Heliyon 6 (2020) e05636

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Product information diffusion model and reasoning process in consumer behavior

Xuehua Sun [a,*], Shaojie Hou [a], Ning Cai [b], Wenxiu Ma [a]

[a] School of Information Technology, HeBei University of Economics and Business, HeBei 050061, PR China
[b] School of Culture and Communication, HeBei University of Economics and Business, HeBei 050061, PR China

ARTICLE INFO

Keywords:
Social network
Product information diffusion
Bayesian network modeling
Consumer behavior reasoning
Mathematical modeling
Stochastic process
Network analysis
Social media
Behavioral economics
International relations
Economics
Psychology
Information science

ABSTRACT

Information diffusion on social media has become a major approach in people's daily communication, and the value contained therein holds great interest for both academic and industrial communities. However, the process of information diffusion is affected by many factors, and the complexity of that process has not been fully explored. Most previous studies have concentrated on the strategies and driving forces in social media operations, as well as the identification of influential seed nodes, yet analyses of consumer behavior choice in the process of information diffusion are rare. Thus, This study proposes a multipoint cross-diffusion model based on MapReduce, which improves the single-point model and can better describe the product information diffusion process. On that basis, a Bayesian network model of product information diffusion was constructed to analyze the associations between factors and consumer behaviors. Moreover, the posterior probability of consumer behavior choice affected by a series of factors in the information forwarding process was considered and analyzed. This study's findings can be used to estimate the posterior probability that users will purchase, forward, or stay silent, thereby predicting the effect of product diffusion and obtaining the quantitative relationships between factors and consumer behavior.

1. Introduction

As consumer-led media and technology—such as mobile intelligent terminals, WeChat, and social media websites—rapidly develop, consumers not only produce large quantities of data online but also spontaneously build their own marketing networks, thus blurring the boundaries between enterprises and consumers. As content providers and information publishers, consumers have built their own media networks. Additionally, enterprises publish abundant product information through various forms of social media—such as network platforms and intelligent media terminals—to attract the attention of potential users. Incentives have been used to encourage consumers to share product information on their self-organized social media, such that consumers are transformed from people who browse product information and purchase products into product promoters and enterprise collaborators. Such partnerships can benefit both enterprises and consumers.

As a social media platform, Facebook is valued at more than 100 billion US dollars. This is mainly because posts by Facebook users, and the information users exchange, contain potential market demand. This represents not only a market resource that enterprises strive to explore but also the commercial value of Facebook as a social media network.

Consumers have their own circles of friends on social media. Presented with massive amounts of data and media advertising, consumers do not always know how to make the right choices. Therefore, for reasons of trust, they prefer to receive recommendations from friends in their communication circles.

Information diffusion is mainly a rumor-diffusion model developed based on the epidemic model. It assumes that the relationship between people is unknown, and this relationship constitutes an invisible network. Therefore, the epidemic model is applicable for studying a global model (e.g., trends and speed in information diffusion) when the specific propagation path is not the focus.

Proposed by Anderson in 1991, the classic epidemic disease model SIR (Susceptible-Infectious-Recovered) (Cheng et al., 2013; Dybiec, 2009) has been widely used and extended, especially in the spread of rumor. Scholars have made many improvements to the SIR model, and these improved models are summarized as follows. The classic rumor-spread model is the DK (Daley and Kendal) model (Daley & Kendal, 1964, 1965), which posits that rumor spread is similar to the

* Corresponding author.
E-mail address: sunxuehua795@126.com (X. Sun).

https://doi.org/10.1016/j.heliyon.2020.e05636
Received 11 April 2020; Received in revised form 5 October 2020; Accepted 27 November 2020

transmission of infectious disease. In the DK model, people are divided into three categories: the ignorant, the spreader, and the terminator. The SSIC model (Tian et al., 2015) can effectively interfere with the spread of rumors on super networks (Denning, 1985) by (1) identifying rumors and isolating them, and (2) improving the openness of rumors and allowing the public to know more about them to weaken their spread. The SEIR model (Xia et al., 2015) considers the attractiveness and ambiguity of rumor content. Mean-field equations are used to characterize dynamics in the SEIR model in homogeneous and heterogeneous networks. In the SEIR model, rumors spread faster in the BA network than in the WS network, while the diffusion scales of rumor spread in the two networks are exactly the opposite. Mean-field equations are used to describe the dynamics of rumor models to better understand the characteristics of rumor spread and analyze the key events in rumor spread in complex networks. A novel SIR model (Wang et al., 2013) was applied to homogeneous/heterogeneous networks. The finding was that rumors spread faster in homogeneous networks than in heterogeneous networks, while the diffusion scales of rumor spread in the two networks are exactly the opposite. Naumov and Tao (2017) added marketing into the standard threshold model of social networks and studied properties of the influence relation in social networks.

In general, rumor spread has been improved using the SIR model, mainly in two aspects: (1) increasing the attributes of the study objects and (2) constructing network structures based on different attributes.

The epidemic disease model and the improved rumor spread model differ from the information diffusion model in networks self-organized by consumers in two ways. First, information is diffused in different ways. In the epidemic disease model, node infection is forced and spontaneous, while information diffusion in a consumer network is voluntary and optional. Second, the targets are different. Scholars study rumor models to suppress and disturb rumor spread in hopes of minimizing its impact. Meanwhile, studies of product information diffusion in consumer networks aim to encourage consumers' forwarding behaviors and maximize their effects (Kempe et al., 2003; He et al., 2012; Li et al., 2013). Therefore, the epidemic model and the improved rumor model based on the epidemic model cannot abstractly represent information dissemination processes in consumer networks.

Moreover, existing information diffusion models are mostly single-point models that do not consider multipoint cross diffusion. In reality, users often use several types of social media, so different social media tend to share some users in common who often spread messages across platforms and groups; thus, cross-diffusion situations often arise. As such, single-point models cannot adequately represent real information diffusion processes. Hence, a multipoint cross-diffusion model improves upon the shortcomings of single-point models and can better describe product information diffusion processes. Moreover, the results obtained by such an approach have great theoretical significance as well as practical value.

This study has conducted a multi-attribute analysis on the information diffusion process. In previous studies, scholars mainly focused on the analysis of the diffusion network structure and took a single influence factor into account. However, information diffusion process is subjected to multiple factors and the complexity is not fully considered. This study investigates information diffusion process from four aspects, namely, network structure, information attribute, users' attributes and the leakage of users' information in transmission process. More specifically, the strong and weak connections in social networks, information overload, controversies and emotions in information, the connection strength among users, the influence of users, the users' transmitting behaviors and the threats to security and privacy are examined in detail; in addition, the effects of the above factors on information diffusion and the influence relations among these factors are also analyzed. This section has provided antecedent analysis for the subsequent modeling of information diffusion process.

The diffusion model of the products' information is established. This study proposes a multi-point cross communication model based on MapReduce so as to spread the information in a wider range. It should be noted that the cross communication among multiple points is not to transmit information simply and blindly and but to realizes the following two functions: (1) filtering the information in one group; (2) sorting the filtered dataset and finding the most interesting and unique subject for diffusion. On that basis, the Bayesian network (BN) model for describing the diffusion of products' information is established, in which the correlations of influence factors and users' behaviors among the propagation of products' information are analyzed. Further, this study explores the model's three attributes, namely, conditional independence attribute, factorization attribute and the minimum independent set attribute, respectively. The established BN model of products' information diffusion and the related features analysis provide a data structure of knowledges representation and inference bases for subsequent inference.

We propose some effective methods for behavioral inference, which refers to acquiring the posterior probability of the consumers' behavior choice in information transmitting process under a series of influencing factors. This study employs sum-product inference algorithm for generating the information diffusion clique tree, then applies influence passing algorithm in the clique tree and acquires the marginal posterior probabilities of the consumers purchasing, transmitting products' information and taking no actions.

The structure of this paper is as follows. The subsequent section describes the construction of the multipoint model of product information diffusion; this includes a multipoint cross-diffusion model based on MapReduce technology and a Bayesian model of product information diffusion. The third part explores consumer behavior reasoning in the process of product information diffusion. The fourth part presents the experiment and the discussion, and the fifth part provides the conclusions.

2. Multipoint model of product information diffusion

In this study, the data used for the multipoint cross-diffusion model were extracted from different social media or servers; thus, the data were distributed. MapReduce can perform the parallel processing of distributed data (Fang et al., 2013); thus, multipoint cross diffusion based on MapReduce was proposed in this study.

As a probabilistic digraph, a Bayesian network can intuitively express people's causal knowledge with digraphs, conduct multifactorial modeling, graphically represent joint probability distribution between random variables, and deal with various uncertain factors. Therefore, in this study a Bayesian network was used to represent the data structure of the product information diffusion network, extract the multiattribute characteristics of the information diffusion process, and model the information diffusion process.

2.1. Multipoint cross diffusion based on MapReduce

Current social media platforms mainly have two forms of information diffusion: diffusion within one point and point-to-point diffusion, where the “point” refers to a social media network (e.g., a WeChat group or QQ group) that has community structures composed of closely related nodes. These relationships are attached to the abovementioned social media and are diffused within a specific small circle. Due to the clustering of the community structure, trust between the nodes is high, and information spreads rapidly within the community.

However, a drawback of such a clustered social media structure is that it hinders large-scale information diffusion. Thus, to spread information in a broader context, some network porters are needed to diffuse information among multiple points. These network porters can be nodes formed by the crossover of multiple social media or seed nodes cultivated by enterprises. Among these, nodes formed by the crossover of multiple social media may be ordinary nodes whose cross-diffusion behavior is random. Nevertheless, the cross-diffusion behavior of seed nodes is mostly deliberate. Therefore, they might not be that closely connected with other nodes within the social media network. However, their

important function is to perform cross diffusion and promote the spread of new information so that information can quickly transfer from a fixed small circle and spread across a wider range.

However, as opposed to blindly forwarding information, the nodes responsible for cross diffusion between multiple points perform two functions: they filter information within the group and sort filtered data sets to determine the theme set of product information diffusion.

2.1.1. Information filtering in multipoint cross diffusion

There is substantial information spread in social media, and data sets are quite large. However, in cross diffusion the nodes aim to focus on one specific data subset and reduce the amount of data to be processed by removing content that does not interest users. The nodes might make a subsequent analysis of the removed content. The data subset might be the most unique and valuable part of the entire data set, and it might also be the part the seed nodes intention to deliver the most. In all of these cases, it is necessary to use MapReduce's parallel extension capability to traverse all the information and determine the needed parts.

Filtering is a mechanism that extracts data from a subset and provides them for subsequent analysis. It is also a method for focusing on the data in a subset that the seed nodes are most interested in. Filtering serves as preparation for subsequent, more valuable actions.

The mapper task performs an evaluation function on each record it receives. Normally, the types of keys/values output by mapper are the same as the input ones since the record has not been changed. Function f was applied to each record, and a Boolean-type value (true or false) was returned. If the function returned true, the current record was saved; otherwise, it was discarded. Mapper outputs keys and values in turn.

2.1.2. Determination of the top sets in multipoint cross diffusion

A relatively small TopK records set was obtained based on ranking the data sets obtained after filtering, regardless of size.

By ranking the filtered themes, the most valuable TopK theme set, or the one satisfying a particular preparation, was found. A ranking function or comparison function was defined to determine which of the two themes had greater communication value from the perspective of the seed nodes. Then, the model could be used to find the most valuable record in the entire data set through MapReduce.

This model used mapper and reducer at the same time. The mapper task found the TopK of the social groups at one specific point in the local community; then, all independent TopK sets were collected in the reducer for the final TopK operation. Because the largest number of data records that mapper outputs is K, and K is usually small, only one reducer is required to handle the final operation. Figure 1 shows the structure of the TopK model of the themes.

2.2. Bayesian network modeling of the product information diffusion process

2.2.1. Determination of random variables in the product information diffusion process

The selection of variables for use in the Bayesian network modeling of the information diffusion process was performed based on three aspects: network structure, information attributes, and user attributes. The variables were defined as follows:

1) Network structure

Cross diffusion: By breaking the clustering within the community structure of social media that hinders the large-scale diffusion of information (Molaei et al., 2020), cross diffusion enables information diffusion in multiple groups.

2) Information attributes

Theme: Product information purposefully created by enterprises is diffused on social media. Additionally, social media users post messages about specific products to form a theme. The themes for cross diffusion were screened by TopK in MapReduce.

Emotion conveyed by each piece of information: this includes disputes over a specific theme (Fu et al., 2019) or thought, including positive and negative information (Zhu et al., 2020).

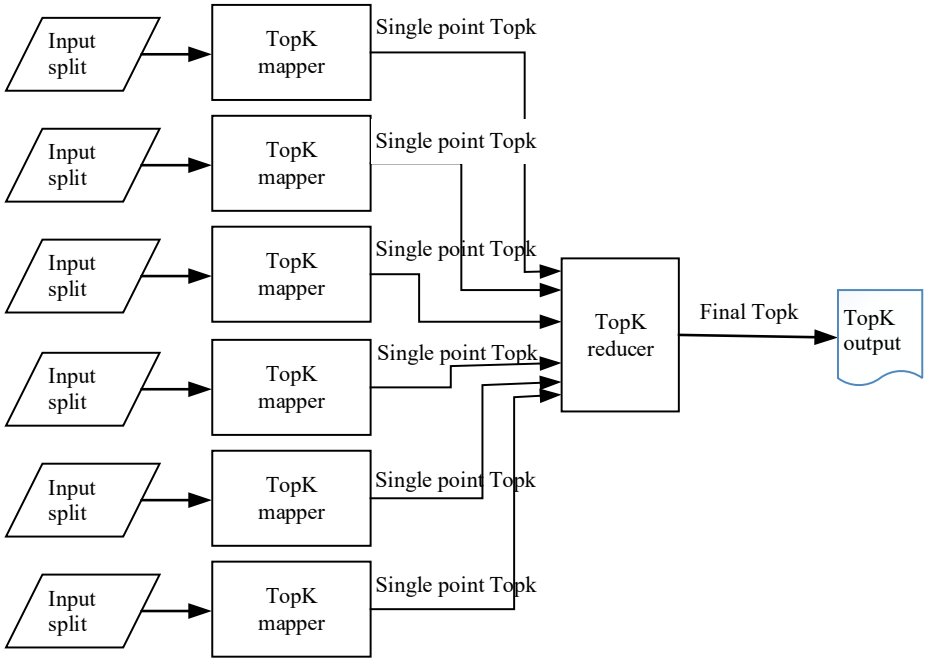

Figure 1. TopK model of the theme.

Origin of information: Social media produce a great deal of real-time content at an incredible rate, and relationships between users can affect users' judgment of information. Meanwhile, the value of the information will in turn affect relationships among users. Therefore, the source of the information is important for information diffusion.

Resonance of users toward the information: The way emotions conveyed by information resonate with users is of great importance for information forwarding.

3) User attributes

User state: This refers to seed nodes, diffusion nodes, and information nodes. Seed nodes have a large number of connections affecting the diffusion of information, and are usually an effective information source in product information diffusion (Li et al., 2019). Their most important function is to inform consumers about new products and trigger large information cascades. Diffusion nodes (Lin and Li, 2021) drive users' forwarding behaviors along with the information cascades triggered by the seed nodes. It is often difficult for information nodes to launch long chain responses, but they have the advantage of quantity and are also the creators of ultimate value.

Connection strength between nodes: This includes strong and weak connections. Weak connections provide a large number of bridges connecting other networks, and abundant weak connections diffuse new information, playing a leading role in information diffusion. Strong connections bind user nodes together through intimate personal relationships. Almost all social relationships occur between intimate friends who are likely to know each other in real life. Strong connections contribute to spreading human behaviors on social media, either online or in real life, and strong connections are more influential among individuals.

User effect: This refers to behaviors or abilities affecting others without obvious compulsory measures or direct orders.

User preference: Whether users are interested in the information they receive.

User activity (Wang et al., 2021): User activity has high- and low-attribute values.

Limited user attention: Whether the information attracts users.

Similarities between users: This refers to the convergence of user attributes in terms of behavior, interest, activity, language, and other aspects. Similar individuals are more likely to be connected than dissimilar individuals.

User behavior: This includes three behaviors: forwarding, purchasing, and becoming information nodes. Forwarding refers to spreading information users receive in their own social media. Purchasing means that users are interested in the product recommended by the information and then purchase it to create value, thus becoming value nodes. Becoming information nodes refers to users seeing information but not responding to it. For the convenience of subsequently analyzing the model features, each variable was represented by a corresponding symbol (Table 1).

2.2.2. Construction of the Bayesian network model of information diffusion

A general way to build a structure is through backward construction. This begins with a variable we are interested in (e.g., user behavior), whose prior probabilities we then try to determine. If the probability is uncertain because it depends on other factors, then other factors will be added as the father nodes of the variable and brought into the network. However, when determining the structure, it should be noted that approximation is difficult to avoid.

For the above variables, we can construct a situation where a variable depends on another variable. There are many weaker effects in addition to the relationships between the above variables. However, if they were all taken into consideration, the network would become very complicated. From the perspective of representation, such networks are difficult to understand and fix, and the parameters are difficult to determine. Moreover, since Bayesian network reasoning is strongly dependent on connection function, the addition of such edges would undoubtedly make the cost of using the network very high. Figure 2 shows the network structure of product information diffusion and its corresponding symbol representation.

Generally speaking, each variable in the model is associated with a conditional probability distribution (CPD). This is used to specify the distribution of the value of this variable under the condition that each joint assignment of its father node is known. For nodes without father nodes, CPD is subject to an empty variable set. Thus, CPD was transformed into a marginal distribution, such as $P(T)$ and $P(N)$. The network structure of information diffusion constituted the Bayesian network $B^{difussion}$ together with CPD.

In $B^{difussion}$, there are two types of special variable nodes, one of which is "S-Strength of Connection." Variable S is referred to as the multiplexer of CPD. In other words, the value of the selected variable is a copy of one of the values of its father nodes.

Packaged CPD variables are another type of special variable node—namely, "N-Near" and "I-Effect," whose values were obtained externally and determined by the values of other variables.

2.3. Characteristics of the Bayesian network of information diffusion

In this study, a Bayesian network was used as the data structure framework to show the diffusion process of product information on social media. A local probability model was combined with this framework to

Table 1. Symbolic representation, type, and value of variables in information diffusion networks.

Name of variable	Symbolic representation	Type	Value
Cross diffusion	D-Cross diffusion	Binary variable	$\{d^0$ to the disadvantage of, d^1 to the advantage of$\}$
Connection	C-Connection	Binary variable	$\{c^0$ hard, c^1 easy$\}$
Theme	T-Themes	Integer variable	$\{t^1, \ldots\ldots, t^k\}$
Emotion conveyed by information	E-Emotion	Binary variable	$\{e^0$ negative, e^1 positive$\}$
Origin of information	O-Origin	Boolean variable	$\{o^0$ no, o^1 yes$\}$
Resonance toward the theme	R-Resonance	Binary variable	$\{r^0$ not friends, r^1 friends$\}$
User state	U-Users	Three-valued variable	$\{u^0$ information node, u^1 diffusion node, u^2 seed node$\}$
Connection strength between nodes	S-Strength of connection	Binary variable	$\{s^0$ weak, s^1 strong$\}$
User effect	I-Effect	Binary variable	$\{i^0$ small, i^1 big$\}$
User preference	F-Preference	Binary variable	$\{f^0$ disinterested, f^1 interested$\}$
User activity	A-Activity	Binary variable	$\{a^0$ low, a^1 high$\}$
Limited user attention	L-Limited attention	Binary variable	$\{l^0$ not attracted, l^1 attracted$\}$
Similarities between users	N-Near	Binary variable	$\{n^0$ low, n^1 high$\}$
User behavior	B-Behavior	Binary variable	$\{b^0$ forward, b^1 purchase, b^2 information node$\}$

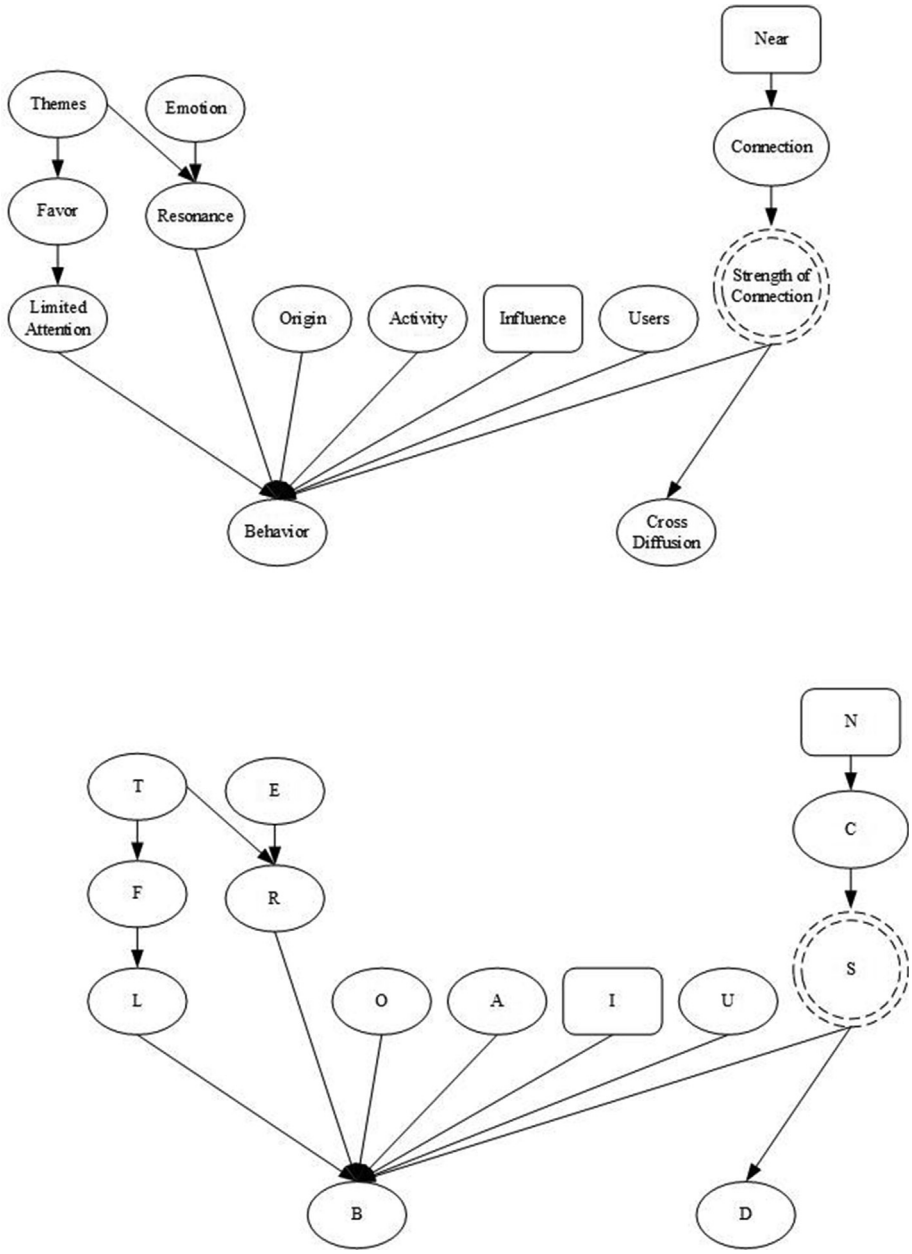

Figure 2. Network structure and symbolic representation $G_{diffusion}$ of product information diffusion.

define joint distribution. Below, a series of characteristics of the Bayesian network of information diffusion are analyzed.

2.3.1. Conditional independence of the model

In $B^{difussion}$, the edges denote direct dependencies. For instance, "whether the user will resonate with the information depends on the content and implied emotions of each message" and "nodes depend on their father nodes" are the semantic core of the Bayesian network. This is the conditional independence hypothesis implied by $B^{difussion}$: when father nodes E and T are given, R is conditionally independent of all nonchild nodes in the network:

$$(R \perp O, U, I | E, T).$$

In other words, once the content and emotion implied by each message are known, whether the user will resonate with the message will not be affected by the information provided by any variable other than the child node. Similarly, "whether it is easy for users to befriend and connect with each other is determined by the similarity between the users." That is, with the father node N given, C is conditionally independent of all other nonchild nodes in the network:

$$(C \perp O, U, I, A | N).$$

To put it another way, under the condition of given N, it is very important to confirm the independence of C and O, U, I, A, which will be used in the follow-up reasoning. That is, once the value of N, the father node of C, is known, the reliability of C will not be affected by any information directly or indirectly related to its father node or other ancestor nodes. However, information about its descendant nodes (e.g., B) can still change our judgment by affecting the reasoning, which comprises the core semantics of the Bayesian network.

2.3.2. Factorization of the model

The Bayesian network of information diffusion is a graph with CPD as the note that defines a joint distribution for the information diffusion

process following Bayesian chain rules. In this section, the $I-map$ feature of $B^{diffusion}$ is analyzed.

Here, $B^{diffusion}$ is taken as an example to prove the mutual transformation characteristic of $I-map$ and factorization.

Proposition 1: Let G be a Bayesian network structure defined on the variable set X and P the joint distribution in the same space. If G is an $I-map$ of P, then P is factorized according to G.

Proof: Assuming that $X_1,...,X_n$ is a variable in X and is a topological order relative to G, the chain rules of the probability are first applied:

$$P(X_1,...,X_n)=\prod_{i=1}^{n}P(X_i|X_1,...,X_n). \tag{1}$$

One of the factors, $P(X_i|X_1,...,X_{i-1})$, is considered. As G is the $I-map$ of P, $(X_i\perp NonDescendants_{x_i}|Pa_{x_i}^G)\subseteq I(P)$. Based on this assumption, all the father nodes of X_i are concentrated in the set $X_1,...,X_{i-1}$. In addition, this set does not contain any descendant node of X_i. Thus,

$$\{X_1,...,X_n\}=Pa_{x_i}\cup Z, \tag{2}$$

where $Z\subseteq NonDescendants_{x_i}$. According to the local independence and decomposition properties of X_i, it can be concluded that $(X_i\perp Z|Pa_{x_i})$. Thus, the following formula is workable:

$$P(X_i|X_1,...,X_{i-1})=P(X_i|Pa_{x_i}). \tag{3}$$

To apply the conversion to $B^{diffusion}$ factorized by the chain rules, the following formula can be obtained:

$$P(T,F,L,E,R,O,U,I,A,N,C,S,B)=P(T)P(F|T)P(L|F)P(E)P(R|T,E)\\P(N)P(C|N)P(S|C)P(B|L,R,O,U,I,A,S). \tag{4}$$

Therefore, the conditional independence assumption contained by $B^{diffusion}$ is that $B^{diffusion}$ is a series of smaller CPDs factorized from P, which is the distribution of an $I-map$. It should be noted that this demonstration is constructive and is a method to factorize construction factors given the conditions of distribution P and graph G.

In general, 2^n-1 independent parameters are needed to specify a joint distribution defined among n binary random variables. If the distribution is factorized according to graph G, while each node of graph G has at most k father nodes, then the number of independent parameters required will be less than $n\cdot 2^k$. In many applications, assumptions can be made about specific zones where the variables mutually affect each other. Although each variable is usually associated with a number of other variables, it often depends only on a small number of other variables. Therefore, in many cases, even when n is very large, k will still be very small. Thus, the number of parameters in the Bayesian network is exponentially smaller than that of joint distribution. This is a major advantage of the Bayesian network.

The conditional independence contained in the Bayesian network and the factorization of distribution factors form the basic relationship between local probability models. Conditional independence implies factorization, and the contrary is true as well—that is, factorization according to G implies relevant conditional independence.

Proposition 2: Let G be a Bayesian network defined on random variable set X, and let P be the joint distribution in the same space. If P is factorized per G, then G is an $I-map$ of P.

Proof: Let P be the probability distribution factorized per $B^{diffusion}$. It should be proven that $I(B^{diffusion})$ works in P. $(L\perp T,R,E|F)$, the independence assumption of L, should be taken into consideration. To prove that it works in P, the following formula should be proven:

$$P(L\perp T,R,E,F)=P(L|F). \tag{5}$$

According to the definition,

$$P(L\perp T,R,E,F)=\frac{P(L,T,R,E,F)}{P(T,R,E,F)}. \tag{6}$$

Based on the chain rules of the Bayesian network, the numerator of the fraction is $P(T)P(F|T)P(L|F)P(E)P(R|T,E)$. By marginalizing the joint distribution, the denominator obtained is

$$\begin{aligned}P(T,R,E,F)&=\sum_L P(T,R,E,F,L)\\&=\sum_L P(T)P(F|T)P(E)P(R|T,E)P(L|F)\\&=P(T)P(F|T)P(E)P(R|T,E)\sum_L P(L|F)\\&=P(T)P(F|T)P(E)P(R|T,E).\end{aligned} \tag{7}$$

The last step was performed because $P(L|F)$ is a distribution defined on L. Thus, the sum is 1, and the following formula can be obtained:

$$\begin{aligned}P(L\perp T,R,E,F)&=\frac{P(L,T,R,E,F)}{P(T,R,E,F)}\\&=\frac{P(T)P(F|T)P(E)P(R|T,E)P(L|F)}{P(T)P(F|T)P(E)P(R|T,E)}=P(L|F).\end{aligned} \tag{8}$$

3. Consumer behavior reasoning in information diffusion

Cooper (1990), for the first time, conducted a formal analysis of the computational complexity of probability reasoning in the Bayesian network. Variants of the variable elimination algorithm were invented independently by multiple teams. An early variant came from the peeling algorithm proposed by Cannings et al. (1976 & 1978), which is a systematic exposition of genetic lineage analysis.

The general problem of probability reasoning in the graph model proposed a local message passing algorithm in the Bayesian network featuring the multiple tree structure. These views provoked the development of various algorithms that are more common, including a series of methods proposed by scholars, such as Schachter (1998) and Dechter (1999), all of these methods end with the variable elimination algorithm.

Following the idea of the multi-tree algorithm, Kim and Pearl (1983) presented a simple approach, namely generating multiple trees with clustering nodes, but the efficiency of the process is low. The sum-product message passing algorithm was developed by Shenoy and Shafer (1990). They described it with a very broad form, which, in addition to the probability graph model, also applies to many factorization models. The sum-product-division method was developed in a series of papers by Lauritzen and Spiegelhalter (1988) and Jensen et al. (1990). Studies in this direction have also produced the theory that takes message passing operations as a re-parameterization implemented on the initial distribution. The sum-product-division algorithm described by Andersen et al. (1989) formed the basis of the Bayesian network system, which then led to the extensive application of this method.

Consumer behavior reasoning used the clique tree data structure to pass information between adjacent cliques. In the Bayesian network of information diffusion described above, consumer behavior reasoning was performed to estimate the posterior probability of user behaviors such as purchasing, forwarding, and staying silent; predict the effect of product information diffusion; and obtain the quantitative relationships between the factors and user behavior.

3.1. Generation of clique tree via sum-product reasoning

Assume

$$P(A,B,C,D)=\varphi_A\varphi_B\varphi_C\varphi_D. \tag{9}$$

The marginal probability distribution of D is

$$P(D)=\sum_C\sum_B\sum_A P(A,B,C,D). \tag{10}$$

It is worth noting that the effect caused by the multiplication and addition of factors is exactly the same as that caused by the multiplication and addition of numbers. Calculating any marginal probability includes calculating the product of all CPDs and summing all variables other than query variables. In general, the task to be completed can be regarded as calculating the value of the following formula:

$$\sum_{z}\prod_{\varphi\in\Phi}\varphi. \tag{11}$$

The limited scope of a factor is the key idea for efficiently calculating this formula; thus, some sum formulas can be "inserted" and executed only on the product of a subset of the factors. The basic idea of the algorithm is to sum one variable at a time. When one variable is summed, all factors associated with it can be multiplied to generate a product factor.

Now, let us return to $B^{difussion}$, the information diffusion network constructed in this study, as shown in Table 2. The Bayesian chain rule of this network was asserted as

$$\begin{aligned} &P(T,F,L,E,R,O,U,I,A,N,C,S,B) = P(T)P(F|T)P(L|F)P(E)P(R|E,T)\\ &\quad P(O)P(U)P(I)P(A)P(N)P(C|N)P(S|C)P(B|L,R,O,U,I,A,S)\\ &= \varphi_T(T)\varphi_F(F,T)\varphi_L(L,F)\varphi_E(E)\varphi_R(R,E,T)\varphi_O(O)\varphi_U(U)\varphi_I(I)\varphi_A(A)\\ &\quad \varphi_N(N)\varphi_C(C,N)\varphi_S(S,C)\varphi_B(B,L,R,O,U,I,A,S). \end{aligned} \tag{12}$$

$P(B)$ was calculated by the above sum-product reasoning method. The following elimination sequence was used:

$$T,F,E,N,C,L,R,O,U,I,A,S.$$

Theorem 1: If $I_{\Phi,\prec}$ is the induction graph of factor set Φ and one specific elimination sequence $\prec$, then the scope of each factor produced in the variable elimination process is a clique in $I_{\Phi,\prec}$. Figure 3 shows the clustering tree educed according to Theorem 1.

In view of the sum-product reasoning process in Table 2, altogether there are 12 factors, $\psi_1,\ldots\ldots\psi_{12}$, whose scopes are shown in Table 2. $\tau_1(F,R,E)$, the effect generated from $\psi(F,T,R,E)$, participated in the calculation of ψ_2. Therefore, there is an edge from C_1 to C_2. Similarly, the factor $\tau_3(L,R)$ was generated according toψ_3 and used to calculate ψ_6. Thus, an edge could be added from C_3 to C_6, with the complete structure shown in Figure 2. The edges in the figure are annotated by direction to indicate the information flow between clusters during the execution of the sum-product algorithm. Each factor in the initial factor set Φ is also correlated with cluster C_i; for example, $\varphi_F(F,T)$ (corresponding to the CPD$P(F|T)$) is correlated with C_1, and $\varphi_B(B,L,R,O,U,I,A,S)$ (corresponding to the conditional probability $P(B|L,R,O,U,I,A,S)$) is correlated with C_6.

By multiplying the existing factors, each step in the sum-product reasoning generated factor ψ_i and eliminated a variable in ψ_i to generate the new factor τ_i; then, τ_i was used to generate another factor. Here, ψ_i was considered as a calculated data structure that carried the "effect" τ_j caused by ψ_j and generated effect τ_i to be used in $\psi_i.\tau_i$. Each intermediate factor in the sum-product reason algorithm could only be used once at most: when φ_i generated ψ_j, it could be removed from factor set Φ and could not be used again. Therefore, the clustering diagram obtained through an implementation of the sum-product reasoning must be a tree. If T is the clustering tree generated by the sum-product reason algorithm in one specific factor set Φ, C_i and C_j are two adjacent clusters, and C_i sends effect τ_i to C_j. Then, the scope of τ_i will exactly be $C_i \cap C_j$.

A digraph induced by effects is a directed tree where all effects flow to the single cluster of a final result, which is referred to as the root of the digraph. Here, the root of the tree is assumed to be at the "top," so all effects sent to the "root" are upward. If C_i is located on the path from C_j to the root, then C_i is at the upstream of or C_j is at the downstream of C_i.

If the clustering tree produced by sum-product reasoning satisfies the running intersection property, then it is a clique tree. For the running intersection property, assume T is a clustering tree on Φ, V_T is the vertex of T, and ε_T is its edge. Whenever there is a variable X that makes $X \in C_i$ and $X \in C_j$, X will also be in every cluster of the only path from C_i to C_j, and T will have the running intersection property.

It is easy to verify that the running intersection property is applicable to the clustering tree shown in Figure 3. For example, B appears in C_6 and C_{12}, so it also appears in the clique on the path between them—that is, C_7, C_8, C_9, C_{10}, and C_{11}. The clustering tree that satisfies the running intersection property is called a clique tree, where a cluster is also referred to as a clique.

3.2. Consumer behavior reasoning in product information diffusion based on a Bayesian network

Behavior reasoning used the clique tree data structure to pass effects between adjacent cliques, send all effects to the cliques that function as the root, and obtain the posterior probability of $P(B)$. The factor set ψ was calculated in the clique tree, and the effects were sent along the edge. Each clique received the incoming effect factors and multiplied them, summed one or more variables, and then sent the effects to another clique.

The variable elimination algorithm in the clique tree was briefly explained. This algorithm was executed by passing the effects in the clique tree. Let T be a clique tree composed of $C_1,\ldots\ldots C_k$ to generate the initial potential energy starting from the multiplication of the factors assigned to each clique. Then, the clique tree data structure was used to

Table 2. One-time execution of the sum-product reasoning task.

Step	Variables eliminated	Factors used	Variables involved	New factors
1	T	F,T,R,E	$\varphi_T(T),\varphi_F(F,T),\varphi_R(R,E,T)$	$\tau_1(F,R,E)$
2	F	L,F,R,E	$\varphi_L(L,F),\tau_1(F,R,E)$	$\tau_2(L,R,E)$
3	E	L,R,E	$\varphi_E(E),\tau_2(L,R,E)$	$\tau_3(L,R)$
4	N	C,N	$\varphi_N(N),\varphi_C(C,N)$	$\tau_4(C)$
5	C	S,C	$\varphi_S(S,C),\tau_4(C)$	$\tau_5(S)$
6	L	B,L,R,O,U,I,A,S	$\varphi_B(B,L,R,O,U,I,A,S),\tau_3(L,R)$	$\tau_6(B,R,O,U,I,A,S)$
7	R	B,R,O,U,I,A,S	$\tau_6(B,R,O,U,I,A,S)$	$\tau_7(B,O,U,I,A,S)$
8	O	B,O,U,I,A,S	$\varphi_O(O),\tau_7(B,O,U,I,A,S)$	$\tau_8(B,U,I,A,S)$
9	U	B,U,I,A,S	$\varphi_U(U),\tau_8(B,U,I,A,S)$	$\tau_9(B,I,A,S)$
10	I	B,I,A,S	$\varphi_I(I),\tau_9(B,I,A,S)$	$\tau_{10}(B,A,S)$
11	A	B,A,S	$\varphi_A(A),\tau_{10}(B,A,S)$	$\tau_{11}(B,S)$
12	S	B,S	$\tau_{11}(B,S),\tau_5(S)$	$\tau_{12}(B)$

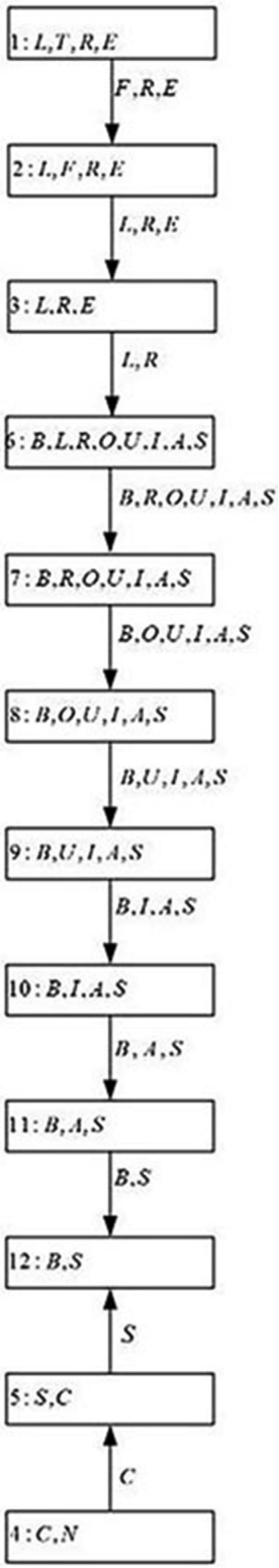

Figure 3. Clustering tree of the sum-product reasoning in Table 2.

pass effects between adjacent cliques and send the effects to the cliques that function as the root.

Each $\varphi \in \Phi$ was assigned to some cliques $a(\varphi)$, and the initial potential energy of C_j was defined as

$$\psi_j(C_j) = \prod_{\varphi:a(\varphi)=j} \varphi. \tag{13}$$

As each factor was assigned to one clique, the formula obtained is

$$\prod_{\varphi} \varphi = \prod_{j} \psi_j. \tag{14}$$

Let C_r be the selected root clique. Variable elimination was implemented in this clique starting from the leaf node of the tree and moving inward. More precisely, for each clique C_i, Nb_i was defined as the index set of the adjacent cliques. Let $p_r(i)$ be the upstream neighbor of i on the path to root clique r. Except for the root clique, each clique C_i performed an effect transfer calculation and sent the effects to its upstream neighbor $C_{P_r(i)}$.

The effect from C_i to C_j can be calculated by the following formula:

$$\xi_{i\to j} = \sum_{C_i - S_{ij}} \psi_i \prod_{k\in(Nb_i-\{j\})} \xi_{k\to i}. \tag{15}$$

It seems that C_i multiplied all the effects coming from its other neighbors with the initial potential energy, which made its scope behind ψ, one factor in this clique. Except for the variables in the cut set between C_i and C_j, it summed all the variables and sent the result as an effect to C_j.

This effect was transferred upward along the tree to the root clique. After the root clique received all the effects, it multiplied these effects with its initial potential energy to gain the result—a factor called belief that was denoted by $B_r(C_r)$. It was represented as follows:

$$\tilde{P}_{\Phi}(C_r) = \sum_{X-C_r} \prod_{\varphi} \varphi. \tag{16}$$

The first step was to generate an initial potential set correlated with different cliques. Initial potential energy $\psi_i(C_i)$ was produced by multiplying the initial potential set with the initial factor designated by clique C_i. For instance,

$$\psi_4(C,N) = \varphi_N(N)\varphi_C(C,N).$$

Cliques containing B (e.g., C_{12}) were selected as the root cliques. Then, the following steps were carried out:

(1) In C_1: T was eliminated by running $\sum_T \psi_1(F,T,R,E)$; the scope of the factor as the result was F,R,E, which was sent to C_2 as $\xi_{1\to2}(F,R,E)$.

(2) In C_2: Define $\beta_2(L,F,R,E) = \xi_{1\to2}(F,R,E)\cdot\psi_2(L,F,R,E)$; the factor on L,R,E was obtained by eliminating F and then sent to C_3 as $\xi_{2\to3}(L,R,E)$.

(3) In C_3: Define $\beta_3(L,R,E) = \xi_{2\to3}(L,R,E)\cdot\psi_3(L,R,E)$; $\xi_{3\to6}(L,R)$, the factor on L,R was obtained by eliminating E.

(4) In C_4: N was eliminated by running $\sum_N \psi_4(C,N)$; the result was taken as the factor $\xi_{4\to5}(C)$ and sent to C_5.

(5) In C_5: Define $\beta_5(S,C) = \xi_{4\to5}(C)\cdot\psi_5(S,C)$; $\xi_{5\to12}(S)$, the factor on S was obtained by eliminating C.

(6) In C_6: Define $\beta_6(B,L,R,O,U,I,A,S) = \xi_{3\to6}(L,R)\cdot\psi_6(B,L,R,O,U,I,A,S)$; $\xi_{6\to7}(B,R,O,U,I,A,S)$, the factor on B,R,O,U,I,A,S was obtained by eliminating L.

(7) In C_7: R was eliminated by running $\sum_R \psi_7(B,R,O,U,I,A,S)$; the result was taken as the factor $\xi_{7\to8}(B,O,U,I,A,S)$ and sent to C_8.

(8) In C_8: O was eliminated by running $\sum_O \psi_8(B,O,U,I,A,S)$; the result was taken as the factor $\xi_{8\to9}(B,U,I,A,S)$ and sent to C_9.

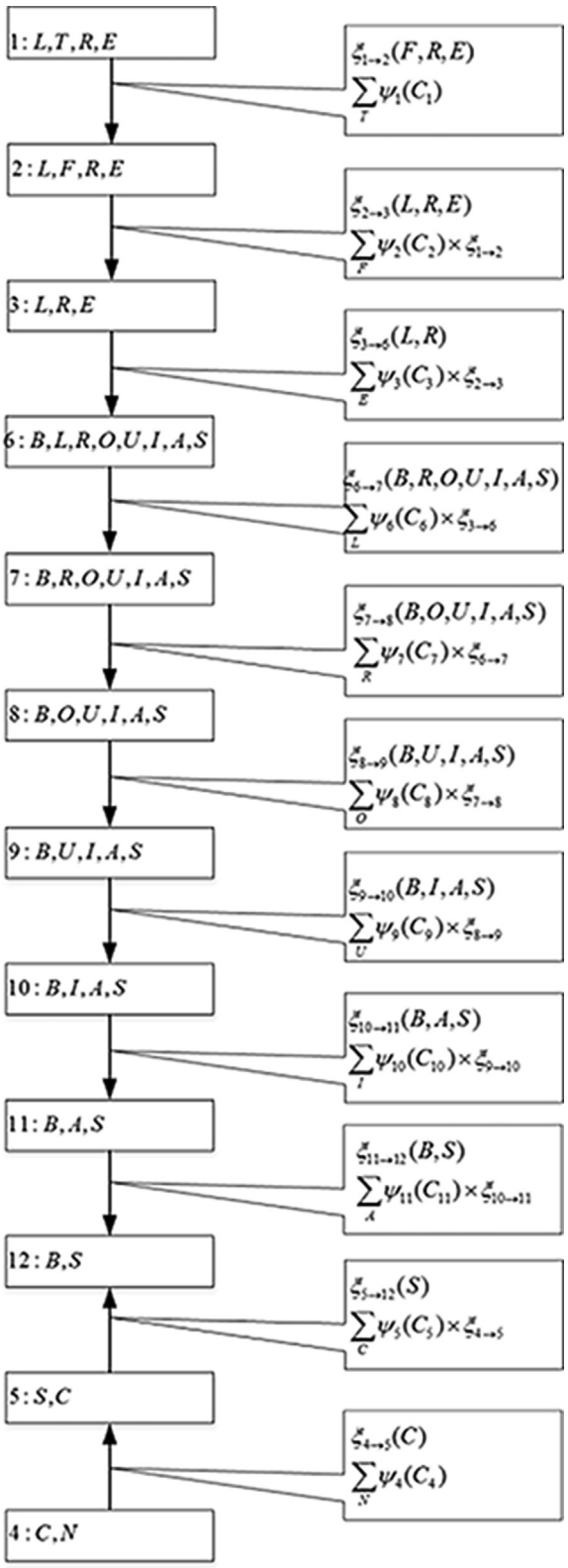

Figure 4. C_{12}, the behavior reasoning of the root clique in the information diffusion clique tree.

(9) In C_9: U was eliminated by running $\sum_U \psi_9(B,U,I,A,S)$; the result was taken as the factor $\xi_{9\to10}(B,I,A,S)$ and sent to C_{10}.

(10) In C_{10}: I was eliminated by running $\sum_I \psi_{10}(B,I,A,S)$; the result was taken as the factor $\xi_{10\to11}(B,A,S)$ and sent to C_{11}.

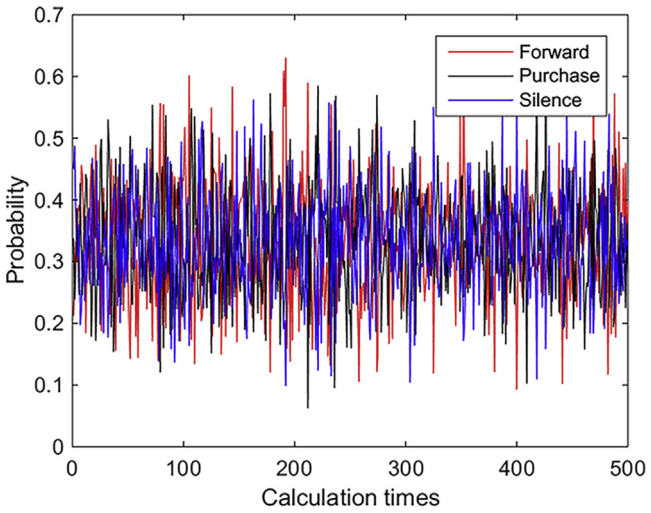

Figure 5. Marginal posterior probability of $P(B)$.

(11) In C_{11}: A was eliminated by running $\sum_A \psi_{11}(B,A,S)$; the result was taken as the factor $\xi_{11\to12}(B,S)$ and sent to C_{12}.

(12) In C_{12}: Define $\beta_{12}(B,S) = \tau_{11\to12}(B,S)\tau_{5\to12}(S)$.

β_{12} was a factor defined on B,S to code the joint distribution $P(B,S)$. With all the CPDs multiplied and all the other variables eliminated, the calculation of $P(B)$ now only requires the sum total of S.

It is worth noting that the clique should collect all incoming effects from its downstream neighbors before sending any effect to its upstream neighbors. If a clique has completed the collection of all the incoming effects, then the clique is ready. Thus, when the algorithm just started, C_1, C_4 was ready, and any calculation related to it could be conducted at any time during the running process. However, C_2 was not ready until it received the incoming effect from C_1. So, for the clique tree whose root is in C_{12}, $C_1C_2C_3C_6C_7C_8C_9C_{10}C_{11}C_4C_5C_{12}$ is the legal execution order. Figure 4 shows the set of effects transferred by the algorithm throughout the executing process.

At this point, the behavior reasoning process has been completed in the $G_{difussion}$ information diffusion process with the influences of many factors; that is, the marginal posterior probability of consumer behavior choice has been calculated—namely, the posterior probability that users will choose purchasing, forwarding, or staying silent.

4. Experiment and discussion

Assume the probability of all influencing factor nodes is randomly generated. Then, the marginal posteriori probability of consumer behavior choice was calculated according to the behavior reasoning algorithm (Figure 5). With the algorithm run 500 times, the mean and variance of user behavior probability of forwarding, purchasing, and becoming information nodes were as follows: $mean(b^0) = 0.3312$, $mean(b^1) = 0.3366$, $mean(b^2) = 0.3322$; $std(b^0) = 0.0271$, $std(b^1) = 0.0281$, $std(b^2) = 0.0275$.

Since the prior probabilities of the influencing factors are random, the result indicating that the posterior probability of consumer behavior choice fluctuates near the mean is reasonable.

5. Conclusion

Based on product information diffusion on social media self-organized by consumers, this study explored consumers' behavior choices in the product information diffusion process. The main contributions include the following.

1. Multiattribute Analysis of Product Information Diffusion. Among the factors affecting the product information diffusion process, those with important value were extracted, and the causal relationship between them was analyzed. Reasonable ranges were assigned to the binary, three-valued, and Boolean variables of those factors to more accurately analyze the complexity of the product information diffusion process.
2. Bayesian Model of Product Information Diffusion. A Bayesian network was used to represent the information diffusion process under the effects of multiple factors. It showed how those factors influenced each other as well as user behavior choice. It also explained the process and structure produced by the product information diffusion network, thus expressing the product information diffusion network in a more accurate form.
3. Consumer Behavior Reasoning in Product Information Diffusion. To obtain the quantitative relationship between factors and user behavior, a quantitative method was used to estimate the posterior probability of user behavior choice in the information diffusion process affected by multiple factors.

Declarations

Author contribution statement

Xuehua Sun: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Shaojie Hou, Ning Cai, Wenxiu Ma: Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by Department of Education of Hebei Province [grant number QN2018259].

Data availability statement

The authors are unable or have chosen not to specify which data has been used.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Andersen, S., Olesen, K., Jensen, F., 1989. HUGIN-a shell for building Bayesian belief universes for expert systems. In: Proc. 11th International Joint Conference on Artificial Intelligence. IJCAI), pp. 1080–1085.

Cannings, C., Thompson, E.A., Skolnick, H.H., 1976. The recursive derivation of likelihoods on complex pedigrees. Adv. Appl. Probab. 8 (4), 622–625.

Cannings, C., Thompson, E.A., Skolnick, M.H., 1978. Probability functions on complex pedigrees. Adv. Appl. Probab. 10 (1), 26–61.

Cheng, J.J., Liu, Y., Shen, B., et al., 2013. An epidemic model of rumor diffusion in online social networks. Eur. Phys. J. B 86, 29.

Cooper, G., 1990. Probabilistic inference using belief networks is NP-hard. Artif. Intell. 42, 393–405.

Daley, D.J., Kendal, D.G., 1964. Epidemic and rumors. Nature 1118.

Daley, D.J., Kendal, D.G., 1965. Stochastic rumors. IMA J. Appl. Math. 42–55.

Dechter, R., 1999. Bucket elimination: a unifying framework for inference. Artif. Intell. 113 (1-2), 41–85.

Denning, P.J., 1985. The science of computing: super networks. Am. Sci. 73 (3), 225–227.

Dybiec, B., 2009. SIR model of epidemic spread with accumulated exposure. Eur. Phys. J. B 67, 377–383.

Fang, Q., Yue, K., Fu, X.D., Wu, H., Liu, W.Y., 2013. A MapReduce-Based Method for Learning Bayesian Network from Massive Data. APWeb, pp. 697–708.

Fu, G.Y., Chen, F., Liu, J.G., Han, J.T., 2019. Analysis of competitive information diffusion in a group-based population over social networks. Physica A 525, 409–419.

He, X., Song, G., Chen, W., Jiang, Q., 2012. Influence blocking maximization in social networks under the competitive linear threshold model. In: SDM, pp. 463–474.

Jensen, F.V., Olesen, K.G., Andersen, S.K., 1990. An algebra of Bayesian belief universes for knowledge-based systems. Networks 20 (5), 637–659.

Kempe, D., Kleinberg, J., Tardos, E., 2003. Maximizing the spread of influence through a social network. In: KDD2003:Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp. 137–146.

Kim, J., Pearl, J., 1983. A computational model for combined causal and diagnostic reasoning in inference systems. In: Proc. 8th International Joint Conference on Artificial Intelligence (IJCAI), pp. 190–193.

Lauritzen, S.L., Spiegelhalter, D.J., 1988. Local computations with probabilities on graphical structures and their application to expert systems. J. Roy. Stat. Soc. B 50 (2), 157–224.

Li, Y., Chen, W., Wang, Y., Zhang, Z.L., 2013. Influence diffusion dynamics and influence maximization in social networks with friend and for relationships. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM2013. ACM, New York, NY, USA, pp. 657–666.

Li, D., Wang, W., Jin, C.L., et al., 2019. User recommendation for promoting information diffusion in social networks. Physica A 534, 121536.

Lin, L.F., Li, Y.M., 2021. An efficient approach to identify social disseminators for timely information diffusion. Inf. Sci. 544, 78–96.

Molaei, S., Zare, H., Veisi, H., 2020. Deep learning approach on information diffusion in heterogeneous networks. Knowl. Base Syst. 189, 105153.

Naumov, P., Tao, J., 2017. Marketing impact on diffusion in social networks. J. Appl. Logic 20, 49–74.

Schachter, R.D., 1998. Bayes-ball: the rational pastime. In: Proc. 14th Conference on Uncertainty in Artificial Intelligence. UAI), pp. 480–487.

Shenoy, P., Shafer, G., 1990. Axioms for probability and belief-function propagation. In: Proc. 6th Conference on Uncertainty in Artificial Intelligence. UAI), pp. 169–198.

Tian, R.Y., Zhang, X.F., Liu, Y.J., 2015. SSIC model :A multi-layer model for intervention of online rumors spreading. Physica A 42, 7181–7191.

Wang, Y.Q., Yang, X.Y., Han, Y.L., Wang, X.A., 2013. Rumor spreading model with trust mechanism in complex social networks. Commun. Theor. Phys. 59, 510–516.

Wang, Y.N., Wang, J., Wang, H.Y., Zhang, R.L., Li, M., 2021. Users' mobility enhances information diffusion in online social networks. Inf. Sci. 546, 329–348.

Xia, L.L., Jiang, G.P., Song, B., Song, Y.R., 2015. Rumor spreading model considering hesitating mechanism in complex social networks. Physica A 437, 295–303.

Zhu, X., Kim, Y., Park, H., 2020. Do messages spread widely also diffuse fast? Examining the effects of message characteristics on information diffusion. Comput. Hum. Behav. 103, 37–47.