Journal of Mathematical Research with Applications
Sept., 2024, Vol. 44, No. 5, pp. 569–582
DOI:10.3770/j.issn:2095-2651.2024.05.001
Http://jmre.dlut.edu.cn

Influencer Identification of Threshold Models in Hypergraphs

Xiaojuan SONG, Xilong QU, Ting WEI, Jilei TAI, Renquan ZHANG*
School of Mathematical Sciences, Dalian University of Technology, Liaoning 116024*, P. R. China*

Abstract This paper mainly studies the influence maximization problem of threshold models in hypergraphs, which aims to identify the most influential nodes in hypergraphs. Firstly, we introduce a novel information diffusion rule in hypergraphs based on Threshold Models and conduct the stability analysis. Then we extend the CI-TM algorithm, originally designed for complex networks, to hypergraphs, denoted as the H-CI-TM algorithm. Secondly, we use an iterative approach to get the globally optimal solutions. The analysis reveals that our algorithm ultimately identifies the most influential set of nodes. Based on the numerical simulations, H-CI-TM algorithm outperforms several competing algorithms in both synthetic and real-world hypergraphs. Essentially, when provided with the same number of initial seeds, our algorithm can achieve a larger activation size. Our method not only accurately assesses the influence of individual nodes but also identifies a set of nodes with greater impact. Furthermore, our results demonstrate good scalability when handling intricate relationships and large-scale hypergraphs. The outcomes of our research provide substantial support for the applications of the threshold models across diverse fields, including social network analysis and marketing strategies.

Keywords hypergraph; threshold model; influence maximization; information diffusion; sub-critical path

MR(2020) Subject Classification 05C65; 60J60; 76R50; 94C15

1. Introduction

For decades, complex networks have served as an effective tool for representing intricate dynamical behaviors within various complex systems [1]. Historically, various complex systems have been effectively described as networks, where nodes represent individuals and edges signify interactions among them [2–4]. Numerous studies have been focused on complex networks [5,6]. However, the complexity inherent in biological and social interactions is often derived from the interactions among three or more entities, which cannot be adequately captured by binary interactions alone [7]. In recent years, hypergraphs have emerged as a representation for capturing higher-order interactions [7,8]. Simple graph is a simplified form of complex networks, which is also the basis of complex network research. The key distinction between hypergraphs and simple

Received October 15, 2023; Accepted December 17, 2023
Supported by the National Natural Science Foundation of China (Grant No. 12371516), the Natural Science Foundation of Liaoning Province (Grant No. 2022-MS-152) and the Fundamental Research Funds for the Central Universities (Grant No. DUT22LAB305).
* Corresponding author
E-mail address: sxjde@mail.dlut.edu.cn (Xiaojuan SONG); zhangrenquan@dlut.edu.cn (Renquan ZHANG)

graphs lies in the fact that in a simple graph, an edge can only connect two nodes, whereas in a hypergraph, a hyperedge can link any number of nodes. The introduction of hyperedges as a topological structure in hypergraphs enhances their flexibility, but it also increases the complexity of the structures and properties under investigation accordingly.

Information diffusion is a pivotal concept in the field of network science, particularly in the context of complex networks [9–11]. The application of information diffusion models extends across a diverse spectrum of domains, encompassing social networks, advertising and marketing strategies, disease transmission research, and search engine optimization [10]. Within social networks, information diffusion models assist in predicting and understanding the process by which information spreads throughout the network, thereby aiding the containment of rumor propagation. In the realm of marketing, these models enable us to anticipate user behavior and consumption decisions, facilitating the development of more effective market strategies. How information spreads from initially active nodes across the network topology into the entire network and how inactive nodes within the network are influenced and activated by their neighboring nodes are questions that have gained significant attention from a growing number of researchers [11,12]. One classic challenge within the realm of information diffusion research is the problem of Influence Maximization (IM) [2,8,9,13]. The primary goal of this problem is to strategically choose a specified number of highly influential seeds from within the network and, in accordance with specific activation rules, maximize the extent of network activation. Literature [14,15] has researched the Influence Maximization problem on hypergraphs.

The Linear Threshold Model (LTM) [16] has a long history, it finds wide applications in social and economic sciences, particularly in studying the spread of infectious diseases [17], information diffusion [18] and other physical processes [19]. Social contagion phenomena can be observed in everyday life, such as the spread of information, the adoption of new products, technologies, or medications, and the dissemination of public opinion, among others. Despite the seemingly disparate nature of these phenomena, they share common characteristics: they all stem from "variations" in the behavior of a small number of individuals (these individuals, influenced by factors such as public information, adopt behaviors distinct from the broader population), and subsequently, through interactions among individuals, these behaviors gain widespread prevalence, transitioning from individual actions to collective behavior. In the context of complex network propagation, the Watts Threshold Model [20] has played a pioneering role. Xu et al. delved into threshold models applied to hypergraphs [21]. They introduced a theoretical framework employing generative function techniques to deduce cascading conditions and the large components of fragile vertices. This approach is better to understand how higher-order interactions affect the process of system collapse. Liu et al. proposed a threshold model of cascading failure on hypergraphs that describes the propagation mechanism of failures among nodes and hyperedges [22]. They posited that when the proportion of failed nodes within a hyperedge exceeds a certain threshold, the entire hyperedge fails, consequently leading to the failure of the remaining nodes. The article employs the ratio of surviving nodes and the ratio of the largest connected component to assess the structural integrity of the system in a steady

state. Literature [23–26] has explored the application of the SIS model in simple graphs and hypergraphs. In this model, each individual can exist in one of two states: inactive or active, and is assigned a randomly determined "threshold". In the fundamental steps of the model, individuals in the inactive state make decisions about their own state by observing the current states of their neighbors. If the proportion of active individuals among their neighbors exceeds a specific threshold, they switch to the active state, otherwise, they remain inactive. In the activation rule investigated in this paper, it is specified that once an individual in the network transitions to the active state, they cannot revert back to the inactive state. In the initial stage of information propagation, a small random selection of nodes are designated as seed nodes and placed in the active state, while the remaining nodes remain inactive. The cascading effect generated by these seed nodes rapidly spreads the information throughout the entire network.

The aim of this paper is to identify the optimal seed set for maximizing the final activation size. In fact, the influence maximization problem is widely considered to be an NP-hard problem and continues to be a challenging issue within the field of complex network science.

2. Model

Activation rules for Threshold Models in simple graphs: In the simple graph model, an inactive node i is activated only if it has at least m_i active neighbors.

Activation rules for Threshold Models in hypergraphs: In the hypergraph model, an inactive node i is activated only if it has at least m_i active neighbors located in the different hyperedges that include node i. The activation process is depicted in Figure 1 (a). In the hypergraph, when $m_i = 2$, at time $T = 0$, there are three active nodes v_1, v_2, v_6. At $T = 1$, nodes v_1 and v_2 are two active neighbors of node v_3, and they are distributed across two different hyperedges, e_1 and e_2, both connected to node v_3, thus causing node v_3 to become activated. When $T = 2$, node v_5 has been actived by nodes v_3 and v_6 in the same way.

High-order Collective Influence in Threshold Models (H-CI-TM): We propose a theoretical framework to analyze the collective influence of nodes in hypergraphs based on threshold models. For a hypergraph with N nodes and M hyperedges, $\{A_{ij}\}_{N\times N}$ is the adjacency matrix of the hypergraph, A_{ij} represents the number of hyperedges that also contain nodes i, j . The vector $\mathbf{n} = (n_1, n_2, \ldots, n_N)$ represents the initial state of the nodes of hypergraph. $n_i = 1$ represents that the node i is selected as the initial seed, otherwise, $n_i = 0$. The initial activation scale in the hypergraph is $q = \sum_{i=1}^{N} \frac{n_i}{N}$. In the process of propagation, each node only has two states: active I and inactive S. To effectively induce cascading effects, the threshold models necessitate a specific quantity of initial seeds. These initial seeds are provided based on the ranking of their hyperdegrees. The propagation follows the activation rules described above until no new node will be activated. The introduction of a state variable v_i, denoted as "final state" of node i, $v_i = 1$ ($v_i = 0$) indicates the node has been actived (inactived). For a simple link $i \to j$, we define the variable $v_{i\to j}$ to describe the state of the node i assuming node j is removed from the hypergraphs. If $n_i = 1$, then $v_{i\to j} = 1$. If $n_i = 0$, $v_{i\to j} = 1$ only if the active neighbors of node

i (except node j) are at least scattered in m_i hyperedges attached to node i. There are $P_{\partial i\backslash j}^{m_i}$ situations for the choice of the m_i nodes from the neighbors (except node j) of node i, $P_{\partial i\backslash j}^{m_i}$ have $\binom{k_i-1}{m_i}$ elements, k_i is the degree of node i, which represents the number of all neighbors of node i. If these m_i neighbors are distributed across at least m_i hyperedges attached to node i, the combination is an effective combination P_{eh}, $P_{eh} = \{P_{eh_1}, P_{eh_2}, \ldots, P_{eh_n}\}$, apparently, P_{eh} is the subset of the $P_{\partial i\backslash j}^{m_i}$, each element $P_{eh_i} = \{P_{h_1}, P_{h_2}, \ldots, P_{h_{m_i}}\}$, $i = 1, \ldots, n$. There are $P_{\partial i}^{m_i}$ choices to choose m_i nodes from the neighbors (include node j) of node i, where effective combinations are denoted as P_{nh}. Simply understood as follows: the difference set between P_{eh} and P_{nh} is the effective combinations containing node j. An explanation of the effective combinations is shown in Figure 1 (b). The neighbors (except node j) of node i are in set $\{k_1, k_2, k_3, k_4\}$, combine them in pairs for a total of six combinations. However, there are only 5 groups satisfying that the nodes in these combinations are distributed in two or more hyperedges, and the valid combination is P_{eh_i}, $i = 1, \ldots, 5$. Since the combination $\{k_3, k_4\}$ is in the same hyperedge e_4, it is not a valid combination.

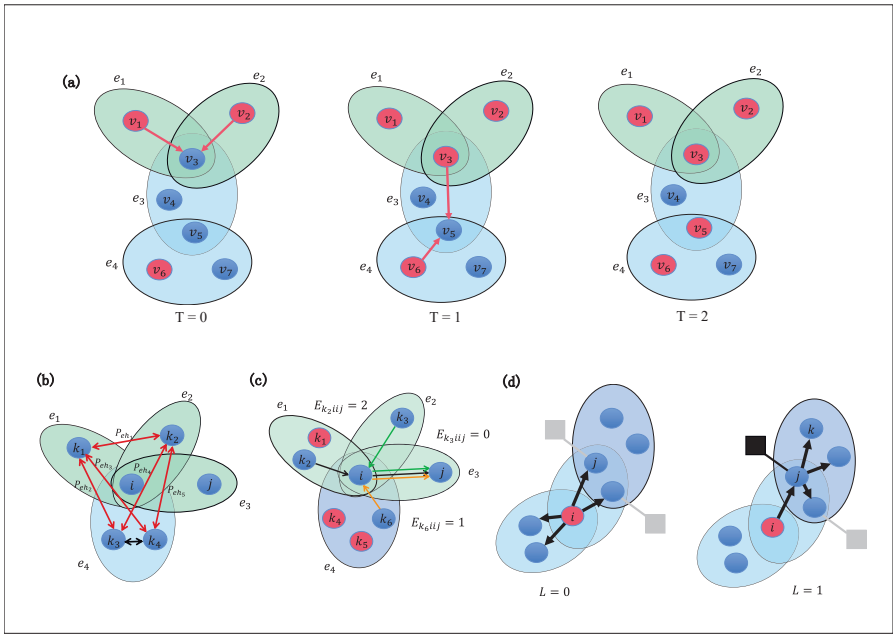

Figure 1 Interpretation about activation rule, effective combination, effective hyperedge, subcritical path in hypergraphs

For tree-like hypergraphs, the Massage Passing Equation can be described as:

$$v_{i\to j} = n_i + (1 - n_i)\Big[1 - \prod_{P_h \in P_{eh}} (1 - \prod_{p \in P_h} v_{p\to i})\Big]. \tag{2.1}$$

The final state of node i can be described as:

$$v_i = n_i + (1 - n_i)\Big[1 - \prod_{P_h \in P_{nh}} (1 - \prod_{p \in P_h} v_{p\to i})\Big]. \tag{2.2}$$

For the X directed link $i \to j$ in the hypergraph, let $v_{\to} = (\cdots, v_{i\to j}, \cdots)^{\mathrm{T}}_{X\times 1}$, then (2.1) can be seen as a nonlinear function of $v_{\to}$:

$$v_{\to} = n_{\to} + \mathbf{Y}(v_{\to}). \tag{2.3}$$

In (2.3), $n_{\to} = (\cdots, n_{i\to j}, \cdots)^{\mathrm{T}}$ where $n_{i\to j} = n_i$ for the directed link $i \to j$. $\mathbf{Y}(v_{\to}) = (\cdots, Y_{i\to j}, \cdots)^{\mathrm{T}}$ where $Y_{i\to j}$ is the nonlinear function of $v_{i\to j}$. A certain number of neighbors are taken from all the neighbors as a combination, and the combination number increases exponentially with the increase of the number of neighbors, which is very adverse to solving the equation, so this paper adopts iteration and linearization methods to solve (2.3). In discrete dynamical systems, (2.3) can be expressed as

$$v_{\to}^{t+1} = \mathbf{n}_{\to} + \mathcal{Z}^{\mathbf{t}}(v_{\to}^{t}) \tag{2.4}$$

with the initial condition $v_{\to}^{0} = \mathbf{n}_{\to}$, where $\mathcal{Z}$ is the Jacobian matrix of the nonlinear function $\mathbf{Y}$. Calculation method of matrix $\mathcal{Z}$ is as follows. First, partial derivation with $v_{\to}$ to $Y_{i\to j}$, then $Y'_{i\to j}(v_{\to}) = (\cdots, \frac{\partial Y_{i\to j}}{\partial v_{k\to l}}, \cdots)$. By (2.1), we know $\frac{\partial Y_{i\to j}}{\partial v_{k\to l}} = 0$ when $l \neq i$. When $l = i, k \neq j$, we have

$$\begin{aligned}\frac{\partial Y_{i\to j}}{\partial v_{k\to i}} = (1-n_i) \prod_{\overline{P_h}\in P_{eh}, k\notin\overline{P_h}} (1 - \prod_{p\in\overline{P_h}} v_{p\to i})\times \\ \sum_{P_h\in P_{eh}, k\in P_h} \Big[\Big(\prod_{p\in P_h\setminus k} v_{p\to i}\Big) \prod_{P_h{}'\neq P_h, k\in P_h{}'} \Big(1 - \prod_{p\in P_h{}'} v_{p\to i}\Big)\Big].\end{aligned} \tag{2.5}$$

The number of effective hyperedges attached to node i with link $k \to i, i \to j$ can be defined as

$$e_{k\to i, i\to j} = \sum_{e_l} \delta\Big(\sum_{k\notin e_l, p\in e_l\setminus j} v_{p\to i}\Big), \tag{2.6}$$

where $\delta(x) = 1$ if $x > 0$, otherwise $\delta(x) = 0$. The term e_l refers to an effective hyperedge attached to node i with link $k \to i, i \to j$ if and only if it satisfies the following three conditions: (a) The hyperedge associated with node i; (b) Node k is not located in this hyperedge; (c) There exists at least one active node within the effective hyperedge, excluding node j. Although (2.5) is complex, it can actually be interpreted using (2.6). When $e_{k\to i, i\to j} \geq m_i$, there is at least one term leading $\prod_{p\in\overline{P_h}} v_{p\to i} = 1$, because of the effective combinations obtained by taking m_i nodes from neighbors (excluding node k), at least one of them whose all nodes are active, then makes $\prod_{p\in\overline{P_h}} v_{p\to i} = 1$, which leads to $\frac{\partial Y_{i\to j}}{\partial v_{k\to i}} = 0$. When $e_{k\to i, i\to j} \leq m_i - 2$, there is $\prod_{p\in P_h\setminus k} v_{p\to i} = 0$, since all effective combinations containing node k contain at least one inactive node p after removing the node k, such that $\prod_{p\in P_h\setminus k} v_{p\to i} = 0$, which also leads to $\frac{\partial Y_{i\to j}}{\partial v_{k\to i}} = 0$. When $e_{k\to i, i\to j} = m_i - 1$, we have $\prod_{p\in\overline{P_h}} v_{p\to i} = 0$ and $\prod_{p\in P_h{}'} v_{p\to i} = 0$. There is only one effective combination which can lead $\prod_{p\in P_h\setminus k} v_{p\to i} = 1$. Therefore, we have $\frac{\partial Y_{i\to j}}{\partial v_{k\to i}} = 1 - n_i$. □

Then we give the definition of "subcritical state" in the hypergraphs. In simple graphs, node i is in a subcritical state if and only if it has $m_i - 1$ active neighbors [27]. However, in hypergraphs, node i is in a subcritical state if and only if it has $m_i - 1$ active neighbors which are located in at least different hyperedges that include node i. Based on the above analysis, the subcritical

variable for linking $k \to i, i \to j$ can be defined as follows:

$$E_{k\to l,i\to j} = \begin{cases} \prod\limits_{n=1}^{m_i-1} \Big(\sum\limits_{k\notin e_l, p\in e_l\setminus j} v_{p\to i} \Big), & \text{conditions}, \\ 0, & \text{otherwise}. \end{cases} \tag{2.7}$$

$E_{k\to l,i\to j}$ is non-zero if it satisfies the following conditions: (a) $i = l$, $k \neq j$, $k \to i$ and $i \to j$ belong to different hyperedges; (b) Node k is inactive; (c) Node i is subcritical. Especially, when $m_i = 2$, $E_{k\to l,i\to j}$ is quantitatively equal to the number of active nodes in this effective hyperedge, and if no effective hyperedge exists, $E_{k\to l,i\to j} = 0$. As is shown in the Figure 1(c). e_1 is the effective hyperedge attached to the node i for link $k_6 \to i, i \to j$, so $e_{k_6\to i,i\to j} = 1$, $E_{k_6\to i,i\to j} = 1$; e_4 is the effective hyperedge attached to the node i for link $k_2 \to i, i \to j$, so $e_{k_2\to i,i\to j} = 1$, $E_{k_2\to i,i\to j} = 2$; e_1 and e_4 are the hyperedges attached to the node i for link $k_3 \to i, i \to j$, so $e_{k_3\to i,i\to j} = 2$, $E_{k_3\to i,i\to j} = 0$. $\mathcal{Z}^{\mathbf{t}} = (\cdots, Y'_{i\to j}(v^{\mathrm{T}}_{\to}), \cdots)^{\mathrm{T}}$ in (2.4) is the $X \times X$ matrix with the link $k \to l, i \to j$:

$$\mathcal{Z}^t_{k\to l,i\to j} = \frac{\partial Y_{i\to j}}{\partial v_{k\to i}}\Big|_{v^t_{\to}}. \tag{2.8}$$

Review the definition of $E_{k\to l,i\to j}$, $\mathcal{Z}^t_{k\to l,i\to j}$ can be written as

$$\mathcal{Z}^t_{k\to l,i\to j} = (1 - n_i)E^t_{k\to l,i\to j}. \tag{2.9}$$

In order to facilitate calculation, $\mathcal{Z}^t_{k\to l,i\to j}$ and $E^t_{k\to l,i\to j}$ can be simplified to $\mathcal{Z}^t_{klij}$ and E^t_{klij} respectively by extending them to N dimensional space. And they can be defined as:

$$\mathcal{Z}^t_{klij} = (1 - n_i)A_{kl}A_{ij}\delta_{il}(1 - \delta_{kj})E^t_{klij}. \tag{2.10}$$

If $i = l$, then $\delta_{il} = 1$, and 0 otherwise. When $i = l, k \neq j$, we have

$$v^1_{i\to j} = n_i + (1 - n_i)A_{ij}\sum_k A_{ki}(1 - \delta_{kj})E^0_{kiij}n_k. \tag{2.11}$$

Similar to the subcritical paths in simple graphs, define the subcritical paths in a hypergraph: for a directed link $i \to j$, if $n_{i_1} = 1, n_{i_2} = 0, \ldots, n_i = 0, E^0_{i_1i_2i_2i_3} \neq 0, \ldots, E^{L-1}_{i_Liij} \neq 0$, then $i_1 \to i_2 \to \cdots \to i_L \to i \to j$ is called a subcritical path of length L in the hypergraph. Figure 1 (d) shows the seed node i contributes to $\|v_{\to}\|$ by a subcritical path of length 0 and 1, where the square represents the nodes in subcritical state. Due to the complex structure of hypergraphs, this paper only studies the subcritical paths of length 1. □

H-CI-TM algorithm: The active population $\|v_{\to}\|$ is defined as follows:

$$\begin{aligned} \|v_{\to}\| &= \sum_{i,j\in\partial i} v_{i\to j} = \sum_{i,j\in\partial i} A_{ij}n_i + \sum_{i,j\in\partial i,k\in\partial j\setminus i} (1 - n_i)A_{ki}A_{ij}E^t_{kiij}n_k \\ &= \sum_{i,j\in\partial i} A_{ij}n_i + \sum_{k\in\partial j\setminus i} n_k \sum_{i,j\in\partial i} (1 - n_i)A_{ki}A_{ij}E^t_{kiij} \\ &= \sum_{i,j\in\partial i} A_{ij}n_i + \sum_i n_i \sum_{j\in\partial i,k\in\partial j\setminus i} (1 - n_k)A_{ik}A_{kj}E^t_{ikkj} \\ &= \sum_{i,j\in\partial i} A_{ij}n_i + \sum_i n_i \sum_{j\in\partial i,k\in\partial j\setminus i} (1 - n_j)A_{ij}A_{jk}E^t_{ijjk} \end{aligned}$$

$$= \sum_{i,j\in\partial i} A_{ij}n_i + \sum_i n_i \sum_{j\in\partial i}(1-n_j)A_{ij} \sum_{k\in\partial j\setminus i} A_{jk}E^t_{ijjk}. \tag{2.12}$$

Based on the CI-TM algorithm on simple graphs, this paper extends it to hypergraphs and proposes the H-CI-TM algorithm. Similar to propagation in the simple graph, maximize $\|v_\rightarrow\|$. When there are no active nodes in the hypergraphs, $\|v_\rightarrow\| = 0$, $\|v_\rightarrow\|$ increases with the increase of active nodes. The goal of this paper is to optimize the selection of nodes to maximize $\|v_\rightarrow\|$ given a certain number of seed nodes. A node's contribution to $\|v_\rightarrow\|$ is defined as the H-CI-TM value (that is, the collective impact if the node is selected as seeds), and the nodes with higher H-CI-TM values have the greater influence in the network. When $L = 0$, the CI-TM value of the node in the simple graph is defined as the degree of the node, that is, the number of neighbors of the node; The H-CI-TM value of the node in the hypergraph is defined as:

$$H - CI - TM_0 = \sum_{j\in\partial i} A_{ij}, \tag{2.13}$$

where A denotes the adjacency matrix of the hypergraph. The 0-th order H-CI-TM value of the node i corresponds to the summation of the i-th row in the adjacency matrix, in which subcritical paths are not involved. Moving forward, when $L = 1$, the H-CI-TM value of the node i is defined as follows:

$$H - CI - TM_1 = \sum_{j\in\partial i} A_{ij} + \sum_{j\in\partial i}(1-n_j)A_{ij} \sum_{k\in\partial j\setminus i} A_{jk}E_{ijjk}. \quad \square \tag{2.14}$$

Algorithm: H-CI-TM algorithm

Input: Initial seed set S_0 (The corresponding hypergraph state is Sta_0), Hypergraph H
Output: Seed node set S
Step 1. Based on Sta_0, calculate the $H - CI - TM_1$ of the inactive nodes;
Step 2. Select i with the largest $H - CI - TM_1$, seed set updated to $S = S_0 \cup \{i\}$;
Step 3. Use S to activate the whole hypergraph until no new nodes are activated, then gets the new hypergraph state Sta_1;
Step 4. Calculate the $H - CI - TM_1$ for inactive nodes within the first and second-layer neighbors of newly activated nodes, select j with the largest $H - CI - TM_1$, then $S = S \cup \{j\}$;
Step 5. Repeat Steps 3–Step 4 until all effective nodes whose hyperdegree higher than one in the hypergraph are activated, then obtain the optimal seed set S.

3. Numerical simulation

To test the performance of the H-CI-TM algorithm, numerical simulation experiments were conducted in both randomly generated hypergraphs and real-world hypergraphs. The generated hypergraphs included Erdős-Rényi (ER) hypergraphs, Scale-free (SF) hypergraphs, and Uniform hypergraphs. Here, "Uniform hypergraphs" refers to hypergraphs in which all hyperedges have

the same cardinality. In this paper, we focus on the propagation scenario where the threshold is set to 2. This means that a node will become active only if it has two or more active neighbors, and these active neighbors are distributed across two or more hyperedges attached to this node. It is worth noting that the hypergraphs used in our experiments are sparse hypergraphs. This is because standard formulations of message-passing equations have a significant limitation: they rely on the assumption that the states of neighbors must be independent of each other, which is true only when a hypergraph is devoid of loops [28]. However, this assumption is idealized, as real-world hypergraphs almost always contain loops. In hypergraphs, as a hyperedge can encompass several nodes and a node can be present in several hyperedges simultaneously, they offer enhanced flexibility compared to simple graphs. However, this increased flexibility also substantially raises the complexity of problem investigation. Therefore, the generated hypergraphs used in this study are all sparse hypergraphs.

To demonstrate the effectiveness of the H-CI-TM algorithm, a comparison was conducted with several other classical ranking algorithms, including High-Hyper Degree (HHD) [29], Page Rank (PR) [30] and Random algorithms. In the model, p represents the proportion of selected seeds relative to the entire hypergraph, while $Q(q)$ represents the activation scale indicating the proportion of active nodes in the hypergraphs, N and M represent the number of nodes and hyperedges in the hypergraphs, respectively. (N, M) represents the scale of the hypergraphs.

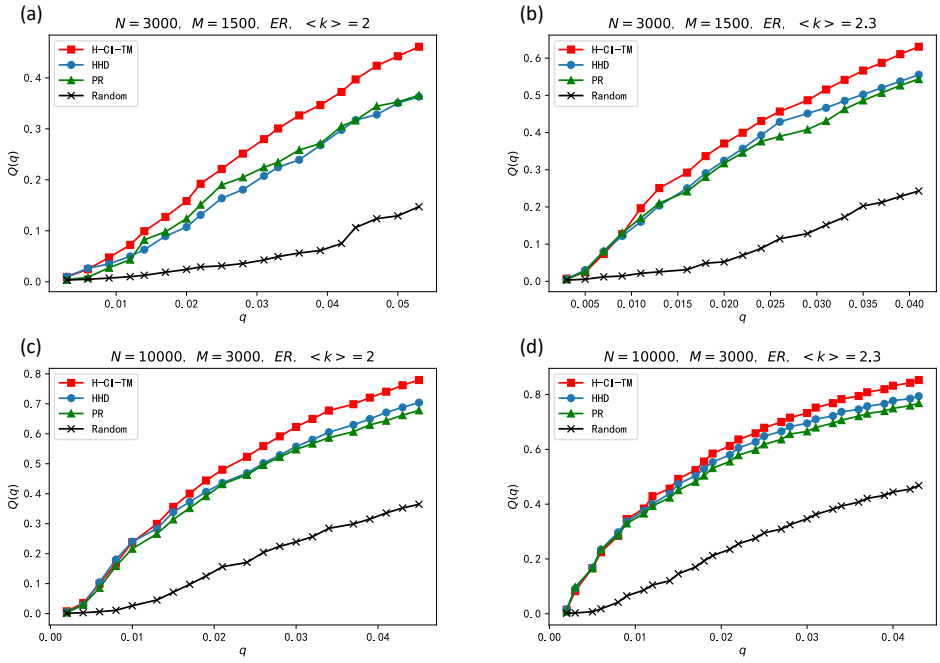

Figure 2 The performance of H-CI-TM algorithm on the ER hypergraphs

3.1. Erdős-Rényi hypergraphs

Firstly, numerical simulations were conducted in classical ER hypergraphs with scales $N =$

3000 and $N = 10000$, featuring average hyperedegree $\langle k \rangle = 2$ and $\langle k \rangle = 2.3$, respectively. As depicted in Figure 2, when selecting the same proportion of seeds to activate the hypergraphs, the H-CI-TM algorithm achieved the highest activation scale. The experimental effect of HHD and PR algorithm is similar, and the seed selected by Random has the worst activation effect. It is noteworthy that in the ER hypergraph experiments with $N = 10000$, $\langle k \rangle = 2.3$, while the H-CI-TM algorithm demonstrates the best experimental performance, the improvement relative to other algorithms is not particularly significant. This could be attributed to the possibility that when the average hyperdegree reaches 2.3, the presence of "loops" within the hypergraphs is already abundant, indicating that the hypergraphs at this stage are no longer characterized by sparse hypergraphs.

3.2. Scale-free (SF) hypergraphs

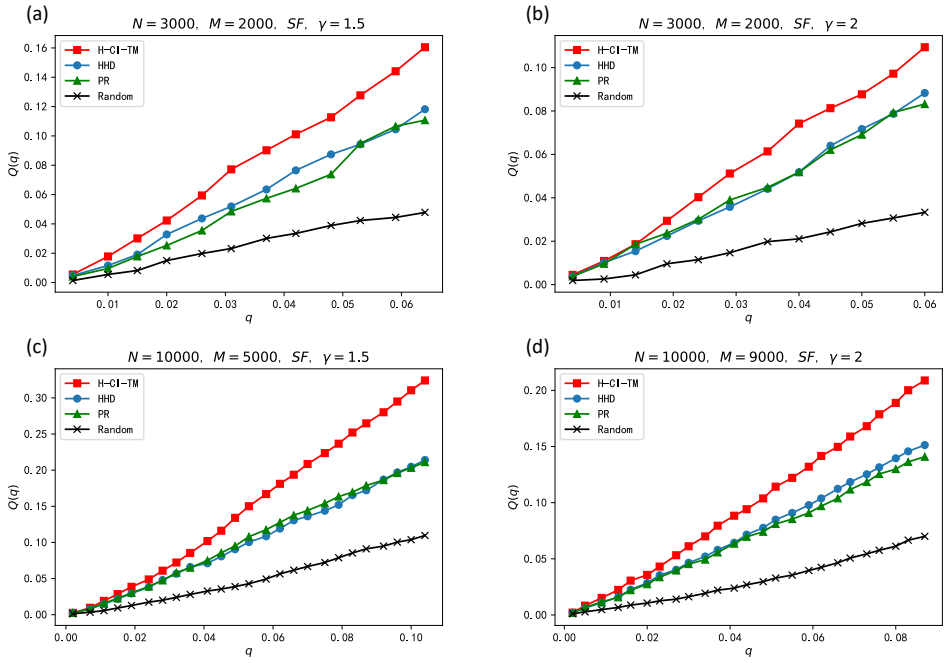

Figure 3 The performance of H-CI-TM algorithm on the SF hypergraphs

Secondly, we conduct numerical simulation in the classical SF hypergraphs, "SF hypergraphs" refers to the hypergraphs in which the hyperdegree sequence of nodes follows a power-law distribution, and the sequence of hyperedge cardinalities also follows a power-law distribution. We run SF hypergraphs of size $N = 3000$ and $N = 10000$, each with power-law indices of $\gamma = 1.5$ and $\gamma = 2$, respectively. As illustrated in Figure 3, when the same proportion of seeds was selected to deactivate the hypergraphs, the H-CI-TM algorithm exhibited the largest activation scale. The experimental effect of HHD and PR algorithm is similar, and the seed selected by Random has the worst activation effect. In the course of experiment, it is found that SF hypergraph takes the shortest time to conduct the experiment. It is because SF hypergraphs are sparser than other

types of hypergraphs, which once again verifies that our algorithm is more suitable for sparse hypergraphs.

3.3. Uniform hypergraphs

Finally, we conducted numerical simulation on uniform hypergraphs. If the cardinality of each hyperedge is identical, equal to d, then the hypergraph is called d-uniform hypergraph. Similar to the other two types of generated hypergraphs, we also ran uniform hypergraphs with sizes $N = 3000$ and $N = 10000$, each having a hyperedge cardinality of $d = 3$ and $d = 5$, respectively. As depicted in Figure 4, the H-CI-TM algorithm consistently exhibited the highest activation scale when selecting the same proportion of seeds. The experimental effect of HHD and PR algorithm is similar, and the seed selected by Random has the worst activation effect. It is worth noting that in the uniform hypergraphs, when the network size is 3000 and the seed ratio is 0.1, the final activation ratio of the seeds selected by H-CI-TM is already over 70%.

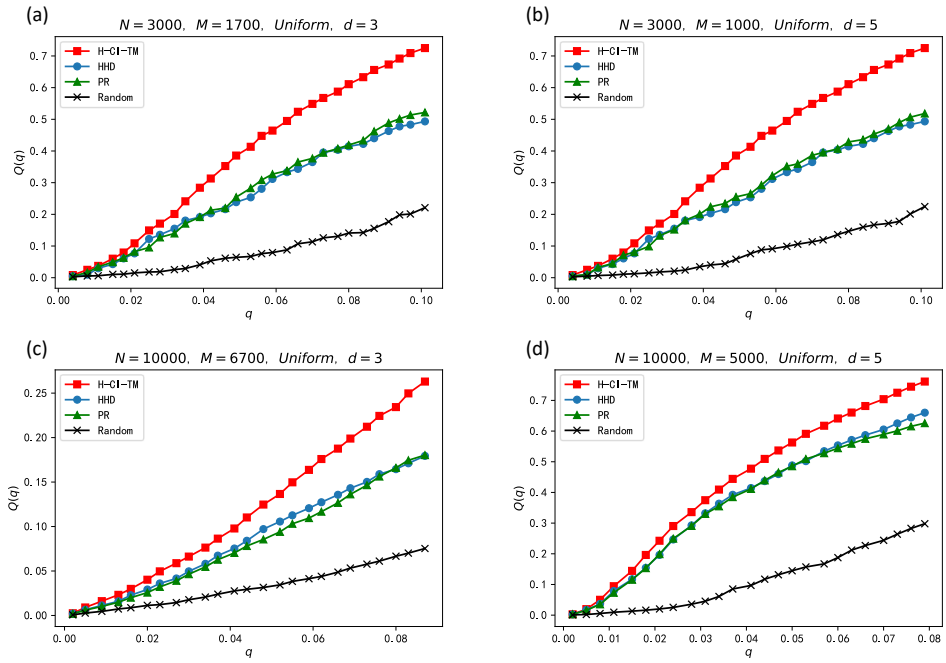

Figure 4 The performance of H-CI-TM algorithm on the uniform hypergraphs

3.4. Real-world hypergraphs

In addition to our experiments on generated hypergraphs, we conducted a validation of our algorithm using real-world hypergraphs datasets across various domains. These real-world hypergraph datasets are shown in the Table 1: N and M represent the number of nodes and hyperedges of the hypergraphs, respectively; $\langle k \rangle$ signifies the mean hyperdegree of the nodes; $\langle d \rangle$ represents the average of the cardinality of the hyperedges; $\langle \text{deg} \rangle$ characterizes the degree of the nodes. The detailed description of each dataset is given as follows:

Cat-edge-algebra-questions dataset (Algebra): This dataset comprises user interactions on a mathematics website, where users who answered or commented on the same algebra problem were grouped into the same hyperedge.

Cat-edge-vegas-bars-reviews (Bars-Rev): In this dataset, we examined user interactions within a mobile application (APP), with users who browsed the same sections or bars being clustered into the same hyperedge.

Senate-committees: This dataset includes the relationship of the political party of the person, and the people belonging to the same political party are in the same hyperedge.

Trivago-clicks: The hypergraph was constructed from training data originating from user interactions on Trivago. In this dataset, each row in the original training dataset corresponds to an action taken by a user during a browsing session on the Trivago platform. Here, nodes within the hypergraph represent accommodations, while hyperedges consist of sets of accommodations for which a user performed a "click-out" action during the same browsing session, indicating that the user was redirected to a partner site.

Cat-edge-geometry-questions dataset (Geometry): This dataset is similar to the Algebra dataset. The nodes represent the users of MathOverflow and hyperedges are sets of users who answerered the same geometry question.

iAF1260b: The data contains nodes representing reaction-based metabolites and hyperedges are sets of metabolites which are applied to a certain reaction.

Hypergraph	N	M	$\langle k \rangle$	$\langle d \rangle$	$\langle \deg \rangle$
Algebra	423	1,268	19.53	1.95	78.9
Bars-Rev	1,234	1,194	9.61	2.1	174.3
Senate-committees	282	315	19	17.5	14.7
Trivago-clicks	171,495	220,758	4	4.2	7
Geometry	580	1193	21.53	1.75	164.79
iAF1260b	1668	2351	5.46	2.67	13.26

Table 1 Summary of real-world datasets

In light of the H-CI-TM algorithm's applicability, which is primarily suited for sparse hypergraphs, it becomes essential to preprocess real-world hypergraph data to transform them into sparser hypergraphs. This preprocessing typically involves several steps, such as the removal of excessively large hyperedges, the elimination of hyperedges containing only two nodes, experimentation on subgraphs, and the pruning of highly connected nodes along with their associated hyperedges. The application of these techniques effectively reduces the density of the hypergraphs. These preprocessing strategies are vital for adapting H-CI-TM to real-world hypergraphs, ensuring its effectiveness and reliability across diverse hypergraph structures. The experimental outcomes for six real-world hypergraphs are presented in Figure 5, showcasing the robust performance of H-CI-TM across all tested real hypergraphs. Notably, in the Senate-committees dataset, the hypergraph generated from the data frequently exhibits large hyperedges, resulting

in an activation scale exceeding 80 percent in the initial stages of the experiment. The results obtained from the other five real hypergraphs align consistently with those from the generated hypergraphs.

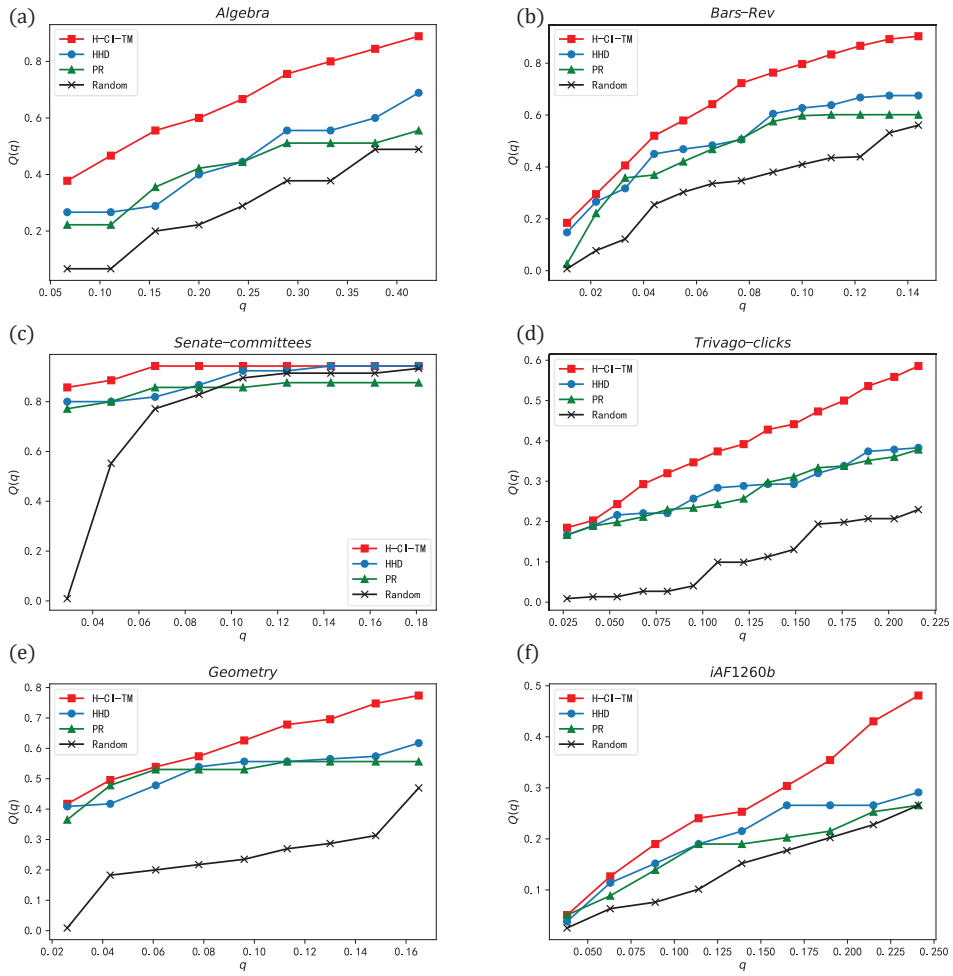

Figure 5 The performance of H-CI-TM algorithm on the real-world hypergraphs

4. Conclusion

Threshold models and the influence maximization problem have been extensively studied due to their wide range of applications, including financial risk assessment, social networks, and protein action networks. This paper focuses on researching the influence maximization problem in hypergraphs based on threshold models. Firstly, we propose a novel threshold model that considers the complex association between nodes and hyperedges in hypergraphs. We then extend the Message Passing theory analysis framework to study the high-order collective influence (H-CI) of nodes in hypergraphs. Subsequently, we design a greedy algorithm (H-CI-TM) to efficiently

search for the most influential set of nodes in hypergraphs based on the Threshold Models. Finally, we validate the effectiveness and superiority of the H-CI-TM algorithm through numerical simulations conducted on synthetic and real-world hypergraphs. Moreover, our findings reveal that, under this activation rule, there is no first-order phase transition phenomenon present in simple graphs. This discovery further underscores the potential applicability of influence maximization based on hypergraph threshold models in realms such as social networks and information dissemination. It introduces novel ideas and approaches for further research and practical implementations in these domains.

It is worth highlighting that our current study primarily concentrates on scenarios where the threshold is set to 2 and the subcritical path is 1. However, there is a need to investigate more general threshold models in hypergraphs in future research. By exploring different threshold values and their effects on influence maximization, we can gain a deeper understanding of the dynamics and behavior of information spread in various hypergraphs. Furthermore, it is important to acknowledge that the simplicity of the threshold models may limit their ability to accurately depict complex scenarios, such as disease transmission dynamics. Therefore, in future studies, we aim to explore and develop a more comprehensive probability model in hypergraphs. This probabilistic approach will allow us to capture the inherent uncertainties and complexities associated with real-world phenomena, enabling a more accurate analysis of information diffusion and influence maximization. By expanding our research to encompass more general threshold models and incorporating a probabilistic framework, we can enhance the applicability and robustness of our findings. This will contribute to a more comprehensive understanding of influence maximization in hypergraphs and potentially offer valuable insights for various domains, including social networks, epidemiology, and marketing strategies.

Acknowledgements We thank the editor and referees for their time and comments.

References

[1] I. IACOPO, P. GIOVANNI, B. ALAIN, et al. *Simplicial models of social contagion*. Nat. Commun., 2019, **10**(1): 2485.

[2] O. A. SICHANI, J. MAHDI. *Influence maximization of informed agents in social networks*. Appl. Math. Comput., 2015, **254**: 229–239.

[3] F. ALTARELLI, A. BRAUNSTEIN, L. D. ASTA, et al. *Large deviations of cascade processes on graphs*. Phys. Rev. E., 2013, **87**: 062115.

[4] F. ALTARELLI, A. BRAUNSTEIN, L. D. ASTA, et al. *Optimizing spread dynamics on graphs by message passing*. J. Stat. Mech. Theory Exp., 2013, **9**: P09011, 24 pp.

[5] K. MAKSIM, K. G. LAZAROS, H. SHLOMO, et al. *Identification of influential spreaders in complex networks*. Nat. Phys., 2010, **6**: 888–893.

[6] Renquan ZHANG, Xiaolin WANG, Sen PEI. *Targeted influence maximization in complex networks*. Phys. D, 2023, **446**: Paper No. 133677, 12 pp.

[7] B. FEDERICO, C. GIULIA, I. IACOPO, et al. *Networks beyond pairwise interactions: structure and dynamics*. Phys. Rep., 2020, **874**: 1–92.

[8] A. ALESSIA, C. GENNARO, S. CARMINE, et al. *Social influence maximization in hypergraphs*. Entropy, 2021, **23**(7): Paper No. 796, 20 pp.

[9] Sen PEI, Xian TENG, S. JEFFREY, et al. *Efficient collective influence maximization in cascading processes with first-order transitions*. Sci-Rep., 2017, **7**(1): 45240.

[10] Qi SUO, Jinli GUO, Aizhong SHEN. *Information spreading dynamics in hypernetworks*. Phys. A., 2018, **495**: 475–487.

[11] S. KAZUMI, M. KIMURA, K. OHARA, et al. *Super mediator–a new centrality measure of node importance for information diffusion over social network*. Inf. Sci., 2016, **329**: 985–1000.

[12] Ang GAO, Ying LIANG, Xiaojie XIE, et al. *Social network information diffusion method with support of privacy protection*. Front. Comput. Sci., 2021, **15**(2): 233–248.

[13] M. FLAVIANO, M. A. HERNAN. *Influence maximization in complex networks through optimal percolation*. Nature (London), 2015, **524**(7563): 65–68.

[14] Ming XIE, Xiuxiu ZHAN, Chuang LIU, et al. *An efficient adaptive degree-based heuristic algorithm for influence maximization in hypergraphs*. Inf Process Manag., 2023, **60**(2): 103161.

[15] M. E. AKTAS, S. JAWAID, I. GOKALP, et al. *Influence Maximization on Hypergraphs via Similarity-based Diffusion*. IEEE International Conference on Data Mining Workshops (ICDMW), 2022, **22**: 1197–1206.

[16] M. GRANOVETTER. *Threshold models of collective behavior*. Am. J. Sociol, 1978, **83**: 1420–1443.

[17] C. H. CHENG, Y. H. KUO, Ziye ZHOU. *Outbreak minimization v.s. influence maximization: an optimization framework*. Med Inform Decis Mak., 2020, **20**(1): 266.

[18] L. SUNGSU, J. INWOO, J. KYOMIN, et al. *Analysis of information diffusion for threshold models on arbitrary networks*. Eur Phys J B., 2015, **88**(8): 201.

[19] Wei ZHANG, Mingsheng HE. *Influence of Opinion Leaders on Dynamics and Diffusion of Network Public Opinion*. ICMSEM., 2013, 139–144.

[20] D. J. WATTS. *A simple model of global cascades on random networks*. Proc. Natl. Acad. Sci. U.S.A., 2002, **99**(9): 5766–5771.

[21] Xinjian XU, Shuang HE, Lijie ZHANG. *Dynamics of the threshold model on hypergraphs*. Chaos (Woodbury, N.Y.), 2022, **32**(2): 023125.

[22] Runran LIU, Chunxiao JIA, Ming LI, et al. *A threshold model of cascading failure on random hypergraphs*. Chaos. Solitons Fractals., 2023, **173**: 113746.

[23] Á. BODÓ, G. Y. KATONA, P. L. SIMON. *SIS epidemic propagation on hypergraphs*. Bull. Math. Biol., 2016, **78**(4): 713–735.

[24] C. V. PEDRO, F. BULLO. *Multigroup SIS epidemics with simplicial and higher-order interactions*. IEEE Trans. Control., 2022, **9**(2): 695–705.

[25] A. DE, F. GUILHERME, P. GIOVANNI, et al. *Social contagion models on hypergraphs*. Phys. Rev. Res., 2020, **2**(2).

[26] B. JHUN, J. MINJAE, B. KAHNG. *Simplicial SIS model in scale-free uniform hypergraph*. J Stat Mech-Theory E., 2019, **12**: 123207.

[27] A. V. GOLTSEV, S. N. DOROGOVTSEV, J. F. F. MENDES. *K-core (bootstrap) percolation on complex networks: critical phenomena and nonlocal effects*. Phys. Rev. E., 2006, **73**: 056101.

[28] A. KIRKLEY, T. C. GEORGE, M. E. J.NEWMAN. *Belief propagation for networks with loops*. Sci. Adv., 2021, **7**(17).

[29] R. ALBERT, H. JEONG, A. L. BARABÁSI. *Error and attack tolerance of complex networks*. Nature., 2000, **406**: 378–382.

[30] S. BRIN, L. PAGE. *The anatomy of a large-scale hypertextual web search engine*. Comput. Netw., 1998, **30**(1-7): 107–117.