TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/12 pp161–179
DOI: 10.26599/TST.2018.9010126
Volume 25, Number 2, April 2020

Propagation History Ranking in Social Networks: A Causality-Based Approach

Zheng Wang, Chaokun Wang*, Xiaojun Ye, Jisheng Pei, and Bin Li

Abstract: Information diffusion is one of the most important issues in social network analysis. Unlike most existing works, which either rely on network topology or node profiles, this study focuses on the diffusion itself, i.e., the recorded propagation histories. These histories are the evidence of diffusion and can be used to explain to users what happened in their networks. However, these histories can quickly grow in size and complexity, limiting their capacity to be intuitively understood. To reduce this information overload, in this paper we present the problem of propagation history ranking. The goal is to rank participant edges/nodes by their contribution to the diffusion. We first discuss and adapt a causal measure, Difference of Causal Effects (DCE), as the ranking criterion. Then, to avoid the complex calculation of DCE, we propose two integrated ranking strategies by adopting two indicators. One is responsibility, which captures the necessity aspect of causal effects. We further give an approximate algorithm, which could guarantee a feasible solution, for this indicator. The other is capability, which captures the sufficiency aspect of causal effects. Finally, promising experimental results are presented to verify the feasibility of the proposed ranking strategies.

Key words: propagation history ranking; causality; social networks; information diffusion

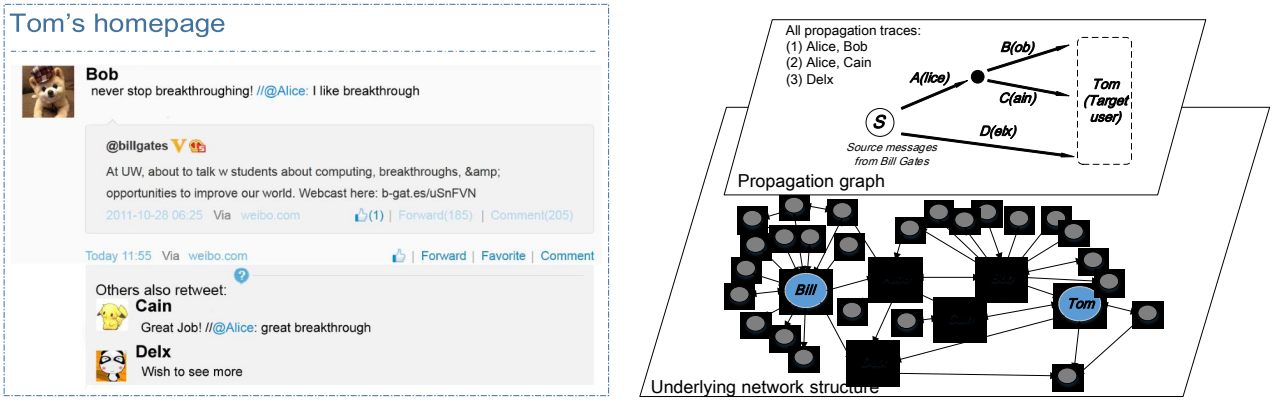
1 Introduction

In the digital age, online social networks have become a key communication platform for millions of internet users. Every day, huge number of posts (tweets, messages) are emerging from and disseminating among online Social Network Sites (SNS)[1]. Usually, users are eager to make sense of the information propagation process around them. For example, a user may receive the same news from several different followees or friends. At this time, the user might want to know why he/she received this news, or what the role was of each user involved in the propagation. Fortunately, propagation histories, which record the diffusion process, are partially provided to users in some online SNS, such as Sina Weibo (the Chinese counterpart of Twitter).

Example 1 (Propagation history of a diffusion on Weibo). Bill Gates posts a message about his speech on Sina Weibo, and Tom receives this news via Bob (one of his followees). The trace of the repost processes, via Alice/Bob, is recorded and illustrated on Tom's homepage (Fig. 1a). In addition, the other two propagation traces are included, i.e., via Alice/Cain and Delx respectively.

People may then be curious to know something "hidden" rather than the superficial diffusion process of the news. For example, does Bob play a key role in propagating the public speech of Bill Gates to Tom? What is Alice's contribution to Tom's reception of the

• Zheng Wang is with the Department of Computer Science and Technology, University of Science and Technology Beijing, China. E-mail: wangzheng@ustb.edu.cn.
• Chaokun Wang, Xiaojun Ye, and Jisheng Pei are with the School of Software, Tsinghua University, Beijing 100084, China. E-mail: chaokun@tsinghua.edu.cn; yexj@tsinghua.edu.cn; pjs07@mails.tsinghua.edu.cn.
• Bin Li is with China Information Technology Security Evaluation Center, Beijing 100085, China. E-mail: lib@itsec.gov.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2018-05-28; revised: 2018-09-07; accepted: 2018-09-28



(a) Screenshot of Tom's homepage on Sina Weibo

(b) Illustration of the propagation history and underlying network structure

Fig. 1 An illustration of the propagation history in Example 1. (a) Sina Weibo reminds Tom of the three ways to receive this message (via Alice and Bob, via Alice and Cain, and via Delx); (b) The propagation graph constructed by the propagation history (top), where every edge stands for a user's retweet behavior; the underlying network structure (bottom), where every node stands for a user and some users are tagged for convenience.

news?

Note that only propagation histories, rather than the underlying network topology, are used to explain information diffusion in this study. An illustration of Example 1 is given at the top of Fig. 1b, where the propagation history is represented as a set of propagation traces (i.e., ordered edge lists). We can see that these traces exactly capture users' repost behaviors in the dynamic flow of information, regardless of the underlying network topology. With these recorded traces, we can intuitively explain this diffusion; e.g., Alice tends to be an important person since there are two propagation traces going through this user.

However, propagation histories can rapidly grow in size and complexity, making them difficult to understand intuitively. To reduce this information overload, in this paper, we present the problem of *propagation history ranking*. The goal is to rank propagation participants (edges/nodes) by their contribution to the diffusion. As such, we put forward our solution from the viewpoint of *causality*[2]. Causality-based reasoning plays a vital role in various human decision making activities (such as legal/business decision making) and has long been a hot topic in philosophy, AI, and areas of cognitive science research. The kind of method we propose here can draw a clearer picture of each participant's contribution to the information propagation process.

Example 2 (Ranking the propagation history in Example 1). For convenience, as illustrated in the top of Fig. 1b, we describe the repost behaviors in a diffusion as "directed edges". The propagation history of this diffusion is then $\{(A, B), (A, C), (D)\}$, where (A, B) represents a propagation trace from A to B, and similarly for (A, C) and (D).

To evaluate the contribution of each of the relevant edges to the diffusion, we show a causal measure: Difference of Causal Effects (DCE) (we explain in Section 3 how these scores are computed). For comparison, we also show a non-causal measure: *"out-degree"*. Table 1 lists the results. As an example, we compare edge B to edge D. Although edge B has a higher out-degree score, edge B makes a limited contribution to the diffusion compared to edge D. Looking at the corresponding propagation history (from the top of Fig. 1b), we can explain this in two ways. Firstly, B lies in a branch path (A, B) of the diffusion, while D lies in a trunk path, where a branch path is one with a high fan out source node and a trunk path is one with a low fan out source node. Secondly, D itself can directly guarantee the diffusion, while B alone cannot. These interpretations are captured in the causal measure DCE generated from the propagation history. In addition, the propagation history size can become very large in practice, and it is critical to reduce this information overload.

Table 1 Different kinds of measures for the diffusion in Example 1.

Edge	Causal measure (DCE)	Non-causal measure (Out-degree)
A	3/8	5
B	1/8	12
C	1/8	2
D	5/8	2

Applications. Propagation history ranking aims to determine the most important or influential edges/nodes on diffusion paths. This approach has a variety of applications; without loss of generality, we discuss two of particular interest.

In online SNS, rumor can disseminate on an unsurpassed scale and can cause great damage[3,4]. The task of identifying the party that is primarily responsible (such as a rumor-monger or key mediator) for the damage has attracted considerable interest recently. Most works[5,6] identify these parties based on network properties, such as the network structure and user profiles. Nevertheless, as most online SNS record the reposting behavior, i.e., propagation histories, of all the users involved, legal judgements based on this evidence are more convincing.

Another interesting application is *network reliability*[7,8]: Given a network where each edge/node has a probability of failure, we are asked to identify the edges/nodes that are necessary to guarantee connectivity. However, in practice, the entire network structure is not always available, e.g., for security reasons[9]. The distribution records of gateways[10,11], which are another kind of propagation history, might be examined as a priority in making diagnoses.

Challenges. Generally, the ground truth of many ranking problems is given by human judgement, such as the relevance score between a query and a web page in information retrieval[12]. However, this would be impractical as a solution to the propagation history ranking problem, due to the large size and complex dependencies of propagation histories.

When human judgement is not reliable, an alternative solution is to use ranking criteria. However, the selection of criteria has not been investigated to date. Specifically, certain characteristics of propagation histories need to be considered to make this selection. Once the ranking criterion has been chosen, the next challenge is how to calculate it efficiently; a naive idea is to make the calculation directly according to its definition, which is not always efficient. Therefore, for practical use, we need to consider the trade-off between accuracy and speed.

Contributions. To the best of our knowledge, this paper is the first to study the propagation history ranking problem in SNS. Our main contributions in this paper are summarized as follows:

(1) We present the propagation history ranking problem in SNS. We further discuss and adapt DCE as the ranking criterion.

(2) We propose two integrated ranking strategies (to avoid the hardness of DCE calculation) by adopting two indicators, *responsibility* and *capability*, to capture the necessity and sufficiency aspects of causal effects, respectively.

(3) We give an approximate algorithm for responsibility calculation in propagation histories, since this problem is generally NP-hard.

(4) We conduct extensive experiments on real-world datasets. These experiments not only illustrate the rationality of the selected ranking criterion DCE but also demonstrate the feasibility of our ranking strategy.

The rest of the paper is organized as follows. We present the propagation history ranking problem in SNS in Section 2. In Section 3, we discuss and adapt DCE as the ranking criterion. We propose two integrated ranking strategies by adopting two indicators (responsibility and capability) in Section 4. In Section 5, we propose an approximate algorithm to calculate responsibility. Experimental results are provided in Section 6. We review the related work in Section 7. Conclusion and outline for future work are given in Section 8.

2 Preliminaries and Problem Statement

In this paper, we restrict our discussion to the information diffusion from one source node to one target node. Without loss of generality, we consider edges as propagation participants. Herein we give a formal definition of propagation history and its associated ranking problem.

Definition 1 (Propagation history of a diffusion). The *propagation history* of a diffusion records all the actual propagation trails of an event ε from the source s to the target τ, and is usually formalized as

$$\Phi_{(\varepsilon,s,\tau)}=\{t^1,\ldots,t^n\}=\{(t_1^1,\ldots,t_{l_1}^1),\ldots,(t_1^n,\ldots,t_{l_n}^n)\},$$

where each t^i is called a *trace*, constructed by l_i ordered edges $(t_1^i,\ldots,t_{l_i}^i)$. We drop the subscript (ε,s,τ) when there is no ambiguity caused by doing so.

Propagation history ranking. Let Φ be the propagation history of a diffusion, and let T be the edge set of Φ. The goal is to rank those edges in T by their contribution to this diffusion.

3 DCE as Ranking Criterion

In this section, we first discuss how to estimate the importance of a particular edge in a propagation history.

We then introduce the selected ranking criterion and finally discuss its rationality and calculation.

Edge importance estimation. The estimation of the importance of a specific edge to the overall diffusion mainly concentrates on two effects: the overall effect if the edge failed, and the overall effect if the edge is non-failed. Considering the likelihood of a successful transmission, these two effects can be estimated as

• $P(\Phi_{x'}=\text{true})$ is the probability of a successful information transmission if edge X is intervened to be failed.

• $P(\Phi_x=\text{true})$ is the probability of a successful information transmission if edge X is intervened to be non-failed.

Note that, subscripts x and x' stand for interventions to set edge X to "non-failed" and "failed", respectively, in randomized experiments[13]. Therefore, $P(\Phi_{x'}=\text{true})$ and $P(\Phi_x=\text{true})$ are two causal effect measures considering how necessary and sufficient the edge is for the diffusion, respectively. Naturally, we can consider the above two effects together, inspired by Ref. [14], as follows.

Definition 2 (DCE). Suppose the propagation history of a diffusion is Φ, and X is a participant edge. The *difference of causal effects* made by edge X in this diffusion is defined as the expression:

$$\text{DCE}(X) = P(\Phi_x=\text{true}) - P(\Phi_{x'}=\text{true}) \quad (1)$$

Intuitively, DCE considers both the sufficiency and necessity of edge X for this diffusion, and this property is formally stated as follows.

Proposition 1 DCE measures both the necessity and sufficiency aspects of causal effects.

We theoretically prove this property in Appendix A. A similar idea is widely adopted in other research areas, such as network reliability[7] and economics[15].

Example 3 (Example 1 continued). The propagation history of the diffusion is $\Phi=\{t^1, t^2, t^3\}$, where $t^1=(A, B)$, $t^2=(A, C)$, and $t^3=(D)$. In this paper, we suppose every participant edge in Φ has the same failure probability of 50%. The $P(\Phi_{x'}=\text{true})$, $P(\Phi_x=\text{true})$, and DCE values of these edges are shown in Table 2 (in the later of this section, we explain how these values are obtained).

Table 2 DCE values of all participant edges in Example 1.

Edge	$P(\Phi_{x'}=\text{true})$	$P(\Phi_x=\text{true})$	DCE
A	1/2	7/8	3/8
B	5/8	3/4	1/8
C	5/8	3/4	1/8
D	3/8	1	5/8

For the comparison of $P(\Phi_{x'}=\text{true})$ values, let us consider edge A and edge B. Intuitively, edge A lies in a trunk path, while edge B lies in a branch path. If edge B wants to enable the diffusion, edge A is indispensable, i.e., A is a necessary condition for B to enable the diffusion. However, B is not a necessary condition for A, because of the existence of edge C. Therefore, the removal of edge A will have a greater negative impact on information transmission than the removal of B (the difference in their $P(\Phi_{x'}=\text{true})$ values is 0.125).

For the comparison of $P(\Phi_x=\text{true})$ values, let us consider edge A and edge D. From the sufficiency aspect of causal effects, edge D makes a greater contribution than edge A (the difference in their $P(\Phi_x=\text{true})$ values is 0.125). Intuitively, this is because D by itself can guarantee the diffusion, while A alone cannot.

Because DCE considers causal effects from two aspects, higher DCE values equate to more important edges. We compare edge A with edge D as an example. Compared to edge A, edge D has a higher DCE value for two reasons. On one hand, the occurrence of D has a higher probability of guaranteeing a successful information transmission than the occurrence of A (the difference in their $P(\Phi_x=\text{true})$ values is 0.125). On the other hand, the absence of D has a higher probability of leading to an unsuccessful transmission than the absence of A (the difference in their $P(\Phi_{x'}=\text{true})$ values is 0.125).

Rationality of DCE. Choosing DCE as the ranking criterion has several advantages. Firstly, DCE adopts the concept of causality rather than probability. Probabilistic measures arise from subjective observations, reflecting what we know or believe about the world, whereas causal measures describe objective physical constraints in the world, revealing more stable relationships[14]. Due to this stability, people prefer to encode knowledge in causal rather than probabilistic structures. In addition, causality is inherently encoded in propagation histories, which record the entire diffusion process[16].

Secondly, DCE is a fine-grained measure which assesses the likelihood of causal effects. This is important because, on the one hand, social networks are large-scale and have complex connection structures; on the other hand, information propagation traces can be seen as random walks in these complex networks. Therefore, the involved participants usually show up in

several significantly diverse propagation traces and tend to make different contribution to the diffusion. Thus, a fine-grained ranking measure is needed to prioritize the significance of these participants.

Thirdly, DCE considers causal effects from two aspects: necessity and sufficiency. These two aspects can be easily observed in propagation histories in SNS. Intuitively, edges in short propagation traces are more likely to be sufficient causes than edges in long traces, while, edges in trunk paths are more likely to be necessary causes than edges in branch paths. Consequently, considering causal effects in both of these two aspects not only matches up with common sense but is also highly suitable for this application.

DCE calculation. According to the definition of DCE, the most direct way to calculate this causal measure is to perform randomized experiments[17]. We briefly describe this procedure as follows. Given the input edge X, we first enforce X to have "non-failed" (or "failed"). Then, in each simulation, we randomly set the other participant edges to be failed or non-failed, and further check if the diffusion would be successful under this new condition. After a great number of simulations, we arrive at a convergent value of $P(\Phi_x = \text{true})$ (or $P(\Phi_{x'} = \text{true})$). Finally, we obtain the DCE value according to Eq. (1).

4 Proposed Ranking Strategies

Although conducting randomized experiments is the preferred method for DCE calculation, this kind of approach must run simulations many times over in order to obtain a convergent value[13, 18]. To avoid this complication, we introduce two ranking indicators (i.e., *responsibility* and *capability*) to capture the intuitive two aspects of DCE. Finally, by considering these two indicators together, we propose two integrated ranking strategies.

4.1 Responsibility

In this subsection, we first introduce the concept of *responsibility*. We then theoretically prove that responsibility can be used to capture the necessity aspect of an edge's causal effects in a diffusion. Finally, we analyze its time complexity.

The concept of responsibility[19] is proposed based on the following definition inspired by Ref. [14].

Definition 3 (Causality in a diffusion). Suppose the edge set of the propagation history is T. Let $t \in T$ be a participant edge, and let $\Gamma \subset T$ be an edge set. t is to be called a cause for the diffusion w.r.t. Γ, if the following two conditions are satisfied:

(1) The diffusion remains successful with $T-\Gamma$, and

(2) After removing Γ, the subsequent removal of t would make the diffusion fail.

Γ is called the contingency set for t.

Although checking causality (i.e., identifying each cause and its related contingency set) is NP-complete in general[20], Ref. [21] gives a PTIME solution for the provenance data[22, 23] of relational databases. This method could be directly applied to the propagation history in SNS. In this paper, we do not discuss this causality checking problem, but only focus on the related ranking problem.

Definition 4 (Responsibility). Suppose the edge set of the propagation history is T, and let $t \in T$ be a participant edge. The *responsibility* of t for this diffusion is

$$\delta_t = \frac{1}{1 + \min_\Gamma |\Gamma|},$$

where Γ ranges over all contingency sets for t.

Example 4. We continue with the propagation history in Example 1. The responsibility of edge A is 1/2, because the smallest contingency set for A is $\{D\}$. Similarly, the responsibility of D is 1/2 with $\{A\}$ being the smallest contingency set. The responsibility of edge B is 1/3, since its smallest contingency set is $\{C, D\}$. Edge C's responsibility is also 1/3 with $\{B, D\}$ as the smallest contingency set.

The responsibility of edge t is determined by the minimum edge set whose removal would make t indispensable for a successful information transmission; this leads to the following proposition.

Proposition 2 Responsibility measures the necessity aspect of causal effects.

Proof The proof is based on a causal measure *Probability of Necessary* (PN), which is defined as the probability that event y would not have occurred in the absence of event x, given that x and y did in fact occur[14]. Therefore, PN measures the necessity aspect of causal effects.

Let X be a participant edge, and let x and x' stand for the propositions "X non-failed" and "X failed", respectively. Let set S contain all propagation traces which go through edge X, with the rest of the traces put into another set (denoted as $\bar{S}$). Let s and $\bar{s}$ stand for the cases in which S and $\bar{S}$ can successfully transmit the information respectively, and let s' and $\bar{s}'$ denote their complements. We could calculate the PN value

of edge X for this diffusion as PN $= P(\bar{s}'s)/[P(\bar{s}x)+P(\bar{s}'s)]$ (more details about this derivation can be found in Appendix B). Suppose every edge has the same failure probability of 50%. Then we get the following equation:

$$\begin{aligned}\text{PN} &= 1/[P(\bar{s}x)/P(\bar{s}'s)+1] = \\ &\quad 1/[P(\bar{s})P(x)/P(\bar{s}'s)+1] = \\ &\quad 1/[0.5\times P(\bar{s})/P(\bar{s}'s)+1] \propto P(\bar{s}'s)/P(\bar{s}).\end{aligned}$$

As the responsibility of X increases, $\bar{S}$ becomes more likely to break ($P(\bar{s}')$ increases and $P(\bar{s})$ decreases). In this case, if $P(s)$ remains the same, PN will increase. Therefore, responsibility has a positive relationship with PN, i.e., responsibility measures the necessity aspect of causal effects. ■

Complexity of responsibility. In theory, to compute the responsibility one must iterate over all contingency sets, i.e., computing responsibility in general is NP-hard[19]. Therefore, we propose an approximate algorithm for general propagation histories, more details of which can be found in Section 5.

4.2 Capability

In this subsection, we first define the concept of *capability*. We then prove that capability can be used to capture the sufficiency aspect of an edge's causal effects in a diffusion. Finally, we analyze its time complexity.

Definition 5 (Capability). Suppose the edge set of the propagation history is T, and let $t \in T$ be a participant edge. The *capability* of t for this diffusion is

$$\rho_t = \frac{1}{\min_{Ł} |\text{st}(Ł)|},$$

where $Ł$ ranges over all propagation traces going through t, and function st($Ł$) returns the edge set of $Ł$.

Example 5. We continue with the propagation history in Example 1. The capability of edge A is 1/2, since it needs edge $\{B\}$ or $\{C\}$ to ensure the diffusion. The capability values of B and C are both 1/2, because they both need $\{A\}$ for successful information transmission. Edge D's capability is 1 because D itself can guarantee the diffusion.

The capability of edge t is determined by the minimum edge set whose addition would make t indispensable for a successful information transmission; this leads to the following proposition.

Proposition 3 Capability measures the sufficiency aspect of causal effects.

Proof The proof is based on a causal measure *Probability of Sufficiency* (PS). As stated in Ref. [14], PS is defined as the probability that enabling x would produce y in a situation where x and y are in fact absent. Therefore, PS measures the sufficiency aspect of causal effects.

Continuing with the same definitions of X, x, x', S, $\bar{S}$, s, $\bar{s}$, s', and $\bar{s}'$ as in the proof of Proposition 2, we can calculate the PS value of edge X for this diffusion as PS $= P(\bar{s}'(s|x)x')/P(\bar{s}'x')$ (more details about this derivation can be found in Appendix C). First of all, since $\bar{S}$ consists of the traces which do not contain edge X, $P(\bar{s})$ and $P(\bar{s}')$ are not affected by X. As ρ_x (the capability of X) increases, $P(s|x)$ increases, i.e., it becomes easier for S to ensure the diffusion. Since $P(\bar{s}')$ and $P(x')$ are not affected by ρ_x, PS will increase when ρ_x increases. Therefore, capability has a positive relationship with PS, i.e., capability measures the sufficiency aspect of causal effects. ■

Complexity of capability. Suppose the propagation history contains N traces and M edges. Using an inverted index, calculating the capability values of all edges can be done in $O(L\times N)$ (with $O(M)$ space complexity), where L is the average length of all propagation traces. Generally speaking, L is a small number according to the concept of six degrees of separation[24]. Therefore, the capability problem has a linear complexity with respect to the number of propagation traces.

4.3 An integrated ranking strategy

We have introduced two ranking indicators, responsibility and capability, to capture the intuitive notion of two aspects to DCE. In this subsection, we show that by combining these two indicators directly, we can arrive at a simple integrated "responsibility-capability" ranking strategy (short for "resp-cap") as follows:

$$\text{score}=\alpha\times\text{fn}(\text{responsibility})+(1-\alpha)\times\text{fn}(\text{capability}) \quad (2)$$

where fn stands for a normalized function calculating the *standard score*[25] (Specifically, $\text{fn}(x) = \frac{x-\mu}{\sigma}$, where μ is the mean of the population and σ is the standard deviation of the population) and $0<\alpha<1$ is a balance factor.

Example 6 (Example of the "resp-cap" ranking). Continuing with the propagation history in Example 1, the results of our method together with responsibility, capability, "resp-cap", and DCE values are listed in Table 3 (note that, to facilitate understanding, we do not use the normalized function in the "resp-

Table 3 "resp-cap" ranking (Eq. (2)) and DCE values in Example 1.

Edge	Responsibility	Capability	"resp-cap"	DCE
A	1/2	1/2	1/2	3/8
B	1/3	1/2	5/12	1/8
C	1/3	1/2	5/12	1/8
D	1/2	1	3/4	5/8

cap" method in this example, and we set $\alpha = 0.5$). As shown, "resp-cap" method gives the sorted result (D, A, B, C). The two highest ranked edges are D and A. Compared to A, edge D is ranked higher because of its higher capability value. The next two edges are B and C, which are less important because of their lower degrees of responsibility and capability. In this simple example, the "resp-cap" method successfully captures the intuition of DCE by considering two aspects of causal effects.

4.4 A light integrated ranking strategy

The proposed "resp-cap" method needs to calculate responsibility and capability values for every participant. Therefore, this method cannot scale to large datasets, since for each participant, even the approximate responsibility calculation algorithm still needs $O(k\times(L\times N+M))$ time (Section 5.2).

In this subsection, we present a light version of "resp-cap" (denoted as "resp-cap*"). The idea is that, since we typically only care about the top K influential participants in a diffusion, we do not need to evaluate all of the participants if we can pre-select some likely candidates. Specifically, we first get the capability values of all participants, and then we select the $r\times K$ highest ranked participants as candidates (r is a small pre-defined constant). We are then required to calculate the responsibility values for only these $r\times K$ candidates. Finally, we adopt Eq. (2) to get the ranking scores of these $r\times K$ candidates so as to identify the top K most influential participants.

Complexity of "resp-cap*" The capability problem has a linear complexity solution (Section 4.2), and we only need to call the approximate responsibility calculation method (a linear time complexity method shown in Section 5.2) $r \times K$ times. Consequently, "resp-cap*" has a linear time complexity, and can scale to large datasets.

Note that these two indicators can be incorporated into other more complex ranking methods, which we leave for future work.

5 An Approximate Algorithm for Responsibility Calculation

Since responsibility is difficult to calculate in general, in this section we propose an approximate algorithm which guarantees a feasible solution. The basic idea behind our method is that we can reduce the responsibility problem to a variant of the classical Set Cover Problem (SCP)[26]. Let us first compare these two problems.

5.1 Responsibility vs. SCP

Suppose the propagation history has been processed into a collection of sets: $\Phi=\{c_1,\ldots,c_n\}$, where c_i is the edge set of the original propagation trace t^i. In other words, each c_i is a subset of the whole participant edge set $T=\{t_1,\ldots,t_m\}$. We assume function $\mathrm{sc}(t_j,\Phi)=\{c_i|t_j\in c_i\wedge c_i\in\Phi\}$ ($\mathrm{sc}(t_j)$ for short), and intuitively $\mathrm{sc}(t_j)$ covers (contains) all the sets (in Φ) containing t_j. Given an edge t, the intuition of SCP is to find the minimum k which satisfies $\mathrm{sc}(t)\cup\{\mathrm{sc}(t_1)\cup\cdots\cup\mathrm{sc}(t_k)\}=\Phi$. Note here we are actually discussing a constrained SCP problem by ensuring the solution of SCP to contain the input edge t. The responsibility problem has the same intuition, and also contains another constraint in that the removal of $\{t_1,\ldots,t_k\}$ must ensure t is still a cause for this diffusion, i.e., $\mathrm{sc}(t_1)\cup\cdots\cup\mathrm{sc}(t_k)\neq\Phi$.

Example 7 (The difference between responsibility and SCP). Continuing with Example 1, the equivalent SCP form of the propagation history is $\Phi=\{c_1,c_2,c_3\}=\{\{A,B\},\{A,C\},\{D\}\}$. We can also get $\mathrm{sc}(A)=\{c_1,c_2\}$, $\mathrm{sc}(B)=\{c_1\}$, $\mathrm{sc}(C)=\{c_2\}$, and $\mathrm{sc}(D)=\{c_3\}$. Obviously, the solutions of this SCP problem are $\{\mathrm{sc}(A),\mathrm{sc}(D)\}$ and $\{\mathrm{sc}(B),\mathrm{sc}(C),\mathrm{sc}(D)\}$.

Taking edge B as an example, for an SCP problem that has a solution containing B, $\mathrm{sc}(B)\cup\{\mathrm{sc}(A),\mathrm{sc}(D)\}$ and $\mathrm{sc}(B)\cup\{\mathrm{sc}(C),\mathrm{sc}(D)\}$ are two answers. Now, let us consider the corresponding responsibility problem. The result $\{C,D\}$ is a contingency set for B, whereas $\{A,D\}$ is not. This is because after removing $\{A,D\}$, the further removal of B will not make this diffusion fail, i.e., $\mathrm{sc}(A)\cup\mathrm{sc}(D)=\Phi$.

5.2 Approximate algorithm for responsibility

If Γ is the selected contingency set for edge t, Γ must satisfy two constraints: (a) after removing all edges in Γ, the diffusion remains but the removal of t would make the diffusion fail; and (b) Γ must be the minimum

set satisfying $\bigcup_{x\in\Gamma} \mathrm{sc}(x)=\Phi-\mathrm{sc}(t)$. Based on these two constraints, we propose a greedy algorithm named *Appresp*; the main aspects of this algorithm are as follows:

(1) Get the covered set $\mathrm{SA}=\mathrm{sc}(t,\Phi)$ and the uncovered set $\mathrm{ST}=\Phi-\mathrm{SA}$.

(2) Choose edge $x \in T-\Gamma$ which satisfies these two rules: (a) $\mathrm{sc}(x,\mathrm{ST})$ contains as many as possible of the sets in ST, and (b) $\mathrm{SA} \neq \mathrm{sc}(x,\mathrm{SA})$.

(3) Add x to Γ, remove $\mathrm{sc}(x,\mathrm{SA})$ from SA, and remove $\mathrm{sc}(x,\mathrm{ST})$ from ST.

(4) Repeat Steps (2) and (3) until ST gets empty, and compute the responsibility of t.

The key aspects of our algorithm are the two rules listed in Step 2. The first rule is a heuristic for ranking the edges to be added to the contingency set. The second rule is a constraint to ensure that t is still a cause of the diffusion (in accordance with Definition 3) after removing the calculated contingency set. This selection guarantees a feasible solution, provided that the propagation history contains no *redundancy* (Redundancy is defined by Ref. [21], i.e., a propagation trace t^i is redundant if there exists another trace t^j whose edge set is a subset of t^i's edge set. After removing all redundancies, the remaining edges are causes (Definition 3). This is also the PTIME solution for the causality checking problem in propagation histories). We formalize this property in Proposition 4 and present the pseudo code in Algorithm 1.

Proposition 4 Appresp guarantees a feasible solution, provided that the propagation history contains no redundancy.

Proof We continue with the definitions of Φ, T, Γ, sc, ST, and SA from Algorithm 1. Suppose the propagation history does not contain any redundant traces. We calculate the responsibility of edge t as an example.

Case A (If Appresp returns a contingency set Γ). Suppose we get the initial covered set $\mathrm{SA}=\mathrm{sc}(t,\Phi)$ and the uncovered set $\mathrm{ST}=\Phi-\mathrm{SA}$. In this case, ST is covered by $\bigcup_{x\in\Gamma} \mathrm{sc}(x,\Phi)$. For simplicity, here "the removal of edge t" refers to removing all sets in $\mathrm{sc}(t,\Phi)$ from both SA and ST. In accordance with our constraint rule, the removal of Γ makes ST empty but cannot render SA empty. In addition, if we first remove all edges from Γ, the further removal of t renders SA empty. Consequently, according to Definition 3, Γ is a feasible contingency set for t.

Case B (If Appresp cannot find a contingency set). Suppose we have gotten temporary results Γ', SA', and ST', when no edge satisfies our constraint rule, i.e., for each left edge $x \in T-\Gamma'$, we get $\mathrm{SA}' = \mathrm{sc}(x,\mathrm{SA}')$. Suppose c_t is the edge set of trace t^t in ST' and c_a is the edge set of trace t^a in SA'. For each edge x in c_t, we will find $\mathrm{sc}(x,\mathrm{SA}')$ contains c_a. Thus, we get $c_t \subseteq c_a$, i.e., t^a is redundant. This is opposite to our hypothesis of non-redundancy.

Therefore, Appresp guarantees a feasible solution for a propagation history without redundancy. ■

Algorithm 1 Appresp

Input: The equivalent SCP form of the propagation history $\Phi = \{c_1,\ldots,c_n\}$;
The total participant edge set $T = \{t_1,\ldots,t_m\}$;
The input edge $t \in T$.
Output: The approximate responsibility of t.

```
1  SA = sc(t, Φ);
2  ST = Φ − SA;
3  Γ ← {};
4  T = T − t;
5  while ST is not EMPTY do
6      find edge x ∈ T, which enables sc(x, ST) to cover the
       sets in ST as many as possible and SA ≠ sc(x, SA);
7      Γ = Γ ∪ x;
8      T = T − x;
9      ST = ST − sc(x, ST);
10     SA = SA − sc(x, SA);
11 end
12 return 1/(|Γ| + 1);
```

Complexity of Appresp. Suppose the propagation history has N traces and M edges, the average length of traces is L, and the corresponding contingency set size is k. On average, the time complexity of Appresp is $O(k\times(L\times N+M))$. Note that, both k and L are usually small numbers, i.e., Appresp has linear time complexity (We verify the small values of k in our experiments, and L is also small according to the concept of six degrees of separation[24]).

Proof Given the input edge t, to calculate its responsibility we need to loop the following steps k times.

(1) Line 6 of Algorithm 1 performs two tasks. Firstly, it calculates $\mathrm{sc}(x,\mathrm{ST})$ for each edge x; we can use an inverted index to speed up this calculation (the time complexity is $O(L\times N)$). It then selects edge x, which satisfies the two rules (the time complexity is at most $O(M)$, since SA is usually small).

(2) Line 9 removes all sets in $\mathrm{sc}(x,\mathrm{ST})$ from ST. The time complexity is $O(N/k)$, since on average we need

to repeat this removal k times to empty ST.

(3) Line 10 does a similar task on SA to that performed on ST in Line 9. Therefore, the time complexity is also $O(N/k)$.

Therefore, the overall time complexity is $O(k \times (L \times N + M + 2 \times N/k)) = O(k \times (L \times N + M))$. ■

Example 8 (Example of Algorithm 1). Continuing with the same propagation history Φ in Example 1, we use edge B as an example and employ Algorithm 1 to compute its responsibility. We first get the covered set SA=sc(B, Φ)=$\{c_1\}$ and the uncovered set ST=$\{c_2, c_3\}$. Then, we find that A cannot be selected because sc$(A, \text{SA}) = \text{SA}$. We can select either of C or D, since both satisfy our constraint rule and have the same priority according to our heuristic rule. Suppose we choose C first and then choose D in the next round. Finally, we get the contingency set $\{C, D\}$, so the responsibility of B is $1/3$.

Unlike the approach of our Appresp method, Qin et al.[27] directly adopted a greedy strategy to solve the corresponding SCP problem, i.e., they ignore the constraint rule in our method. However, without this rule, the greedy strategy cannot guarantee a feasible solution for the responsibility problem (as shown in Example 7). We will show this property and compare the two methods in presenting the results of our experiments.

6 Experimental Evaluation

In this section, we aim to show two major contributions of this paper by reporting experimental results. First, we reiterate the rationality of the selected ranking criterion DCE using several real-world datasets. Second, we show the effectiveness of the proposed two ranking strategies ("resp-cap" and "resp-cap*") by answering the following three questions. Q1: Can the two indicators (i.e., responsibility and capability) partly capture the intuition of causal effects? Q2: Can the integrated ranking strategy "resp-cap" capture two aspects of causal effects and thus improve performance? Q3: Is the light version of the integrated ranking strategy "resp-cap*" effective and efficient?

Dataset. We use the real-world SNAP ego-Facebook dataset[28], containing 4039 nodes and 88 234 undirected links. From this network, we generate three propagation history datasets, and explain the corresponding diffusion phenomena by ranking participant edges. We first remove redundant parts of these propagation history datasets, resulting in one small and two large datasets. Table 4 shows the details of these three trimmed datasets.

FB-Sample is a small propagation history dataset generated as follows: (1) we sample ego-Facebook with the Re-Weighted Random Walk strategy (an unbiased sampling method)[29, 30] and obtain a small-scale network; and (2) with this sampled network, we enumerate all simple paths from source (node 0) to target (node 197) to produce a propagation history dataset.

FB-Walk-0-197 (start node 0 and target node 197) and FB-Walk-158-146 (start node 158 and target node 146) are two large propagation history datasets with different start and target nodes. These two datasets are generated as follows: (1) we start many random walks from the source node in the raw network; and (2) we record a random walk path (as a propagation trace) if it reaches the target node in a certain limited number of steps.

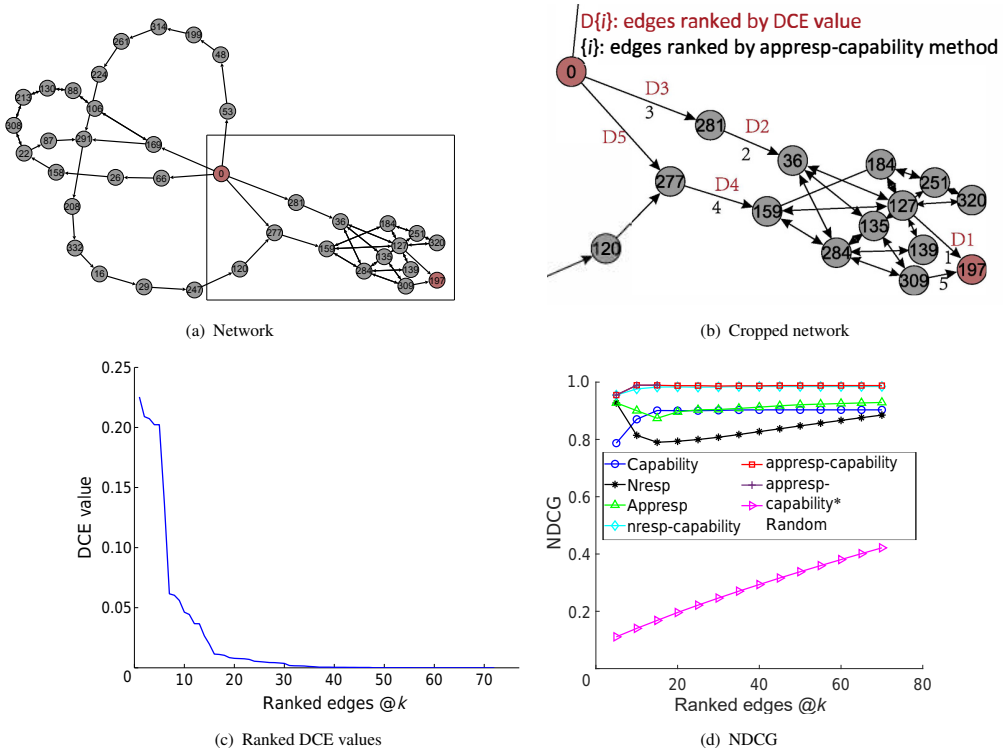
To obtain the ground truth of the ranking, we run randomized experiments to get DCE values on four servers (with 8 cores and 32 GB memory) for 100 – 400 hours for each dataset (To get convergent values, we need to conduct randomized experiments with many repetitions (around 10^9). Moreover, larger propagation history datasets require even more repetitions.)

6.1 Rationality of DCE

In this subsection, we use these three propagation history datasets to illustrate the rationality of DCE. The network constructed by FB-Sample is shown in Fig. 2a. We can intuitively identify some critical edges in this network. However, this does not work when the data size grows. We can take FB-Walk-0-197 and FB-Walk-158-146 as an example; as shown in Figs. 3a and 4a, the networks constructed by these two datasets are large and complex, which makes it difficult to understand them directly. In practice, therefore, the ground truth of the propagation history ranking problem in SNS cannot be manually established.

Table 4 Propagation histories generated from ego-Facebook.

	Source →Target	Number of edges	Number of traces	Len (trace)
FB-Sample	$0 \to 197$	72	442	$4-21$
FB-Walk-0-197	$0 \to 197$	1052	349	$1-18$
FB-Walk-158-146	$158 \to 146$	1024	213	$2-18$



(a) Network (b) Cropped network

(c) Ranked DCE values (d) NDCG

Fig. 2 Causality-based ranking of FB-Sample. The source node is 0, and the target node is 197. A directed edge stands for a directed message propagation between two nodes. (a) Network constructed by the traces in FB-Sample; (b) Cropped version of the network; (c) Ranked DCE values of all participant edges; and (d) Normalized Discount Cumulative Gain (NDCG).

Here we continue to explain the intuition of DCE on FB-Sample. We show the ranked DCE values of FB-Sample in Fig. 2c and provide more details in Table 5. We can see that DCE estimates the causal effects of edge X by computing $P(\Phi_x) - P(\Phi_{x'})$ (the decline in the probability of information being successfully transmitted if the edge is removed).

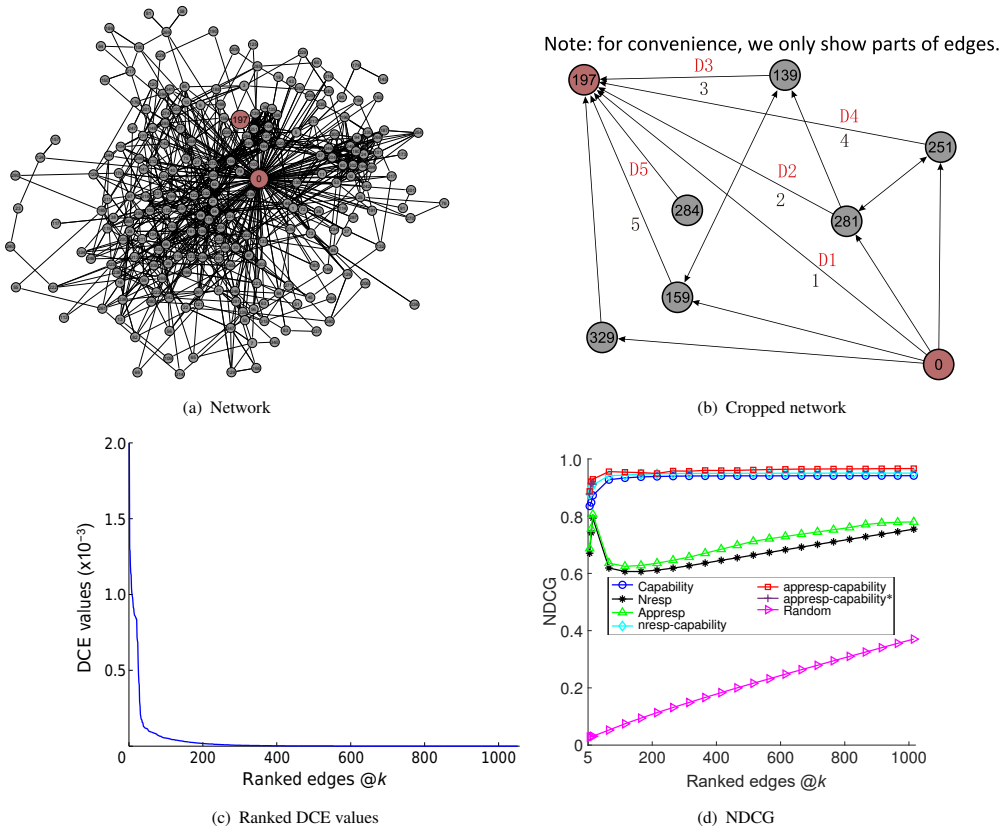
We compare edges 127→197 and 309→197 as an example. Edge 127→197 makes the greatest contribution to this transmission, because its removal would cause the largest decline of information transmission probability (from 0.3441 to 0.1194). Comparatively speaking, edge 309→197 is less important (ranked 6th). Looking at the corresponding network (Fig. 2b), we can explain this from the perspectives of both necessity and sufficiency. From the necessity perspective, these two edges can construct a min-cut edge set, so their nonexistence would cause the transmission to fail. From the sufficiency perspective, edge 127→197 is more important because it lies in the shortest path from source to target, whereas 309→197 does not, i.e., edge 127→197's occurrence equals a higher probability for a successful transmission.

Since the results of the other networks show the same trend, we do not give further accounts of the rationality of DCE on FB-Walk-0-197 and FB-Walk-158-146. However, we show the networks constructed by these histories and their distributions of DCE values in Figs. 3 and 4, from which we can also validate the rationality of DCE.

To summarize, DCE is a fine-grained measure which is not only able to evaluate causal effects but also to consider the effects from the viewpoints of both necessity and sufficiency. Therefore, we select DCE as the ranking criterion for propagation history ranking in SNS.

6.2 Ranking quality

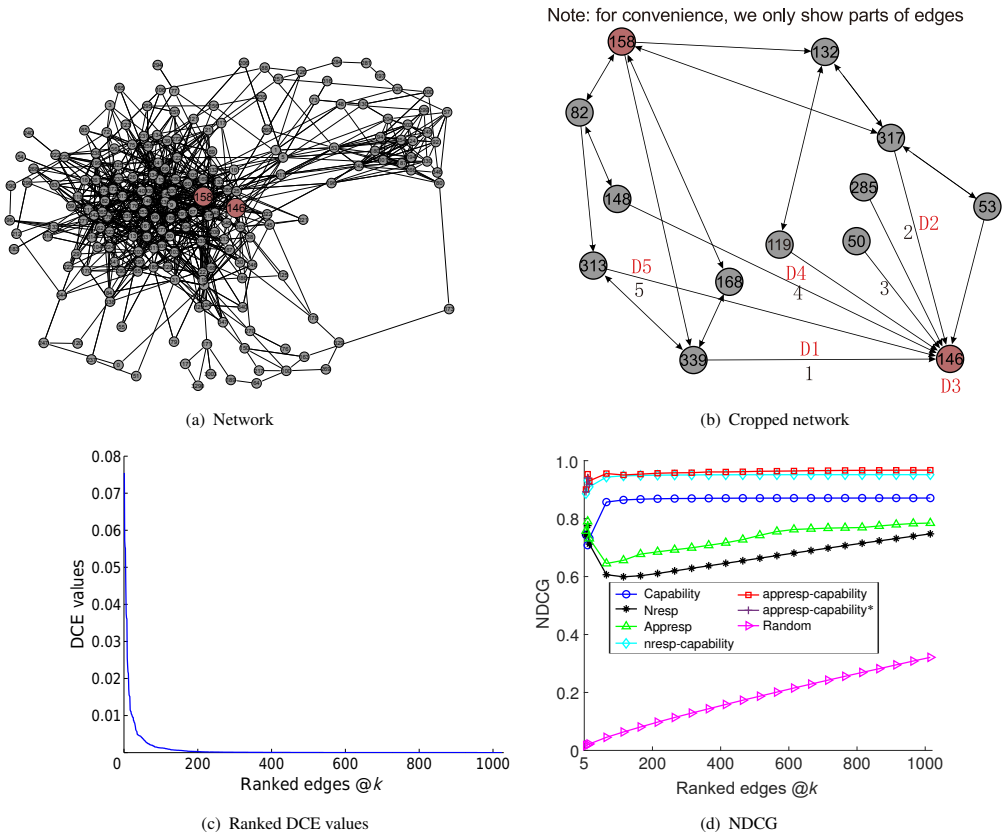
Evaluation metric. Since there is no existing standard evaluation metric for this problem, we first recall the ranking evaluations used in information retrieval, where evaluation metrics can be classified into two categories. The first category is designed for situations of binary relevant notions, such as *Mean Average Precision*[31]. In this situation, each ranking item is simply labeled as relevant or non-relevant in relation to the input queries. The second category is designed for situations characterized by non-binary notions of relevance, such as NDCG[32].



(a) Network (b) Cropped network

(c) Ranked DCE values (d) NDCG

Fig. 3 Causality-based ranking of FB-Walk-0-197. (a) Network constructed by the traces in FB-Sample; (b) Cropped version of the network; (c) Ranked DCE values of all participant edges; and (d) NDCG.



(a) Network (b) Cropped network

(c) Ranked DCE values (d) NDCG

Fig. 4 Causality-based ranking of FB-Walk-158-146. (a) Network constructed by the traces in FB-Sample; (b) Cropped version of the network; (c) Ranked DCE values of all participant edges; and (d) NDCG.

Table 5 Causal effects and DCE values of some edges in FB-Sample.

No.	Edge	$P(\Phi_x$=true)	$P(\Phi_{x'}$=true)	DCE
1	127→197	0.3441	0.1194	0.2247
2	281→36	0.3363	0.1275	0.2088
3	0→281	0.3360	0.1277	0.2083
4	277→159	0.3337	0.1300	0.2037
5	0→277	0.3331	0.1305	0.2027
6	309→197	0.2999	0.1634	0.1365
⋮				
10	135→309	0.2548	0.2086	0.0462
⋮				
15	184→127	0.2414	0.2223	0.0192

As discussed in Section 3, we adopt DCE (a non-binary measure) values as ground truth, so we use NDCG as the evaluation metric in our experiments. NDCG is a popular evaluation metric following two rules: (1) highly related edges are more useful than marginally relevant ones; and (2) lower ranked edges are less valuable for users, since they are less likely to be examined. The NDCG value of a ranking list at a particular rank position n is defined as

$$\mathrm{NDCG}_n = Z_n \left(\mathrm{rel}_1 + \sum_{i=1}^{n} \frac{\mathrm{rel}_i}{\log_2 i} \right),$$

where rel_i is the graded rating of the i-th edge in the ranking list, and Z_n is a normalization constant to make the perfect list obtain NDCG score of 1. Note that DCE values are used as the graded ratings ($\{\mathrm{rel}_i\}$) in all experiments.

Ranking strategies. With the proposed two indicators, we compare the following seven different ranking strategies.

(1) Capability: Ranking by capability value (Section 4.2);

(2) Appresp: Ranking by responsibility value calculated by Appresp (Section 5);

(3) Nresp: Ranking by responsibility value calculated by Nresp (Qin et al.'s method[27]);

(4) appresp-capability: The integrated "resp-cap" ranking (Section 4.3) with responsibility calculated by Appresp;

(5) nresp-capability: The integrated "resp-cap" ranking (Section 4.3) with responsibility calculated by Nresp;

(6) appresp-capability*: The light version of "resp-cap" ranking (Section 4.4) with responsibility calculated by Appresp; and

(7) Random: Ranking randomly.

For all three integrated ranking methods, we set the parameter $\alpha = 0.5$. For appresp-capability*, we set $K = 15$ and $r = 3$, i.e., we only want to identify the top 15 important participants in a diffusion, and we pre-select $3 \times 15 = 45$ likely candidates. In order to evaluate the ranking quality of each method, three steps were followed: (1) we first get its ranking result, (2) we generate 1000 permutations of this result by shuffling edges with the same ranking score, and (3) we use the mean NDCG of these permutations as this method's performance.

(1) Comparing quality. We evaluate the seven ranking strategies. Figures 2d, 3d, and 4d show the NDCG results at different ranking positions. Table 6 shows the details. In addition, we illustrate the top-5 ranked edges obtained by DCE values and our appresp-capability method in Figs. 2b, 3b, and 4b. Table 7 shows the details of these top-5 ranked edges.

Our first observation is that our integrated ranking methods (i.e., appresp-capability and nresp-capability) successfully capture the intuition of DCE. For instance, as shown in the cropped versions of networks (Figs. 2b, 3b, and 4b), our appresp-capability method successfully identifies the most important edges in all three propagation history datasets. The nresp-capability method can do the same (although we do not show its result in the paper, we can validate its success by its NDCG results). Therefore, our integrated methods

Table 6 NDCG results.

Method	FB-Sample			FB-Walk-0-197			FB-Walk-158-146		
	NDCG5	NDCG10	NDCG15	NDCG5	NDCG10	NDCG15	NDCG5	NDCG10	NDCG15
Capability	0.7870	0.8703	0.9009	0.8354	0.8476	0.8726	0.7442	0.7083	0.7319
Nresp	0.9274	0.8143	0.7900	0.6706	0.7417	0.7933	0.7435	0.7758	0.7122
Appresp	0.9277	0.9004	0.8739	0.6872	0.7579	0.8061	0.7575	0.7903	0.7311
nresp-capability	0.9548	0.9762	0.9827	0.8769	0.9019	0.9100	0.8900	0.9369	0.9146
appresp-capability	**0.9549**	**0.9892**	0.9889	**0.8871**	**0.9213**	**0.9291**	**0.9004**	**0.9534**	**0.9303**
appresp-capability*	0.9548	0.9890	**0.9898**	0.8777	0.9153	0.9122	0.8926	0.9404	0.9178
Random	0.1110	0.1405	0.1682	0.0277	0.0299	0.0310	0.0165	0.0194	0.0223

Table 7 Top 5 ranked edges in three propagation history datasets obtained by DCE (ground truth) and our appresp-capability method. (Here † means its ranking score is equal to the preceding one.)

Top-id	FB-Sample		FB-Walk-0-197		FB-Walk-158-146	
	DCE	appresp-capability	DCE	appresp-capability	DCE	appresp-capability
1	127→197	127→197	0→197	0→197	339→146	339→146
2	281→36	281→36†	281→197	281→197	317→146	317→146†
3	0→281	0→281†	139→197	139→197†	50→146	50→146
4	277→159	277→159†	251→197	251→197†	148→146	148→146†
5	0→277†	309→197	284→197	159→197†	313→146	313→146†

achieve a high ranking accuracy. Taking NDCG@5 as an example, the integrated methods achieve a ranking accuracy of 90%−95%. In addition, their ranking accuracies are very stable even when the rank position increases.

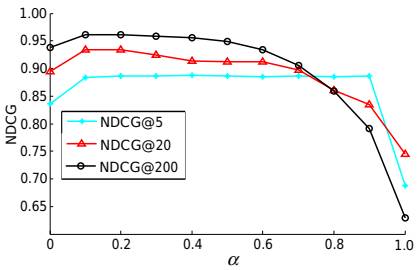
Our second observation is that our integrated ranking methods significantly outperform unintegrated ones (i.e., Capability, Appresp, and Nresp). Again taking NDCG@5 as an example, integrated methods outperform unintegrated ones by 10%−25%. This improvement persists even when the rank position increases. These first two observations demonstrate that our integrated strategy can capture two aspects of causal effects and thus improve performance by combining these two indicators. This summary answers Q2 posed above.

Our third observation is that appresp-capability*, the light version of appresp-capability, is very effective. Although, the performance of appresp-capability* is not the highest, this method still performs much better than unintegrated ones (i.e., Capability, Appresp, and Nresp). Furthermore, appresp-capability* performs better than the integrated method nresp-capability in all three datasets. These experimental results show the effectiveness of appresp-capability* and thus answer Q3 posed above (the efficiency of this method is validated in the scalability experiment).

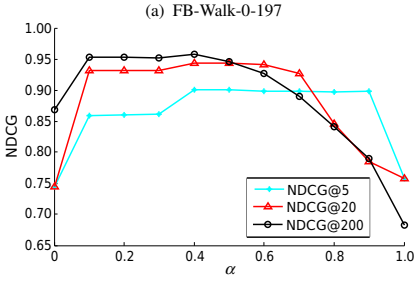
The final observation is that ranking either by capability or by responsibility (Appresp or Nresp) alone can achieve passable accuracy. Again taking NDCG@5 as an example, the ranking accuracies of unintegrated methods are around 75%. This is consistent with our theoretical analysis that responsibility and capability can evaluate causal contribution in two different aspects. Consequently, these two indicators together can partly capture the intuition of DCE, thus answering Q1 posed above.

(2) Effect of balance factor α. In our integrated "resp-cap" ranking strategy, there is a parameter α that controls the balance of consideration of causation between necessity and sufficiency. We test different values of α in our appresp-capability method on both of the two large propagation history datasets. Figure 5 shows the results. We can see that (1) combining appresp and capability values does increase ranking performance; and (2) although the results fluctuate, performance is stable and preferable with α set to around 0.5. These observations suggest that we should consider the necessity and sufficiency of causation equally, which matches our common sense and daily experience.

(3) Appresp vs. Nresp. From Figs. 2d, 3d, and 4d, we can see that our Appresp method outperforms Nresp in both the unintegrated and integrated strategies. We can explain this from Figs. 6 and 7, which show the size distributions of the contingency sets calculated by these two methods. The results of Nresp are highly centralized, which is caused by its greedy strategy. With this strategy, the results are highly influenced by those

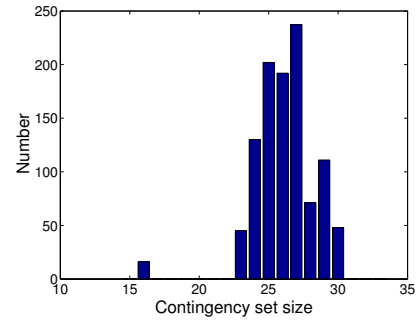


(a) FB-Walk-0-197

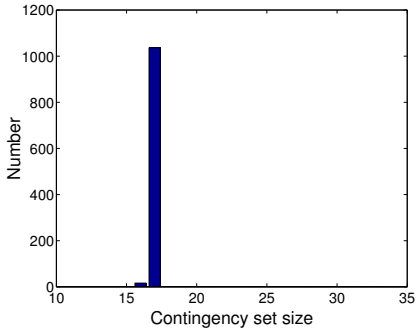


(b) FB-Walk-158-146

Fig. 5 Effect of changing α in our appresp-capability method.

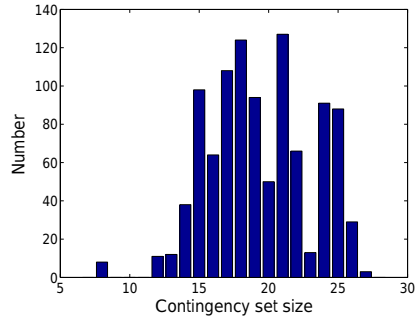


(a) Appresp

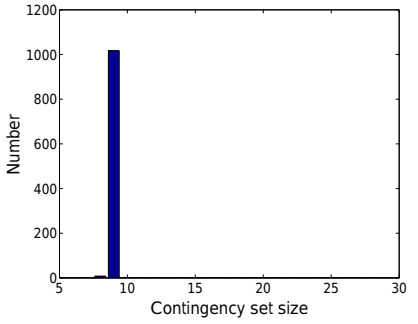


(b) Nresp

Fig. 6 Size distributions of the contingency sets calculated by different methods on FB-Walk-0-197.



(a) Appresp



(b) Nresp

Fig. 7 Distributions of the contingency sets calculated by different methods on FB-Walk-158-146.

edges that are involved in more traces. An instance of this can be found in Example 7.

In contrast, the results of Appresp are decentralized. We think this is because Appresp ensures causal relationships, i.e., it guarantees a feasible contingency set. This property allows its results to avoid being too highly influenced by edges involved in many traces. Therefore, our Appresp method can handle complex propagation histories in practical applications.

(4) Scalability. We not only test the scalability of the Appresp, Nresp, and Capability methods, but also test the proposed ranking strategy appresp-capability and its light version appresp-capability*. We use five different datasets with varied propagation edge sizes and trace sizes. Three of these are the propagation histories already used in Section 6.2, while the other two are newly generated propagation history datasets called LP1 and LP2. Both are generated similarly to FB-Walk-0-197. Specifically, LP1 contains 6×10^3 edges and 10^5 traces, while LP2 contains 10^5 edges and 5×10^5 traces. Figure 8 shows the results.

For the responsibility calculation, we can see that both Appresp and Nresp methods have the same efficiency level. The small extra time cost of Appresp is used to ensure causal relationships and, considering the superior performance of Appresp, this cost is worthwhile for a feasible solution to responsibility calculation.

Compared to the responsibility calculation, our Capability method is much more efficient. This chimes with our analysis that the capability indicator involves straightforward computation. Therefore, this indicator is always recommended for its good performance and high efficiency.

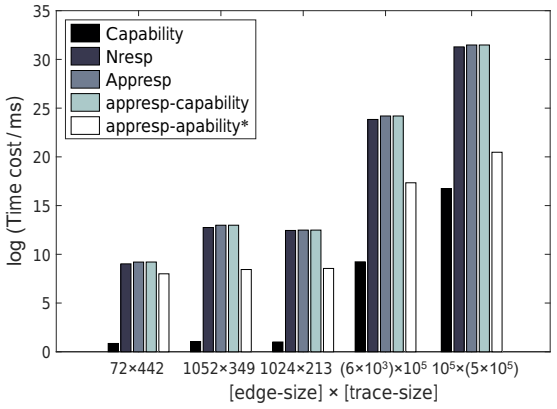


Fig. 8 Scalability of Capability, Nresp, Appresp, appresp-capability, and appresp-capability*. Note 72×442 (FB-Sample), 1052×349 (FB-Walk-0-197), 1024×213 (FB-Walk-158-146), $(6\times10^3)\times10^5$ (LP1), and $10^5\times(5\times10^5)$ (LP2) stand for the sizes of five propagation history datasets. Note also that the y axis is in log scale (e.g., $y = 5$ means the time cost is 2^5 ms).

Comparing the two integrated ranking strategies, appresp-capability* is much more efficient than appresp-capability. Taking the LP2 dataset as an example, appresp-capability* operates approximately 2000 times faster than appresp-capability. Consequently, this light version can be considered an efficient ranking method.

7 Related Work

7.1 Diffusion in SNS

Social network analysis is a hot research topic over recent years[33–35], and information diffusion analysis is one of the most important problems in this field[36–38]. This topic has many interesting applications, such as the identification of influencers[39,40], the maximization of influence[41,42], and the identification of hot topics[43,44]. Most works analyze diffusion problems from the viewpoint of network structure under different information propagation models, such as the susceptible infected model[45] and the independent cascade model[46].

In this paper, we analyze diffusion based on propagation histories rather than network structure. To the best of our knowledge, this is the first effort to understand diffusion phenomena based on propagation histories.

7.2 Causality

The classical account of counterfactual causality (if event X had not occurred, event Y would not have occurred) goes back to Hume in 1739[2]. Lewis[47] analyzed it in a formal way. Recently, a rigorous definition of causality, *"Actual Causality"*, was initiated by Halpern and Pearl[14,48]. Using this definition, Chockler and Halpern[19] introduced the degree of responsibility to evaluate the contribution of each cause. Recently, Meliou et al.[21,49] first studied the complexity of causality and responsibility problems in relational databases. Qin et al.[27,50] analyzed the causality problem for conjunctive queries with inequalities. Lian and Chen[51] focused on causality in the probabilistic queries in uncertain databases. We refer to Refs. [52,53] for expositions of causality theory.

In this paper, we employ DCE as the ranking criterion for the propagation history ranking problem in SNS. DCE is a fine-grained causal measure which not only matches our common sense but is also well suited to the SNS scenario. To avoid complex calculations, we further propose an integrated ranking strategy to capture the intuition of DCE.

7.3 Edge/node ranking in networks

Edge/node ranking is a traditional problem in a number of research areas, including information retrieval[54,55] and network reliability[7]. In the information retrieval domain, most works (e.g., Page Rank[12] and HITS[56]) rank nodes (pages) by considering (1) the relevance of pages with respect to input queries, and (2) the network topology. In the network reliability domain, edges/nodes are ranked in terms of their overall importance to *reliability*[8,57]. Here, reliability is a measure of the connectivity of a network where each edge/node has a probability of failure.

In these works, the importance of each edge/node is rated mainly based on network architecture. In contrast, for the propagation history ranking problem the importance of each edge/node is evaluated by its causal contribution to the actual information diffusion process, regardless of the underlying network structure.

8 Conclusion and Future Work

This study establishes a connection between the diffusion process that has been extensively studied in social network analysis and the causality theory that has been heavily studied in artificial intelligence. Specifically, this study presents the propagation history ranking problem in SNS, and proposes a solution from the viewpoint of causality. In this study, we first show the difference between propagation histories and network structure. Then we introduce the causal measure DCE as the ranking criterion. Due to the hardness of DCE calculation, we propose two integrated ranking strategies by introducing the indicators "responsibility" and "capability", which capture the necessity and sufficiency aspects, respectively, of causal effects. Furthermore, for the responsibility calculation, we design an approximate algorithm that can guarantee a feasible solution for general propagation histories. Finally, extensive experiments demonstrate the feasibility and advantages of our approach.

There are some interesting directions for future work. Firstly, in this paper, we study the propagation history ranking problem only in the simplest diffusion case: information diffusion from one source node to one target node. However, there are more complicated

diffusion cases, such as one-to-many or many-to-many diffusion, that have not yet been studied. Secondly, although this paper only focuses on information diffusion in SNS, the proposed methods can be generalized to other domains. For future work, we plan to extend our method to other applications, such as network reliability and model checking. Finally, causality-based reasoning and management is a useful tool to explore the potential value of big data, which points to other worthwhile future research directions.

Appendix

A DCE as Ranking Criterion

We first review Pearl's definition of *Probability of Necessity and Sufficiency* (PNS)[14], which has been proved to consider causal effects from both the necessity and sufficiency aspects. We then prove Proposition 1 by showing that PNS and DCE are equivalent in the scenario of propagation history ranking.

Definition 6 (PNS). Suppose the propagation history of a diffusion is Φ, and X is a participant edge. Let x and x' stand for the propositions "X non-failed" and "X failed", respectively. The *probability of the necessity and sufficiency* of edge X for diffusion Φ is defined as the expression:

$$\text{PNS} \triangleq P(\Phi_x = \text{true}, \Phi_{x'} = \text{false}) \tag{3}$$

In this definition, Φ_x = true stands for that enabling edge X leads to a successful information transmission, and $\Phi_{x'}$ = false stands for the removal of X makes this transmission fail. Thus, PNS considers causal effects from both the necessity and sufficiency aspects; a rigorous proof can be found in Ref. [14].

Proof of Proposition 1 We continue with the same definitions of Φ, X, x, and x' in Definition 6. Let set S contain all propagation traces which go through edge X, with the rest of the traces is put into another set (denoted as $\bar{S}$). Let s and $\bar{s}$ stand for the cases that S and $\bar{S}$, respectively, can successfully transmit the information, and let s' and $\bar{s}'$ denote their complements. We first get the following two properties.

Property 1 X does not affect $\bar{s}$ and $\bar{s}'$:

$P(\bar{s}_x) = P(\bar{s}_{x'}) = P(\bar{s})$ and $P(\bar{s}'_x) = P(\bar{s}'_{x'}) = P(\bar{s}')$.

Property 2 Since all traces in S go through X, we have $P(s_{x'}) = 0$.

Consider the definition of DCE (Definition 2):

$$\begin{aligned}
\text{DCE}(X) &= P(\Phi_x{=}\text{true}) - P(\Phi_{x'}{=}\text{true}) = \\
&P\big((s+\bar{s})_x\big) - P\big((s+\bar{s})_{x'}\big) = \\
&P(s_x + \bar{s}_x) - P(s_{x'} + \bar{s}_{x'}) = \\
&P(s_x + \bar{s}_x) - P(\bar{s}_{x'}) = \qquad \text{(Property 2)} \\
&P(s_x + \bar{s}) - P(\bar{s}) = \qquad \text{(Property 1)} \\
&P(s_x) + P(\bar{s}) - P(s_x, \bar{s}) - P(\bar{s}) = \\
&P(s_x) - P(s_x, \bar{s}) = \\
&P(s_x)\big(1 - P(\bar{s}|s_x)\big) = \\
&P(s_x)P(\bar{s}'|s_x) = \\
&P(s_x, \bar{s}').
\end{aligned}$$

Consider the definition of PNS (Definition 6):

$$\begin{aligned}
\text{PNS}(X) &= P(\Phi_x = \text{true}, \Phi_{x'} = \text{false}) = \\
&P\big((s+\bar{s})_x, (s+\bar{s})'_{x'}\big) = \\
&P\big((s+\bar{s})_x, (\bar{s})'_{x'}\big) = \qquad \text{(Property 2)} \\
&P(s_x + \bar{s}_x, \bar{s}'_{x'}) = \\
&P(s_x + \bar{s}, \bar{s}') = \qquad \text{(Property 1)} \\
&P(s_x, \bar{s}'). \qquad \text{(only } \bar{s}'\text{=true satisfies this)}
\end{aligned}$$

Combining the above two results, we get the following equation:

$$\text{DCE}(X) = \text{PNS}(X) = P(\Phi_x{=}\text{true}) - P(\Phi_{x'}{=}\text{true}) \tag{4}$$

Note that a similar equation has also been proved by Pearl[14], when an elaborate complex assumption (*exogeneity* and *monotonicity*) holds true. As an extension, we prove it from the viewpoint of information propagation.

As shown in Eq. (4), PNS and DCE can be seen as the same causal measure in the problem of propagation history ranking. Therefore, we successfully prove Proposition 1. ■

B PN Calculation

We first review Pearl's definition of *Probability of Necessity*[14]. Then, we show how to calculate it from the propagation history.

Definition 7 (PN). Continuing with the same definitions of Φ, X, x, and x' in Definition 6, the *probability of the necessity* of edge X for diffusion Φ is defined as the expression:

$$\text{PN} \triangleq P(\Phi_{x'}{=}\text{false} | X{=}\text{true}, \Phi{=}\text{true}).$$

Proposition 5 Continuing with the same definitions of X, x, x', s, $\bar{s}$, s', and $\bar{s}'$ in the proof of Proposition 1 found in Appendix A, the PN value of edge X (for diffusion Φ) can be estimated as the expression:

$$\text{PN} = P(\bar{s}'s)/[P(\bar{s}x) + P(\bar{s}'s)] \tag{5}$$

Proof To calculate PN, we need to find a situation where X=true and Φ=true did in fact occur, but where the removal of X will produce Φ=false. In the first step, we find the situation $(\bar{s} \vee \bar{s}'s) \wedge x$ ensures that X=true and Φ=true did in fact occur. The next step is to find a situation (based on the first step) where the removal of X will produce Φ=false. Here $s\bar{s}'$ satisfies, because the removal of X will cause $s \to s'$ (since all traces in S

go through edge X). Consequently, we get the following equation:

$$\begin{aligned}\text{PN} &= P(\Phi_{x'}{=}\text{false}|X{=}\text{true}, \Phi{=}\text{true}) = \\ &P\big(s\bar{s}' \wedge x|[(\bar{s}+\bar{s}'s)\wedge x]\big) = \\ &P\big(s\bar{s}'x \wedge (\bar{s}+\bar{s}'s)x\big)/P\big((\bar{s}+\bar{s}'s)x\big) = \\ &P(s\bar{s}'\bar{s}x + \bar{s}'sx)/P(\bar{s}x+\bar{s}'sx) = \\ &P(s\bar{s}'\bar{s}x+\bar{s}'sx)/[P(\bar{s}x)+P(\bar{s}'sx)-P(\bar{s}x\bar{s}'sx)]= \\ &P(s\bar{s}'\bar{s}x + \bar{s}'sx)/[P(\bar{s}x) + P(\bar{s}'sx)] = \\ &P(\bar{s}'sx)/[P(\bar{s}x) + P(\bar{s}'sx)] = \\ &P(\bar{s}'s)/[P(\bar{s}x) + P(\bar{s}'s)]. \end{aligned}$$ ■

C PS Calculation

We first review Pearl's definition of *Probability of Sufficiency*[14]. Then, we show how to calculate it from the propagation history.

Definition 8 (PS). Continuing with the same definitions of Φ, X, x, and x' in Definition 6 in Appendix A, the *probability of the sufficiency* of edge X for diffusion Φ is defined as the expression:

$$\text{PS} \triangleq P(\Phi_x{=}\text{true}|X{=}\text{false}, \Phi{=}\text{false}).$$

Proposition 6 Continuing with the same definitions of Φ, X, x, x', s, $\bar{s}$, s', and $\bar{s}'$ in the proof of Proposition 1 in Appendix A, the PS value of edge X (for diffusion Φ) can be estimated as the expression:

$$\text{PS} = P\big(\bar{s}'(s|x)x'\big)/P(\bar{s}'x') \tag{6}$$

Proof To calculate PS, we need to find a situation where X=false and Φ=false shows up at first, but where changing $x' \to x$ would produce Φ=true. This situation corresponds to $s'\bar{s}' \wedge x'$, i.e., X=false and Φ=false both occur. Then, we change $x' \to x$, and cause Φ=false $\to$ Φ=true. So we get

$$\begin{aligned}\text{PS} = \; & P(\Phi_x = \text{true} \mid X = \text{false}, \Phi{=}\text{false}) = \\ & P(\bar{s}'s_x x'|\bar{s}'s'x') = \\ & P(\bar{s}'s_x x' \wedge \bar{s}'s'x')/P(\bar{s}'s'x') = \\ & P(\bar{s}'s_x s'x')/P(\bar{s}'s'x') = \\ & P(\bar{s}'s_x x')/P(\bar{s}'x'). \qquad (x' \text{ causes } s')\end{aligned}$$

According to the definition of intervention[14], $P(y_x)$ equals $P(y|x)$ in the causal model of propagation history, because each participant edge X is not affected by any other cause. Consequently, we could get

$$\text{PS} = P\big(\bar{s}'x'(s|x)\big)/P(\bar{s}'x').$$ ■

Acknowledgment

This work was supported in part by the Fundamental Research Funds for the Central Universities (No. FRF-TP-18-016A1), China Postdoctoral Science Foundation Funded Project (No. 2018M640066), and the National Natural Science Foundation of China (No. 61872207).

References

[1] D. Yin and Y. Shen, Location- and relation-based clustering on privacy-preserving social networks, *Tsinghua Sci. and Technol.*,vol. 23, no. 4, pp. 453–462, 2018.

[2] D. Hume, *A Treatise of Human Nature*. London, UK: John Noon, 1739.

[3] F. Chierichetti, S. Lattanzi, and A. Panconesi, Rumor spreading in social networks, in *Automata, Languages and Programming*, S. Albers, A. Marchetti-Spaccamela, Y. Matias, S. Nikoletseas, and W. Thomas, eds. Springer, 2009, pp. 375–386.

[4] J. Kostka, Y. A. Oswald, and R. Wattenhofer, Word of mouth: Rumor dissemination in social networks, in *Structural Information and Communication Complexity*, A. A. Shvartsman and P. Felber, eds. Springer, 2008, pp. 185–196.

[5] C. T. Li, S. D. Lin, and M. K. Shan, Finding influential mediators in social networks, in *Proc. 20^{th} Int Conf Companion on World Wide Web*, Hyderabad, India, 2011, pp. 75–76.

[6] D. Shah and T. Zaman, Rumors in a network: Who's the culprit? *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5163–5181, 2011.

[7] L. B. Page and J. E. Perry, Reliability polynomials and link importance in networks, *IEEE Trans. Reliab.*, vol. 43, no. 1, pp. 51–58, 1994.

[8] D. R. Shier, *Network Reliability and Algebraic Structures*. New York, NY, USA: Clarendon Press, 1991.

[9] K. H. Xu, Y. X. Guo, L. K. Guo, Y. G. Fang, and X. L. Li, Control of photo sharing over online social networks, in *Proc. 2014 IEEE Global Communications Conf*, Austin, TX, USA, 2014, pp. 704–709.

[10] K. A. Harrison and A. H. Smith, Some new species and distribution records of Rhizopogon in North America, *Can. J. Bot.*, vol. 46, no. 7, pp. 881–899, 1968.

[11] Y. M. Tan, V. Venkatesh, and X. H. Gu, Resilient self-compressive monitoring for large-scale hosting infrastructures, *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 3, pp. 576–586, 2013.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, The PageRank citation ranking: Bringing order to the web, http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf, 1999.

[13] R. A. Fisher, *Statistical Methods for Research Workers*. Genesis Publishing Pvt Ltd, 1925.

[14] J. Pearl, *Causality*: *Models*, *Reasoning and Inference*. Cambridge, UK: Cambridge University Press, 2000.

[15] J. Y. Campbell, A. W. C. Lo, and A. C. MacKinlay, *The Econometrics of Financial Markets*. Princeton, NJ, USA: Princeton University Press, 1997.

[16] J. Cheney, Causality and the semantics of provenance, arXiv preprint arXiv: 1004.3241, 2010.

[17] I. Hacking, Telepathy: Origins of randomization in experimental design, *Isis*, vol. 79, no. 3, pp. 427–451, 1988.

[18] R. L. Wasserstein, Monte Carlo: Concepts, algorithms, and applications, *Technometrics*, vol. 39, no. 3, p. 338, 1997.

[19] H. Chockler and J. Y. Halpern, Responsibility and blame: A structural-model approach, *J. Artif. Intell. Res.*, vol. 22,

pp. 93–115, 2004.

[20] T. Eiter and T. Lukasiewicz, Complexity results for structure-based causality, *Artif. Intell.*, vol. 142, no. 1, pp. 53–89, 2002.

[21] A. Meliou, W. Gatterbauer, K. F. Moore, and D. Suciu, The complexity of causality and responsibility for query answers and non-answers, *Proc. VLDB Endow.*, vol. 4, no. 1, pp. 34–45, 2010.

[22] P. Buneman, S. Khanna, and W. C. Tan, Why and where: A characterization of data provenance, in *Proc. 8th Int. Conf. Database Theory*, London, UK, 2001, pp. 316–330.

[23] I. Altintas, O. Barney, and E. Jaeger-Frank, Provenance collection support in the Kepler scientific workflow system, in *Proc. Int. Provenance and Annotation Workshop*, Chicago, IL, USA, 2006, pp. 118–132.

[24] S. Milgram, The small world problem, *Psychol. Today*, vol. 2, no. 1, pp. 60–67, 1967.

[25] G. Strang, *Introduction to Applied Mathematics*. Wellesley, MA, USA: Wellesley-Cambridge Press, 1986.

[26] V. Chvatal, A greedy heuristic for the set-covering problem, *Math. Oper. Res.*, vol. 4, no. 3, pp. 233–235, 1979.

[27] B. Qin, S. Wang, X. F. Zhou, and X. Y. Du, Responsibility analysis for lineages of conjunctive queries with inequalities, *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 6, pp. 1532–1543, 2014.

[28] J. McAuley and J. Leskovec, Learning to discover social circles in ego networks, in *Proc. 25th Int. Conf. Neural Information Processing Systems*, 2012, pp. 539–547, 2012.

[29] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, Walking in Facebook: A case study of unbiased sampling of OSNs, in *Proc. 2010 IEEE INFOCOM*, San Diego, CA, USA, 2010, pp. 1–9.

[30] M. J. Salganik and D. D. Heckathorn, Sampling and estimation in hidden populations using respondent-driven sampling, *Sociol. Methodol.*, vol. 34, pp. 193–239, 2004.

[31] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA, USA: MIT Press, 2005.

[32] K. Järvelin and J. Kekäläinen, IR evaluation methods for retrieving highly relevant documents, in *Proc. 23rd Ann. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Athens, Greece, 2000, pp. 41–48.

[33] J. Scott, *Social Network Analysis*. Thousand Oaks, CA, USA: SAGE Publications Ltd, 2017.

[34] Z. Wang, X. J. Ye, C. K. Wang, Y. X. Wu, C. P. Wang, and K. W. Liang, RSDNE: Exploring relaxed similarity and dissimilarity from completely-imbalanced labels for network embedding, in *Proc. 32nd AAAI Conf Artificial Intelligence*, New Orleans, LA, USA, 2018, pp. 475–482.

[35] Q. Wang, Z. Wang, and X. J. Ye, Equivalence between line and matrix factorization, arXiv preprint arXiv: 1707.05926, 2017.

[36] S. Taylor and P. A. Todd, Understanding information technology usage: A test of competing models, *Inf. Syst. Res.*, vol. 6, no. 2, pp. 144–176, 1995.

[37] Z. Wang, C. K. Wang, J. S. Pei, X. J Ye, and P. S. Yu, Causality based propagation history ranking in social networks, in *Proc. 25th Int. Joint Conf. Artificial Intelligence*, New York, NY, USA, 2016, pp. 3917–3923.

[38] X. Meng, G. Xu, T. Guo, Y. Yang, W. Shen, and K. Zhao, A novel routing method for social delay-tolerant networks, *Tsinghua Sci. and Technol.*, vol. 24, no. 1, pp. 44–51, 2019.

[39] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, Identification of influential' spreaders in complex networks, *Nat. Phys.*, vol. 6, no. 11, pp. 888–893, 2010.

[40] Z. Wang, C. K. Wang, J. S. Pei, and X. J Ye, Multiple source detection without knowing the underlying propagation model, in *Proc. 31st AAAI Conf. Artificial Intelligence*, San Francisco, CA, USA, 2017, pp. 217–223.

[41] W. Chen, Y. J. Wang, and S. Y. Yang, Efficient influence maximization in social networks, in *Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Paris, France, 2009, pp. 199–208.

[42] S. Chen, J. Fan, G. L. Li, J. H. Feng, K. L. Tan, and J. H. Tang, Online topic-aware influence maximization, *Proc. VLDB Endow.*, vol. 8, no. 6, pp. 666–677, 2015.

[43] J. Kleinberg, Bursty and hierarchical structure in streams, *Data Min. Knowl. Discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[44] M. Cataldi, L. Di Caro, and C. Schifanella, Emerging topic detection on twitter based on temporal and social terms evaluation. in *Proc. 10th Int. Workshop on Multimedia Data Mining*, Washington, DC, USA, 2010, p. 4.

[45] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and Its Applications*. London, UK: Charles Griffin & Company Ltd., 1975.

[46] S. Bikhchandani, D. Hirshleifer, and I. Welch, A theory of fads, fashion, custom, and cultural change as informational cascades, *J. Polit. Econ.*, vol. 100, no. 5, pp. 992–1026, 1992.

[47] D. Lewis, Causation, *J. Philos.*, vol. 70, no. 17, pp. 556–567, 1973.

[48] J. Y. Halpern and J. Pearl, Causes and explanations: A structural-model approach. Part I: Causes, *Br. J. Philos. Sci.*, vol. 56, no. 4, pp. 843–887, 2005.

[49] C. Freire, W. Gatterbauer, N. Immerman, and A. Meliou, The complexity of resilience and responsibility for self-join-free conjunctive queries, *Proc. VLDB Endow.*, vol. 9, no. 3, pp. 180–191, 2015.

[50] B. Qin, S. Wang, and X. Y. Du, Efficient responsibility analysis for query answers, in *Proc. 18th Int. Conf. Database Systems for Advanced Applications*, Wuhan, China, 2013, pp. 239–253.

[51] X. Lian and L. Chen, Causality and responsibility: Probabilistic queries revisited in uncertain databases, in *Proc. 22nd ACM Int. Conf. Information & Knowledge Management*, San Francisco, CA, USA, 2013, pp. 349–358.

[52] S. Kleinberg and G. Hripcsak, A review of causal inference for biomedical informatics, *J. Biomed. Inf.*, vol. 44, no. 6, pp. 1102–1112, 2011.

[53] J. Y. Li, L. Liu, and T. D. Le, *Practical Approaches to Causal Relationship Exploration*. Springer, 2015.

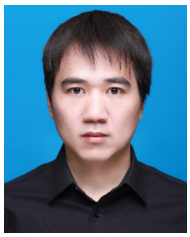
[54] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983.

[55] W. B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs, NJ,

USA: Prentice Hall, 1992.

[56] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, *J . ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[57] C. J. Colbourn, *The Combinatorics of Network Reliability*. New York, NY, USA: Oxford University Press, 1987.



Zheng Wang received the PhD degree from Tsinghua University in 2018. Currently, he is an assistant professor at Department of Computer Science and Technology, University of Science and Technology Beijing, China. His research interests include social network analysis, data management, and artificial intelligence.

Chaokun Wang received the BS, MS, and PhD degrees from Harbin Institute of Technology, China in 1997, 2000 and 2005, respectively. He joined the School of Software at Tsinghua University in February 2006, where currently he is an associate professor. His current research interests include social network analysis, graph data management, and music computing.

Xiaojun Ye received the BS degree from Northwest Polytechnical University, China in 1987, and the PhD degree from INSA Lyon, France, in 1994. Currently, he is a professor at School of Software, Tsinghua University, China. His research interests include cloud data management, data security and privacy, and database system testing.

Jisheng Pei received the PhD degree from Tsinghua University in 2018. His research interests include data provenance and business process trace clustering.

Bin Li received the PhD degree from Peking University, in 2010, and BEng and MS degrees from Huazhong University of Science and Technology, in 1993 and 2001, respectively. Currently, he works in China Information Technology Security Evaluation Center. His current research interests include information security, risk assessment, and cloud security.