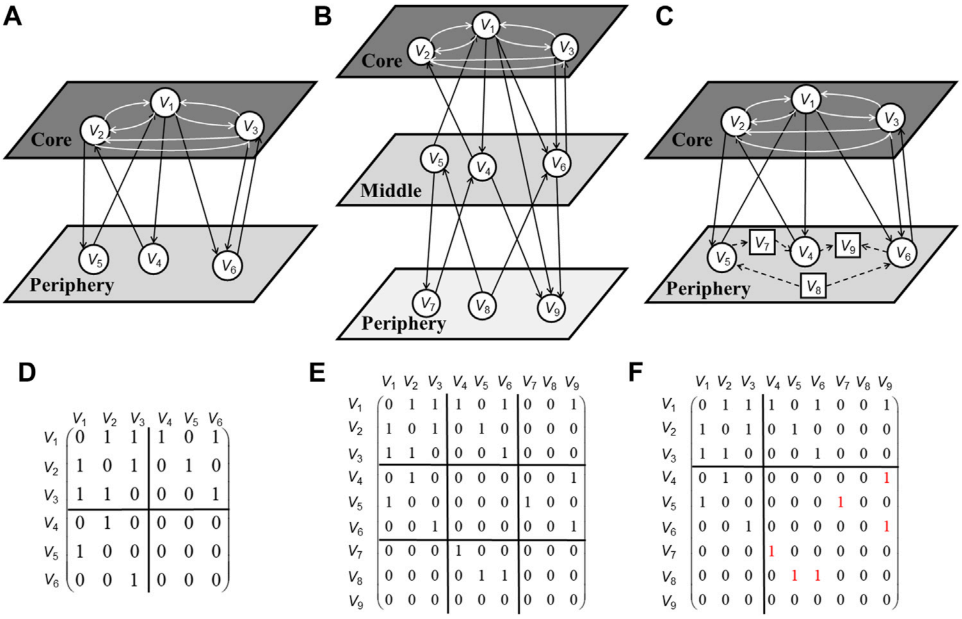
KEYWORDS

bank network, network structure, China banking system, core–periphery model, core–middle–periphery model

Introduction

The network structure plays a critical role in network resilience and risk transmission (Allen and Gale [1], Cassar et al. [2], Battiston et al. [3], Hu et al. [4]). The interbank market is the first step in the transmission of monetary policy to the rest of the financial system and the real economy, which could be mapped by the banking network where the banks and their interconnectedness are set as the nodes and links. The bilateral relationships between banks (Fang et al. [5]) play an important role in the China banking system. The interconnectedness of banks in the real world does not necessarily correlate with the size of their assets (Martinez et al. [6]), and the lending relationship that constitutes the banking network is crucial to liquidity management and risk contagion in the interbank market (Angelini [7]; Furfine [8]; Iori et al. [9]). A dense and complex network of bank liability can easily transmit risk to the whole market, giving rise to systemic risk.

The US federal funds (Bech and Atalay [10]) and the Austrian interbank lending networks (Boss et al. [11]) have the small-world properties. Empirical studies find that



$$\begin{array}{c|ccc|ccc} & V_1 & V_2 & V_3 & V_4 & V_5 & V_6 \\ \hline V_1 & 0 & 1 & 1 & 1 & 0 & 1 \\ V_2 & 1 & 0 & 1 & 0 & 1 & 0 \\ V_3 & 1 & 1 & 0 & 0 & 0 & 1 \\ \hline V_4 & 0 & 1 & 0 & 0 & 0 & 0 \\ V_5 & 1 & 0 & 0 & 0 & 0 & 0 \\ V_6 & 0 & 0 & 1 & 0 & 0 & 0 \end{array}$$

$$\begin{array}{c|ccc|ccc|ccc} & V_1 & V_2 & V_3 & V_4 & V_5 & V_6 & V_7 & V_8 & V_9 \\ \hline V_1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ V_2 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ V_3 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline V_4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ V_5 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ V_6 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline V_7 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ V_8 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ V_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

$$\begin{array}{c|ccc|cccccc} & V_1 & V_2 & V_3 & V_4 & V_5 & V_6 & V_7 & V_8 & V_9 \\ \hline V_1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ V_2 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ V_3 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline V_4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ V_5 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ V_6 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ V_7 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ V_8 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ V_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

FIGURE 1
Illustration of the core–periphery and core–middle–periphery models. **(A)** shows the schematic diagram of the idealized core-periphery structure. **(D)** shows the adjacency matrix corresponding to **(A)**. **(B)** gives the schematic diagram of the idealized core–middle–periphery model with three core nodes, three middle, and three periphery nodes. **(E)** shows the adjacency matrix corresponding to **(B)**. **(C)** shows a possible result of modelling the market structure in **(B)** by the core–periphery model without the middle layer. **(F)** shows the adjacency matrix corresponding to **(C)** with red "1"s indicating the errors in the periphery layer.

the interbank market networks in the United States (Soramaki [12]), Japan (Inaoka et al. [13]), Austria (Boss et al. [14]), Brazil (Edson et al. [15]), and Mexico (Martinez et al. [6]) approximately obey the power–law degree distributions. Borgatti et al. [16] proposed the core–periphery (CP) model to regenerate the empirical interbank system with a dense cohesive core and a sparse periphery (as shown in the diagonal blocks **1** and **0** in Figure 1D). Furthermore, Craig et al. [17] investigated the connection pattern of the nodes in core and periphery layers (as indicated in the row-regular and column-regular off-diagonal blocks in Figure 1D) and empirically found that the German interbank market network showed a core–periphery structure. For the UK interbank credit exposure network, Langfield et al. [18] empirically found the core-periphery structure, which is also observed in the Italian overnight interbank lending network [19]. Brassil and Nodari [20] presented a density-based CP model to analyze the Australian interbank overnight lending network, which is consistent with the empirical data more closely than the traditional CP model. Yang et al. [21] investigated risk contagion in the China banking system by using the core–periphery structure. By adjusting the density of the core network, Xing et al. [22] presented an improved CP network model to regenerate the lending network in terms of the balance sheet, which could generate a lending network with a more closely connected core subnetwork.

All the aforementioned studies suppose that the two-layer core-periphery analytical framework captures the evolution mechanism of the banking system. Apart from the above studies, there are relatively few studies on the true stratification structure of the China banking system. One reason is the limitation of available underlying microdata. The other reason is the neglect of the real-world complex lending relationships between banks in liquidity distribution, which may contribute to the inaccuracy of depicting the banking network as a typical core-periphery structure. Actually, a special tiering structure existed in China's financial system and the banking system in China could be naturally divided into at least three layers from the market function perspective and regulatory perspective. The core banks generally are flush with liquidity, whereas the periphery banks tackling with long-term funding gap lack a stable relationship with the core banks. The banks in the middle layer help to finance the periphery banks and improve the liquidity redistribution between the core and the periphery in the network. Inspired by the above ideas, we propose a three-layer network model, namely the core-middle-periphery (CMP) model, to analyze the Chinese banking system.

In this study, we empirically investigate the liability network between banks in China spanning 2008–2019, and the statistical results show that the average path length of the network lies in the interval of [2, 3], suggesting that most of the nodes are more than two steps away from each other, so the two-tier core-periphery structure may not be sufficient to portray the complex market hierarchy. The neighbor connectivity and degree distribution also indicate existence of a large number of medium-degree nodes between large-degree nodes and small-degree nodes in the network. In the CMP model, the nodes belonging to the core layer are connected tightly with each other, and the nodes in the middle layer bridge the funding gap between the nodes in the core and periphery layers. We do not put strict internal structural constrain on the banks in the middle layer and encourage more lending activities from the middle to the periphery.

According to the definition of the CMP model, we present an algorithm to separate China bank nodes into three layers correspondingly. The empirical results show that the CMP model captures the crucial characteristic of the China banking structure. Moreover, compared with the traditional CP model, the CMP model detects a narrower subset of core bank nodes, which is conducive to detecting core bank groups in the China banking system more accurately.

Figure 1A shows the schematic diagram of the idealized core-periphery structure proposed by Craig et al. [17]. Figure 1D shows the adjacency matrix $\mathbf{A}_{\mathbf{C-P}}$ corresponding to Figure 1A. The arrow of a link indicates the direction of credit exposure, and we set the presence or absence of a link by 1 or 0. Node V_1, V_2, and V_3 belong to the core subset **C**, which could connect with each other. Node V_4, V_5, and V_6 belong to the periphery subset **P**. They do not lend to each other, so the periphery block of $\mathbf{A}_{\mathbf{C-P}}$ is a square matrix of 0. The periphery nodes only trade with the core nodes, and each core node borrows from and lends to at least one periphery node. Accordingly, to simplify the algorithm procedure, Craig et al. [17] set the off-diagonal blocks of idealized $\mathbf{A}_{\mathbf{C-P}}$ as "row regular" with at least one link in every row and "column regular" with at least one in every column, as indicated in Figure 1D.

Figure 1B gives the schematic diagram of the idealized CMP model with three core nodes, three middle, and three periphery nodes. Figure 1E shows the adjacency matrix $\mathbf{A}_{\mathbf{C-M-P}}$ corresponding to Figure 1B, where node V_1, V_2, and V_3 belong to the core layer subset **C**, which connect with each other, so the core block of $\mathbf{A}_{\mathbf{C-M-P}}$ is a square matrix (ignoring the zero diagonal as we exclude self-loops, which indicates that banks could not lend to themselves). Node V_4, V_5, and V_6 belong to the middle layer subset **M**, and the links between middle nodes are unconstrained; In other words, they may either lend to or not lend to each other. Node V_7, V_8, and V_9 belong to the periphery subset **P**, which only have connections with nodes in the middle layer and have no connections among themselves. Therefore, the periphery block of $\mathbf{A}_{\mathbf{C-M-P}}$ is a square matrix of 0. The remaining blocks depict the links between the core, middle and periphery nodes. Each core node borrows from and lends to at least one node in the middle or periphery layers, whereas each node in the middle layer borrows from and lends to at least one node in the periphery layer; Accordingly the related blocks are "row regular" or "column regular", as indicated in Figure 1E. Moreover, it is much more preferable if nodes in the middle subset lend out to more nodes in the periphery subset, which may facilitate the financing of the periphery layer.

Figure 1C,F show that compared with modeling by the idealized CMP model, if we model the China banking network by the CP model without the middle layer, node V_1, V_2, and V_3 will still belong to the core bank subset **C**, whereas node V_4, V_5, V_6, V_7, V_8, and V_9 are all classified into the periphery subset **P**. In this study, we argue that the CP model is insufficient to portray the complex tiering structure with more than two layers. It should be noted that although in this example errors only arise from the periphery, there are cases that errors arise from both the core and periphery layers.

Statistical properties of empirical data

Network definition

Based on bilateral liability data between China banks, the China banking network is given as following: $\mathbf{G}_{\mathbf{real}}(\mathbf{V}, \mathbf{E})$ consists of $|\mathbf{V}| = N$ nodes and $|\mathbf{E}| = M$ unweighted-directed edges, where $\mathbf{V} = \{v_1, v_2, \ldots, v_N\}$ denotes the set of bank nodes and $\mathbf{E} = \{e_{ij}|i, j \in \mathbf{N}\}$ illustrates the lending relationships between banks. The topology of the China banking network can also be represented by the adjacency matrix $\mathbf{A}_{\mathbf{real}} = \{a_{ij}\}_{N\times N}$, where $a_{ij} = 1$ means a link going from node v_i to node v_j (i.e., bank v_i lends to bank v_j); otherwise, $a_{ij} = 0$.

Data description

In this study, we collect bilateral liability data between China banks from 1 January 2008 to 31 December 2019 to construct a bank network. In this study, we construct the lending network for each year, and finally give an ensemble network.

Empirical results

In this study, we start by providing a set of network metrics (for definition and calculation of metrics refer to Barabasi [23]) for the China banking system in order to analyze its statistical regularities and topological properties, which may affect the resilience and risk contagion of the network.

The results of the metrics (Table 1) show that the average clustering coefficient (ACC) of the network is significantly higher than that of a random network of the same dimension, and the average path length (APL) is significantly lower than $\ln N$ while close to $\ln\ln N$, suggesting that the network may have thick-tailed characteristics and follow small-world properties (Barabasi [23], Manoj et al. [24]), indicating that the bank nodes are able to lend to each other through a small number of steps. The average neighborhood degree $\langle k \rangle$ expands rapidly as the market developed over years. However, the APL, distance (D), and density (S) stay small, suggesting that the network does not become sparser when it becomes larger.

From a regulatory perspective, there exists a special tiering structure in China's financial system. The large-size banks generally are flush with liquidity, whereas small and medium-size banks tackling with long-term funding gaps. However, owing to its high credit risk, asymmetric market information, and disparity of lending volume between large and small-size banks, a portion of small and medium-size banks lack stable lending relationships with large-size banks. Therefore, the banking system in China is naturally divided into three layers: the first layer is consist of the central bank's open market operation primary dealers and money market makers, which form the core of the banking system; the second layer is consist of qualified financial institutions with sound internal control, which are the main participants of the banking system; the third layer is the small and medium-sized financial institutions that rely on funding relationships with the first two tiers of financial institutions. Therefore, we present the CMP model to regenerate the China banking system with three layers.

Further statistical results show that nodes with large-degree have smaller clustering coefficients and larger betweenness, which indicates that as the intermediary property of large-degree nodes strengthens, the probability of lending among their "neighbors" decreases and thus the "cluster community" is reduced. The APL of the network lies in the interval of (2, 3), indicating that most of the nodes are more than two steps away from each other, suggesting that the two-layer core-periphery structure may not be sufficient to regenerate the hierarchal structure. From the statistical results of the China banking network, the small-degree nodes or node with fewer neighbors tend to connect with nodes with a similar degree, which lead to positive assortativity. On the other hand, when the node degree is large, they tend to trade with small-degree nodes, leading to negative assortativity. Such a phenomenon suggests that bank nodes may contain three subsets with different properties. The degree distribution also shows that, although the network degree distribution has a fat-tailed feature with more small-degree nodes and fewer large-degree nodes, a large number of medium-degree nodes exist in the network, which implies a more complex market hierarchy (Brassil and Nodari [20]).

Inspired by the aforementioned analysis, this study constructs the CMP model containing three subsets of bank nodes and proposes an algorithm to divide the bank nodes into the corresponding three layers. The total error function and the transition probability matrix show that the CMP model and the algorithm could capture the evolution mechanism more accurately.

Model

Traditional idealized core–periphery model

The idealized CP model [17] sets the bank liability relationships as directed networks and defines the ideal CP network $\mathbf{G_{C-P}(V,E)}$ consisting of two subsets of nodes, namely, the core bank layer $\mathbf{C} = \{c_1, c_2, \ldots, c_{N_c}\}$ and the periphery bank layer $\mathbf{P} = \{p_1, p_2, \ldots, p_{N_p}\}$, where $c_{N_c} \in \{2, 3 \ldots, N-1\}$, $N = N_c + N_p$. The idealized CP model's adjacency matrix is $\mathbf{A_{C\text{-}P}} = \{a_{ij}\}_{N\times N}$, in which $a_{ij} = 1$ denotes node v_i lending to v_j, whereas $a_{ij} = 0$ indicates no lending relationship between node v_i and v_j.

For any given N_c, an idealized CP model (see Figure 1A,D) is constructed as follows: Step 1, a number of banks(N_c) in the core layer lend to and borrow from each other; Step 2, the remaining ($N - N_c$) periphery banks do not lend to or borrow from each other; Step 3, each core bank lends to at least one periphery bank node; Step 4, each core bank borrows from at least one periphery bank node.

The adjacency matrix $\mathbf{A_{C-P}}$ of idealized CP structure can be represented as four blocks, namely, $\mathbf{A_{CC}}$, $\mathbf{A_{PC}}$, $\mathbf{A_{CP}}$, and $\mathbf{A_{PP}}$ (Eq. 1). The core submatrix $\mathbf{A_{CC}}$ is an $N_c{}^*N_c$ matrix of ones with zero diagonals, representing the existence of directed links between all core nodes. Periphery banks do not transact with each other, so the submatrix $\mathbf{A_{PP}}$ is an $(N - N_c)^*(N - N_c)$ matrix of zeros. Since the most restricted form of the off-diagonal matrix is a 1-block, which is the upper limit of relationships between core and periphery and too rare to be found in the real-world network, Craig et al. [17] set the threshold of detecting the core-periphery relationship by adopting the off-diagonal blocks in the idealized core–periphery model as row regular and column regular. To be more specific, $\mathbf{A_{CP}}$ (core lending to periphery) block is a row regular (RR) submatrix, suggesting that it contains at least one link in every row. Similarly, since each core bank borrows from at least one periphery bank, the $\mathbf{A_{PC}}$ submatrix is a column regular (CR) matrix with at least one in every column. Figure 1D is an example of the adjacency matrix of the CP model with $N = 6$ and $N_c = 3$.

$$\mathbf{A_{C-P}} = \begin{pmatrix} \mathbf{A_{CC}} & \mathbf{A_{CP}} \\ \mathbf{A_{PC}} & \mathbf{A_{PP}} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{RR} \\ \mathbf{CR} & \mathbf{0} \end{pmatrix}. \tag{1}$$

The $\mathbf{A_{C-P}}$ is the benchmark for evaluating whether there is a core–periphery structure in the adjacency matrix $\mathbf{A_{real}}$ of the observed network $\mathbf{G_{real}}(\mathbf{V}, \mathbf{E})$. As the real-world networks are unlikely to match the ideal theoretical CP structure exactly, the objective of the empirical analysis is to apply an algorithm to find the optimal set of $\mathbf{C}$ in $\mathbf{A_{real}}$, which achieves the best structural match between $\mathbf{A_{real}}$ and $\mathbf{A_{C-P}}$ which both contain N_c core nodes. To be more specific, for any given assignment of banks into $\mathbf{C}$ and $\mathbf{P}$ subset, the structural inconsistencies between $\mathbf{A_{real}}$ and its nearest tiering model $\mathbf{A_{C-P}}$ could be measured by an error function. The error describes differences in each block between an ideal model and the real-world network. The total error is the result of summing up and standardized errors in all the blocks between an ideal model and the real-world network. The error function is the equation used to obtain the total error.

In order to measure the difference between the empirical network structure $\mathbf{A_{real}}$ and the idealized traditional CP model $\mathbf{A_{C-P}}$, firstly we define the difference $\mathbf{E}$ comprising of four elements, in which the sum of all missing links (outside the diagonal) in the core layer is defined as $E_{\mathbf{CC}}$, the cumulative value of all observed links among the periphery layer is defined as $E_{\mathbf{PP}}$, and off-diagonal errors are calculated by $E_{\mathbf{CP}}$ and $E_{\mathbf{PC}}$, respectively. To be more specific, for the off-diagonal blocks, a zero row in $\mathbf{A_{CP}}$ indicates that a core bank does not lend to any of the $(N - N_c)$ banks in the periphery layer, which is not consistent with the defined feature of core banks, resulting in an error of $(N - N_c)$*1 in this row, and errors in each row add up to $E_{\mathbf{CP}}$. Similarly, a zero column in $\mathbf{A_{PC}}$ shows that this core bank does not borrow from any periphery, resulting in an error of $(N - N_c)$*1 in this column, and errors in each column are summed up to $E_{\mathbf{PC}}$.

The aggregated errors in each of these blocks are thus given by the following sums:

$$\mathbf{E} = \begin{pmatrix} N_c \times (N_c - 1) - \sum_{i\in\mathbf{C}} \sum_{j\in\mathbf{C}} a_{ij} & (N - N_c) \sum_{i\in\mathbf{C}} max\left\{0, 1 - \sum_{j\in\mathbf{P}} a_{ij}\right\} \\ (N - N_c) \sum_{j\in\mathbf{C}} max\left\{0, 1 - \sum_{i\in\mathbf{P}} a_{ij}\right\} & \sum_{i\in\mathbf{P}} \sum_{j\in\mathbf{P}} a_{ij} \end{pmatrix}. \tag{2}$$

The aggregated error measures the difference between $\mathbf{A_{real}}$ and $\mathbf{A_{C-P}}$ by adding up the $E_{\mathbf{CC}}$, $E_{\mathbf{PP}}$, $E_{\mathbf{CP}}$, and $E_{\mathbf{PC}}$. As subset $\mathbf{C}$ is an internally tightly connected small community, that is, a subset of intermediary bank subset $\mathbf{I}$ in which each node has at least one outgoing link and one incoming link. The total error η_{C-P} of the model is obtained by aggregating and standardizing errors in the following way:

$$\eta_{\mathbf{C-P}} = \frac{E_{\mathbf{CC}} + E_{\mathbf{PP}} + E_{\mathbf{CP}} + E_{\mathbf{PC}}}{\sum_i \sum_j a_{ij}} = \frac{E_{\mathbf{CC}} + E_{\mathbf{PP}}}{\sum_i \sum_j a_{ij}}. \tag{3}$$

Take Figure 2 as an example, the missing link between V_1 and V_2 is an error among core banks($E_{\mathbf{CC}}$); the existing link between V_4 and V_6 is an error among periphery banks($E_{\mathbf{PP}}$); and the missing links from V_3 to V_4, V_5, and V_6 generate an error($E_{\mathbf{CP}}$) in the off-diagonal blocks. The difference between idealized model and real-world network in Figure 2 is calculated as total error which equals to 5 ($\eta_{\mathbf{C-P}} = E_{\mathbf{CC}} + E_{\mathbf{CP}} + E_{\mathbf{PP}} = 1 + (N - N_c)*1 + 1 = 1 + 3*1 + 1 = 5$).

We could develop algorithms to minimize the $\eta_{\mathbf{C-P}}$ in order to find the optimal division solution. The optimal solution $\mathbf{R}^*$ is the result with the optimal set of cores which generates the smallest distance to the idealized CP model of the same dimension. $\mathbf{\Gamma}$ is the set of all possible solution set $\mathbf{R}$:

$$\mathbf{R}^* = \arg\ \min \eta_{\mathbf{C-P}}(\mathbf{R}) = \{\mathbf{R} \in \mathbf{\Gamma} | \eta(\mathbf{R}) \le \eta(r), \forall r \in \mathbf{\Gamma}\}. \tag{4}$$

Idealized core–middle–periphery model

By introducing the middle layer, this study develops the CMP model. The idealized CMP network $\mathbf{G_{C-M-P}}(\mathbf{V{,}E})$ consists of core layer subset $\mathbf{C} = \{c_1, c_2, \ldots, c_{N_c}\}$, middle layer subset $\mathbf{M} = \{m_1, m_2, \ldots, m_{N_m}\}$, and periphery layer subset $\mathbf{P} = \{p_1, p_2, \ldots, p_{N_p}\}$, with N_c, N_m, and N_p as number of nodes in $\mathbf{C}$, $\mathbf{M}$ and $\mathbf{P}$ respectively. The middle banks, which do not have internal transaction constraints and may lend to or not lend to each other freely, are supposed to improve connections between layer $\mathbf{C}$ and $\mathbf{P}$, assuming the role of expanding the financing sources of banks in the periphery layer $\mathbf{P}$.

To be more specific, for any given N_c and N_m, an idealized CMP structure $\mathbf{G_{C-M-P}}(\mathbf{V{,}E})$ is defined as:

(a) All the N_c core banks lend with each other, and all the $(N - N_c - N_m)$ periphery banks do not lend with each other;
(b) Each bank in the core layer lends to (some) non-core banks, and each bank in the core layer borrows from (some) non-core banks;
(c) Each bank in the middle layer lends to (some) periphery banks, and each bank in the middle layer borrows from (some) periphery banks;
(d) Each bank in the middle layer lends to as many banks in the periphery layer as possible.

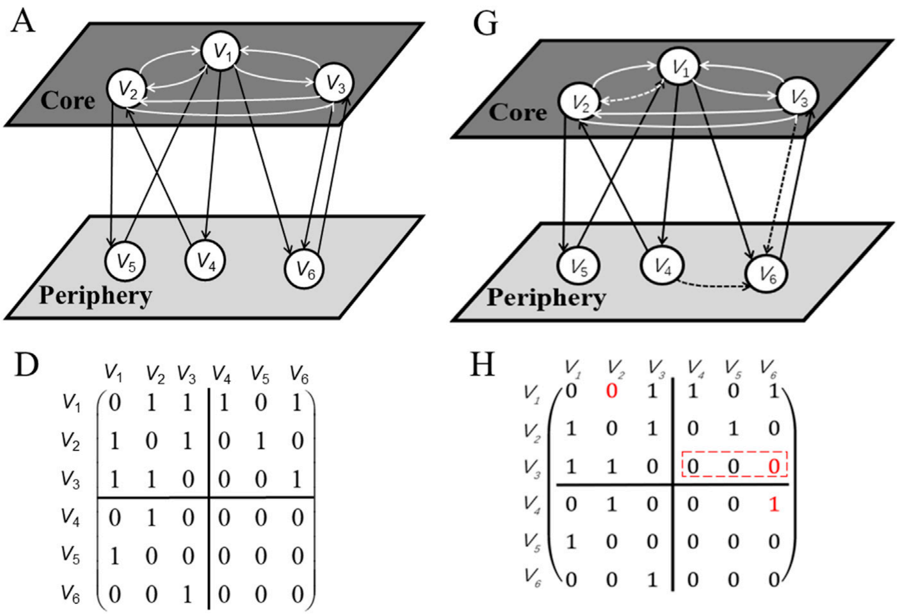


FIGURE 2
(Color online) Example of calculating error score between idealized CP model $\mathbf{A}_{\mathbf{C-P}}$ **(A,D)** and real-word network $\mathbf{A}_{\mathbf{real}}$ **(G,H)**.

According to feature (a)–(c), we define the intermediary bank subset $\mathbf{I} = \{i_1, i_2, \ldots, i_z\}$, in which banks acting both as lenders and borrowers. Intermediary nodes cannot be classified into the subset **C** if they do not lend to and borrow from non-core nodes. Meanwhile, intermediary nodes that are not classified in the core subset also cannot be classified into the subset **M** if they do not lend to and borrow from periphery nodes.

1) CMP adjacency matrix $\mathbf{A}_{\mathbf{C-M-P}}$. The adjacency matrix of idealized CMP model is

$$\mathbf{A}_{\mathbf{C-M-P}} = \begin{pmatrix} \mathbf{A}_{\mathbf{CC}} & \mathbf{A}_{\mathbf{CM}} & \mathbf{A}_{\mathbf{CP}} \\ \mathbf{A}_{\mathbf{MC}} & \mathbf{A}_{\mathbf{MM}} & \mathbf{A}_{\mathbf{MP}} \\ \mathbf{A}_{\mathbf{PC}} & \mathbf{A}_{\mathbf{PM}} & \mathbf{A}_{\mathbf{PP}} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \mathbf{A}_{\mathbf{CM}} & \mathbf{A}_{\mathbf{CP}} \\ \mathbf{A}_{\mathbf{MC}} & \mathbf{A}_{\mathbf{MM}} & \mathbf{A}_{\mathbf{MP}} \\ \mathbf{A}_{\mathbf{PC}} & \mathbf{A}_{\mathbf{PM}} & \mathbf{0} \end{pmatrix}. \quad (5)$$

In the idealized CMP adjacency matrix $\mathbf{A}_{\mathbf{C-M-P}}$, the core bank block $\mathbf{A}_{\mathbf{CC}}$ is a 1-block and the periphery bank block $\mathbf{A}_{\mathbf{PP}}$ is a 0-block, which are exactly the same as those of the CP adjacency matrix $\mathbf{A}_{\mathbf{C-P}}$.

For the off-diagonal blocks, we also introduce the "row-regular" and "column-regular" patterns. There is at least one inward link and one outward link between each core bank and the non-core banks, presenting as at least one non-zero element in each row of the $N_c{*}(N - N_c)$ matrix of sub-blocks containing $\mathbf{A}_{\mathbf{CM}}$ and $\mathbf{A}_{\mathbf{CP}}$. Similarly, there is at least one inward link and one outward link between each middle bank and the periphery bank layer, that is at least one non-zero element exists in each column of the $(N - N_c){*}N_c$ matrix of blocks containing $\mathbf{A}_{\mathbf{MC}}$ and $\mathbf{A}_{\mathbf{PC}}$. Accordingly, the submatrix $\mathbf{A}_{\mathbf{MP}}$ is row regular and $\mathbf{A}_{\mathbf{PM}}$ is column regular. According to the role of the middle bank nodes, there is no internal constraint on the $\mathbf{A}_{\mathbf{MM}}$. The example in Figure 1E shows one idealized CMP adjacency matrix with $N = 9$, $N_c = 3$, and $N_m = 3$. It should be noted that in $\mathbf{G}_{\mathbf{C-M-P}}(\mathbf{V,E})$, for any given N_c and N_m, we can depict a series of idealized CMP models with different $\mathbf{A}_{\mathbf{C-M-P}}$, featuring various patterns of off-diagonal blocks which all satisfy the "row regular" and "column regular" pattern.

2) Error function. In order to find the three layers, say **C**, **M** and **P**, from the empirical China banking network, it is natural to minimize the total error between $\mathbf{A}_{\mathbf{real}}$ and $\mathbf{A}_{\mathbf{C-M-P}}$ of the same dimension as the optimization function.

Moreover, we treat the $\mathbf{A}_{\mathbf{MP}}$ differently based on the fact that middle banks' lending to periphery banks helps to compensate for the missing links between the core and periphery banks, which help to close the funding gap of periphery banks as well as improve redistribution of liquidity in the banking system. As more lending relationships from middle to periphery is desirable but not compulsory in the model, it is not feasible to design the idealized block $\mathbf{A_{MP}}$ as a 1-block, which means that missing links from middle to periphery would be punishment items and contribute to a higher total error. In addition, it is much more complicated to design a variable $\mathbf{A_{MP}}$ in the idealized model with more than one lending relationship from middle to periphery. To solve this problem, we calculate the number of links indicating $\mathbf{M}$'s lending to $\mathbf{P}$ as the reward term to emphasize the role of the middle layer in terms of providing liquidity to the periphery layer and to simplify the searching procedure. The more lending links, the smaller the total error is.

Compared with the idealized CMP model, we develop an algorithm to minimize the total error function. Since it is complicated to obtain the optimized $\mathbf{C}$, $\mathbf{M}$ and $\mathbf{P}$ subsets simultaneously, we design a two-step algorithm to simplify the searching procedure.

Step1: Select the intermediary set $\mathbf{I} = \{i_1, i_2, \ldots, i_z\}$ from all bank set of the $\mathbf{G_{real}}$ network, where $\mathbf{Z} \in \{1, 2, \ldots, N\}$;
Step2: Filter out the core set $\mathbf{C}$ by using the following method. Given a determined core size N_c, where $N_c \in \{1, 2, \ldots, Z\}$, searching for the optimal core set $\mathbf{C}$ in $\mathbf{I}$ using simulated annealing by minimizing the total error. Then the remaining banks are set into the bank set $\mathbf{T}$=$\{t_1,t_2,\ldots,t_{N-Nc}\}$;
Step3: Filter out the middle set $\mathbf{M}$ by using the following method. After fixing each core set result $\mathbf{C}$ in Step 2, filter out the middle bank set $\mathbf{M}$ from subset $\mathbf{T}$. To be specific, after choosing each middle size N_m, where $N_m \in \{1, 2, \ldots, Z - N_c\}$, we search for the optimal middle set $\mathbf{M} \in \{\mathbf{I} \backslash \mathbf{C}\}$ to minimize the total error and group the remaining banks into $\mathbf{P}$.
Step4: Based on the total error calculated by the $\mathbf{C}$, $\mathbf{M}$, and $\mathbf{P}$ allocation in Step 3, find another candidate core size $N_c \in \{1, 2, \ldots, Z\}$ and repeat Step 2 to search for the optimal core set $\mathbf{C}$. Then repeat Step 3 to search for the optimal middle set $\mathbf{M}$. Stop the iteration when the difference of total error for timestep $t - 1$ and t is smaller than the threshold and output the final set of $\mathbf{C}$, $\mathbf{M}$ and $\mathbf{P}$.

The total error function $\eta_{\mathbf{C-M-P}}$ is defined as the standardized difference between idealized CMP model $\mathbf{A_{C-M-P}}$ and real-world network $\mathbf{A_{real}}$. According to the two main steps of the algorithm, the total error is divided into two parts, namely, E_1 and E_2 which are defined as follows.

$$\mathbf{E}_1 = \begin{pmatrix} N_c \times (N_c - 1) - \sum_{i\in\mathbf{C}} \sum_{j\in\mathbf{C}} a_{ij} & (N - N_c) \sum_{i\in\mathbf{C}} max\left\{0, 1 - \sum_{j\in\mathbf{T}} a_{ij}\right\} \\ (N - N_c) \sum_{j\in\mathbf{C}} max\left\{0, 1 - \sum_{i\in\mathbf{T}} a_{ij}\right\} & \sum_{i\in\mathbf{T}} \sum_{j\in\mathbf{T}} a_{ij} \end{pmatrix}, \tag{6}$$

$$\mathbf{E}_2 = \begin{pmatrix} 0 & p \sum_{i\in\mathbf{M}} max\left\{0, 1 - \sum_{j\in\mathbf{P}} a_{ij}\right\} \\ p \sum_{j\in\mathbf{M}} max\left\{0, 1 - \sum_{i\in\mathbf{P}} a_{ij}\right\} & \sum_{i\in\mathbf{P}} \sum_{j\in\mathbf{P}} a_{ij} \end{pmatrix}. \tag{7}$$

For each determined core size N_c and corresponding potential sets of $\mathbf{C}$ in Step 2, we continue to select potential node set $\mathbf{M}$ in Step 3. After selecting each potential node set $\mathbf{M}$, we are able to calculate the corresponding $E_{\mathbf{MM}}$, $E_{\mathbf{MP}}$, $E_{\mathbf{PM}}$, and $E_{\mathbf{PP}}$, where $E_{\mathbf{PP}}$ is defined as the number of connected links between periphery banks, $E_{\mathbf{MP}}$ denotes the number of missed relationships in terms of middle banks subset lending to periphery banks, and $E_{\mathbf{PM}}$ measures the number of missed relationships in terms of middle banks subset borrowing from periphery banks. The middle bank block error $E_{\mathbf{MM}}$ is zero as there is no constrain for node-set $\mathbf{M}$. In order to enhance the importance of $\mathbf{M}$ in the error calculation process, we add a reward term $E'_{\mathbf{MP}}$ (Eq. 8) into the error function as follows.

$$E'_{\mathbf{MP}} = -\sum_{i\in M} \sum_{j\in P} a_{ij}. \tag{8}$$

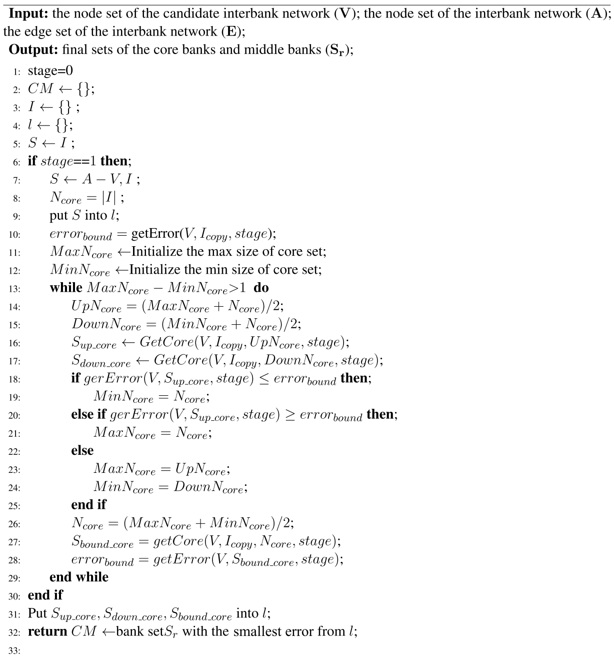
Then the total error function $\eta_{\mathbf{C-M-P}}$ is given as follows.

$$\eta_{\mathbf{C-M-P}} = \frac{E_{\mathbf{CC}} + E_{\mathbf{CT}} + E_{\mathbf{TC}} + E_{\mathbf{TT}} + E_{\mathbf{MP}} + E_{\mathbf{PM}} + E_{\mathbf{PP}} + E'_{\mathbf{MP}}}{\sum_i \sum_j a_{ij}}. \tag{9}$$

The optimal bank layer classification result $\mathbf{R}^*$ can be obtained by minimizing the total error function as follows

$$\begin{aligned} \mathbf{R}^* &= \arg\min \eta_{\mathbf{C-M-P}}(\mathbf{R}) \\ &= \left\{\mathbf{R} \in \mathbf{\Gamma} | \eta_{\mathbf{C-M-P}}(\mathbf{R}) \leq \eta_{\mathbf{C-M-P}}(r), \forall r \in \mathbf{\Gamma}\right\}. \end{aligned} \tag{10}$$

The algorithm pseudo-code is given as follows (Algorithms 1-3). θ_{stage} represents the threshold of number of intermediary nodes in each step.



```
Input: the node set of the candidate interbank network (V); the node set of the interbank network (A);
the edge set of the interbank network (E);
Output: final sets of the core banks and middle banks ($S_r$);
stage=0
$CM \leftarrow \{\}$;
$I \leftarrow \{\}$ ;
$l \leftarrow \{\}$;
$S \leftarrow I$ ;
if $stage$==1 then;
    $S \leftarrow A - V, I$ ;
    $N_{core} = |I|$ ;
    put $S$ into $l$;
    $error_{bound}$ = getError($V, I_{copy}, stage$);
    $MaxN_{core}$ ←Initialize the max size of core set;
    $MinN_{core}$ ←Initialize the min size of core set;
    while $MaxN_{core} - MinN_{core} > 1$ do
        $UpN_{core} = (MaxN_{core} + N_{core})/2$;
        $DownN_{core} = (MinN_{core} + N_{core})/2$;
        $S_{up\_core} \leftarrow GetCore(V, I_{copy}, UpN_{core}, stage)$;
        $S_{down\_core} \leftarrow GetCore(V, I_{copy}, DownN_{core}, stage)$;
        if $gerError(V, S_{up\_core}, stage) \leq error_{bound}$ then;
            $MinN_{core} = N_{core}$;
        else if $gerError(V, S_{up\_core}, stage) \geq error_{bound}$ then;
            $MaxN_{core} = N_{core}$;
        else
            $MaxN_{core} = UpN_{core}$;
            $MinN_{core} = DownN_{core}$;
        end if
        $N_{core} = (MaxN_{core} + MinN_{core})/2$;
        $S_{bound\_core} = getCore(V, I_{copy}, N_{core}, stage)$;
        $error_{bound} = getError(V, S_{bound\_core}, stage)$;
    end while
end if
Put $S_{up\_core}, S_{down\_core}, S_{bound\_core}$ into $l$;
return $CM$ ←bank set$S_r$ with the smallest error from $l$;
```

Algorithm 1. The heuristic process to get an optimized CMP combination.

```
Input: the node set of the interbank network ($V$); the set of intermediary banks ($I_{copt}$); the number of core banks ($N_{core}$); current stage ($stage$);
Output: detected sets of core or middle nodes ($S$)
$S_r \leftarrow$Sample a random set from $I_{copt}$ with size $N_{core}$;
$S \leftarrow S_r$;
if $stage$==1 then;
    $S \leftarrow A - V, S_r$ ;
    $error_{min} \leftarrow getError(V, S_r, stage)$;
    $T_{min} = l, T_0 = N, t = 0$;
    $T = T_0/(l + t)$;
    while $T \geq T_{min}$ do
        sample $num_{switch}$ from $[1, coreNum]$,and get $S_{new}$ replace $num_{switch}$ nodes in $S_r$ from $S$;
        $error_{new} \leftarrow getError(V, s_{new}, stage)$;
        if $error_{new}$<$error_{min}$, then
            $error_{min} = error_{new}$;
            $S_r = S_{new}$;
        else
            $p = exp(-(error_{new} - error_{min}/T)$;
            sample $r$ from $[0, 1]$;
            if $r$<$p$, then
                $error_{min} = error_{new}$;
                $S_r = S_{new}$;
            end if
            $t = t + l$;
            $T = T_0/(l + t)$;
        end if
    end while
end if
if $stage$==0 then;
    $S_M$=C_M_P_Detector($V - s, A, E, 1$);
    $S \leftarrow S, S_M$;
end if
return $S$
```

Algorithm 2. The simulated annealing optimizer to get a candidate core with fixed core size.

```
Input: the node set of the interbank network ($V$);the set contains detected sets of core or middle nodes ($S$); the core bank set of the interbank network ($C$); current stage ($stage$);
Output: current error ($error$);
if $stage == 0$, then
    Get set $\mathbf{T} = \{\mathbf{t}|\mathbf{t} \notin \mathbf{C}, \mathbf{t} \in \mathbf{V}\}$;
    $error = E_1$// Eq. 7;
else
    Get set $C, M$ from $S$, get set $\mathbf{P} = \{\mathbf{p}|\mathbf{p} \notin \mathbf{M}, \mathbf{t} \in \mathbf{V}\}$ ;
    $error = E_{\mathbf{C-M-P}}$// Eq. 10;
end if
return $error$
```

Algorithm 3. Calculation of total error function.

Results

In addition to the total error, this study uses a transition matrix to track the evolution properties of the nodes in different layers. For the bank set sequence $\mathbf{X} = \{X_1, X_2, \ldots, X_t\}$, where X_t denotes the set of states of each bank at moment t, and its corresponding state space $\mathbf{S} = \{c, m, p, Exit, New\}$ denotes the node status of the core bank, middle bank, periphery bank, exit bank, and new bank, respectively. We defined the transition matrix $\mathbf{H}(X_t = s|X_{t-1} = s') = Q(s)/Q(s')$ as probability of banks moving from state s' at moment $t-1$ to state s at moment t, and $Q(s)$ denotes the number of nodes with state s. The transition matrixes of the CP model and the CMP model are as follows:

$$\mathbf{H_{C-P}}(\mathrm{s}|\mathrm{s}') = \begin{pmatrix} H(c|c') & H(p|c') & H(Exit|c') \\ H(c|p') & H(p|p') & H(Exit|p') \\ H(c|New') & H(p|New') & H(Exit|New') \end{pmatrix}, \tag{11}$$

where $\sum_s(s|c') = \sum_s(s|p') = \sum_s(s|New') = 1$.

$$\mathbf{H_{C-M-P}}(\mathrm{s}|\mathrm{s}') = \begin{pmatrix} H(c|c') & H(m|c') & H(p|c') & H(Exit|c') \\ H(c|m') & H(m|m') & H(p|m') & H(Exit|m') \\ H(c|p') & H(m|p') & H(p|p') & H(Exit|p') \\ H(c|New') & H(m|New') & H(p|New') & H(Exit|New') \end{pmatrix}, \tag{12}$$

where $\sum_s(s|c') = \sum_s(s|m') = \sum_s(s|p') = \sum_s(s|New') = 1$.

We defined diagonal elements $H(c|c')$, $H(m|m')$, and $H(p|p')$ in the transition matrix as stability probability W_c, W_m, and W_p, respectively, measuring the probability of nodes remaining unchanged in the core, middle and periphery layers, respectively, for different timesteps. $\bar{W}_c$, $\bar{W}_m$, and $\bar{W}_p$ are defined as the average stability probability of the core, middle and periphery subsets for the entire 12 years. Other elements in the transition matrix are deemed as transition probability, reflecting the probability of nodes moving to other layers. Low stability probability combined with high transition probability indicates that the bank nodes play different roles in different timesteps, which is not consistent with the China scenario. In reality, for most of the banks, the role that each bank plays is relatively stable in the banking system.

In China banking network, the average stability transition probability shows the following results:

$$\mathbf{\bar{H}_{C-M-P}}(\mathrm{s}|\mathrm{s}') = \begin{pmatrix} 0.78 & 0.15 & 0.07 & 0.00 \\ 0.02 & 0.76 & 0.18 & 0.04 \\ 0.02 & 0.33 & 0.60 & 0.05 \\ 0.00 & 0.64 & 0.36 & 0.00 \end{pmatrix}, \tag{13}$$

$$\mathbf{\bar{H}_{C-P}}(\mathrm{s}|\mathrm{s}') = \begin{pmatrix} 0.50 & 0.49 & 0.01 \\ 0.31 & 0.64 & 0.05 \\ 0.15 & 0.85 & 0.00 \end{pmatrix}. \tag{14}$$

Figure 3A shows the empirical error function results of the China banking system. The average total error function of the CMP model is 7%, much smaller than the 20% which is obtained by the CP model. Furthermore, one may find that the total errors of the CMP model range from 3% to 11%, whereas those of the CP model range from 15% to 31%.

Second, in terms of the number of banks in different layers, the ratios of the core, middle and periphery banks under the CMP model remain stable between 2008 and 2019. Along with the gradual increase of the number of total nodes due to the market development, the numbers of the core, middle, and periphery banks also increase accordingly. Meanwhile, the annual average ratio of core, middle, and periphery banks stays stable (Figure 3B). Specifically, the ratio of core banks stays within the narrow interval of (4%, 9%), suggesting a small and tight core subset over the sample period. In comparison, under the CP model the ratio of core banks fluctuates from 30% to 44% and the ratio of periphery banks fluctuates from 57% to 69% over the sample period, reflecting broader and unstable subsets.

Third, according to the evolution properties of the transition probability, we find that the tiering structure in the CMP model (Figure 3C) is more stable than that of the CP model. Specifically, the composition of the core bank subset remains remarkably

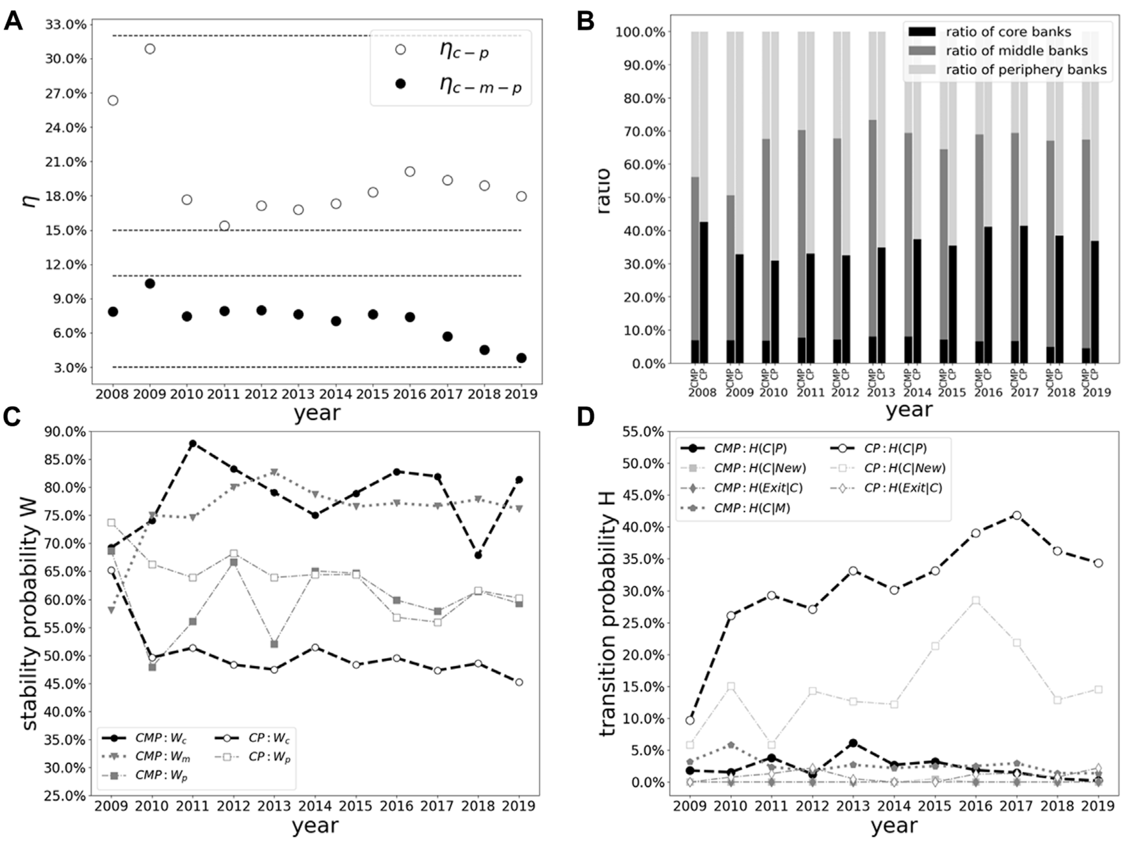


FIGURE 3
(A) Total errors η, **(B)** ratio of different layers, **(C)** stability probability W and **(D)** transition probability H obtained from the China bankinig network by the CMP model and the CP model.

stable. Under the CMP model, the annual averages of stability probability $\bar{W}_c, \bar{W}_m$ and $\bar{W}_p$ are 78%, 76% and 60%, respectively, indicating a relatively high probability of the core, middle, and periphery banks remaining in the same layer. Specifically, W_c falls within the range of (69%, 88%), which indicates that the categorizing of core banks is relatively stable. In reality, the role of

TABLE 1 Network metrics for the China banking system.

Year	*APL*	*D*	*ACC*	⟨*k*⟩	*S*
2008	2.55	6	0.20	44	0.06
2009	2.90	9	0.19	35	0.04
2010	2.38	7	0.23	58	0.06
2011	2.29	6	0.25	80	0.07
2012	2.32	7	0.24	82	0.07
2013	2.21	5	0.27	111	0.08
2014	2.18	6	0.26	126	0.09
2015	2.32	6	0.22	117	0.07
2016	2.24	6	0.22	153	0.07
2017	2.20	5	0.23	180	0.07
2018	2.31	6	0.20	160	0.06
2019	2.36	7	0.19	157	0.05
Average	2.35	6	0.23	109	0.07

banks in the core layer is very stable, which rarely changes compared with non-core banks. Therefore, compared with the CP model, the CMP model could capture the evolution properties more accurately.

Fourth, as can be seen in Figure 3D, for the CMP model, the average probabilities of a middle and a periphery bank becoming a core bank are as low as 2%. A core bank does not exit the market, and it is difficult for a new entrant to become a core bank in the following year. All these experimental results are reasonable in terms of explaining the real-world banking system structure. In comparison, according to the empirical result of the CP model, a core bank has a 1% probability of exiting the market, whereas a new entrant and a periphery bank have a 15% and 31% probability of becoming a core bank respectively, which is extremely high comparing with the CMP model. Moreover, a core bank obtained by the CP model has a 49% probability of becoming a periphery bank in the next year which is inconsistent with the fact that the banks in the core layer play an important role in the banking system and have stable characteristics.

Generally speaking, the China banking system exhibits an evident and stable three-layer structure, which is consistent with the proposed CMP model.

Conclusion and discussions

In this study, we present the CMP model to analyze the evolution properties of the China banking network, in which banks are nodes and the existence of lending relationships are directed links. We find that this network shows characteristics commonly found in other empirical networks, such as the small-world phenomenon and fat-tailed degree distribution. However, as China banking system possesses a special three-layer structure, we extend the CP model into the CMP model by adding middle nodes to improve the connection between large-degree nodes and small-degree nodes, so as to better describe the real-world network structure.

The total error function and the stable transition probability show that China banking network is prone to approximate a CMP structure with smaller total error scores, tighter core subset, and stronger structural stability over time rather than a CP structure.

During the sample period, the average total error of optimal CMP structure fitting on the annual China banking network comes to less than 10% of network links, decreasing by 64% from the 20% average total error under the CP model, indicating that the CMP model explains the real network better.

For the CMP model, the ratios of the core, middle, and periphery banks under the CMP model remain stable between 2008 and 2019, featuring a tight core set and a large middle set. In comparison, the sizes of core and periphery banks subets under the CP model are volatile, suggesting that a portion of middle modes are likely to be classified into core banks.

Furthermore, an in-depth analysis of banks in each layer shows that the average stability probability of the core bank subset $(\bar{W}_c)$ increases from 50%(under the CP model) to 78% by adopting the CMP model, contributing to a more accurate categorizing of core node subset with higher intertemporal stability. The average stability probability $(\bar{W}_m)$ of the middle bank layer in the CMP model also features a high level of 76%, whereas that of the periphery bank set is 60% $(\bar{W}_p)$.

In terms of the transition probabilities, a core bank does not exit the market, and it is of quite a low probability for a middle bank, a periphery bank, or a new entrant to become a core bank in the following year under the CMP model, which is consistent with the intuitive understanding of real-world banking system scenario. In contrast, in the CP model, a core bank may exit the market although the probability of this scenario is as low as 1%. Meanwhile, a new entrant and a periphery bank have a 15% and 31% probability of becoming a core bank, respectively.

To sum up, compared with the CP model, the CMP model could regenerate the stable core, middle and periphery structure for China banking network.

In future work, the following issues will be examined. This study lacks a solid theoretical foundation of a three-tier structure which aims to answer the following question: Is a three-tier structure superior to a multi-layer structure in arguing for a Chinese banking network? Whether no internal constraint of middle nodes is rationale and sufficient? What is the optimal size of middle banks since a loose constraint on middle nodes and tight constraint on periphery nodes may lead to a large portion of nodes being dropped into the middle community so as to minimize the total error? What is the exact theoretical relationship between middle nodes, core, and periphery nodes? How to measure the effectiveness of middle nodes' intermediary role in liquidity distribution? Is there a better way besides adding a reward item to encourage more intermediary behavior of the middle nodes? We would go deep into a theoretical discussion of these three-layer structures in the future.

Moreover, as the size of the network increases, the computation complexity would increase greatly, making it difficult to obtain a global optimum. How to develop fast optimization algorithm through theoretical analysis will be the next research topic. Whether the two-step algorithm could be improved and optimized is also an interesting question. From a network evolution perspective, we present the annual transition matrix to depict stability probability and transition probability. A solid and stable multilayer network is a multiple dimension problem as opposed to the transition of core and periphery nodes in a two-layer model, as more layers may generate complex transition scenarios which are more complicated to analyze and monitor. The role of the core, middle, and periphery may affect the network resilience, and it is challenging to discuss the three-layer network resilience on a time-varying basis.

In the future we plan to explore these aforementioned problems further, laying a much solid foundation for the observed three-tier structure.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Materials; further inquiries can be directed to the corresponding author.

Author contributions

NC and JL designed and performed the research. DT and YO performed the computations, NC, JL, YC, DT, YO, and MJ wrote the manuscript.

Funding

JL is supported by the National Natural Science Foundation of China (Grant Nos. 72171150, 61773248, 92146004, and 71901144).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Allen F, Gale D. Financial contagion[J]. *J Polit Economy* (2000) 108(1):1–33. doi:10.1086/262109

2. Cassar A, Duffy N. *Contagion of financial crises under local and global network [M]*. Berlin, Germany: Springer US (2002).

3. Battiston S, Caldarelli G, May RM, Roukny T, Stiglitz JE. The price of complexity in financial networks[J]. *Proc Natl Acad Sci U S A* (2016) 113(36): 10031–6. doi:10.1073/pnas.1521573113

4. Hu Z, Shen Z, Cao S, Podobnik B, Yang H, Wang WX, et al. Locating multiple diffusion sources in time varying networks from sparse observations[J]. *Sci Rep* (2018) 8(2685). doi:10.1038/s41598-018-20033-9

5. Fang H, Wang Y, Wu X. *The collateral channel of monetary policy: Evidence from China*. NBER Working Papers (2020).

6. Martinez-Jaramillo S, Alexandrova-Kabadjova B, Bravo-Benitez B, Solorzano-Margain JP. An empirical study of the Mexican banking system's network and its implications for systemic risk[J]. *J Econ Dyn Control* (2014) 40:242–65. doi:10.1016/j.jedc.2014.01.009

7. Angelini P. *Liquidity and announcement effects in the euro area[R]*. Rome, Italy: Bank of Italy (2002). p. 451.

8. Furfine C. Interbank exposures: Quantifying the risk of contagion[J]. *J Money, Credit Banking* (2003) 35(1):111–28. doi:10.1353/mcb.2003.0004

9. Iori G, Masi GD, Precup OV, Gabbide G, Caldarelli G. A network analysis of the Italian overnight money market[J]. *Econ Dyn Control* (2008) 32(1):259–78. doi:10.1016/j.jedc.2007.01.032

10. Bech ML, Atalay E. The topology of the federal funds market[J]. *Physica A: Stat Mech its Appl* (2010) 389(22):5223–46. doi:10.1016/j.physa.2010.05.058

11. Boss M, Elsinger H, Summer M, Thurner S. Network topology of the interbank market[J]. *Quantitative Finance* (2004) 4(6):677–84. doi:10.1080/14697680400020325

12. Soramaki K, Bech ML, Arnold J, Glass RJ, Beyeler WE. The topology of interbank payment flows[J]. *Physica A: Stat Mech its Appl* (2007) 379(1):317–33. doi:10.1016/j.physa.2006.11.093

13. Inaoka H, Takayasu H, Shimizu T, Ninomiya T, Taniguchi K. Self-similarity of banking network[J]. *Physica A: Stat Mech its Appl* (2004) 339(3-4):621–34. doi:10.1016/j.physa.2004.03.011

14. Boss M, Elsinger H, Summer M, Thurner S. Network topology of the interbank market[J]. *Quantitative Finance* (2004) 4(6):677–84. doi:10.1080/14697680400020325

15. Edson B, Cont R. *The Brazilian interbank network structure and systemic risk [R]*. Brasília, Federal, Brazil: Central Bank of Brazil (2010). p. 219.

16. Borgatti SP, Everett MG. Models of core/periphery structures[J]. *Social Networks* (2000) 21(4):375–95. doi:10.1016/s0378-8733(99)00019-2

17. CraigVon Peter G, von Peter G. Interbank tiering and money center banks [J]. *J Financial Intermediation* (2014) 23(3):322–47. doi:10.1016/j.jfi.2014.02.003

18. Langfield S, Liu Z, Ota T. Mapping the UK interbank system[J]. *J Banking Finance* (2014) 45:288–303. doi:10.1016/j.jbankfin.2014.03.031

19. Fricke D, Lux T. Core–periphery structure in the overnight money market: Evidence from the e-MID trading platform[J]. *Comput Econ* (2015) 45(3):359–95. doi:10.1007/s10614-014-9427-x

20. Brassil A, Nodari G. *A density-based estimator of core/periphery network structures: Analysing the Australian interbank market[R]*. Australia: RBA Research Discussion Papers (2018).

21. Yang H, Hu M. Risk contagion of China's interbank markets based on Core-Periphery network[J]. *J Manage Sci China* (2017) 20(10):44–56. (in Chinese).

22. Xing J, Guo Q, Liu J. Interbank network estimation based on local clustering features[J]. *J Univ Electron Sci Technol China* (2021) 2021(08). (in Chinese).

23. Barabási AL. Network science[J]. *Philosophical Trans R Soc A* (2013) 371(1987):2012.

24. Manoj BS, Abhishek C, Rahul S. *Complex networks: A networking and signal processing perspective[M]*. Whitby, ON, Canada: Pearson (2018).

TYPE Original Research
PUBLISHED 15 November 2022
DOI 10.3389/fphy.2022.1032913

OPEN ACCESS

EDITED BY
Jianguo LIU,
Shanghai University of Finance and Economics, China

REVIEWED BY
Jichang Zhao,
Beihang University, China
Cheng-Jun Wang,
Nanjing University, China

*CORRESPONDENCE
Haibo Hu,
hbhu@ecust.edu.cn

[†]These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Social Physics,
a section of the journal
Frontiers in Physics

RECEIVED 31 August 2022
ACCEPTED 27 October 2022
PUBLISHED 15 November 2022

CITATION
Xu N, Lin Q, Hu H and Li Y (2022), The effect from elimination mechanism on information diffusion on entertainment programs in Weibo.
Front. Phys. 10:1032913.
doi: 10.3389/fphy.2022.1032913

The effect from elimination mechanism on information diffusion on entertainment programs in Weibo

Nannan Xu[†], Qiaoting Lin[†], Haibo Hu* and Ying Li

Department of Management Science and Engineering, East China University of Science and Technology, Shanghai, China

Information diffusion in social media has attracted the wide attention of scholars from diverse disciplines. In real life, many offline events can cause online diffusion of relevant information, and the relation between the characteristics of information diffusion and offline events, as well as the diffusion differences corresponding to different phases of offline events have been studied. However, the effects of offline events on information diffusion are not well explored. In this paper, we study the influence of a popular and multi-phase talent show with elimination mechanism on relevant information diffusion. We find that elimination mechanism has significant influence on the features of information diffusion, and elimination results have a negative effect on followers' emotional tendency. Elimination results also significantly affect the topics discussed by users. Besides elimination results have a negative effect on participants' popularity, but do not affect the followers' loyalty to program participants. This study not only reveals the effects of offline events on online information diffusion, but also provides approaches for studying the online diffusion of similar offline events.

KEYWORDS

information diffusion, social media, talent show, elimination mechanism, effect estimation

1 Introduction

The emergence and development of online social networks and social media have not only changed the way people make friends, but also change the way of information acquisition and diffusion. Users are not only receivers of information, but also producers and disseminators of it. In recent years, information diffusion has attracted the attention of scholars from different fields, and significant progress has been made in empirical, modeling and prediction research [1–4].

In fact, the information posted by users on social media is often closely related to events occurring in the offline real world. Some events have far-reaching impact, and some last for a long time, thus attracting the attention of numerous users and triggering extensive discussions. In certain events, especially political ones, offline events and online discussions can influence each other, creating online-offline interactions. The correlation between offline events and corresponding online discussions provides a new scenario for information diffusion research.

Some researchers study the influence of TV series on related information diffusion. Since TV series usually update once every week, the related discussion is periodic impulsive. Fu et al proposed an impulsive susceptible-infected-removed (SIR)-like model to reproduce the periodic impulsive feature [5]. The influence of offline sports events on online user behavior has also been explored. Chung et al examined #BoycottNFL, an online connective action created to discontinue support of the National Football League, and found that associated offline trigger events affect the diversity of actors participating in connective action and fostering interactions between actors and online communities of diverse backgrounds [6].

The influence of offline vicious events on users' online behavior has been extensively studied. For instance, Burnap et al studied the terrorist event in Woolwich, London in 2013 and built models to predict information flow size and survival using data from Twitter. They found that the number of offline press reports relating to the event published on the day the tweet is posted is a significant predictor of size [7]. Zhou mined the users' behaviors in four emergency events from microblogs to reveal their behavior preferences, and found that users' behaviors in emergencies are related to their own interests and economic status [8]. Some studies explored the communication dynamics in social networks/media during or after natural disasters. For example, Kim and Hastak explored patterns created by the aggregated interactions of users on Facebook during responses in the 2016 Louisiana flood [9], and Pourebrahim et al investigated the communication dynamics on Twitter during Hurricane Sandy in 2012 [10]. The studies can help emergency agencies develop better operation strategies for a disaster mitigation/relief plan.

The COVID-19 pandemic is the most influential public health event in recent years which has caused profound social and economic impacts. Shen et al analyzed posts related to COVID-19 on Weibo, a popular Twitter-like social media site in China, to predict COVID-19 case counts in mainland China [11]. During COVID-19, unreliable information or fake news spread on the Internet. Using Twitter messages, Gallotti et al assessed the risks of the spread of information of questionable quality during the early stages of COVID-19 epidemics [12]. Vaccines are an important means to contain the large-scale spread of COVID-19. Hu et al investigated public opinion and perception on COVID-19 vaccines in the United States with Twitter data and found the rising confidence and anticipation of the public towards vaccines [13]. The COVID-19 not only affects people's physical health, but also their mental health. Through the analysis on Twitter, it was found that Australians' mental health signals, quantified by sentiment scores, have a shift from pessimistic (early pandemic) to optimistic (middle pandemic). However, the signals progressively recess towards a more pessimistic outlook (later pandemic) [14]. Using social media data from Twitter and Weibo, Wang et al found that COVID-19 outbreaks cause steep declines in expressed sentiment globally [15]. They also found moderate to no effects of lockdown policies on expressed sentiment.

Political events have also attracted the attention of scholars because of their extensive and profound influence on people's life. High-impact public protests [16–28] or election campaigns [29–32] and related online discussions often interact, such as the Arab Spring [23, 27], the Spanish indignados movement [17, 18, 21, 24], the Occupy Wall Street movement [21, 22, 24] and the 2016 United States Presidential Election [32]. During the Arab Spring movement, social media activity in Twitter correlates with subsequent large-scale decentralized coordination of protests [23]. For the Spanish indignados movement, social media are the main tools for informing and mobilizing [19], and there are four types of users (influentials, hidden influential, broadcasters, and common users) in Twitter and they play different roles in the growth of the protest [18]. During the Occupy Wall Street movement, Twitter users generated a loosely connected hub-and-spoke network, suggesting that information is likely to be organized by several central users in the network and that these users bridge small communities [22]. Only a very small minority of tweets refer to protest organization and coordination issues [21]. During the 2016 United States Presidential Election, individuals are more active in interacting with similar-minded Twitter users ("echo chambers" effect), and the aggressive use of Twitter bots, coupled with the fragmentation of social media and the role of sentiment, could enhance political polarization.

Recently the causal impact of offline or online events on information diffusion in social media has also been studied. Yu et al examined the effect of the online 16 Days Campaign on the changes in public discussions of the MeToo in Twitter by applying the state-space model, and found that there are significantly more discussions in MeToo after the launch of the campaign [33]. Leveraging difference-in-difference (DID) method, Balawi et al investigated the impact of the United Airlines crisis on three dimensions of customer relationship management efforts on social media, and found that the brand crisis increases informativeness efforts but reduces timeliness and attentiveness efforts [34]. Falavarjani et al studied the causal relation between real world activities and emotional expressions of users in social media based on a quasi-experimental design, and found that users' offline activities impact their online affective expressions, both of emotions and moods [35].

The researches on the correlation between offline events and online discussions rely on event details and social media data, and vital conclusions have been obtained. However, the effects [36, 37] of offline events on online discussions as well as the underlying mechanisms still have not been well explored. Besides in real life, some offline events can last for a long time and show significant multiple stages over time. Few studies have explored

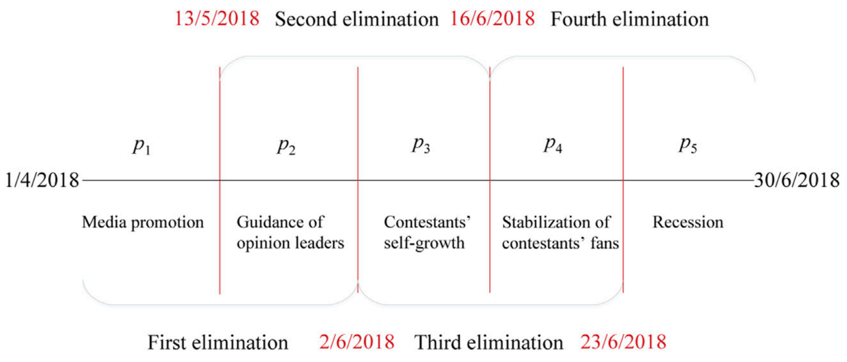


FIGURE 1
Different phases of information diffusion.

the impact of such events on online discussions, of which results will deepen our understanding of how offline events affect social media. In this paper, take an influential and multistage entertainment program for example, we will study the effects of offline events on related information diffusion in social media by collecting publicly available data, and try to fill the research gaps.

2 Data set

The public pays much attention to entertainment programs in their leisure time. Compared with the social news and political events, the influence of entertainment programs is long-term and moderate [38]. The TV talent show "Produce 101" (first aired on 21 April 2018, and last aired on 23 June 2018) is one of the most influential entertainment programs in the Chinese mainland in recent years. The 101 contestants participated in the training and assessment of singing and dancing. According to the results of multiple rounds of voting and elimination, the final winners were determined.

Specifically, in the first elimination round, 101 contestants competed based on the audience's votes in the official voting channel, with 55 contestants advancing and the rest eliminated. In the second round, 36 were promoted and 19 eliminated. In the third round, 22 were promoted and 14 eliminated. In the fourth round, i.e. the final round, the top 11 votes won.

The talent show was presented in the form of live TV. In addition to watching the program, the audience also published relevant posts on social media, which makes it a typical event combining online and offline. We collected publicly available posts with the topic of "Produce 101" from Sina Weibo, Chinese largest online microblogging platform, including the original posts and reposts. The time span of the data set is from 1 April 2018 to 30 June 2018 and it includes the information of microblog release time, content, poster's encrypted ID number, gender, and city. After data cleaning, there are 33,522,289 original posts and 56,267,334 reposts in the data set.

We choose this program for research for two reasons. First, this program has a large popularity, a wide range of audience, and is representative. Second, the program has a relatively long time span, with four rounds of elimination, making it possible to study the information diffusion in different phases of the program.

3 Program progress

As shown in Figure 1, we divide the data set according to the four rounds of elimination. The phase before the first elimination can be that for media promotion. In the early stage of the program preparation and broadcast, the publicity of the official account is crucial to ensure the authority and credibility of the program information. In addition, some audiences learned about the program from various channels in the early stage and spontaneously became the propagandists of the program. Official account and spontaneous users published and spread program information online, built initial nodes of information diffusion network, and introduced and connected more users by virtue of their own social networks and open platforms.

The phase between the first and second elimination can be the guidance stage of opinion leaders. Opinion leaders played an important role in the diffusion process. By releasing relevant highlights in the program and by virtue of their social capital and position advantages in social networks, they could reach more users at a faster speed and in a wider range, and finally expand the audience.

The phase between the second and third elimination can be the stage of contestants' self-growth. With the broadcast of the program, the audience increased, ordinary users gradually became fans, contestants accumulated their own fan groups, and the increase of audience also made the information diffusion scale expand again. At this phase, the increase of users could not be all from the influence of the program, but more from the attractiveness of the contestants.

The phase between the third and fourth elimination can be the stabilization stage for the contestants' fans. In the previous stages, the fans kept increasing, and in this stage, the fan groups could tend to be stable. Some users gradually lost interest in the program and no longer paid attention to it, while the remaining ones could be more loyal.

The phase after the program can be the recession one. The popularity of talent shows decreased with its end. For the audience, the novelty of and the enthusiasm for the program is limited. Without the real-time discussion, the related posts in social media could gradually decrease.

4 Characterizing information diffusion

Elimination mechanism is commonly used in competitive programs, which means that the discussion of users on Weibo may change with elimination results, and the indicators of posts related to contestants with different results after eliminations can also change. In preliminary research we have found that for the two groups of promoted and eliminated contestants, there exists interaction between information on them, and different interactions can occur at different stages. From the three perspectives of posts, users and contestants, we study the influence of elimination mechanism on the indicators involved in each perspective.

4.1 Posts

According to the repost relation in microblogs, each original post and its reposts can form an information diffusion tree, which represents the diffusion path of the original post and can be viewed as directed networks where the nodes without in-degree are root nodes or seed nodes, the nodes without out-degree are leaf nodes or passive nodes, and the nodes with both in-degree and out-degree are viral nodes.

In the paper, at the post level, we utilize four basic indicators to measure the diffusion capacity of original microblogs from different perspectives [39]. Let the numbers of seed nodes, viral nodes and passive nodes be N_s, N_v, and N_p respectively, and the total number of nodes is $N = N_s + N_v + N_p$ which represents the diffusion scale and reflects the overall influence. The branches of diffusion trees refer to the forwarding chains, with one branch starting from the root node and ending with a leaf node. According to this definition, the number of branches N_b is equal to leaf node number N_p. Let the length of each branch be d_i $(i = 1, 2, \ldots, N_b)$, and the maximum length $D = \max_i d_i$ from the leaf nodes to the root node in a tree, a measure of vertical diffusion, represents the diffusion depth. Nodes with the same distance to the root node are called nodes with the same level. The number of nodes w_i $(i = 1, 2, \ldots, D)$ at each level is the level width, and the maximum level width $W = \max_i w_i$ represents the diffusion width reflecting horizontal influence. In a diffusion tree, the time difference between the last reposting time of the original post in the data set and the release time of the post represents the active time of the post (measured in hours). We obtain the four indicators of each diffusion tree.

4.2 Users

At the user level, microblogs posted by users usually have specific emotions and belong to specific topics. Thus, we use two indicators, emotional tendency and topic category, to describe the content of microblogs.

The emotional tendency of microblogs refers to the degree of positive or negative emotion expressed in posts, which is usually measured by a single value. The positive values indicate positive emotion, while negative values negative emotion, and the absolute values indicate the degree of tendency. In this study, we utilize an authoritative simplified Chinese affective lexicon ontology to obtain emotional tendency of microblogs [40, 41]. The lexicon divides emotions into seven categories (happy, good, surprise, anger, sad, fear, and disgust) and 21 subcategories, which are labeled with polarity, part of speech, and emotional intensity for each emotional vocabulary. The emotional intensity label has levels 1, 3, 5, 7, and 9. Level 1 has the lowest intensity, and level 9 indicates the highest intensity. To better assess the emotions of microblogs, we comprehensively consider the combination and order of emotional words, adverbs of degree and negative words. The detailed calculation methods are shown in the Supplementary Material. The accuracy of emotion analysis for the lexicon ontology is 0.79, and its effectiveness has been demonstrated in many studies [42, 43] and it has been extensively applied in the emotion analysis of short texts [44, 45].

We obtain the emotional tendency of microblogs and find that the mean tendency value of all original posts is 1.85, the mean value of original posts with positive emotion is 3.27, and the mean value of original posts with negative emotion is -1.79. Overall the sentiment of microblogs in the data set is positive.

We use Biterm Topic Model (BTM) [46], a topic classification model suitable for short texts, to evaluate the topic of each original microblog. BTM integrates word co-occurrence information into latent Dirichlet allocation (LDA) to solve the problem of inferring topics from large-scale short texts. For the classification performance of Chinese short texts, the accuracy of BTM is close to 0.7. BTM has many advantages over some previous topic classification methods, and recently has been widely used in (Chinese) short text topic classification [47, 48].

We first set the number of topics to 50, and except the topics that cannot be specifically identified, we manually classify the others into six ones by topic merging according to the characteristics of entertainment programs. The posts with endorsement topic are on followers' behaviors. Contestants' fans or other users could increase the popularity of their supporters through microblogs, and increase ranking by soliciting votes, inviting clicks and other behaviors. The posts with stage performance topic are on the stage performance of contestants in the program. The posts with praise and encouragement topic are to praise the participants to achieve the purpose of publicity. The posts with contestant activities topic are on the activities of the participants outside the program. The posts with program publicity topic are on the publicity released by program producer to improve the popularity of the program. The posts with program criticism topic are on users' negative evaluation of the program, including doubts about or objections to program editing, competition fairness, elimination results, etc.

4.3 Contestants

We defined two indicators to characterize contestants, i.e., the mean number of mentions $< cnt_{d,s}>$ of contestant s on some day d and the losing rate $outRatio_{d+1,s}$ of contestant s's active fans on the next day d+1.

Specifically, if a user posts original microblogs that only mention a certain contestant on some day, the user is considered to be the active follower of that contestant on that day. For each contestant we obtain the number of daily active followers $dau_{d,s}$ and the number of next day active followers $dau_{d+1,s}$. According to the posts published by each user every day, we obtain the number of times $cnt_{d,s,u}$ that user u mentions contestant s on day d (similarly, only the microblogs that mention only one contestant will be considered). Then, the mean number of daily mentions $< cnt_{d,s}>$ of each contestant can be obtained from the ratio of the total number of daily mentions of each contestant to the number of daily active followers of each contestant, i.e., $\langle cnt_{d,s}\rangle = \sum_u cnt_{d,s,u}/dau_{d,s}$, which reflects the loyalty of contestants' followers.

We define the retention number $r_{d+1,s}$ of followers of contestant s on the next day d+1 as the number of users who post microblogs that only mention that contestant on day d and d+1, namely the intersection of the active followers of day d and d+1. For each contestant, we define the number of followers who lose activity on day d+1 as the number of users who post microblogs that only mention that contestant on day d but not on day d+1, i.e. the difference $out_{d+1,s} = dau_{d,s} - r_{d+1,s}$ between the number of active followers on day d and the number of retained followers on day d+1. Finally, we define the losing rate $outRatio_{d+1,s}$ of each contestant's active followers on day d+1 as the ratio of the number of followers who lose activity on day d+1 to the number of active followers on day d, i.e. $outRatio_{d+1,s} = out_{d+1,s}/dau_{d,s} = (dau_{d,s} - r_{d+1,s})/dau_{d,s}$ which reflects contestants' popularity.

5 Effect estimation

5.1 Methods

In competitive programs, different participants will face different results after elimination. Some participants will be eliminated and lose the qualification to continue to participate in the following program, while others will be promoted and continue to participate in the program. Since the offline process is closely related to the online discussion, we will discuss the influence of elimination mechanism on the information diffusion related to the two types of contestants.

There are two dimensions for the elimination competitions, one is before and after eliminations, and the other is the promoted and eliminated contestants. If elimination matches are considered as an intervention, the significance of treatment effect after intervention can determine the existence of causality. We use the DID method to estimate the treatment effect of elimination matches.

According to the counterfactual reasoning, a group of samples similar to the treatment group is selected as the control group to obtain the results without treatment, and the difference between them is the treatment effect of the event. In the paper, the observation value of the treatment group is the diffusion performance of the relevant microblogs of the eliminated contestants in an elimination match, while the observation value of the control group is that of the promoted contestants in the same elimination match.

In the DID method, for the treatment and control groups, the first difference is the difference between them before elimination and the difference between them after elimination. The mean difference between the treatment group and the control group before elimination is $E(y_{it}|Treat_i = 1, Period_t = 0) - E(y_{it}|Treat_i = 0, Period_t = 0)$, and the mean difference after elimination is $E(y_{it}|Treat_i = 1, Period_t = 1) - E(y_{it}|Treat_i = 0, Period_t = 1)$. $Treat_i$ is the dummy variable of treatment: one represents elimination, and 0 no elimination. $Period_t$ is the dummy variable for the occurrence of events: 0 means before the occurrence of events, i.e., before eliminations, and one means after the occurrence of events, i.e., after eliminations. Y_{it} represents the observed value of the explained variable. For the treatment and control groups, the second difference is the difference between the difference between the two groups after the eliminations and the difference before the eliminations, i.e. the treatment effect of the elimination events. The average treatment effect on the treated (ATT) for an elimination match is $E(y_{it}|Treat_i = 1, Period_t = 1) - E(y_{it}|Treat_i = 0,$

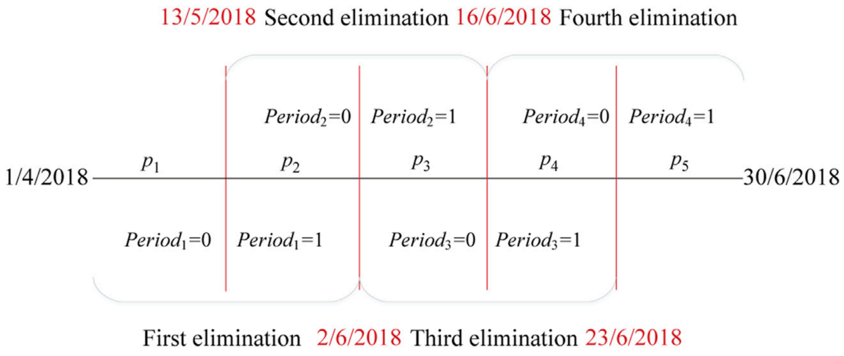

13/5/2018 Second elimination 16/6/2018 Fourth elimination
Period₂=0
Period₂=1
Period₄=0
Period₄=1
1/4/2018
p₁
p₂
p₃
p₄
p₅
30/6/2018
Period₁=0
Period₁=1
Period₃=0
Period₃=1
First elimination 2/6/2018 Third elimination 23/6/2018

FIGURE 2
Different periods of the program.

$Period_t = 1) - E(y_{it}|Treat_i = 1, Period_t = 0) + E(y_{it}|Treat_i = 0, Period_t = 0)$.

The DID method estimates the treatment effect by constructing a regression model with interaction terms, and the basic model is shown in Eq. 1:

$$Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Period_t + \beta_3 Treat_i \cdot Period_t + e_{it} \quad (1)$$

Let $Period_t = 0$, $Treat_i = 0$, and the observed value of the control group before the event is $\hat{Y}_{it} = \beta_0$; let $Period_t = 0$, $Treat_i = 1$, and the observed value of the treatment group before the event is $\hat{Y}_{it} = \beta_0 + \beta_1$; let $Period_t = 1$, $Treat_i = 0$, and the observed value of the control group after the event is $\hat{Y}_{it} = \beta_0 + \beta_2$; let $Period_t = 1$, $Treat_i = 1$, and the observed value of the treatment group after the event is $\hat{Y}_{it} = \beta_0 + \beta_1 + \beta_2 + \beta_3$. The difference between the two groups before the event can be expressed by β_1, and the difference between the two groups after the event can be expressed by β_1+β_3, thus the treatment effect of the event is β_3, i.e. causality can be determined by the coefficient of the interaction term.

5.2 Data partitioning

As shown in Figure 2, there are four elimination matches for the talent show, and the dates are May 13, June 2, June 16, and June 23. The four elimination matches divide the entire data set into five parts, and the treatment effect of each elimination match is discussed separately. The DID method is usually used in policy research, and the time span of the data studied is often very long. Considering some periodic fluctuations, data are usually studied in periods of a year or a month. There are often multiple periods of data before the implementation of policies. However, the time span of the entertainment program studied is less than 2 months, and the duration of each phase is shorter, even only a week. Thus, in the paper, only the two phases before and after the elimination match are considered in each analysis. For example, for the first elimination match, only the microblogs in p_1 and p_2 phases are considered, where p_1 is the phase before the elimination match, i.e., $Period_1 = 0$, and p_2 is the phase after the elimination match, i.e., $Period_1 = 1$.

To ensure the reasonability of the DID regression model, we put forward two assumptions. The first is stable unit treatment value assumption (SUTVA). It is difficult to rigorously test this hypothesis, and we adopt the interpretation method used by Weiler et al. to give reasons that SUTVA can be true in our study [49]. SUTVA consists of three aspects. 1) Individual independence assumption. According to the rules of the program, all contestants participate as individuals, not as multi-person teams, thus the results of the program will only affect individuals and their related posts, i.e., the elimination result of a contestant only affects the contestant's related posts, and there is no interaction between contestants. 2) The assumption of single treatment. There are only two results of elimination and promotion. The eliminated contestants, regardless of their specific ranking in the eliminated group, will leave the stage. The promoted contestants, regardless of their specific ranking in the promoted group, will compete on the stage until the next elimination. In other words, the treatment effect of elimination matches on eliminated contestants is the same, and the difference of contestants' rankings has no additional effect on the characteristics of online information diffusion associated with them. 3) The assumption of no interference in posts. The program studied has a wide range of audience. Users only discussed the status of the contestants they followed in their posts, and there may be no correlation between posts published by different users, i.e., different microblogs do not affect each other.

The second is parallel trend assumption, i.e., other factors have the same influence on the information diffusion characteristics of microblogs associated with eliminated and promoted contestants. That is, in the absence of elimination, the trend of the mean characteristics of the treatment group and the control group is parallel over time. As explained above, in the paper, we set one period before and after the elimination match in each analysis, and the DID method for one period of data may not be able to carry out the parallel trend test. However, this assumption can be satisfied with propensity score matching (PSM) which will be discussed in section 6.1.

6 Results

Previous studies have revealed the correlation between the performance of information diffusion and information topics [50], emotions expressed in texts [51], and user attributes [52]. In fact, in some cases, user attributes are not related to diffusion performance [27]. This study focuses on online interactions in an entertainment program context, and traditional cues such as user gender, or location can become irrelevant to information diffusion on contestants.

Recently textual data have been applied to causal inference studies, such as the influence of collective sentiment expressed in social media on stock market or cryptocurrency prices [53] or the causal effects of brevity of tweets on their success by controlled experiments [54]. Text characteristics can also be dependent variables. For example, Egami et al. studied how awareness about an individual's criminal history affects attitudes toward immigration using a survey experiment [55]. Besides texts can be confounders in causal analyses and Roberts et al. used text analysis to control for this type of confounding [56]. In this section we use microblog features as dependent variables or confounders to study the effect from elimination mechanism on information diffusion.

It is noteworthy that the entertainment program studied was presented in the form of live TV. Some users can watch the program on TV and then published relevant posts on social media. There are also some users who were informed of the program's progress through other channels rather than TV, and it is hard to obtain channel information from post texts. In this paper, offline means the program was presented on TV, we focus on the influence of elimination mechanism on the information diffusion, and may not focus on the channels through which users got information about the program. Besides since the user ID numbers in the data set have been encrypted, some of their important attributes, including centrality indices characterizing user influence, are unavailable, and we have applied the PSM method to try to address the endogeneity issue caused by user influence and inertia.

6.1 Posts

We estimate the treatment effect of elimination on the structural features of information diffusion trees in different phases by taking each elimination match as a time point. The original microblog content corresponding to each diffusion tree is analyzed, the microblogs related to eliminated contestants in each time period are taken as the treatment group, and the ones related to promoted contestants are taken as the control group. Eq. 1 can be further written as

$$\begin{aligned} Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Period_t + \beta_3 Treat_i \cdot Period_t + \gamma X_{it} \\ + \beta_4 Days + e_{it} \end{aligned} \tag{2}$$

where Y_{it} is the characteristics of information diffusion trees, such as diffusion scale (ln (*size*)), depth, width (ln (*width*)) and active time, and *Days* is the fixed effect refined to every day. We use also Chinese affective lexicon ontology to obtain the fine-grained emotions of each microblog which include happy, good, surprise, anger, sad, fear, and disgust, and the detailed calculation methods are also presented in the Supplementary Material. Control variable X_{it} contains content features of microblogs, such as emotional tendency, fine-grained emotions and topics.

Selection bias and the endogeneity problem among microblogs on the eliminated contestants may exist. The microblogs on eliminated and promoted contestants can differ substantially, meaning that they may not be directly comparable. Besides there are three confounding variables, i.e. post emotional tendency, fine-grained emotions and discussion topics. To study the influence of the results of elimination matches on the characteristics of information diffusion, we use PSM to match the control variables which is performed before the DID regression.

PSM converts multidimensional confounders according to the corresponding function (for example, logistic regression) to a one-dimensional propensity score p_{score}, which means the probability of the sample being treated, i.e. $p_{score}(X_i = x) = P(Treat_i = 1 \mid X_i = x)$, where p_{score} includes confounders. This method fits a probability to each sample in control group, and samples in treatment group find the one that is closest to their own in the control group. Specifically, PSM selects only those microblogs on promoted contestants who closely resemble microblogs on the eliminated ones. The objective of this approach can be thought of as "finding an artificial twin" that closely resembles a sample in treatment group. PSM solves the problem of sample matching between the treatment and the control group and can mitigate the endogeneity problem by controlling confounders [57]. After matching, the distribution of observable features of both groups is balanced, i.e. $E(X_i|Treat_i = 1, p_{score}(X_i)) = E(X_i|Treat_i = 0, p_{score}(X_i))$.

PSM can guarantee the homogeneity between control group and treatment group and the establishment of the assumption of

TABLE 1 DID regression results for diffusion tree characteristics for four elimination matches.

	Diffusion depth			Active time			Diffusion width			Diffusion scale		
	Promoted (1)	Eliminated (2)	Difference (2)–(1)	Promoted (1)	Eliminated (2)	Difference (2)–(1)	Promoted (1)	Eliminated (2)	Difference (2)–(1)	Promoted (1)	Eliminated (2)	Difference (2)–(1)
First												
Before	1.539	1.608	0.069**	402.880	292.226	-110.654***	1.154	1.249	0.096*	1.192	1.307	0.115**
After	1.908	1.757	-0.151***	427.109	404.184	-22.925	1.944	1.550	-0.394***	1.988	1.597	-0.392***
DID (β_3)			-0.220***			87.729***			-0.489***			-0.506***
N	3,050	3,050		3,050	3,050		3,050	3,050		3,050	3,050	
Second												
Before	1.595	1.507	-0.088	125.995	123.731	-2.264	1.575	1.437	-0.138*	1.620	1.477	-0.143*
After	1.714	1.542	-0.172**	137.613	137.581	-0.031	1.971	1.511	-0.460***	2.025	1.588	-0.437***
DID (β_3)			-0.084			2.232			-0.322***			-0.294**
N	1,868	1,868		1,868	1,868		1,868	1,868		1,868	1,868	
Third												
Before	1.623	1.616	-0.007	89.133	80.809	-8.324**	1.586	1.669	0.082	1.640	1.731	0.091*
After	1.658	1.670	0.011	88.268	96.573	8.305	1.536	1.795	0.258**	1.603	1.873	0.270***
DID (β_3)			0.019			16.629**			0.176			0.179
N	2,832	2,832		2,832	2,832		2,832	2,832		2,832	2,832	
Fourth												
Before	1.521	1.448	-0.073***	36.494	51.098	14.603***	1.334	1.265	-0.069*	1.388	1.309	-0.079*
After	1.621	1.665	0.044	86.201	92.312	6.110***	1.298	1.397	0.099**	1.389	1.487	0.098**
DID (β_3)			0.117***			-8.493***			0.168***			0.177***
N	4,910	4,910		4,910	4,910		4,910	4,910		4,910	4,910	

Significance: ***$p< 0.01$, **$p< 0.05$, *$p< 0.1$.

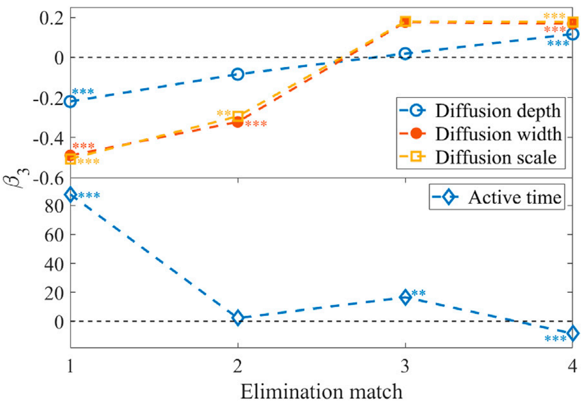


FIGURE 3
The β_3 values with significance level for DID regressions for diffusion tree characteristics.

long-term trend consistency to a certain extent. Then the matched samples are used for DID to ensure the applicability of DID method.

Table 1 shows the results of the DID regression model (please see Supplementary Material for summary statistics and distributions of variables, correlation matrices and VIF tests for independent variables, PSM results, and complete regression results). The difference column before elimination is the estimation for β_1, the difference column after elimination is the estimation for β_1+β_3, and finally the difference column corresponding to DID is the estimation for β_3. Figure 3 shows the values of β_3 with significance level indicated. We find that the first and fourth elimination matches have a significant effect on diffusion depth, the first, third and fourth elimination ones have a significant effect on active time, and the first, second and fourth elimination ones have a significant effect on diffusion width and scale.

The first and last elimination matches have a significant effect on the metrics of information diffusion trees. In the first elimination match, the audience does not have a deep understanding of the contestants, the eliminated contestants do not have enough time to show themselves and gain high popularity, and they leave the stage in a hurry. The microblogs related to the eliminated contestants may also gradually become silent, and the attention of ordinary audiences will be shifted to the promoted contestants. Elimination results have a negative effect on diffusion depth, width and scale of posts on eliminated contestants (β_3 = -0.220 for depth, β_3 = -0.489 for width, and β_3 = -0.506 for scale). Even in the second elimination match, elimination results still have a negative effect on diffusion width and scale. On the contrary, the elimination results have a positive effect on the active time of posts on eliminated contestants (β_3 = 87.729). One possible reason is that the promoted contestants are still active on the stage, and the original microblogs related to them emerge every day. Users tend to forward the latest related microblogs, and the forwarding frequency of past posts can reduce, thus the active time of the original microblogs is short. While the eliminated contestants lose the opportunity to perform on stage, the number of new posts related to the contestants can decrease, and some users choose to forward the past posts related to them, which may increase the mean active time of the posts.

At different stages of the program, the elimination matches result in different treatment effects, which may be related to the external factors associated with contestants in offline events. In the final elimination, both the eliminated and the final winners have accumulated a large number of followers over the course of the program. The treatment effect of elimination on diffusion depth, width and scale of the microblogs on eliminated contestants is positive (β_3 = 0.117 for depth, β_3 = 0.168 for width, and β_3 = 0.177 for scale). The possible reason is that the eliminated contestants leave the stage earlier, have access to industry resources faster than the winning contestants, and are known to users outside the program. After the final round, users can forward more new posts on eliminated contestants than the old ones, the active time of posts decreases, and finally the final match has a negative effect on it (β_3 = -8.493).

6.2 Users

Elimination matches not only have a significant impact on the diffusion tree characteristics of the original microblogs, but have a certain impact on the emotional tendency of users to contestants and the topics they talk about on contestants. We study the effect of elimination matches at the user level. Since information diffusion features cannot affect content features, it is unnecessary to match them when constructing the regression models. First, we explore the influence of elimination matches on users' emotional tendency, and Eq. 3 gives the DID regression model:

$$Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Period_t + \beta_3 Treat_i \cdot Period_t + \gamma Topic_{it} + \beta_4 Days + e_{it} \tag{3}$$

where $Topic_{it}$ is the topic classification of microblogs which is the control variable and is represented by a dummy variable. In this context PSM is used to match topic categories.

Table 2 shows the impact of elimination matches on users' emotional tendency (please see Supplementary Material for more information on variables, PSM results and complete regression results), and Figure 4 shows the values of β_3 with significance level indicated. In the first two matches, elimination results have no significant effect on users' emotional tendency, but in the latter two, elimination results have significant negative effect on

TABLE 2 DID regression results for sentiment tendency for four elimination matches.

		Promoted (1)	Eliminated (2)	Difference (2)–(1)	$p>\|t\|$	N
First						
	Before	1.477	1.468	-0.009	0.925	3,785
	After	2.133	2.119	-0.014	0.918	2,315
	DID			-0.005	0.977	
	N	3,050	3,050			
Second						
	Before	0.744	0.812	0.068	0.580	2079
	After	0.596	0.874	0.278**	0.046	1,657
	DID			0.209	0.265	
	N	1,868	1,868			
Third						
	Before	0.315	0.484	0.169*	0.059	4,772
	After	1.127	-0.398	-1.525***	0.000	892
	DID			-1.693***	0.000	
	N	2,832	2,832			
Fourth						
	Before	0.457	0.778	0.321***	0.000	5,496
	After	1.137	0.633	-0.504***	0.000	4,324
	DID			-0.825***	0.000	
	N	4,910	4,910			

Significance: ***$p< 0.01$, **$p< 0.05$, *$p< 0.1$.

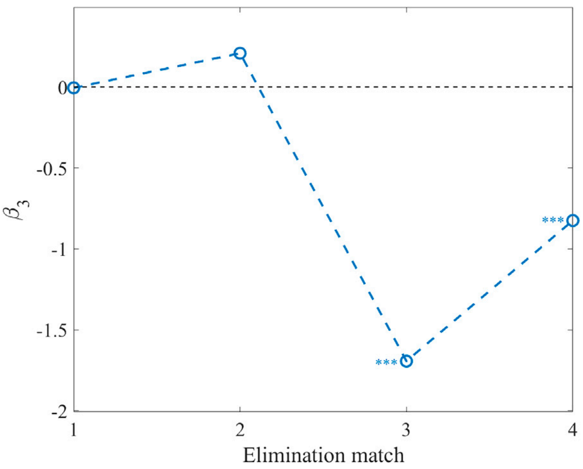


FIGURE 4
The β_3 values with significance level for DID regressions for sentiment tendency.

emotional tendency. The possible reason is that in the early stage, users do not have a deep understanding of the contestants and take a wait-and-see attitude. They might just be ordinary audience of the program, rather than fans of the contestants. Therefore, there could be less of a gap in expectations and less emotional volatility when the elimination results were announced. As can be seen from the p-value, there is almost no difference between the treatment group and the control group in the first elimination match, no matter before or after the competition, and the treatment effect of the elimination match is almost zero. In the second elimination match, some users began to support their favorite contestants, and the emotional tendency to eliminated contestants increases ($p < 0.05$), but the difference before and after the match is not significant.

The topics discussed by users in social media are also closely related to the program, and the elimination results may also affect the topics discussed by users. Eq. 4 shows the DID logistic regression model for the effect:

$$\ln[p(Y_{it}=1)/(1-p(Y_{it}=1)] \\ = \beta_0 + \beta_1 Treat_i + \beta_2 Period_t + \beta_3 Treat_i \cdot Period_t \\ + \gamma Sentiment_{it} + \beta_4 Days + e_{it} \tag{4}$$

where Y_{it} is the topic category which is a binary classification variable and is represented by a dummy variable. As control variables, $Sentiment_{it}$ includes emotional tendency and fine-grained sentiment. PSM is also used to match the control variables.

Table 3 gives the regression results, with the values in each column representing estimates of the differences for each topic (please see Supplementary Material for more

TABLE 3 DID regression results for post topics for four elimination matches.

	Endorsement	Stage performance	Program criticism	Program publicity	Praise and encouragement	Contestant activities
First						
Before	0.026*	0.011	0.011	-0.013	-0.014	-0.014
After	-0.053***	0.065***	0.013	-0.025*	-0.020	-0.001
DID	-0.079***	0.053**	0.002	-0.012	-0.006	0.012
N	6,100	6,100	6,100	6,100	6,100	6,100
Second						
Before	-0.019	0.025	0.006	0.017	0.033*	-0.053***
After	-0.048**	0.005	-0.010	0.008	-0.049**	0.077***
DID	-0.028	-0.020	-0.016	-0.009	-0.082***	0.130***
N	3,736	3,736	3,736	3,736	3,736	3,736
Third						
Before	0.036***	-0.032**	0.000	-0.002	0.024*	-0.024**
After	-0.098***	0.070**	0.000	-0.005	-0.061**	0.060***
DID	-0.134***	0.102***	-0.000	-0.002	-0.085***	0.084***
N	5,664	5,664	5,664	5,664	5,664	5,664
Fourth						
Before	0.065***	-0.060***	0.009**	0.002	0.036***	-0.043***
After	-0.018	0.054***	-0.017***	-0.038***	-0.008	0.013
DID	-0.084***	0.113***	-0.026***	-0.041***	-0.044**	0.056***
N	9,820	9,820	9,820	9,820	9,820	9,820

Significance: ***$p<$ 0.01, **$p<$ 0.05, *$p<$ 0.1.

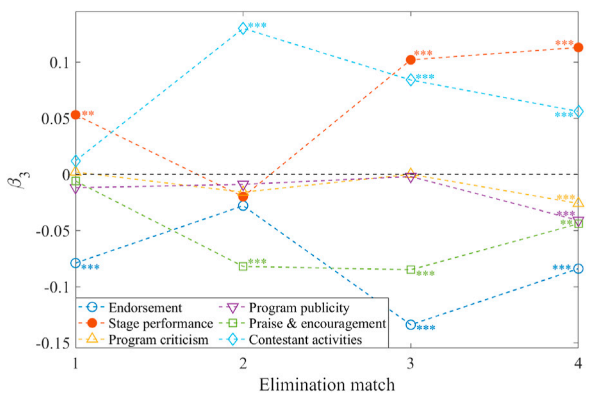


FIGURE 5
The β_3 values with significance level for DID regressions for post topics.

information on variables, PSM results and complete regression results), and Figure 5 shows the values of β_3 with significance level indicated. We find that the early elimination results have a significant impact on the topics of endorsement and stage performance. The elimination results of the middle stage have significant influence on the topics of endorsement, stage performance, praise and encouragement and contestant activities, and the late stage has significant influence on all topics.

From the perspective of post topics, the discussion on the endorsement topic decreases significantly after several eliminations. The results of the elimination matches are closely related to the audience's vote, and Weibo is one of the important platforms to support and attract users to vote. After the elimination matches, the result has been decided, and thus the discussion on eliminated contestants on the topic of calling for vote and support decreases. However, the discussion on the topic of stage performance increases significantly after the elimination matches. The stage performance in the elimination competition attracts much attention. After the elimination, users could discuss the performance of each contestant in combination with the result of the competition. By contrast, users may pay more attention to the performance of the eliminated contestants to analyze the reasons for elimination. The program criticism and propaganda topics are related to the program itself, and there is no significant relationship with most elimination results. In the last elimination match, the posts published by users on program criticism and propaganda topics for eliminated contestants decrease significantly.

TABLE 4 DID regression results for user behavioral characteristics for four elimination matches.

	$outRatio_{d+1,s}$ (popularity)			$<cnt_{d,s}>$ (loyalty)		
	Promoted (1)	Eliminated (2)	Difference (2)–(1)	Promoted (1)	Eliminated (2)	Difference (2)–(1)
First						
Before	0.569	0.563	-0.006	1.416	1.458	0.042
After	0.443	0.485	0.042***	-0.037	0.016	0.053
DID			0.048***			0.011
N	2,965	1,768		2,965	1,768	
Second						
Before	0.634	0.590	-0.044***	1.912	1.878	-0.034
After	0.499	0.575	0.076***	1.754	1.786	0.032
DID			0.120***			0.066
N	2,341	1,416		2,341	1,416	
Third						
Before	0.706	0.717	0.011	1.897	1.938	0.041
After	0.662	0.798	0.135***	1.836	1.882	0.046
DID			0.124***			0.005
N	1,575	1,016		1,575	1,016	
Fourth						
Before	0.715	0.741	0.026***	1.922	1.854	-0.068*
After	0.695	0.826	0.130***	2.349	2.108	-0.241*
DID			0.104***			-0.173
N	864	821			864	821

Significance: $^{***}p< 0.01$, $^{**}p< 0.05$, $^{*}p< 0.1$.

About the eliminated contestants, posts on praise and encouragement topic in the last three elimination matches drop significantly. The possible reasons are, on the one hand, because some contestants are eliminated, users can have negative emotions and the posts on praise and encouragement topic reduce. On the other hand, after each elimination, the competition between the contestants weakens, reducing the discussion on praise and encouragement topic to some extent. Users' discussion on the topic of contestant activities increases significantly in the last three elimination rounds. After each elimination, the eliminated contestants quit the program and could carry out personal activities without following the relevant regulations of the program. As a result, the discussion on contestants' activities increases.

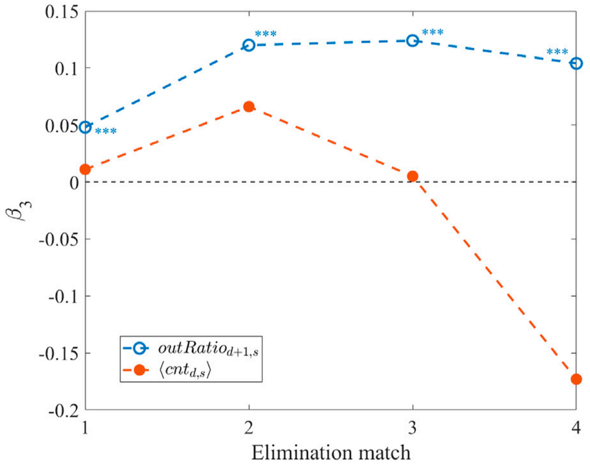


FIGURE 6
The β_3 values with significance level for DID regressions for user behavioral characteristics.

6.3 Contestants

The impact of elimination matches on information diffusion in social media is not only reflected at the levels of posts and users, but also at the level of contestants. Elimination matches may change the behavior pattern of contestants' followers when they posted microblogs related to the contestants. Based on the original microblogs, we obtain the losing rate $outRatio_{d+1,s}$ of each contestant's active followers on day $d+1$ and the mean number of daily mentions $< cnt_{d,s}>$ of each contestant, and Eq. 5 gives the DID regression model:

$$Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Period_t + \beta_3 Treat_i \cdot Period_t + \beta_4 Days + e_{it} \tag{5}$$

where Y_{it} is $outRatio_{d+1,s}$ or $< cnt_{d,s} >$. At the contestant level, the behavioral characteristics of users are concerned with individual contestant rather than individual post, thus the information diffusion characteristics and content characteristics cannot be quantified, and the PSM is not used.

Table 4 shows the regression results (please see Supplementary Material for more information on variables and complete regression results), and Figure 6 shows the values of β_3 with significance level indicated. We find that for all four elimination matches, the results of elimination have a significant positive impact on the eliminated contestants' $outRatio_{d+1,s}$, i.e., weakening their popularity, but have no significant impact on their $< cnt_{d,s} >$, that is, they do not affect the loyalty of contestants' followers.

Elimination matches increase the $outRatio_{d+1,s}$ of the eliminated contestants. According to regression results, the influence of the first elimination ($\beta_3 = 0.048$) is lower than those of the subsequent eliminations, and the contestants' follower group may form in the second elimination. In the first elimination, contestants could have less followers and the change of the losing rate is small. In the fourth elimination competition, although $< cnt_{d,s} >$ of eliminated contestants is significantly lower than that of promoted contestants after elimination, this trend has appeared before the competition, and finally, the elimination match has no significant effect on $< cnt_{d,s} >$ of eliminated contestants.

7 Heterogeneity analysis and robustness test

Users of different characteristics may have different reactions to the program studied. Given that demographic information on users is often unavailable on Weibo platform due to specific profile settings or not filling in information, we perform the heterogeneity analysis to examine the variation of effects for users of different genders (the missing rate of gender attribute is low) and obtain the DID regression results (due to space limitation, we present the results in Supplementary Material).

We find that for the effect of elimination results on diffusion tree characteristics, overall the DID regression results for female users are qualitatively consistent with those in Table 1. For male users, elimination results never have a positive effect on diffusion depth, width and scale, which means that elimination results have a negative effect or no significant effect on the metrics of posts by male users on eliminated contestants.

For the effect of elimination results on sentiment tendency, we find that the DID regression results are qualitatively consistent with those in Table 2 both for female and male users. In the last two elimination matches, elimination results have a negative effect on followers' emotional tendency. For the effect of elimination results on post topics, we find that overall the DID regression results for female users are also qualitatively consistent with those in Table 3. For male users, most regression results are insignificant, which means that elimination results have no significant effect on the topics of posts by male users on eliminated contestants. However, the signs of significant results are consistent with those in Table 3.

We also perform robustness test by supplementing variables. Specifically, we added two control variables on user characteristics to the regression equations, i.e., users' gender and region (province level) where they are located. The DID regression results are also presented in the Supplementary Material, and we find that the conclusions in the paper still hold which indicates the robustness of the conclusions.

8 Conclusion

In this paper, we study the effect of an influential and multi-phase entertainment program on related information diffusion and explore the underlying mechanisms. We find that elimination mechanism significantly influences the features of information diffusion trees, and elimination results negatively affect followers' emotional tendency. Elimination results also negatively affect the topics on endorsement and praise and encouragement discussed by users, and positively affect the topics on stage performance and contestant activities. Besides elimination results negatively affect participants' popularity, but do not affect the followers' loyalty to participants. The methods of this study are generalizable to some extent. Except entertainment events with multiple rounds of elimination, the approach in this paper could apply to research on information diffusion of offline events in different domains, for instance, sports events with multiple rounds of elimination or multi-phase political events.

There are several limitations for the paper. We divide the contestants into the promoted and eliminated groups to estimate the treatment effect of elimination match. In fact, even for contestants in the same group, elimination matches may have different effects. Besides, the conclusions of this paper may lack universality. Considering the different participants, audiences, and contexts of events, different conclusions may emerge, and we need to study more events or scenarios in detail and comparatively. Further there can be mutual influence between online discussion and offline events, which can cause the problem of reverse causality. Reverse causality can cause endogeneity problems which are mainly caused by the four reasons: omitted variables, sample selection bias/self-selection bias, reverse causality, and measurement error, and the problems can be dealt with in a number of ways, such as instrumental variable (IV), Heckman model, fixed effects model, DID, regression discontinuity, and PSM. In the paper we utilize PSM to address the issue. However, PSM only controls the influence of measurable confounders, does not fundamentally solve the endogeneity problem caused by selection bias or omitted variables, and also can not solve the problem of reverse causality. Generally, IV method can address the four endogeneity problems, and for the reverse causality IV method is an option. All of these give potential directions for future further research.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization, HH, YL, and NX. Methodology, QL, NX, HH, and YL. Software, QL and NX. Visualization, NX. Formal analysis, NX, QL, and HH. Writing, NX, QL, and HH.

Funding

The study was partially supported by the National Natural Science Foundation of China (Grant Nos. 61973121 and 71971082).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2022.1032913/full#supplementary-material

References

1. Zhang ZK, Liu C, Zhan XX, Lu X, Zhang CX, Zhang YC. Dynamics of information diffusion and its applications on complex networks. *Phys Rep* (2016) 651:1–34. doi:10.1016/j.physrep.2016.07.002

2. Wang X, Lan Y, Xiao J. Anomalous structure and dynamics in news diffusion among heterogeneous individuals. *Nat Hum Behav* (2019) 3:709–18. doi:10.1038/s41562-019-0605-7

3. Zhou B, Pei S, Muchnik L, Meng X, Xu X, Sela A, et al. Realistic modelling of information spread using peer-to-peer diffusion patterns. *Nat Hum Behav* (2020) 4:1198–207. doi:10.1038/s41562-020-00945-1

4. Xie J, Meng F, Sun J, Ma X, Yan G, Hu Y. Detecting and modelling real percolation and phase transitions of information on social media. *Nat Hum Behav* (2021) 5:1161–8. doi:10.1038/s41562-021-01090-z

5. Fu P, Zhu A, Fang Q, Wang X. Modeling periodic impulsive effects on online TV series diffusion. *PLoS ONE* (2016) 11(9):e0163432. doi:10.1371/journal.pone.0163432

6. Chung TL, Johnson O, Hall-Phillips A, Kim K. The effects of offline events on online connective actions: An examination of #BoycottNFL using social network analysis. *Comput Hum Behav* (2021) 115:106623. doi:10.1016/j.chb.2020.106623

7. Burnap P, Williams ML, Sloan L, Rana O, Housley W, Edwards A, et al. Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Soc Netw Anal Min* (2014) 4:206. doi:10.1007/s13278-014-0206-4

8. Zhou Q. Detecting the public's information behaviour preferences in multiple emergency events. *J Inf Sci* (2022) 016555152110277. doi:10.1177/01655515211027789

9. Kim J, Hastak M. Social network analysis: Characteristics of online social networks after a disaster. *Int J Inf Manage* (2018) 38:86–96. doi:10.1016/j.ijinfomgt.2017.08.003

10. Pourebrahim N, Sultana S, Edwards J, Gochanour A, Mohanty S. Understanding communication dynamics on twitter during natural disasters: A case study of Hurricane Sandy. *Int J Disaster Risk Reduction* (2019) 37:101176. doi:10.1016/j.ijdrr.2019.101176

11. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland China: Observational infoveillance study. *J Med Internet Res* (2020) 22(5):e19421. doi:10.2196/19421

12. Gallotti R, Valle F, Castaldo N, Sacco P, De Domenico M. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nat Hum Behav* (2020) 4:1285–93. doi:10.1038/s41562-020-00994-6

13. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing public opinion towards COVID-19 vaccines with Twitter data in the United States: Spatiotemporal perspective. *J Med Internet Res* (2021) 23(9):e30854. doi:10.2196/30854

14. Wang S, Huang X, Hu T, Zhang M, Li Z, Ning H, et al. The times, they are a-changin': Tracking shifts in mental health signals from early phase to later phase of the COVID-19 pandemic in Australia. *BMJ Glob Health* (2022) 7:e007081. doi:10.1136/bmjgh-2021-007081

15. Wang J, Fan Y, Palacios J, Chai Y, Guetta-Jeanrenaud N, Obradovich N, et al. Global evidence of expressed sentiment alterations during the COVID-19 pandemic. *Nat Hum Behav* (2022) 6:349–58. doi:10.1038/s41562-022-01312-y

16. Van Laer J. Activists online and offline: The Internet as an information channel for protest demonstrations. *Mobilization: Int Q* (2010) 15(3):347–66. doi:10.17813/maiq.15.3.8028585100245801

17. González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y. The dynamics of protest recruitment through an online network. *Sci Rep* (2011) 1:197. doi:10.1038/srep00197

18. González-Bailón S, Borge-Holthoefer J, Moreno Y. Broadcasters and hidden influentials in online protest diffusion. *Am Behav Scientist* (2013) 57(7):943–65. doi:10.1177/0002764213479371

19. Fernandez-Planells A, Figueras-Maz M, Pàmpols CF. Communication among young people in the #spanishrevolution: Uses of online-offline tools to obtain information about the #acampadabcn. *New Media Soc* (2014) 16:1287–308. doi:10.1177/1461444814530097

20. Varol O, Ferrara E, Ogan CL, Menczer F, Flammini A. Evolution of online user behavior during a social upheaval. In: *Proceedings of the 2014 ACM conference on web science*. New York: ACM Press (2014). p. 81–90.

21. Theocharis Y, Lowe W, van Deth JW, García-Albacete G. Using twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Inf Commun Soc* (2015) 18:202–20. doi:10.1080/1369118X.2014.948035

22. Park SJ, Lim YS, Park HW. Comparing twitter and YouTube networks in information diffusion: The case of the "Occupy Wall Street" movement. *Technol Forecast Soc Change* (2015) 95:208–17. doi:10.1016/j.techfore.2015.02.003

23. Steinert-Threlkeld ZC, Mocanu D, Vespignani A, Fowler J. Online social networks and offline protest. *EPJ Data Sci* (2015) 4:19. doi:10.1140/epjds/s13688-015-0056-y

24. González-Bailón S, Wang N. Networked discontent: The anatomy of protest campaigns in social media. *Social Networks* (2016) 44:95–104. doi:10.1016/j.socnet.2015.07.003

25. Vasi IB, Suh CS. Online activities, spatial proximity, and the diffusion of the Occupy Wall Street movement in the United States. *Mobilization: Int Q* (2016) 21(2):139–54. doi:10.17813/1086-671x-22-2-139

26. Ahmed S, Jaidka K, Cho J. Tweeting India's nirbhaya protest: A study of emotional dynamics in an online social movement. *Soc Move Stud* (2017) 16: 447–65. doi:10.1080/14742837.2016.1192457

27. Venkatesan S, Valecha R, Yaraghi N, Oh O, Rao HR. Influence in social media: An investigation of tweets spanning the 2011 Egyptian revolution. *MIS Q* (2021) 45: 1679–714. doi:10.25300/misq/2021/15297

28. Morales PR, Cointet JP, Froio C. Posters and protesters. *J Comput Soc Sci* (2022). doi:10.1007/s42001-022-00163-x

29. Aragón P, Kappler KE, Kaltenbrunner A, Laniado D, Volkovich Y. Communication dynamics in twitter during political campaigns: The case of the 2011 Spanish national election. *Policy Internet* (2013) 5:183–206. doi:10.1002/1944-2866.poi327

30. Xu WW, Sang Y, Blasiola S, Park HW. Predicting opinion leaders in Twitter activism networks: The case of the Wisconsin recall election. *Am Behav Scientist* (2014) 58(10):1278–93. doi:10.1177/0002764214527091

31. Segesten AD, Bossetta M. A typology of political participation online: How citizens used Twitter to mobilize during the 2015 British general elections. *Inf Commun Soc* (2017) 20:1625–43. doi:10.1080/1369118X.2016.1252413

32. Gorodnichenko Y, Pham T, Talavera O. Social media, sentiment and public opinions: Evidence from #Brexit and #USElection. *Eur Econ Rev* (2021) 136:103772. doi:10.1016/j.euroecorev.2021.103772

33. Yu X, Mashhadi A, Boy J, Nielsen RC, Hong L. Causal impact model to evaluate the diffusion effect of social media campaigns. In: *Proceedings of the 20th European conference on computer-supported cooperative work* (2022).

34. Balawi RA, Hu Y, Qiu L. Brand crisis and customer relationship management on social media: Evidence from a natural experiment from the airline industry. *Inf Syst Res* (2022). doi:10.1287/isre.2022.1159

35. Falavarjani SAM, Jovanovic J, Fani H, Ghorbani AA, Noorian Z, Bagheri E. On the causal relation between real world activities and emotional expressions of social media users. *J Assoc Inf Sci Technol* (2021) 72:723–43. doi:10.1002/asi.24440

36. Pearl J. Causal inference in statistics: An overview. *Stat Surv* (2009) 3:96–146. doi:10.1214/09-ss057

37. Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A. A survey on causal inference. *ACM Trans Knowl Discov Data* (2021) 15(5):1–46. Article 74. doi:10.1145/3444944

38. Si M, Cui L, Guo W, Li Q, Liu L, Lu X, et al. A comparative analysis for spatio-temporal spreading patterns of emergency news. *Sci Rep* (2020) 10:19472. doi:10.1038/s41598-020-76162-7

39. Li H, Xia C, Wang T, Wen S, Chen C, Xiang Y. Capturing dynamics of information diffusion in SNS: A survey of methodology and techniques. *ACM Comput Surv* (2023) 55(1):1–51. Article No.: 22. doi:10.1145/3485273

40. Xu L, Lin H, Pan Y, Ren H, Chen J. Constructing the affective lexicon ontology. *J China Soc Scientific Tech Inf* (2008) 27(2):180–5.

41. Chen J, Lin H. Constructing the affective commonsense knowledgebase. *J China Soc Scientific Tech Inf* (2009) 28(4):492–8.

42. Ren G, Hong T. Investigating online destination images using a topic-based sentiment analysis approach. *Sustainability* (2017) 9:1765. doi:10.3390/su9101765

43. Liu SM, Chen JH. A multi-label classification based approach for sentiment classification. *Expert Syst Appl* (2015) 42:1083–93. doi:10.1016/j.eswa.2014.08.036

44. Ye Y, Long T, Liu C, Xu D. The effect of emotion on prosocial tendency: The moderating effect of epidemic severity under the outbreak of COVID-19. *Front Psychol* (2020) 11:588701. doi:10.3389/fpsyg.2020.588701

45. Qiu J, Xu L, Wang J, Gu W. Mutual influences between message volume and emotion intensity on emerging infectious diseases: An investigation with microblog data. *Inf Manage* (2020) 57:103217. doi:10.1016/j.im.2019.103217

46. Cheng X, Yan X, Lan Y, Guo J. Btm: Topic modeling over short texts. *IEEE Trans Knowl Data Eng* (2014) 26:2928–41. doi:10.1109/tkde.2014.2313872

47. Shi L, Song G, Cheng G, Liu X. A user-based aggregation topic model for understanding user's preference and intention in social network. *Neurocomputing* (2020) 413:1–13. doi:10.1016/j.neucom.2020.06.099

48. Li X, Wang Y, Zhang A, Li C, Chi J, Ouyang J. Filtering out the noise in short text topic modeling. *Inf Sci* (2018) 456:83–96. doi:10.1016/j.ins.2018.04.071

49. Weiler M, Stolz S, Lanz A, Schlereth C, Hinz O. Social capital accumulation through social media networks: Evidence from a randomized field experiment and individual-level panel data. *MIS Q* (2022) 46:771–812. doi:10.25300/misq/2022/16451

50. Romero DM, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In: *Proceedings of the 20th international conference on world wide web*. New York: ACM Press (2011). p. 695–704.

51. Fan R, Zhao J, Chen Y, Xu K. Anger is more influential than joy: Sentiment correlation in Weibo. *PLoS ONE* (2014) 9(10):e110184. doi:10.1371/journal.pone.0110184

52. Choudhury MD, Sundaram H, John A, Seligmann DD, Kelliher A. *Birds of a Feather": Does user homophily impact information diffusion in social media?* (2010). Available at: https://arxiv.org/abs/1006.1702 (Accessed October 6, 2022).

53. Keskin Z, Aste T. *Information-theoretic measures for non-linear causality detection: Application to social media sentiment and cryptocurrency prices* (2019). Available at: https://arxiv.org/abs/1906.05740 (Accessed October 18, 2022).

54. Gligorić K, Anderson A, West R. Causal effects of brevity on style and success in social media. *Proc ACM Hum Comput Interact* (2019) 3:1–23. Article No.: 45. doi:10.1145/3359147

55. Egami N, Fong CJ, Grimmer J, Roberts ME, Stewart BM. *How to make causal inferences using texts* (2018). Available at: https://arxiv.org/abs/1802.02163 (Accessed October 20, 2022).

56. Roberts ME, Stewart BM, Nielsen RA. Adjusting for confounding with text matching. *Am J Polit Sci* (2020) 64:887–903. doi:10.1111/ajps.12526

57. Imbens GW, Rubin DB. *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press (2015).

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 17 April 2023
DOI 10.3389/fphy.2023.1017251

Mechanism of supply chain coordination based on price discount with privacy protection in one-supplier-one-buyer system

Cui-Hua Xie[1], Jing-Chun Zhan[2], Le-Tian Zeng[2] and Shao-Yong Han[1,3,4]*

[1]School of Economics and Management, Wenzhou University of Technology, Wenzhou, China, [2]School of Computing and Business, Universiti Teknologi Malaysia, Johor Bahru, Malaysia, [3]Postdoctoral Scientific Research Workstation, Bank of Zhengzhou, Zhengzhou, China, [4]School of Information Engineering and Technology, Changzhou Vocational Institute of Industry Technology, Changzhou, China

Check for updates

OPEN ACCESS

EDITED BY
Libo Zhang,
Southwest University, China

REVIEWED BY
Rende Li,
University of Shanghai for Science and Technology, China
Jinjing Shi,
Central South University, China

*CORRESPONDENCE
Shao-Yong Han,
✉ hanshaoyong@zuaa.zju.edu.cn

RECEIVED 11 August 2022
ACCEPTED 03 April 2023
PUBLISHED 17 April 2023

CITATION
Xie C-H, Zhan J-C, Zeng L-T and Han S-Y (2023), Mechanism of supply chain coordination based on price discount with privacy protection in one-supplier-one-buyer system.
Front. Phys. 11:1017251.
doi: 10.3389/fphy.2023.1017251

It is of great economic significance to optimize the total cost and improve the performance of the supply chain. In this paper, we assume that the market demand is random, and the seller and the buyer share information and make decisions together. We analyze the optimal joint order quantity under probabilistic demand and design the quantity discount model and profit distribution mechanism. Under a certain quantity discount mechanism and profit distribution strategy, both the seller and the buyer can reduce costs. The quantity discount model and profit distribution mechanism designed require supply chain members to share information. In order to protect the privacy of members and improve the willingness of supply chain members to share information, we designed a privacy protection joint ordering policy protocol and privacy protection quantity discount policy based on Secure multiparty computation technology. Then, the joint ordering strategy, the privacy-preserving joint ordering strategy, and quantity discount protocol are numerically simulated. The numerical simulation results show that the privacy-preserving quantity discount coordination mechanism designed by us can reduce the cost of supply chain members to varying degrees and effectively protect the shared information of supply chain members. This work is helpful to the research of cost optimization of the system in complex supply chain systems.

KEYWORDS

supply chain management, quantity coordination, privacy protection, secure multiparty computation, probabilistic demand

1 Introduction

In the social economic network, it is of great economic significance to optimize the total cost and improve the performance of the supply chain. Supply chain literature considers the one-supplier, one-buyer system as the basic building block [1], and in the socio-economic system, the seller can be a manufacturer or wholesaler, and the buyer can be a distributor who faces random market demands. In the traditional mode, the buyer and the seller make decentralized decisions, both based on maximizing their own interests. The buyer usually chooses economic order quantity (EOQ) as his optimal order quantity, however, for the seller, the buyer's order quantity is different each time, and the buyer's order time is also uncertain. In order to cope with the uncertainty of order demand, the seller needs to maintain a high inventory, therefore, will inevitably sell to the buyer at a higher price. This leads to the overall high cost and low efficiency of the supply chain system. The efficiency improvement of supply chain systems has become a hot research topic.

The price discount strategy originated from Monahan's research in 1984 [2]. He assumed that the market demand was constant and had nothing to do with the product price, and the buyer's order uses a lot-for-lot model, finding that the seller can change the buyer's order quantity through price discounting strategies to increase profits. Later, many scholars began to apply the price discount strategy in the performance optimization of the supply chain system. At first, scholars assumed that the market demand was constant, and then progressed to the situation that the market demand was random.

Many supply chain cost optimization strategies require supply chain members to share information [3–7]. However, supply chain members may use the shared private cost information [8]. This may cause the supply chain members to lose their competitive advantages and cause them many hidden dangers. For example, in a supply chain cooperation system, the downstream enterprises providing their own private information to the upper will enhance the authority of the upper in the supply chain, making the downstream enterprises at a disadvantage in the negotiations and losing the profit advantage. Although information sharing is the key to achieving enterprise cooperation, driven by the pursuit of individual interests, enterprises may make individual optimal choices that are contrary to the overall optimal. Even, information leakage exists in the supply chain system [9], which will lead to various fraud events, resulting in the loss of enterprise funds.

Privacy information protection [10] and information security [11] have been paid more and more attention, and its related technologies have also been greatly developed, such as blockchain technology [12, 13], secure multi-party computing, and so on. Secure multiparty computation (SMC) originated from Yao Qizhi's millionaire problem in 1982 [14], and was discussed in detail and systematically by Micali et al. [15]. SMC provides a framework for computing partners, mainly studying how to design secure computing contract functions without a trusted third party. SMC has attracted researchers' attention. Scholars began to study the application of secure multi-party computing to privacy protection in information-sharing scenarios.

In this paper, we are interested in the mechanism of supply chain coordination based on price discount and privacy protection in a one-supplier-one-buyer system, and the reasons are, on one hand, quantity discount mechanism for joint-ordering in a one-supplier-one-buyer system has yet not been reported, on the other hand, the privacy protection of information sharing in price discount mechanism using SMC technology has not been studied. The main contributions and significance of this paper are summarized as follows:

1) In the classic one-supplier-one-buyer supply chain system, the buyer's order adopts the economic order quantity mode; the buyer and seller make independent decisions, and the cost of the overall supply chain system is high. We assume that the market demand is random, and the seller and the buyer share information and make decisions together like two departments of the same company. We analyze the optimal joint order quantity under probabilistic demand and design the quantity discount model and profit distribution mechanism. Under a certain quantity discount mechanism and profit distribution strategy, both the seller and the buyer can reduce costs.
2) The quantity discount model and profit distribution mechanism designed require supply chain members to share information. However, after all, the seller and the buyer are independent companies. On one hand, they may not want the other party to know their private information; On the other hand, even though both parties are willing to share information, they are afraid to share information because they are worried about the harm caused by information leakage. In order to protect the privacy of members and improve the willingness of supply chain members to share information, we designed a privacy protection joint ordering policy protocol and privacy protection quantity discount policy based on SMC technology. It is implemented without using intermediaries and does not disclose the private information of members. Then, the joint ordering strategy, the privacy-preserving joint ordering strategy and the quantity discount protocol are numerically simulated. The numerical simulation results show that the privacy-preserving quantity discount coordination mechanism designed by us can reduce the cost of supply chain members to varying degrees and effectively protect the shared information of supply chain members.

The rest of the paper is organized as follows. Section 2 introduces the work of other researchers related to this paper; Section 3 describes the basic model based on EOQ; In section 4, We designed the improved model based on price discount and quantity coordination; We designed a privacy protection joint ordering policy protocol and privacy protection quantity discount policy based on SMC technology in Section 5; some simulations are performed to validate the effectiveness and feasibility of the proposed mechanism of supply chain coordination based on price discount and privacy protection in Sections 6, 7 is conclusion and discussion. The construction of the basic sub-protocol oblivious transfer (OT) and the content of the privacy protocol are introduced in Supplementary Appendix S1–S4, where Supplementary Appendix S1 (protocol 1: an oblivious transfer protocol) and Supplementary Appendix S2 (protocol 2: a secure two-party real product protocol) are the basic protocols of cryptography and are the basis for constructing Supplementary Appendix S3, S4.

2 Related work

2.1 Economic order quantity

In a supply chain system, in the order decision of the seller, how to determine the quantity of raw materials ordered for the production of certain products is a difficult problem; which batch can obtain the best investment benefit is an important issue. Economic order quantity (EOQ) balances the purchase cost and storage cost accounting, which achieves the best order quantity with the lowest total inventory cost. Mokhtari [16] presented an EOQ model to optimize the total system cost. In the context of uncertain demands, Braglia et al. [17] studied the stochastic periodic-review joint replenishment problem (JRP). Tayebi et al. [18] formulated the joint order (1, T) policy with Poisson demands while ensuring reduced supply chain costs. Güler et al. [19] considered the JRP when the holding cost rate and demand rate are private information and presented a mechanism to allocate costs in the JRP. That quantity coordination strategy can improve the supply chain performance of traditional decentralized system.

2.2 Price discount policy

In the 1990s, scholars began to use quantity-based price discount strategy to achieve supply chain coordination [20–22]. Weng [20] assumed that demand is elastic and affected by price, and found that quantity discount can effectively stimulate the increase of market demand and ensure Pareto Optimality. Under the condition of price elasticity of demand, Gao et al. [21] studied the problem of determining price discount in a supply chain contract composed of one buyer and one seller. Munson et al. [22] studied the overall profit maximization problem of the three-level chain (supplier-manufacturer-retailer) supply chain system. These studies assume that market demand is a constant or a decreasing function of product price. Some scholars further assume that market demand is a random variable, and study the cost optimization problem of supply chain system [23].

2.3 Secure multiparty computing

The research of SMC is mainly aimed at how to calculate a contract function safely without a trusted third party, which is the password basis for many applications [15] such as electronic voting, threshold signatures, and electronic auctions. The application of SMC is a possible mean to solve private information preserving problems [24], which has now become a subfield of cryptography [25].

Scholars began to study the application of SMC in the supply chain system. Atallah et al. [26] proposed the secure supply chain collaboration (SSCC) protocol for capacity allocation while preserving parties' private information. Clifton et al. [27] proposed a secure protocol for swapping loads while preserving trucking companies' private information, but they did not explicitly consider benefit sharing. Xie et al. [28] addressed SMC in the context of joint ordering under deterministic demand to minimize total supply chain expected costs. Pibernik et al. [29] described a privacy-preserving protocol for determining the EOQ with stochastic benefit sharing under deterministic demand with any private (cost and capacity) information preservation. Yang et al. [30] proposed a blockchain-based secure multi-party computation architecture for data sharing. Wang et al. [31] explored a novel approach to support energy storage sharing with privacy protection, based on privacy-preserving blockchain and secure multi-party computation.

2.4 Oblivious transfer protocol

The oblivious transfer (OT) protocol is a basic protocol in cryptography that enables the receiver of a service to obtain messages input by the sender of the service inadvertently, thus protecting the privacy of the receiver from the sender. Long et al. [32] proposed a privacy protection method based on server-assisted reverse oblivious transfer, which includes the protocol of a cloud server and can calculate the result of encrypting the sensing data to avoid fully trusting the sensing platform. Wang et al. [33] proposed a casual transmission protocol and a private set intersection protocol to protect the privacy of users. Based on smart contracts and OT, Li et al. [34] proposed a privacy-preserving big data exchange scheme that allows buyers and sellers to complete transactions independently and fairly without involving any third-party middleman.

3 Basic model based on EOQ

The classical *EOQ* model was created by Harris [35]. Based on the assumptions of the classical model, the basic model assumptions in this study are as follows:

a. The research object of this study is a two-level supply chain, and the current status of the supply chain is assumed to be balanced [36].
b. The seller makes the product, and the unit production cost is constant.
c. The supply capacity of the seller is much greater than the demand of the buyer, so the out-of-stock cost can be ignored.
d. When the market demand tends to be stable, the demand follows the normal distribution, and the buyer's demand expectation is *D.*
e. The buyer is a price taker in a free competitive market, and he can accept the shortage in the market. The buyer uses EOQ to determine the quantity of each purchase, and its ordering strategy uses (s, Q) strategy.
f. The seller's unit order preparation cost consists of two parts: the order processing cost and the production preparation cost.
g. The lead time of the buyer's order is constant.

The notations adopted in this paper are presented in Table 1.

According to the previous assumption, the buyer's order lead time is constant, so the demand in the lead time is only related to the demand quantity. Assuming that the buyer's order amount is Q each time and the product price given by the seller is w, the buyer's total annual cost is:

$$TC_b(Q, w) = wD + \frac{S_b D}{Q} + \left(\frac{Q}{2} + k\delta\right)h_b + B_b \delta G(k) D/Q \qquad (1)$$

where wD is the acquisition cost, $\frac{S_b D}{Q}$ is the ordering cost, $(\frac{Q}{2} + k\delta)h_b$ is the inventory holding cost, and $B_b \delta G(k) D/Q$ is the expected penalty and opportunity cost.

In Formula 1, $\delta G(k)$ represents the expected shortages [37]:

$$G(k) = \int_k^\infty (u-k)\frac{1}{\sqrt{2\pi}} exp\left(-\frac{u^2}{2}\right)du \qquad (2)$$

TABLE 1 Parameter definitions.

Parameter	Meaning of parameter
D	The buyer's expected demand
B_b	The buyer's unit shortage cost
S_b	The buyer's ordering setup cost
K	The buyer's safety factors in the (S,Q) policy
δ	Standard deviation of demand during lead time
w	The seller's wholesale price before applying the discount
α	The proportion obtained by the buyer when allocate the cost saved through coordination between the buyer and the seller
w_1	The seller's wholesale price after applying the discount
h_s	The seller's holding cost
S_s	The seller's ordering setup cost
C	The seller's production cost
Q^*	The optimal order quantity of buyer when seller and buyer make decentralized decision
Q^j	The buyer's order quantity when the joint cost is the lowest
D	The buyer's expected demand for the product
h_b	The buyer's unit holding cost for the product
TC	The joint cost of supply chain system when the seller and the buyer make joint decision

In Formula 2, $G(k)$ represents the distribution function of standard normal variables.

According to Formula 1, the optimal order quantity of the buyer can be obtained:

$$Q^* = \sqrt{2D[S_b + B_b\delta G(k)]/h_b} \tag{3}$$

So, the buyer's total annual cost is:

$$TC_b(Q^*, w) = wD + k\delta h_b + \sqrt{2D[S_b + B_b\delta G(k)]/h_b} \tag{4}$$

Then, the cost of the seller needed to be studied. We have assumed that the buyer's order amount is Q each time. When the buyer's annual demand is D, the buyer needs to order D/Q times in a year. So the buyer needs to place an order with the seller every Q*365/D days. For such order flow, the seller's production quantity should be a multiple of the buyer's order quantity within 1 year. In order to facilitate the account, we define the total cost of the seller as the sum of the cost of production, the fixed cost, and the inventory holding cost minus the sales return. Therefore, the total annual cost of the seller is expressed as:

$$TC_s(Q, w) = CD + \frac{h_s Q}{2} + \frac{S_s D}{Q} - wD \tag{5}$$

Let Q_s^* be the order quantity that the seller expects from the buyer to minimize the seller's cost. We can get $Q_s^* = \sqrt{2S_s D/h_s}$. By substituting Q_s^* into (5), we can infer the annual total cost of the seller.

$$TC_s(Q_s^*, w) = (C - w)D + \sqrt{2h_s S_s D} \tag{6}$$

When the buyer adopts EOQ ordering mode, the Seller's cost is as follows: [by substituting Q^* into (5)]

$$TC_s(Q^*, w) = (C - w)D + \left(\frac{S_s}{[S_b + B_b\delta G(k)]} + h_s/h_b\right)\sqrt{D[S_b + B_b\delta G(k)]h_b/2} \tag{7}$$

Comparing Formulas 6, 7, we can find that $TC_s(Q^*, w) \geq TC_s(Q_s^*, w)$, and the equation is established when $\frac{S_s}{[S_b+B_b\delta G(k)]} = h_s/h_b$.

4 The improved model based on price discount and quantity coordination

According to the basic model, in the case of decentralized decision-making, each member of the supply chain system makes decisions from the perspective of maximizing their own interests, and the strategies of the buyer and the seller are prone to conflict, resulting in high transaction costs for each member.

If the seller and buyer in the system can cooperate, share information with each other, and make joint decisions together, just like two departments in the same large company, their respective costs may be reduced in this case. Based on this idea, we first study the optimal joint order quantity of seller and buyer under probabilistic demand, and then design the quantity discount and profit distribution mechanism.

4.1 Optimal joint-ordering quantity under probabilistic demand

When the seller and the buyer share information and make joint decisions, the overall cost of the system should be the sum of the costs of the seller and the buyer. We use TC to express the joint cost.

The buyer's cost is shown in Formula 1, and the seller's cost is shown in Formula 5, $TC = TC_b(Q,w) + TC_s(Q,w)$, so we can obtain:

$$TC = \frac{[S_b + S_s + B_b\delta_L G(k)]D}{Q} + \frac{(h_b + h_s)Q}{2} + h_b k\delta + CD \tag{8}$$

When the joint cost TC takes the minimum value, that is, calculate the first derivative of TC, the order quantity Q^j of the buyer can be calculated:

$$Q^j = \sqrt{2D[S_b + S_s + B_b\delta G(k)]/(h_b + h_s)} \tag{9}$$

Now, the minimum annul joint cost $TC(Q^j)$ of the system is as follows:

$$TC(Q^j) = \sqrt{2D(h_b + h_s)[S_b + S_s + B_b\delta G(k)]} + h_b k\delta + CD \tag{10}$$

When the seller and the buyer make decentralized decision, the buyer's total annual cost is $TC_b(Q^*, w)$ (Formula 4), and the seller's total annual cost is $TC_s(Q^*, w)$ (Formula 7), then, the sum of the total annual costs of the seller and the buyer is $TC(Q^*) = TC_b(Q^*, w) + TC_s(Q^*, w)$.

Comparing the expressions $TC(Q^*)$ and $TC(Q^j)$, it is easy to get

$$TC(Q^*) \geq TC(Q^j) \tag{11}$$

That is to say, when the buyer orders with the order quantity T under the joint decision, the overall cost of the supply chain system is less than the sum of the respective costs under the decentralized decision of the buyer and the seller. However, for the buyer, when he chooses the order quantity Q^j of joint decision rather than the optimal order quantity Q^* of decentralized decision, his cost will increase, as $TC_b(Q^*, w) \leq TC_b(Q^j, w)$. Therefore, the buyer is unwilling to use the order quantity of joint strategy.

The reason for this situation is that under the joint strategy, the cost reduced by the seller is greater than the cost increased by the buyer, that is $TC_s(Q^j, w) - TC_s(Q^*, w) < TC_b(Q^*, w) - TC_b(Q^j, w)$.

In order to encourage the buyer to increase the order quantity of independent decision to the order quantity of joint decision, the seller needs to provide price discount to compensate the buyer for the increased cost. Suppose that the price provided by the seller to the buyer decreases from w to w_1, and at the same time he requires the buyer to increase the order quantity from Q^* to Q'. Then, only when the cost of the buyer is lower than the cost without price discount will he accept the price discount strategy. Therefore, there is the following constraint:

$$TC_b(Q', w_1) \leq TC_b(Q^*, w) \tag{12}$$

Formula 12 can be converted to:

$$wD - S_b\left(\frac{D}{Q'} - \frac{D}{Q^*}\right) - \left(\frac{Q'}{2} - \frac{Q^*}{2}\right)h_b - B_b\delta G(k)\left(\frac{D}{Q'} - \frac{D}{Q^*}\right) \geq w_1 D \tag{13}$$

Then, we can infer that under the price discount strategy, the seller's wholesale price w_1 has a maximum value.

$$w_1^{max} = w - \left\{\frac{h_b(Q' - Q^*)}{2} + [S_b + B_b\delta G(k)]\left(\frac{D}{Q'} - \frac{D}{Q^*}\right)\right\}\Big/D \tag{14}$$

Similarly, for the seller, he hopes that after implementing the price discount strategy, his cost cannot increase, that is, the following condition should be met:

$$TC_s(Q', w_1) \leq TC_s(Q^*, w) \tag{15}$$

Formula 15 can be rewritten as

$$\frac{h_s(Q' - Q^*)}{2} + S_s\left(\frac{D}{Q'} - \frac{D}{Q^*}\right) + wD \leq w_1 D \tag{16}$$

Then, we can infer that under the price discount strategy, the seller's wholesale price w_1 has a minimum value.

$$w_1^{min} = w - \left[S_s\left(\frac{D}{Q^*} - \frac{D}{Q'}\right) - \frac{h_s(Q' - Q^*)}{2}\right]\Big/D \tag{17}$$

Now, the optimal joint order quantity can benefit both the buyer and the seller without increasing the cost of either party.

We have the following proposition that describes the amount of cost saved:

Proposition 1. Under the joint strategy, the supply chain cost is $TC(Q^j)$, and under the decentralized decision, the supply chain cost is $TC(Q^*)$, which satisfies:

$$TC(Q^*) - TC(Q^j) = D \times (w_1^{max} - w_1^{min}) \tag{18}$$

4.2 Profit distribution and quantity discounts design

According to the previous proposition, when the seller and the buyer adopt a joint strategy, the overall cost saved by the supply chain system is $D(w_1{}^{max} - w_1{}^{min})$. In order to promote cooperation between the seller and the buyer, it is necessary to ensure that their respective costs under the joint strategy are lower than those of the previous independent decisions. Therefore, after the cooperation between the seller and the buyer, it is necessary to reasonably allocate the overall saved cost of the supply chain system to the seller and the buyer.

So we design such an implementation strategy, allocate the cost saved by the whole supply chain system, the proportion obtained by the buyer is α, and the proportion obtained by the seller is $1 - \alpha$. Here, $\alpha \in (0, 1)$ is a random number, named coordination factor. In fact, α means the allocation of the saved costs. If the buyer in the supply chain is stronger than the seller, for example, the buyer has the right to speak and decide, the buyer will save more costs, α will increase and be close to 1.

The cost savings allocated to the buyer is $\alpha D(w_1^{max} - w_1^{min})$, and the cost savings allocated to the seller is $(1 - \alpha)D(w_1^{max} - w_1^{min})$. The implementation strategy can be expressed by the following proposition:

Proposition 2. To encourage the buyer to increase the independent decision-making order quantity to equal the joint order quantity Q^j, the seller changes the sales price from w to w_1, and the quantity discount provided by the seller can be expressed as

$$w_1 = w_1^{max} - \alpha(w_1^{max} - w_1^{min}) \tag{19}$$

Where w_1^{max} and w_1^{min} are given in Eqs 14, 17.

5 Joint ordering strategy and quantity discount design with privacy protection

We assume that the seller and the buyer share information and make decisions together like two departments of the same company. Under a certain quantity discount mechanism and profit distribution strategy, both the seller and the buyer can reduce costs. However, after all, the seller and the buyer are independent companies. On one hand, they may not want the other party to know their private information; On the other hand, even though both parties are willing to share information, they are afraid to share information because they are worried about the harm caused by information leakage. Therefore, a mechanism is needed to realize secure information sharing. In this section, we apply SMC protocols to joint ordering policy and quantity discount design with privacy protection under probabilistic demand.

5.1 Privacy preserving joint-ordering policy protocols

To calculate the minimum joint cost under the joint ordering strategy, the buyer and the seller need to provide the total annual cost when making independent decisions. Therefore, the information that both parties need to provide and obtain is as follows:

5.1.1 Inputs

The buyer supplies $TC_b(Q,w) = wD\ +\frac{S_b D}{Q} + (\frac{Q}{2} + k\delta)h_b +$ $B_b\delta G(k)D/Q$, where S_b, h_b, k, B_b, δ, and $G(k)$ are the buyer's private (cost and capacity) information.

The seller supplies $TC_s(Q,w) = CD + \frac{h_s Q}{2} + \frac{S_s D}{Q} - wD$,where h_s, S_s, and C are the seller's private (cost and capacity) information.

5.1.2 Outputs

The partners learn $Q^j = \sqrt{2D[S_b + S_s + B_b\delta G(k)]/(h_b + h_s)}$ with any private (cost and capacity) information preservation.

5.1.3 Assumptions

The formula $Q^j = \sqrt{2D[S_b + S_s + B_b\delta G(k)]/(h_b + h_s)}$ is public information.

Therefore, the buyer and seller's goals are to compute the formula for Q^j while preserving their private information.

In computer science, formulas are often represented by circuits. So, we construct a circuit for the computation of Q^j, which is displayed in Figure 1.

In Figure 1, in the circuit, on the top, the red values denote the private part of the seller's input, and on the left, the red values denote those of the buyer.

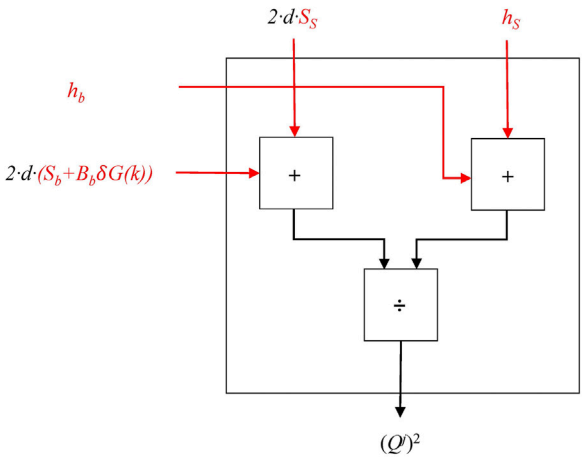


FIGURE 1
Circuit for computation of privacy preserving joint-ordering.

5.1.4 Protocol steps

The buyer holds two values $(h_b, S_b + B_b\delta G(k))$, and the seller holds two values (h_s, S_s).

The common goal is to compute $[S_b + S_s + B_b\delta G(k)]/(h_b + h_s)$ because $2D$ is public information.

Step 1 The buyer generates U_1 (random number), and the seller generates U_2 (random number).

Step 2 The buyer and the seller use Secure two-party add-product protocol (Supplementary Appendix S3), the buyer obtains d_1, and the seller obtains d_2, where $d_1 + d_2 = (U_1 + U_2)(S_b + B_b\delta G(k) + S_s) \times 2D$.

Step 3 The buyer and the seller use secure two-party add-product protocol, which are as follows:

Inputs: the buyer has two reals (x_1, y_1), and the seller has two reals (x_2, y_2).

Outputs: the buyer obtains r_1, and the seller obtains r_2, where $r_1 + r_2 = (x_1 + x_2)(y_1 + y_2)$.

The detailed demonstration process is in Supplementary Appendix S3.

The buyer receives n_1, and the seller receives n_2, where $n_1 + n_2 = (U_1 + U_2)(h_b + h_s)$.

Step 4 The buyer sends n_1 to the seller, the seller computes $n = n_1 + n_2$, and the seller sends n to the buyer.

Step 5 The buyer and the seller independently compute $s_1 = \frac{d_1}{n}$, $s_2 = \frac{d_2}{n}$, and s_1, s_2 obey the equation:

$$s_1 + s_2 = \frac{d_1 + d_2}{n_1 + n_2} = (S_b + B_b\delta G(k) + S_s) \times 2D/(h_b + h_s) = (Q^j)^2$$

5.1.5 Analysis of the protocol

- Information disclosure:

The security in the second (third) step is guaranteed by secure two-party add-product protocol (Supplementary Appendix S3). The independent computation in the 5th step is private. Next, the security of computation in the 4th step is discussed.

The buyer and the seller know the following equation:

$$d_1 + d_2 = (U_1 + U_2)(S_b + B_b\delta G(k) + S_s) \times 2D \tag{20}$$

$$n = n_1 + n_2 = (U_1 + U_2)(h_b + h_s) \tag{21}$$

For the buyer (the seller), there are 5 unknown reals: d_2, U_2, n_2, h_s, S_s $(d_1, U_1, n_1, h_b, S_b + B_b\delta G(k))$. Neither party can know the secret input of another participant.

- Computational complexity:

The protocol used two times Secure two-party add-product protocol (Supplementary Appendix S3).

- Communication complexity:

The protocol only requires communication between the buyer and the seller; both sides know the value of Q^j.

5.2 Privacy preserving quantity discounts design

To implement the quantity discount, that is, to let the seller adjust the sales price from w to w_1, the buyer and the seller need to provide the total annual cost when making their own decisions, and then inform both parties of the quantity discount information w_1, which are as follows:

5.2.1 Inputs

The buyer supplies $TC_b(Q,w) = wD + \frac{S_b D}{Q} + (\frac{Q}{2} + k\delta)h_b + B_b \delta G(k)D/Q$,where S_b, h_b, k, B_b, δ, and $G(k)$ are the buyer's private information.

The seller supplies $TC_s(Q,w) = CD + \frac{h_s Q}{2} + \frac{S_s D}{Q} - wD$,where h_s, S_s, and C are the private information of the seller.

Coordination factor α, where $\alpha \in (0,1)$ is a random number, which is determined by the bargaining power of both sides, the seller generates a random number α.

5.2.2 Outputs

The seller and buyer learn the quantity discount w_1 while preserving their private information.

5.2.3 Assumptions

The seller and the buyer's goal is to compute w_1 with stochastic quantity discounts because Q^j is public information.

We reformulate the stochastic quantity discounts to $w_1 = w_1^{max} - \alpha(w_1^{max} - w_1^{min})$,where $\alpha \in (0,1)$ is a random number.

$$\because w_1^{max} = w - \left\{\frac{h_b(Q' - Q^*)}{2} + [S_b + B_b\delta G(k)]\left(\frac{D}{Q'} - \frac{D}{Q^*}\right)\right\} \div D$$

$$w_1^{min} = w - \left[S_s\left(\frac{D}{Q^*} - \frac{D}{Q'}\right) - \frac{h_s(Q' - Q^*)}{2}\right] \div D$$

$$\therefore w_1 = w_1^{max} - \alpha(w_1^{max} - w_1^{min})$$

$$= w_1^{max} - \alpha(w_1^{max} - w) - \alpha \times \frac{S_s}{Q^*} - \alpha \times \frac{h_s Q^*}{2D}$$

$$+ \alpha \times \left(\frac{S_s}{Q'} + h_s Q' \middle/ 2D\right)$$

Therefore, w_1 is only a function that requires inputs $(w_1^{max}, w_1^{max} - w, \frac{1}{Q^*}, Q^*)$ from the buyer and $(-\alpha, -\alpha \times S_s, -\alpha \times \frac{h_s}{2D}, \alpha \times (\frac{S_s}{Q'} + h_s Q'/2D))$ from the seller.

Because w and Q^j (public information) are known to the buyer, then the buyer can dependably compute $(w_1^{max}, w_1^{max} - w, \frac{1}{Q^*}, Q^*)$.

Because α is a random number, that is, in contrast, determined by the bargaining power of both sides, the seller generates a random number α.

The seller can compute $(-\alpha, -\alpha \times S_s, -\alpha \times \frac{h_s}{2D}, \alpha \times (\frac{S_s}{Q'} + \frac{h_s Q'}{2D}))$.

Where,

$w_1^{max} - \alpha(w_1^{max} - w) - \alpha \times \frac{S_s}{Q^*} - \alpha \times \frac{h_s Q^*}{2D} + \alpha \times (\frac{S_s}{Q'} + \frac{h_s Q'}{2D})$

$$= \left(w_1^{max}, w_1^{max} - w, \frac{1}{Q^*}, Q^*, 1\right) \times \left(1, -\alpha, -\alpha \times S_s, -\alpha \times \frac{h_s}{2D}, \alpha \right.$$

$$\left. \times \left(\frac{S_s}{Q'} + h_s Q' \middle/ 2D\right)\right)$$

The flowchart of privacy preserving quantity discounts design is shown in Figure 2.

In Figure 2, the buyer independently computes vector $\boldsymbol{X}$, and the seller independently computes vector $\boldsymbol{Y}$ based on Q^j and α. The red values denote the private information of the buyer and the seller. The buyer and the seller determine the allocation of overall reduced costs, which is determined by the bargaining power of both parties to the contract, and the seller generates a random number α. Based on the foundation of OT, the calculation framework of privacy preserving quantity discounts design is given in Figure 2.

5.2.4 Protocol steps

Step 1 The buyer and the seller use privacy-preserving optimal joint-ordering quantity protocols, and the seller obtain Q^j.

Step 2 The buyer and the seller determine the allocation of overall reduced costs, which is determined by the bargaining power of both parties to the contract, and the seller generates a random number α.

Step 3 The buyer independently computes vector $\boldsymbol{X} = (w_1^{max}, w_1^{max} - w, \frac{1}{Q^*}, Q^*, 1)$, and the seller independently computes vector $\boldsymbol{Y} = (1, -\alpha, -\alpha \times S_s, -\alpha \times \frac{h_s}{2D}, \alpha \times (\frac{S_s}{Q'} + \frac{h_s Q'}{2D}))$.

Step 4 The buyer and the seller using secure two-party real product protocol (Supplementary Appendix S2), the buyer obtains $u = \boldsymbol{X} \times \boldsymbol{Y}^T + v$, and the seller obtains v, where the letter T stands for 'transpose'.

5.2.5 Analysis of protocol

- Information disclosure:

Protocol 5.1 guarantees security in the first step. The independent computation in the 2nd and 3rd steps is secure. Secure two-party real product protocol (Supplementary Appendix S2) guarantees security in the fourth step.

- Computational complexity:

The protocol uses secure two-party add-product protocol (Supplementary Appendix S3) twice and Secure two-party real product protocol (Supplementary Appendix S2) once.

- Communication complexity:

The protocol requires communication between the buyer and the seller only; both sides of the calculation know the value of w_1.

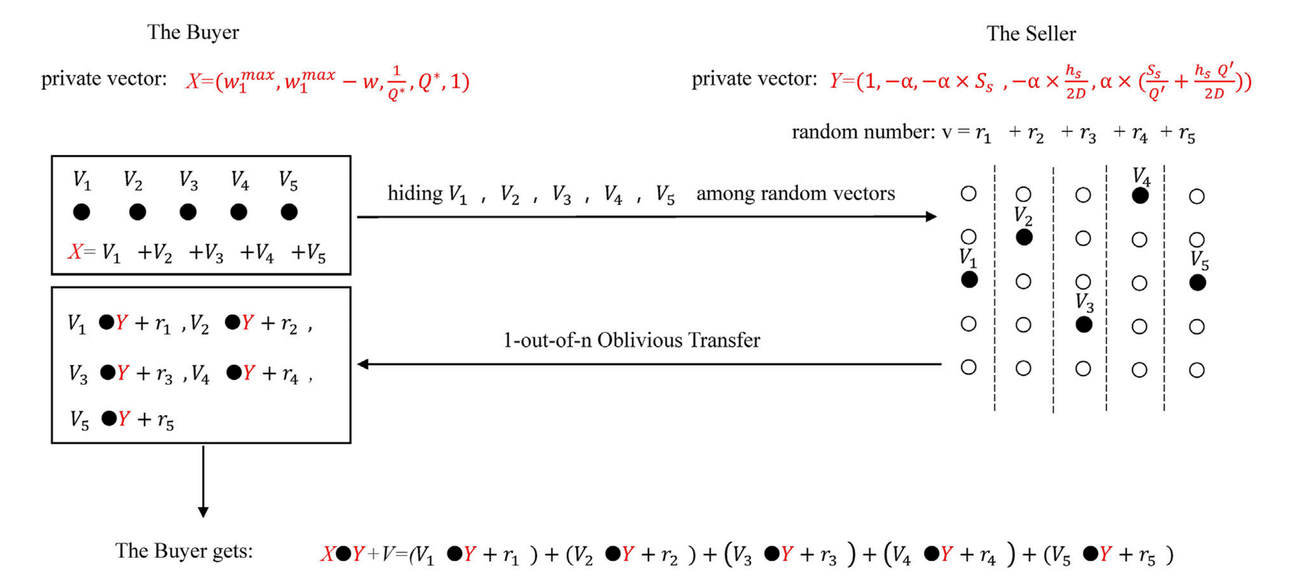

The Buyer
private vector: X=$(w_1^{max}, w_1^{max} - w, \frac{1}{Q^*}, Q^*, 1)$
The Seller
private vector: Y=$(1, -\alpha, -\alpha \times S_s, -\alpha \times \frac{h_s}{2D}, \alpha \times (\frac{S_s}{Q'} + \frac{h_s Q'}{2D}))$
random number: $v = r_1 + r_2 + r_3 + r_4 + r_5$
V_1 V_2 V_3 V_4 V_5
X= V_1 +V_2 +V_3 +V_4 +V_5
hiding V_1, V_2, V_3, V_4, V_5 among random vectors
$V_1 ●Y + r_1$, $V_2 ●Y + r_2$, $V_3 ●Y + r_3$, $V_4 ●Y + r_4$, $V_5 ●Y + r_5$
1-out-of-n Oblivious Transfer
The Buyer gets:
$X●Y+V$=$(V_1 ●Y + r_1) + (V_2 ●Y + r_2) + (V_3 ●Y + r_3) + (V_4 ●Y + r_4) + (V_5 ●Y + r_5)$

FIGURE 2
Flowchart of privacy preserving quantity discounts design.

6 Numerical simulation

In this section, we carry out a numerical simulation on joint ordering, joint ordering for privacy protection, and quantity discount for privacy protection.

6.1 Numerical simulation of joint ordering strategy

According to the previous conclusions, the annual costs of the buyer and the seller under independent decision-making and joint decision-making should be numerically simulated and then compared. The details are as follows.

6.1.1 Steps

First, assume the target stock-out probability $p(k)$ determined by the buyer is 0.1 (0.1 is randomly selected, and other values are also acceptable).

Second, as $G(k)$ represents the distribution function of standard normal variables, query standard normal distribution function table, and $k = 1.28$, and $G(k) = 0.048$ can be obtained. Other parameters of the buyer and the seller, such as D, S_b, h_b, B_b, etc., are listed in Table 2. Assume that the cost allocation mechanism negotiated by the buyer and the seller is $\alpha = 0.4$.

Third, no price discount is considered in Table 3. The results from the model based on quantity coordination are shown in Table 4. A negative cost for the seller means that it is his profit.

TABLE 2 Numerical value of simulation parameters.

Parameter	D	S_b	h_b	B_b	K	δ	w	S_s	h_s	C
The buyer	3,000	200	30	60	1.28	80	50			
The seller								900	10	16

6.1.2 Results

Comparing the seller's pricing and the buyer's order under the above two conditions reveals that when adopting the quantity coordination strategy, the buyer's order quantity increases significantly, and the costs of both parties are reduced. These two benefits greatly reduce the costs of both buyers and sellers, and the supply chain achieves efficient operations and a win-win outcome.

Then, adjust the value of α to observe the cost change of the supply chain system, and the results are listed in Table 5. From Table 5, it can be inferred that the strength of negotiation between buyers and sellers determines the flow of overall profit savings in the supply chain, but the total cost savings are fixed.

Furthermore, from Table 5, we extract the cost saving ratio of all parties in the supply chain system and plot it in the coordinate system, as shown in Figure 3.

In Figure 3, the abscissa α represents the proportion obtained by the buyer when allocate the cost saved through coordination between the buyer and the seller; and the vertical coordinate represents the proportion of cost savings (%). The blue bar chart represents the buyer, the red bar chart represents the seller, and the black bar chart represents the overall supply chain. For the seller, a negative cost means that it is his profit.

When the buyer makes an independent decision, the optimal ordering strategy of individual cost is adopted. The ordering quantity is 243, the purchase price is 50, the cost is 160,364, the ordering cost of the seller is −89,673, and the total supply chain cost is 70,690. When the buyer and the seller use privacy-preserving joint-ordering policy protocols, the order quantity is 423, and the buyer gets 40% of the cost saved by the supply chain. The order price is 49.26, the buyer's cost is 159,292.5, the seller's cost is −91,282.02,

TABLE 3 Basic model without price discount.

	Q	w	w^{min}	w^{max}	Cost	Cost savings	Proportion of savings %
The buyer	243.1	50	50	50	160364.5	----	----
The seller					−89677.29	----	----
Joint cost					70687.22	----	

TABLE 4 The improved model of quantity coordination.

	Q	w	w^{min}	w^{max}	Cost	Cost savings	Proportion of savings %
The buyer	423	49.26	48.72	49.62	159293.8	1,070.74	0.6
The seller					−91283.41	1,606.12	−1.76
Joint cost					68010.34	2,676.87	3.79

TABLE 5 Cost change under different α ($w^{min} = 48.72$, $w^{max} = 49.62$).

α	w	The buyer cost	Cost savings	Proportion of savings %	The seller cost	Cost savings	Proportion of savings %	Joint cost savings
0.01	49.61	160337.7	26.77	0.02	−92327.39	2,650.10	−2.96	2,676.87
0.3	49.35	159561.4	803.06	0.50	−91551	1873.81	−2.09	2,676.87
0.4	49.26	159292.5	1,072	0.6	−91282.02	1,608.11	−1.76	2,676.87
0.5	49.17	159026.1	1,338.43	0.83	−91015.73	1,338.43	−1.49	2,676.87
0.6	49.08	158758.4	1,606.12	1.00	−90748.04	1,070.75	−1.19	2,676.87
0.7	48.99	158490.7	1873.81	1.17	−90480.35	803.06	−0.89	2,676.87
0.8	48.90	158223	2,141.50	1.34	−90212.66	535.37	−0.59	2,676.87
0.99	48.73	157714.4	2,650.10	1.65	−89704.06	26.77	−0.03	2,676.87

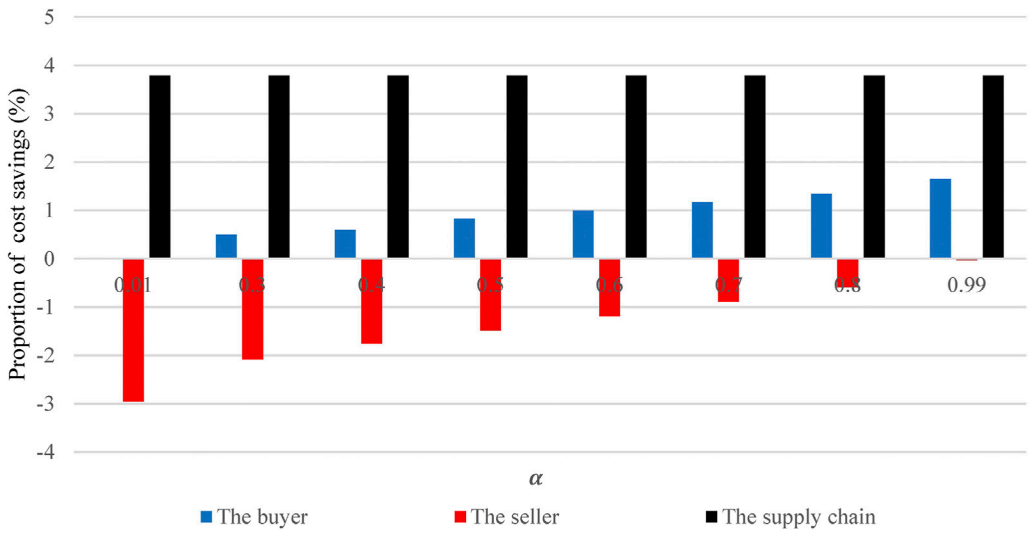


FIGURE 3
Cost saving ratio of all parties in the supply chain system.

and the total supply chain cost is 68,010.48. Through collaborative ordering, the buyer's cost, the seller's cost, and the total supply chain's cost all decrease.

Figure 3 shows the comparative analysis of the cost savings ratio between buyers and sellers under different profit distribution ratios. As can be seen from Figure 3, the total cost savings are fixed; the larger the α is, the more cost savings in the supply chain will flow to buyers; the smaller the α is, the more the cost savings in the supply chain flow to selling.

6.2 Numerical simulation of joint ordering for privacy protection

Taking the data in Table 2 as an example, the following calculates the privacy-preserving joint-ordering policy protocols.

6.2.1 Steps

Step 1 The buyer uses $B_b, \delta, G(k)$ and private information S_b to independently compute private information $S_b^* = 2D \times (S_b + B_b\delta G(k))$, and $H_b^* = h_b$. The seller uses D and private information S_s, h_s to independently compute private information $S_s^* = 2D \times S_s$, $H_s^* = h_s$. The buyer holds two values (1772832, 30), and the seller holds two values (5400000, 10).

Step 2 The buyer generates U_1 (random number 1024), and the seller generates U_2 (random number 2560). The buyer holds two values $(S_b^* = 2D \times (S_b + B_b\delta G(k)), H_b^* = h_b)$, and the seller holds two values $(S_s^* = 2D \times S_s, H_s^* = h_s)$. The seller uses Secure two-party add-product protocol (Supplementary Appendix S3), and the calculation steps and principles are reported in Supplementary Appendix S3. The buyer obtains d_1 (10675782451.2), and the seller obtains d_2 (15031647436.8), where

$$d_1 + d_2 = (U_1 + U_2)(S_b + B_b\delta G(k) + S_s) \times 2D = 2507429888.$$

Step 3 The buyer and the seller use secure two-party add-product protocol (Supplementary Appendix S3), and the calculation steps and principles and principles are reported in Supplementary Appendix S3. The buyer obtains n_1 (86016), and the seller obtains n_2 (57344), where

$$n_1 + n_2 = 86016 + 57344 = (U_1 + U_2)(h_b + h_s) = 143360.$$

Step 4 The buyer sends n_1 to the seller, the seller computes $n = n_1 + n_2 = 143360$, and the seller sends n to the buyer.

Step 5 The buyer and the seller independently compute $s_1 = \frac{d_1}{n} = 74468.35$, and $s_2 = \frac{d_2}{n} = 104852.5$, and s_1, s_2 obey the equation:

$$\begin{aligned} s_1 + s_2 &= \frac{d_1 + d_2}{n_1 + n_2} = 179320.9 \\ &= \frac{(U_1 + U_2)(S_b + B_b\delta G(k) + S_s) \times 2D}{(U_1 + U_2)(h_b + h_s)} = 179320.9 \end{aligned}$$

so, we can obtain $Q^j = \sqrt{179320.9}$.

6.2.2 Analysis

- Information disclosure:

The security in the second and third step is guaranteed by secure two-party add-product protocol (Supplementary Appendix S3). The independent computation in the 5th step is private. Then, the security of computation in the 4th step is discussed.

The buyer and the seller know the Equations 20, 21, for the buyer (the seller), there are 5 unknown reals:d_2, U_2, n_2, h_s, S_s (d_1, U_1, n_1, h_b, $S_b + B_b\delta G(k)$. Neither party can know the secret input of another participant.

- Computational complexity:

The protocol used two times secure two-party add-product protocol (Supplementary Appendix S3).

- Communication complexity:

The protocol only requires communication between the buyer and the seller; both sides know the value of Q^j.

6.3 Numerical simulation of quantity discount for privacy protection

The following calculates the privacy-preserving quantity discount.

6.3.1 Steps

Step 1 The buyer and the seller use privacy-preserving joint-ordering policy protocols, and the buyer and the seller obtain $Q^j = \sqrt{179320.9} = 423$.

Step 2 The buyer and the seller determine the allocation of overall reduced costs, and the seller generates a random number $\alpha = 0.4$.

Step 3 The buyer independently computes vector:

$$\boldsymbol{X} = \left(w_1^{max}, w_1^{max} - w, \frac{1}{Q^*}, Q^*, 1 \right) = (49.62, -0.38, 0.0041, 243, 1)$$

where

$w_1^{max} = w - \left\{\frac{h_b(Q' - Q^*)}{2} + [S_b + B_b\delta G(k)]\left(\frac{D}{Q'} - \frac{D}{Q^*}\right)\right\} \div D = 49.62$

The seller independently computes vector:

$$\begin{aligned} \boldsymbol{Y} &= \left(1|, -\alpha, -\alpha \times S_s, -\alpha \times \frac{h_s}{2D}, \alpha \times \left(\frac{S_s}{Q'} + \frac{h_s Q'}{2D} \right) \right. \\ &= (1, -0.4, -360, -0.00067, 1.333)) \end{aligned}$$

where $w_1^{min} = w - [S_s\left(\frac{D}{Q^*} - \frac{D}{Q'}\right) - \frac{h_s(Q' - Q^*)}{2}] \div D = 48.72$

Step 4 The buyer and the seller use secure two-party real product protocol (Supplementary Appendix 2), and the calculation steps and principles are presented in Supplementary Appendix S4, The buyer obtains $u = \boldsymbol{X} \times \boldsymbol{Y}^T + v = 49.26 + 10.34 = 59.60$, and the seller obtains $v(10.34)$, where the letter T stands for 'transpose'.

6.3.2 Analysis

- Information disclosure:

Secure two-party real product protocol guarantees security in the first step. The independent computation in the 2nd and 3rd steps is secure. Secure two-party real product protocol (Supplementary Appendix S2) guarantees security in the fourth step.

- Computational complexity:

The protocol uses secure two-party add-product protocol (Supplementary Appendix S3) twice and Secure two-party real product protocol (Supplementary Appendix S2) once.

- Communication complexity:

The protocol requires communication between the retailer and the seller only; both sides of the calculation know the value of w_1.

6.4 Global analysis

In brief, there is no information sharing, that is, when the buyer makes independent decisions, he can not get a discount subsidy, and the order cost is very high at this time; under the perfect information sharing, the use of collaborative ordering can reduce various costs; however, due to fear of private information leakage, perfect information sharing cannot be carried out in reality. Collaborative ordering under perfect information sharing through SMC was realized, and all costs of the supply chain system were reduced, which was further verified by the numerical simulation.

7 Conclusion and discussion

In this paper, we assume that the market demand is random, and the seller and the buyer share information and make decisions together like two departments of the same company. We analyze the optimal joint order quantity under probabilistic demand, and design the quantity discount model and profit distribution mechanism. Under a certain quantity discount mechanism and profit distribution strategy, both the seller and the buyer can reduce costs. The quantity discount model and profit distribution mechanism designed require supply chain members to share information. In order to protect the privacy of members and improve the willingness of supply chain members to share information, we designed a privacy protection joint ordering policy protocol and privacy protection quantity discount policy based on SMC technology. Then, the joint ordering strategy, the privacy-preserving joint ordering strategy and the quantity discount protocol are numerically simulated. The numerical simulation results show that the privacy-preserving quantity discount coordination mechanism designed by us can reduce the cost of supply chain members to varying degrees and effectively protect the shared information of supply chain members.

Our research is based on the classic one buyer and one seller supply chain system, and the proposed joint ordering strategy and quantity discount design with privacy protection have a certain practical significance, which is helpful to the research of cost optimization of the system in complex supply chain systems. But there are several limitations. First, the shared information discussed in this paper is all quantitative information. There are still a lot of qualitative information to be shared in supply chain collaborative optimization. Whether supply chain collaborative optimization can make cooperative decisions under the protection of qualitative information deserves further study. Second, enterprises participating in collaborative optimization of supply chain under the protection of private information share their own information, but different private information shared by enterprises will bring different benefits to collaborative optimization. The rational distribution mechanism should be to distribute the value of collaborative optimization reasonably according to private information. Therefore, how to distribute the additional benefits of collaborative optimization reasonably according to the utility of information is the direction that needs further research. Third, there are more buyers or multitier supply chain structures in reality, under these complex circumstances, the joint ordering strategy with privacy protection and quantity discount scheme need to be designed and solved urgently.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

C-HX provided this topic; S-YH wrote and guided the manuscript. S-YH, J-CZ, C-HX, and L-TZ discussed and modified the manuscript. All authors contributed to the manuscript and approved the submitted version.

Funding

This research is supported by the following projects: The Zhejiang Philosophy and Social Sciences Foundation (Grand No. 19NDJC145YB); The Zhejiang Province Soft Science Research Program in China under Grant 2022C25034; The Wenzhou Social Science Association Project (key project): 22WZX01; The Zhejiang First-class Undergraduate Curriculum Project in 2022 (970); The Zhejiang Curriculum Ideological and Political Demonstration Grass-roots Teaching Organization Project in 2022(55); The Doctoral fund project of Changzhou Vocational Institute of Technology under Grant BS202213101007.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2023.1017251/full#supplementary-material

References

1. Jin MZ. *Game theoretic analysis for supply chain coordination: Modeling electronic intermediation, coalition formation, and capacity reservation*. [dissertation]. Bethlehem: Lehigh University (2001).

2. Monahan JP. A quantity discount pricing model to increase vendor profits. *Manag Sci* (1984) 30(6):720–6. doi:10.1287/mnsc.30.6.720

3. Li JL, Liu LW. Supply chain coordination with quantity discount policy. *Int J Prod Econ* (2006) 101(1):89–98. doi:10.1016/j.ijpe.2005.05.008

4. Mesmer-Magnus JR, DeChurch LA. Information sharing and team performance: A meta-analysis. *J Appl Psychol* (2009) 94(2):535–46. doi:10.1037/a0013773

5. Zhang LQ, Liu HF, Cai Z. Addressing the consensus on information sharing in CPFR information systems: Insights from manufacturer–retailer dyads. *Int J Prod Res* (2022) 60(11):3569–88. doi:10.1080/00207543.2021.1926569

6. Fu K, Wang C, Xu JY. The impact of trade credit on information sharing in a supply chain. Omega:. *Int J Manag Sci* (2022) 110:102633. doi:10.1016/j.omega.2022.102633

7. Wijewickrama MKCS, Chileshe N, Rameezdeen R, Ochoa JJ. Information sharing in reverse logistics supply chain of demolition waste: A systematic literature review. *J Clean Prod* (2021) 280:124359. doi:10.1016/j.jclepro.2020.124359

8. Lee HL, Whang S. Information sharing in a supply chain. *Int J Tech Manag* (2000) 1:79–93. doi:10.1504/ijmtm.2000.001329

9. Wu JJ, Xu H. Information leakage and financing decisions in a supply chain with corporate social responsibility and supply uncertainty. *Sustainability* (2021) 13(21): 11917. doi:10.3390/su132111917

10. Rockwern B, Johnson D, Sulmasy LS. Health information privacy, protection, and use in the expanding digital health ecosystem: A position paper of the American college of physicians. *Ann Intern Med* (2021) 174(7):994–8. doi:10.7326/m20-7639

11. Deb R, Roy S. A software defined network information security risk assessment based on pythagorean fuzzy sets. *Expert Syst Appl* (2021) 183:115383. doi:10.1016/j.eswa.2021.115383

12. He B, Li RD, Wang SS, Liu JG. A social mobilized inspection system against external damage of power grid based on block chain technology. *J Phys Conf Ser* (2020) 1453:012104. doi:10.1088/1742-6596/1453/1/012104

13. Ugochukwu NA, Goyal SB, Arumugam S. Blockchain-based IoT-enabled system for secure and efficient logistics management in the era of IR 4.0. *J Nanomater* (2022) 2022:1–10. doi:10.1155/2022/7295395

14. Yao AC. Protocols for secure computations. In: Proceeding of the 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982); 1982 November 3-5; Chicago, IL, USA. IEEE (1982). p. 160–4.

15. Micali S, Goldreich O, Wigderson A. How to play any mental game. In: Proceedings of the Nineteenth ACM Symp. on Theory of Computing, STOC; May 1987; New York, NY, USA. New York: ACM (1987). p. 218–29.

16. Mokhtari H. Economic order quantity for joint complementary and substitutable items. *Mathematics Comput Simulation* (2018) 154:34–47. doi:10.1016/j.matcom.2018.06.004

17. Braglia M, Castellano D, Song D. Distribution-free approach for stochastic Joint-Replenishment Problem with backorders-lost sales mixtures, and controllable major ordering cost and lead times. *Comput Operations Res* (2017) 79:161–73. doi:10.1016/j.cor.2016.11.002

18. Tayebi H, Haji R, Jeddi BG. Joint order (1, T) policy for a two-echelon, single-item, multi-retailer inventory system with Poisson demand. *Comput Ind Eng* (2018) 119: 353–9. doi:10.1016/j.cie.2018.04.009

19. Güler K, Körpeoğlu E, Şen A. Design and analysis of mechanisms for decentralized joint replenishment. *Eur J Oper Res* (2017) 3:992–1002. doi:10.1016/j.ejor.2016.11.029

20. Weng ZK. Modeling quantity discounts under general price-sensitive demand functions: Optimal policies and relationships. *Eur J Oper Res* (1995) 86(2):300–14. doi:10.1016/0377-2217(94)00104-k

21. Gao JJ, Wang YJ, Guo YJ, Zhao XD. Pareto optimization problems in supply chain contract under elastic demand. *J Syst Manag* (2002) 11(1):36–40. doi:10.3969/j.issn.1005-2542.2002.01.008

22. Munson CL, Rosenblatt MJ. Coordinating a three-level supply chain with quantity discounts. *IIE Trans* (2001) 33:371–84. doi:10.1080/07408170108936836

23. Li JL, Liu LW. Two mechanisms of supply chain coordination based on price discount under probabilistic demand. *Chin J Manag Sci* (2005) 13(3):37–43. doi:10.1007/s11769-005-0030-x

24. Bayatbabolghani F, Blanton M. Secure multi-party computation. In: Proceedings of the CCS '18: 2018 ACM SIGSAC Conference on Computer and Communications Security; 2018 October 15–19; Toronto, Canada. New York: Association for Computing Machinery (2018). p. 2157–9.

25. Balasubramanian K, Rajakani M. Secure multiparty computation. In: *Algorithmic strategies for solving complex problems in cryptography*. Hershey, Pennsylvania: IGI global (2017). p. 154–66.

26. Atallah MJ, Elmongui HG, Deshpande V, Schwarz LB. Secure supply-chain protocols. In: Proceedings of the IEEE International Conference on E-Commerce, 2003; 2003 Jun 24-27; Newport Beach, CA, USA. IEEE (2003). p. 293–302.

27. Clifton C, Iyer A, Cho R, Jiang W, Kantarc LM, Vaidya J. An approach to securely identifying beneficial collaboration in decentralized logistics systems. *Manufacturing Serv Operations Manag* (2008) 10:108–25. doi:10.1287/msom.1070.0167

28. Xie CH, Zhong WJ, Zhang YL. A study of privacy preserving joint-ordering policy. In: Proceedings of the 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing; 2008 Oct 12-14; Dalian, China. IEEE (2008). p. 1–4.

29. Pibernik R, Zhang Y, Kerschbaum F, Schröpfer A. Secure collaborative supply chain planning and inverse optimization–The JELS model. *Eur J Oper Res* (2011) 208: 75–85. doi:10.1016/j.ejor.2010.08.018

30. Yang YH, Wei LJ, Wu J, Long CN. Block-smpc: A blockchain-based secure multi-party computation for privacy-protected data sharing. In: Proceedings of the 2020 The 2nd International Conference on Blockchain Technology; March 12–14, 2020; Shanghai Jiao Tong University, Huazhong University of Science and Technology, China. New York. New York: Association for Computing Machinery (2020). p. 46–51.

31. Wang N, Chau SC, Zhou Y. Privacy-preserving energy storage sharing with blockchain and secure multi-party computation. *ACM SIGENERGY Energ Inform Rev* (2021) 1:32–50. doi:10.1145/3508467.3508471

32. Long H, Zhang S, Zhang Y, Zhang L, Wang L. A privacy-preserving method based on server-aided reverse oblivious transfer protocol in MCS. *IEEE Access* (2019) 7: 164667–81. doi:10.1109/access.2019.2953221

33. Wang X, Kuang X, Li J, Li J, Chen X, Liu Z. Oblivious transfer for privacy-preserving in VANET's feature matching. *IEEE Trans Intell transportation Syst* (2020) 22(7):4359–66. doi:10.1109/tits.2020.2973738

34. Li T, Ren W, Xiang Y, Zheng X, Zhu T, Choo KR, et al. Faps: A fair, autonomous and privacy-preserving scheme for big data exchange based on oblivious transfer, ether cheque and smart contracts. *Inf Sci* (2021) 544:469–84. doi:10.1016/j.ins.2020.08.116

35. Harris FW. How many parts to make at once. *Operations Res* (1990) 38:947–50. doi:10.1287/opre.38.6.947

36. Venegas BB, Ventura JA. A two-stage supply chain coordination mechanism considering price sensitive demand and quantity discounts. *Eur J Oper Res* (2018) 2: 524–33. doi:10.1016/j.ejor.2017.06.030

37. Silver EA, Pyke DF, Peterson R. Inventory management and production planning and scheduling. *Wiley New York* (1998) 3:722–3. doi:10.1016/S0278-6125(99)90116-4

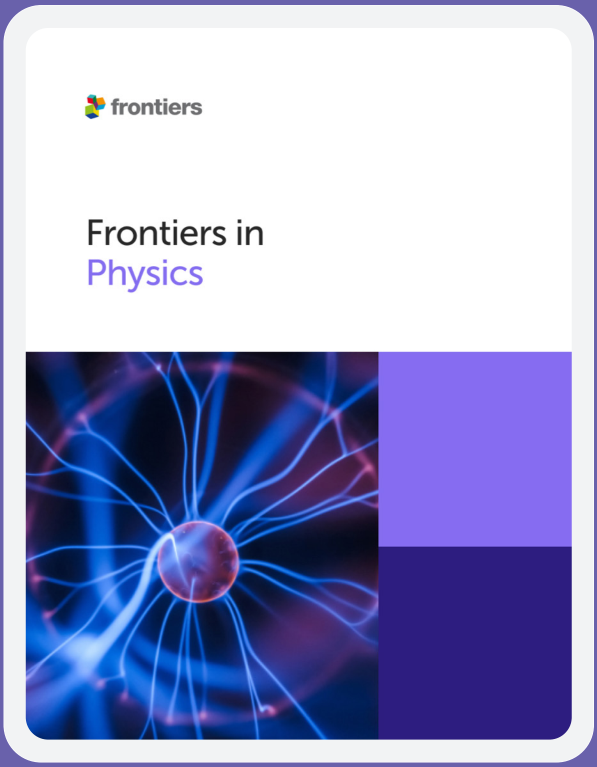
Frontiers in Physics

Investigates complex questions in physics to understand the nature of the physical world

Addresses the biggest questions in physics, from macro to micro, and from theoretical to experimental and applied physics.

Discover the latest Research Topics

See more →


frontiers

Frontiers in Physics

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact