International Journal of Data Science and Analytics (2022) 13:185–198
https://doi.org/10.1007/s41060-021-00293-x

REGULAR PAPER

Rumor containment in peer-to-peer message sharing online social networks

Rohit Tripathi[1] · Shilpa Rao[2]

Received: 28 March 2021 / Accepted: 6 November 2021 / Published online: 21 January 2022

Abstract

Rumors in online social networks (OSNs) create social chaos, financial losses, and endanger property, which makes rumor containment an important issue. We consider an OSN in which the users communicate via private peer-to-peer messages. We consider the proposed peer-to-peer linear threshold (PLT) and peer-to-peer independent cascade-variant (PICV) models for information diffusion in OSNs, which are variants of the classic IC and LT models, respectively. To combat the rumor spread in the OSN with peer-to-peer message sharing, we employ blocking and positive information diffusion strategies. While in blocking strategy, few users of the OSN called the blocked seed nodes are blocked from spreading the rumor, in positive information diffusion strategy, correct information is introduced into few users of the OSN called positive seed nodes. The positive seed nodes further spread the correct information to other users with time. For a given time-period called the rumor-relevance interval, we determine average number of rumor-influenced nodes for the random, the max-degree, the greedy, the proximity heuristic, and the proposed proximity-weight-degree (PWD)-based containment seed node selection schemes for both blocking and positive information diffusion strategies for PLT and PICV models. We compare the effect of the rumor-relevance interval duration and number of seed nodes on the average number of rumor-influenced nodes for different seed selection algorithms. Our experimental results show that proximity-weight-degree-based seed selection algorithm performs on par with the high-complexity greedy scheme.

1 Introduction

Online social networks (OSNs) have been increasingly used for information exchange and dissemination. Few examples of widely used social networks are Twitter, Facebook, Whatsapp, and LinkedIn[2,3]. On the one hand, circulating correct and timely information has social and economic benefits; on the other hand, spreading misinformation and rumors leads to social instabilities and insecurities [4]. For example, the social media rumors on swine flu in 2009 and on hurricane Sandy in 2012 created social anxiety [5,6].

A part of this work is published at International Conference on Communication Systems and Networks (COMSNETS), Bangalore, India, January 2020 [1].

✉ Rohit Tripathi
rohit@iiitg.ac.in

Shilpa Rao
shilpa@iiitg.ac.in

1 Department of CSE, IIIT Guwahati, Guwahati, India

2 Department of ECE, IIIT Guwahati, Guwahati, India

The independent cascade (IC) and linear threshold (LT) models are mostly used to model information diffusion process in OSNs. Open communication models such as posts on Facebook, Twitter, and LinkedIn can be modeled by IC and/or LT models. Every user attempts to send information with a certain probability and independently to each of his uninformed neighbors in the IC model. In the LT model, an uninformed user accepts an information only if he receives it from a pre-determined fraction of his neighbors. In both models, it is assumed that once a user adopts the *accurate* or *rumor* information, the user does not change his viewpoint. Also, a newly informed user tries to circulate the information to his neighbors.

Peer-to-peer models Private peer-to-peer online messages and email posts cannot be represented by IC or LT models. This is because both IC and LT models are broadcast models in which a node sends the message to all its neighbors at a time instance as. Peer-to-peer message, on the other hand, can be sent by a node to only one of its neighbors in a given time instance. Therefore, alternate models are needed to represent peer-to-peer message diffusion in OSNs.

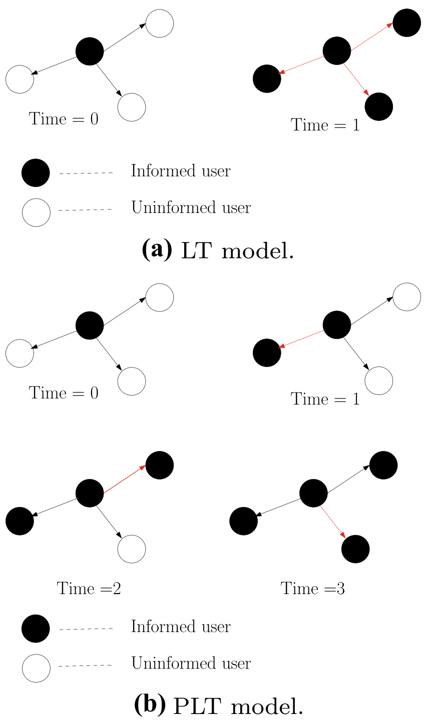


Fig. 1 Example of LT and PLT models having 4 user in which one of them is an informed user at time =0. In this example, a user accepts information if one of its neighbors informs it. The directed edge colored red between two user indicates an attempt made by a user to inform its neighbor

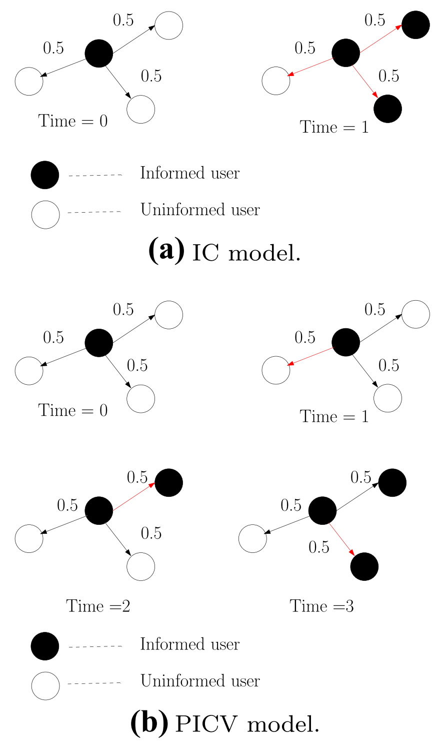


Fig. 2 Example of IC and PICV models having 4 users in which one of them is an informed users at time =0. In this example, a user sends information to its neighbor successfully with probability 0.5. The directed edge colored red between two users indicates an attempt made by a user to inform its neighbor

In this paper, we have presented new models for peer-to-peer communications and presented solutions. The contributions of our work are described as follows:

- **Formulation of a novel peer-to-peer LT (PLT) model** The LT model being a broadcast model is not suitable to represent the peer-to-peer messaging process. Therefore, we present a variant of the LT model that captures online peer-to-peer messaging and call it the peer-to-peer LT model. In this model, at a given instant of time, each user can send the information to only one of his neighbors from which the user has not obtained the same information. If a user receives the same message from a pre-determined fraction of his neighbors, he adopts the message. In this model, a user does not pass the message to his neighbors, which have sent him the message. This approach is in line with practice, as a receiver does not forward the same message to the senders. Information propagation in PLT model is not as fast when compared to the classic LT model because any user sends information to only one of his neighbor in each time-step. The information propagation in LT and PLT models is illustrated in Fig.1. As shown in the figure, the users in the LT model are informed by time=1. In the PLT model, the user are informed by time =3.
- **Formulation of PIC-variant (PICV) model** The PIC model proposed in [4] assumes that the user who has adopted the message tries to pass the message to its neighbors from whom it has received the message as well. As in practice, a user generally does not resend the adopted message to senders of the message, we modify the PIC model in [4]. In our modified PIC model, in a given time-step each user attempts to send the information to one of his neighbors from which the user has not obtained it. Like the PIC model, a receiver adopts the message with a certain probability. We call this variant of the PIC model as the PIC-variant (PICV) model. An example for information diffusion in IC and PICV models is shown in Fig. 2. As shown in the figure, information propagates slowly in the PICV model when compared to the IC model.
- **Design and performance analysis of a novel low-complexity algorithm and evaluation of baseline algorithms for rumor containment by accurate information diffusion in PICV and PLT models:** As PICV and PLT models are novel to this paper, to the best of our knowledge rumor containment algorithms are not designed for these models. In this paper, both for the PLT and PICV models, we consider rumor containment for a specific time period after the rumor starts. This time period is called the *rumor-relevance interval*. The

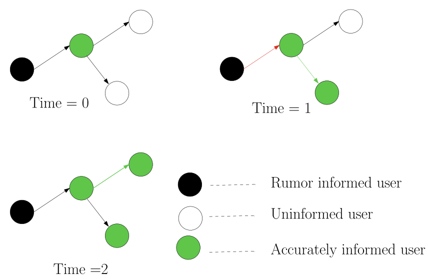


Fig. 3 Example of a PLT model having 4 users in which one of them is a rumor user and one is a accurately informed user at time =0. In this example, a node accepts information if one of its neighbors informs it. The directed edge colored red between two users indicates an attempt made by a rumor user to inform its neighbor. The directed edge colored green between two users indicates an attempt made by an accurate user to inform its neighbor

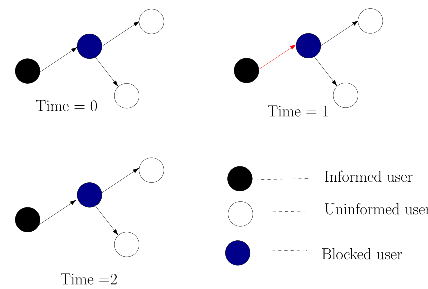


Fig. 4 Example of a PLT model having 4 users in which one of them is an rumor user and one is a blocked user at time =0. In this example, a user accepts information if one of its neighbors informs it. The edge colored red between two users indicates an attempt made by a user to inform its neighbor

notion of rumor-relevance interval is applicable in scenarios where a rumor affects only for certain duration and therefore needs to be curbed for that duration. For example, the rumor of successful candidate in an election is valid only until the election result is declared.
We consider different selection schemes such as (i) random selection, (ii) greedy selection, (iii) max-degree based selection, (iv) proximity heuristic based selection, and (v) the proposed proximity-weight-degree (PWD)-based selection to select the users to circulate correct information for rumor containment for the time period T in the OSN. For each of the selection scheme, we calculate the average number of rumor-affected nodes in time T. We study the effect of the number of nodes spreading accurate information on the average number of rumor-affected users for each of the user selection schemes. Further, we study the variation in average number of rumor-affected users with the rumor-relevance interval. Figure 3 illustrates accurate information spreading strategy for rumor containment in the PLT model. The accurately informed user does accept information from rumor node and thereby stops rumor from propagating further.

– **Design and performance analysis of a novel low-complexity algorithm and evaluation of baseline algorithms for rumor containment by blocking in PICV and PLT models:** We consider the blocking strategy for rumor containment and investigate the random, greedy, max-degree, proximity heuristic and proposed proximity-weight-degree-based selection schemes to select the blocked seeds for rumor containment in peer-to-peer models. For rumor-relevance interval T, we determine the average number of rumor-affected nodes and study the effect of number of blocked nodes and time period T on the number of rumor-affected nodes in the OSN. Figure4 illustrates blocking strategy for rumor containment in the PLT model. The blocked node does accept information from rumor node and thereby stops rumor from propagating further.

Organization of the paper: The related works are given in Sect. 2. The OSN model and the P2P information diffusion process are discussed in Sect. 3. The rumor containment techniques in P2P OSN models are discussed in 4. We discuss the experimental results in Sect. 5. We state the conclusions and future works in Sect. 6.

2 Related works

The spread of rumors in OSNs can be curbed either by *blocking specific links and users* from circulating the rumor [7,8] or by *spreading accurate information* [9,10]. In [8], for the IC model, greedy algorithm is used to block the users and thereby contain the rumor. In [11], it is assumed that nodes cannot be blocked beyond a threshold of time. The paper addresses the problem of minimizing the average of rumor-affected nodes such that any node in the network is not blocked beyond its tolerance levels. In [7], to minimize the spread of rumor, OSN links to be blocked are determined. For LT rumor propagation model, approximation algorithms to minimize the rumor influence by removing the links are discussed in [12].

Rumor detection and arresting its spread in OSNs is matter of paramount importance due to its negative effects [13]. In [14], user features such as number of followers, number of reposts, and others are used to identify rumor-infected users in microblogging sites. A game-theoretic approach for rumor detection wherein the rumor source and the detector are modeled as players is discussed in [15].

For the IC model, in order to minimize the total number of rumor-affected users, correct information is provided to a set of users in the OSNs, which further spread it in the network[9]. [4] discusses game-theoretic schemes to determine the users to spread correct information for rumor

Table 1 Comparison with related papers

Paper	Advantage	Disadvantage	Relation (similarity and difference) with the reference
Doerr et al. [13]	Reference paper explains how rumor spreads very quickly in social networks.	Reference paper doesn't discuss rumor containment method.	Similarity: Our work illustrates rumor spread in OSNs. Difference: We also provide rumor-containment methods
Kimula et al. [7]	Reference paper blocks links for rumor containment.	The model used in the reference does not work for peer-to-peer networks.	Similarity: We have used blocking strategy for rumor containment. Difference: We have blocked nodes instead of links for rumor containment.
Wang et al. [8]	Reference paper blocks node for rumor containment	The model used in the reference paper does not work for peer-to-peer networks	Similarity: We have used node blocking strategy for rumor containment. Difference: We have studied effects of node blocking for rumor containment in peer-to-peer models.
Budak et al. [9]	Reference paper spreads positive information for rumor containment	The model used in the reference paper does not work for peer-to-peer networks	Similarity: We have studied effects of spreading positive information for rumor containment. Difference: We have considered peer-to-peer model.
He et al. [18]	Uses combination of two method: blocking and spreading positive information for rumor containment.	Considers the Susceptible, Infected, Recovered, and Dead Model for nodes.	Similarity: We have used blocking and positive information spread for rumor containment. Difference: We have considered peer-to-peer models.
Tong et al. [4]	Reference paper proposed a basic information diffusion model for peer-to-peer networks	The reference paper has not proposed the PICV and PLT model.	Similarity: We have considered peer-to-peer models in our paper. Difference: We have proposed the new PICV and PLT models.

minimization for the PIC model. Further, for the LT model, the set of users to whom accurate information needs to be injected for rumor minimization is described in [10]. In[16], a multi-feature rumor diffusion in IC model is considered, wherein a user accepts a rumor if weighted sum of the features of rumor exceeds a certain threshold. An algorithm to select users to spread correct information is devised, and it is shown that the performance of the proposed algorithm is close to the greedy algorithm.

The scenario that rumor loses its significance after a certain time is considered in [17]. The optimal seed nodes to be selected to spread the true information, which minimize the rumor-infected users within the time frame, are determined. [18] uses both rumor blocking and truth spreading techniques for rumor containment in a cost-constrained rumor containment setting. In [18], it is shown that spreading the truth when the rumor is in its nascent phase and blocking rumor in the later stages is effective when the positive information diffusion cost is lesser than blocking cost. In [19], a sampling based algorithm for the problem of selecting users, which spread correct information, for rumor containment for the IC model for a specified time-frame is considered. It is shown that performance of the proposed algorithm is close to the greedy algorithm. The problem of seed node selection for rumor containment in general is NP hard as shown by many papers [4,10,19–21].

[4] proposes the peer-to-peer IC (PIC) model to capture peer-to-peer communication in OSNs. In this model, in a time step, a user tries to send information to one of his neighbors with a certain probability. Unlike the IC model, in the PIC model, at the most one of the user's neighbors receives the message in a time stance. Hence, information diffusion happens at a slower rate in the PIC model when compared to the IC model.

As we already mentioned, private peer-to-peer online messages and email posts cannot be represented by IC or LT models; therefore, we have presented two new models PLT and PICV (a variation of the model presented by [4]). We have presented seed selection algorithms for the models using both positive information diffusion and blocking rumor nodes. Table 1 provides a concise comparison of our work with the existing literature. This paper is an extended version of an already published conference paper [1].

3 P2P Information diffusion models in OSN

We first set up the notation used in the paper and then describe the diffusion models.

Notation We represent a directed graph as $G = (V, E)$, where V is the set of vertices and E is the set of edges of the graph G. The vertices in the set V, which are not present in set $S \subset V$ are denoted by the set $V \setminus S$. The set of in-neighbors and out-neighbors of vertex v, are denoted by $N^{\text{in}}(v) = \{u : (u, v) \in E\}$ and $N^{\text{out}}(v) = \{u : (v, u) \in E\}$, respectively. Similarly, the in and out degrees of vertex v are denoted by $\deg^{\text{in}}(v)$ and $\deg^{\text{out}}(v)$. We denote the null set by ϕ and denote cardinality of set A by $|A|$. In this paper, the terms vertices and nodes are used interchangeably.

We represent an OSN by a directed graph $G = (V, E)$, where the vertices correspond to the users and the edges correspond to the links between the users. The information propagation process occurs in discrete time units. A vertex is called *active* if it has accepted the information and *inactive* if it is uninfluenced by information. An activated vertex stays in the state forever [10]. The PLT and PICV information diffusion models are described in the following.

3.1 PLT Model

As shown in Fig. 5, in this model, every edge between the nodes is associated with an edge weight, which indicates the influence of the active source node on the destination node.

The edge weight $w_{uv} \geq 0$ of an edge (u, v) represents the influence of vertex u on vertex v. Similar to the classic LT model [22], we consider $w_{uv} = 0$ for $(u, v) \notin E$ and $\sum_{u \in \text{N}^{\text{in}}(v)} w_{uv} \leq 1$. Every vertex samples a threshold θ_v, which is uniformly distributed in [0, 1]. It is assumed that positive seed selection schemes do not know these threshold values.

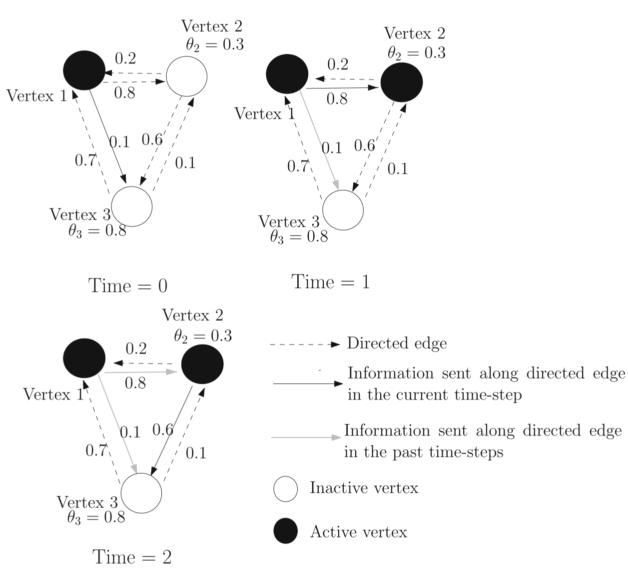


Fig. 5 Information propagation in an OSN with three nodes and one active node as per the PLT model

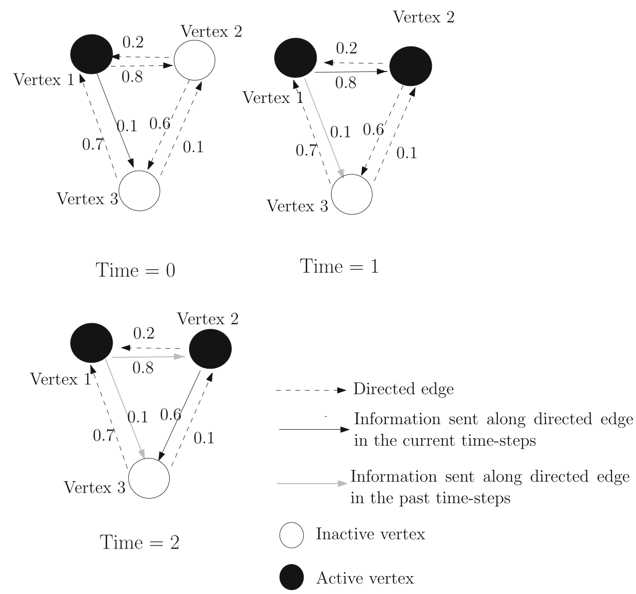


Fig. 6 Information diffusion in a network with three vertices and one active vertex as per the PICV model

In this model, in any time-step, an active node sends the message to one of its neighbors from whom it has not obtained the same message. The neighbors are chosen by the vertex in a random order to send the message. An inactive node v gets activated if the sum of the weights of the incoming edges from which it has received the information is greater than the threshold θ_v.

3.2 PICV model

In this model, an active node activates its neighbor successfully with a certain probability. Let p_{uv} denote the probability that node u activates node v successfully. Clearly, $0 \leq p_{uv} \leq 1$ and $p_{uv} = 0$ for $(u, v) \notin E$. p_{uv} is called the *activation probability* of the incoming edge (u, v) to node v.

As shown in Fig. 6, in the PICV model, in any time-step, an active node sends the information to one of its neighbor provided that the neighbor has not sent the same message to the given node. The order in which neighbors are chosen by the node to send the information is random. The PICV model is similar to the PIC model [4] with a minor difference that in the PICV model, a node does not attempt to pass the message to its neighbors from whom it has received the same message.

4 Rumor containment in P2P message sharing OSNs

Problem statement We consider the problem of rumor containment in PLT and PICV models. In the OSN, the rumor agent injects misinformation in a set of vertices called rumor seed nodes to spread misinformation in OSN. A vertex is called *rumor-active* if it has been influenced by the rumor. To contain the rumor, positive information diffusion and blocking techniques are employed. In the positive information diffusion technique, certain nodes are selected to spread correct information to counteract the rumor spread in the OSN. In blocking technique, certain nodes of the OSN are blocked from spreading the rumor message.

In this paper, we discuss various algorithms and evaluate their performance for selecting the nodes to spread positive information when positive information diffusion technique is used to contain rumor during the rumor-relevance interval. We also study the performance of various algorithms for selecting the nodes to be blocked to contain rumor spread in the OSN during the rumor relevance interval.

More precisely, we explain the problem as follows. Let the set of vertices that are influenced by the rumor agent at time $t = 0$ be denoted by $S^- \subset V$. At $t = 0$, a set of vertices $S_C \subset V$ be the set of containment seed nodes. Let K be the budget of selecting containment seed nodes. Clearly, $|S_C| \leq K$. Let $\sigma(S_C)$ denote the average number of rumor-active nodes after a time period T for the containment seed set S_C. Our goal is to design algorithms to determine the rumor containment seed set S_C of size not exceeding K that minimizes $\sigma(S_C)$. We determine this seed set for positive information diffusion and blocking strategies for PLT and PICV models.

We first describe the positive information diffusion and blocking in detail in this section. We later discuss the rumor containment algorithms in Sect.4.1.

- *Positive Information Diffusion:* In the positive information diffusion technique, a set of vertices called the positive seed set are injected with positive information, which further spread the positive information to the inactive nodes in the OSN. An inactive node becomes *positively active* if it has accepted accurate information. We use the terms accurate information and positive information interchangeably.
  At any time-step, if both rumor-active neighbor and positively active neighbor successfully pass their messages to an inactive node, the node becomes rumor-active due to the inherent negative bias of humans [10]. For PICV model, this corresponds to the scenario in which both rumor-active and positively active neighbors are successful in their activation attempts. For the PLT model, this happens when the total edge weights of incoming rumor and the total edge weights of incoming positive information links of node v both exceed θ_v in the same time-step.
  The positive information diffusion technique example in PLT model is shown in Fig. 7. In this example, we observe that positive active vertex is not affected by the rumor vertex influence. Further, the positive active vertex successfully spreads the positive information to inactive vertex. However, the positive active vertex cannot pass positive information to rumor-active vertex.

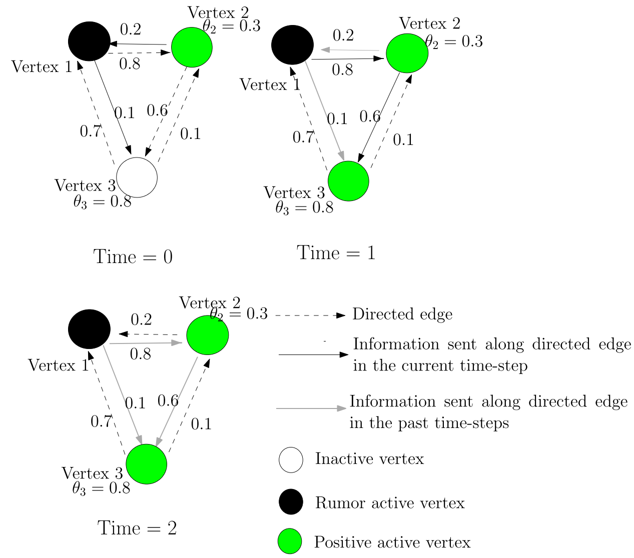


Fig. 7 Rumor and positive information diffusion in OSN with three vertices, one rumor active vertex, and one positive active vertex as per PLT model

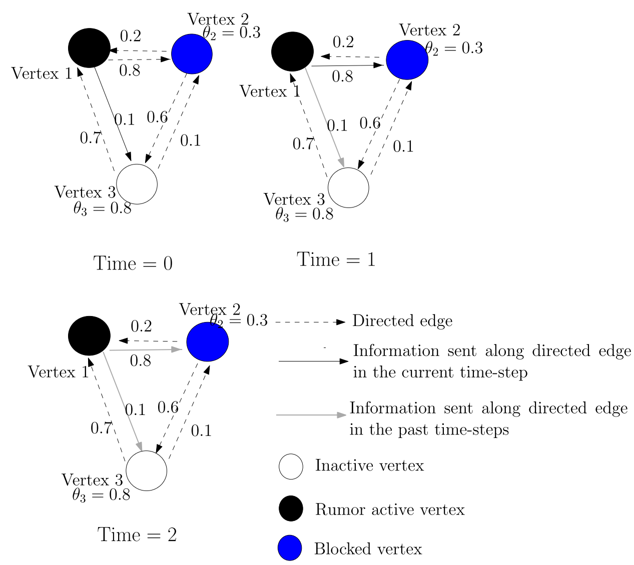


Fig. 8 Rumor diffusion and blocking in OSN with three vertices, one rumor active vertex, and one blocked vertex as per PLT model

- *Blocking:* In blocking, a set of vertices called the blocked seeds in the OSN are blocked from spreading the received messages. The blocking strategy example for PLT model is shown in Fig. 8. In this example, as vertex 2 is blocked, it neither gets activated by the rumor nor spreads the rumor message to vertex 3.

Comparison between positive information diffusion and blocking Positive information diffusion method spreads accurate information in the OSN to counteract the rumor spread, whereas blocking method blocks certain nodes in the OSN from spreading the rumor. Hence, positive information diffusion could be viewed as an active rumor containment method and blocking can be viewed as a passive rumor containment measure. When positive information diffusion measure is used, we find that competing messages—accurate and rumor—flow through the OSN. However, with blocking method only single message—rumor—flows in the OSN. In comparison with blocking, in positive information diffusion as seed nodes are injected with positive information, they both block the rumor message from spreading further as well as circulate accurate information in the OSN. However, the shortcoming of positive information diffusion is that it has higher number of the number of messages flowing in the OSN when compared to blocking. The number of messages will scale up considerably for large OSNs.

4.1 Containment seed selection algorithm

The problem of seed node selection for rumor containment in general is NP hard as shown by many papers [4,10,20,21]. For both PICV and PLT models, we consider random, max degree, greedy, proximity heuristic and the proposed proximity-weight degree containment seed selection schemes and evaluate the average number of rumor-active nodes after a time period T. If positive information diffusion technique is employed, the selected containment seed nodes will be provided with positive information. If blocking technique is employed, the selected containment seed nodes obtained will be blocked.

We describe the different containment seed selection algorithms in the following

1. *Random:* The set S_C, where $|S_C| = \min(K, |V \setminus S^-|)$, of containment seed nodes is randomly chosen from $V \setminus S^-$. The complexity of random selection is O(K) because it picks one node randomly for at the most K times.
2. *Max-degree:* The set S_C, where $|S_C| = \min(K, |V \setminus S^-|)$, of containment seed nodes, which have the maximum out-degree, are chosen from $V \setminus S^-$. The OSN graph is represented as an adjacency matrix with dimension $|V| \times |V|$. The max-degree selection scheme firstly computes the out-degree of every node, which is of $O(|V|^2)$. Next, the nodes are sorted in the increasing order of their degree, which is of $O(|V| \log |V|)$. At the most top K nodes of this list are selected, which involves $O(K)$ complexity. Clearly, max-degree is of $O(|V|^2)$.
3. *Proximity Heuristic:* This simple heuristic scheme is discussed in [10]. In this scheme, all the inactive out-neighbors of the rumor nodes are sorted in the decreasing order of their incoming edge weights for the PLT model and in the decreasing order of the activation probabilities of the incoming edges for the PICV model. If a node has multiple in-neighbors as rumor nodes, then the edge weights/activation probability of all such edges are added. The top K nodes are selected as the seed nodes. The number of out-neighbors of a nodes is O($|V|$), and to find all the out-neighbors of rumor nodes, the time taken is O($|S^-||V|$). The out-neighbors have to be sorted, and this takes O($|V| \log |V|$) time. Therefore, the time complexity is the maximum of O($|V| \log |V|$) and O($|S^-||V|$). This means that when the number of rumor seed nodes is very small ($|S^-|$ is very small), the complexity is O($|V| \log |V|$) and when it is very large ($|S^-| \simeq |V|$), the complexity is O($|V|^2$).
4. *Greedy:* The greedy selection algorithm to select S_C, where $|S_C| = \min(K, |V \setminus S^-|)$, is given in Algorithm 1, which is similar to [10]. The algorithm runs for at the most K iterations. S_C is initialized to ϕ. One inactive node is added to the set of containment seed nodes in every iteration. The inactive seed node, which along with already selected containment seeds minimizes the average number of rumor-active nodes for time T, is added in an iteration to the positive seed set. The average number of rumor-active nodes is computed by Monte Carlo simulations. Hence, the complexity of the algorithm is $O(|V| K \mathcal{T}(\sigma(S_C \cup \{v\})))$, where $\mathcal{T}(\sigma(S_C \cup \{v\}))$ is the complexity of calculating $\sigma(S_C \cup \{v\})$, $v \in V \setminus (S^- \cup S_C)$ via simulations. As for every sample path of the simulation, information propagates over at the most $|E|$ edges, computing the number of rumor-active nodes per sample path is of the order $O(|E|)$. If the simulations are averaged across R sample paths, the order of $\mathcal{T}(\sigma(S_C \cup \{v\}))$ is $O(R|E|)$. Hence, greedy algorithm is of the order $O(RK|V||E|)$ [23].

Algorithm 1: Greedy ($G(V, E)$, K, S_C, T)

initialize $S_C = \phi$
for $i = 1$ *to* K **do**
  select $u = \arg\max_{v \in V \setminus (S^- \bigcup S_C)} \sigma(S_C \bigcup \{v\})$
  $S_C = S_C \bigcup \{u\}$
return S_C

We provide a detailed theoretical proof that the performance of greedy algorithm is within the factor $(1 - 1/e)$ optimal algorithm for the PICV model, where e is the base of natural algorithm. To this end, we first show that $\sigma(.)$ to be submodular and monotonic for the PICV model. Next, we apply the result on performance guarantee of greedy algorithm for monotonic and submodular function[24].

In the following, we state the submodularity and monotonicity properties of a function[22].

a. *Submodular:* For all elements v and all pair of sets S and T, where $S \subseteq T$, function $f(.)$ is submodular if the following condition is satisfied

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T) \quad (1)$$

b. *Monotone* : For all elements v and all sets S, function $f(.)$ is submodular if the following condition is satisfied.

$$f(S \cup \{v\}) \geq f(S) \quad (2)$$

We now prove $\sigma(.)$ to be submodular and monotonic for the PICV model.

Let $\sigma^0(S)$ denote the non-rumor active nodes after rumor relevance interval for the PICV model, where set S is the containment seed set. Minimizing the number of rumor active nodes $\sigma(S)$ is equivalent to maximizing non-rumor active nodes. In PICV mode, a node u activates v with probability p_{uv}. This activation can be viewed as an outcome of tossing a coin with a bias p_{uv} for a pair of vertices u and v. A random subgraph of $G(V, E)$ is generated in advance by considering *live* edges, which have got activated by the coin toss. This subgraph is called a *live arc graph*[25]. Any path in this live arc graph is called as a *live-edge path.* Let $Pr(G_L)$ denote the probability that live arc graph G_L is generated from the set of all possible live arc graphs $\mathcal{G}$. Let $\sigma^0_{G_L}(S)$, which is a deterministic quantity, denote the non-rumor active nodes for live arc graph G_L.

For positive information diffusion strategy, rumor containment seed nodes stop rumor influence from reaching a node x if a live-edge path from rumor activated nodes to node x that is blocked either by containment seed node directly or blocked by nodes positively activated by information diffusion from containment seed nodes. For blocking strategy, containment seed nodes stop rumor influence from reaching a node x if a live-edge path from rumor activated nodes to node x is blocked by containment seed node only. Consider containment seed set $S \cup \{v\}$, $v \in V \setminus (S^- \cup S)$. If a live-edge path from rumor nodes is blocked, it could be due to seed set S or seed node v or by both.

Lemma 1 *For the PICV model, $\sigma(.)$ is monotonic.*

Proof Consider a containment seed set S and a live arc graph G_L. Let us add a node v is added to the seed set S. As $\sigma^0_{G_L}(S \cup \{v\})$ denotes nodes that are reached by either S or v and $\sigma^0_{G_L}(S)$ denotes the nodes only reachable from S, $\sigma^0_{G_L}(S \cup \{v\}) \geq \sigma^0_{G_L}(S)$. As this is true for every $G_L \in \mathcal{G}$, $\sigma^0(.)$ and thereby $\sigma(.)$ is monotonic. □

Lemma 2 *For the PICV model, $\sigma(.)$ is submodular.*

Proof Let $R(v, G_L)$ denote all the nodes that can be reached by set v for live arc graph G_L.

Consider $\sigma^0_{G_L}(S \cup \{v\}) - \sigma^0_{G_L}(S)$. This denotes the nodes that are present in $R(v, G_L)$ but not in $R(S \cup \{v\}, G_L)$. For $S \subseteq T$, this number is greater than or equal to the nodes that are present in $R(v, G_L)$ but not in $R(T \cup \{v\}, G_L)$. Hence, $\sigma^0_{G_L}(S \cup \{v\}) - \sigma^0_{G_L}(S) \geq \sigma^0_{G_L}(T \cup \{v\}) - \sigma^0_{G_L}(T)$ and $\sigma^0_{G_L}$ is submodular.

From [22], $\sigma^0(S) = \sum_{G_L \in \mathcal{G}} Pr(G_L)\sigma^0_{G_L}(S)$.

As a nonnegative linear combination of submodular functions is submodular, $\sigma^0(.)$ is submodular. Hence, $\sigma(.)$ is submodular. □

Theorem 1 *For the PICV model, the greedy algorithm performs close to optimal algorithm within the factor $(1 - 1/e)$ where e is the base of natural algorithm.*

Proof We first state the result on greedy algorithm provided by [24] in the following.

Let us consider a nonnegative, montone, submodular function, function f and a set S^* such that $S^* = \arg\max_{|S| \leq k} f(S)$. Let set S_G be an initially empty set. Elements are added to set S_G one at time such that the added element provides the maximum marginal increase in the function value. Elements are added till S_G obtains a size k. Then, $f(S_G) \geq (1 - 1/e) f(S^*)$, where e is the base of natural logarithm.

Lemmas 1 and 2 and the above statement imply Theorem 1.

Theorem 2 *For the PLT model, the greedy algorithm performs close to optimal algorithm within the factor $(1 - 1/e)$ where e is the base of natural algorithm.*

Proof For the LT model, using live-arc graph argument, it is proved that the influence maximization is monotone and submodular[26]. The live-arc graphs generated for the PLT model are the same as that of the LT model. Therefore, $\sigma^0(.)$ is submodular and monotonic for the PLT model. Hence, $\sigma(.)$ is also submodular and monotonic for the PLT model.

Using the result on greedy algorithm [24] and submodularity and monotonicity of $\sigma(.)$ for the PLT model, we find that the greedy algorithm performs close to optimal algorithm within the factor $(1 - 1/e)$ where e is the base of natural algorithm. □

5. *Proximity-weight-degree :* We propose a low-complexity algorithm called proximity-weight-degree algorithm in which the containment seed node set is chosen on the basis of the proximity of the inactive nodes to the rumor nodes, edge weights of the links from the rumor nodes for the PLT model and the activation probabilities of the links from the rumor nodes in the PICV model, and the out-degree of the inactive nodes. The proximity-weight-degree-based selection algorithm is described in Algorithm 2.

There are two main motivations behind choosing a particular node as seed node. The first motivation is that a node that receives more influence from rumor nodes is more likely to get infected with rumor and should be preferred as a seed node. The second motivation is that a node that has the maximum number of inactivated out-neighbors is more like to spread rumor to more nodes if gets infected and should be preferred as a seed node. In the proposed scheme, we combine both the motivations using a height function. In this algorithm, first neighbors of rumor-active nodes are searched in lines $5-7$. Any inactive out-neighbor v of the rumor node has a height $H(v)$. $H(v)$ captures the influence of the edge weights of rumor links on v for the PLT model and the activation probabilities of the links from the rumor nodes in the PICV model and its inactive out-neighbors. Node-height calculation is given in lines $9-17$ of Algorithm 2. A node's height is $\sum_v$ (incoming edge-weights from rumor active nodes) $\times$(number of its inactive out-neighbors plus one) for the PLT model. For the PICV model, a node's height is given by $\sum_v$ (activation probabilities of the links from rumor active nodes) $\times$ (number of its inactive out-neighbors plus one).

The top K nodes with maximum height are chosen as nodes for the set S_C. If the set S_C thus formed has size less than K, the rest of nodes for set S_C are chosen randomly from the set of inactive nodes. This is shown in line 20 of the algorithm. The complexity is $O(|V|^2)$ because there are two nested for loops as shown in Algorithm 2 and each loop runs for at the most total number of vertices.

Algorithm 2: PWD($G(V, E), K, S_C, DiffusionModel$)

1 initialize $S_C = \phi$
2 initialize $H(v) = 0,\ \forall v \in V$
3 initialize $Y = \phi$
4 **for** $u \in S^-$ **do**
5   **for** $v \in \mathrm{N}^{\mathrm{out}}(u)$ **do**
6   **if** $v \notin S^-$ **then**
7       $Y = Y \cup \{v\}$
8   **if** *Diffusion Model == PLT Model* **then**
9         $H(v) = H(v) + w_{uv}$
10   **else if** *Diffusion Model == PICV Model* **then**
11         $H(v) = H(v) + p_{uv}$
12
13
14 **for** $u \in Y$ **do**
15   $d = 0$
16   **for** $v \in \mathrm{N}^{\mathrm{out}}(u)$ **do**
17   **if** $v \notin S^-$ **then**
18       $d = d + 1$
19
20   $H(u) = H(u) \times (d + 1)$
21 **for** $i = 1$ *to* K **do**
22   select $u = \arg\max_{v \in V \setminus (S^- \bigcup S_C)} H(v)\ S_C = S_C \bigcup \{u\}$;
23 $S' = \mathrm{RandomSelect}(V \setminus (S_C \bigcup S^-), K - |S_C|)$;
24 $S_C = S_C \bigcup S'$
25 **return** S_C;

Comparison of containment seed selection algorithms Random seed selection algorithm is a simple linear complexity algorithm, which requires no information of regarding the OSN features such degree, edge weights, and network diameter. However, it fails to effectively utilize the available network knowledge for better rumor containment. Max degree seed selection algorithm is of polynomial time (quadratic time) complexity and is based on the intuition that selecting seed nodes with higher degree has a better chance of rumor containment as the seed nodes are connected to a larger of nodes in the OSN. Unlike random selection, max-degree selection requires the knowledge of the nodes' degrees of the OSN.

Proximity heuristic algorithm is of quadratic complexity and is based on the intuition that selecting the containment nodes closer to the rumor seed nodes has a better chance of curtailing rumor spread. This is because the containment seed nodes will have their influence in the proximity of the rumor seed nodes and try to stop the rumor diffusion at the source. In the proximity heuristic algorithm, the containment seed nodes are the out-neighbors of the rumor nodes, which have higher incoming edge weights/probabilities from rumor seed nodes. In comparison with random and max-degree selection, proximity heuristic algorithm requires information of the rumor seed nodes and their neighborhood.

In case of greedy algorithm, every containment seed node is selected by simulating the rumor diffusion process through the OSN across several sample paths. Hence, it is a high-complexity polynomial time algorithm. Greedy works on the intuition that simulation of the rumor diffusion process provides a better understanding for effective containment seed selection. In comparison with random, max-degree and proximity heuristic, greedy algorithm requires the knowledge of the entire OSN and not just the network graph properties.

PWD algorithm is a quadratic complexity algorithm which selects the containment seed nodes from the set of out-neighbors of the rumor seed nodes. In addition to incoming edge weights/probabilities from the rumor seed nodes to this set of nodes, the number of inactive out-neighbors of this set of nodes is also considered in selecting the containment seed nodes. This is based on the intuition that choosing containment nodes that have higher number of inactive out-neighbors and are closer to the rumor nodes helps in curtailing the rumor spread near the rumor sources and much earlier in time. When compared to random and max-degree selection algorithm, PWD algorithm requires knowledge of the neighborhood of the rumor seed nodes. When compared to the greedy algorithm, it is much faster as it based on the local sub-graph knowledge and not extensive simulations.

Table 2 Description of datasets

	email-enron-only	soc-firm-high-tech	ca-GrQc	facebook
Nodes	143	33	5242	4039
Edges	623	91	14496	88234
Average degree	8.71	5.52	5.53	43.69
Average clustering coefficient	0.434	0.453	0.530	0.606

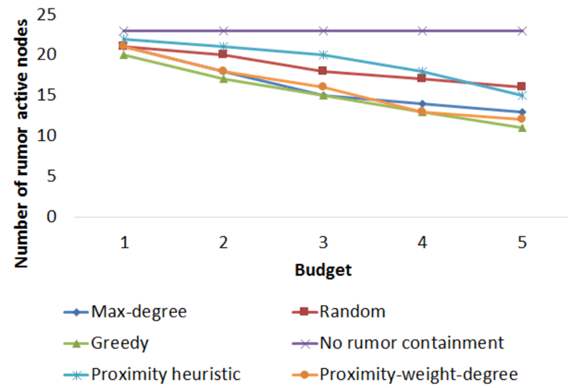
5 Experimental results

Monte Carlo simulations are executed for 1000 sample paths for the soc-firm-high-tech dataset and the email-enron-only dataset and 100 sample paths for the remaining datasets [10] to determine the performance of the algorithms given in Sec. 4. We use real-world datasets: email-enron-only, soc-firm-hi-tech, ca-GrQc (General Relativity and Quantum Cosmology collaboration network) and facebook [27] for the simulations. The dataset description is given in Table 2. For the PLT model, we assume that the edge weight $w_{uv} = \frac{1}{\deg^{in}(v)}\ \forall\ u, v \in V$ similar to [10]. For the PICV model, we consider the activation probability of edge (u, v) to be 0.1.

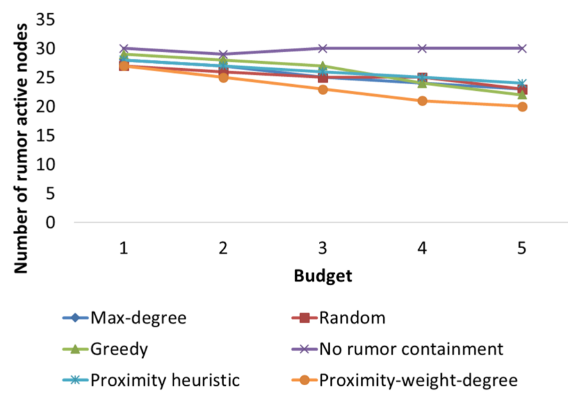
Between any two nodes which are directly connected (say u and v), there are two directed edges (from u to v and from v to u). The rumor seed nodes are chosen as per max-degree algorithm [4].

The simulations in Sect. 5.1 show the effect of positive seed selection algorithms on rumor containment in PLT model. The simulations in Sect. 5.2 show the effect of blocking of nodes on rumor containment in PLT model. Similarly, Sect. 5.3 shows the effect of positive seed selection algorithms on rumor containment in PICV model and Sect. 5.4 shows the effect of blocking of nodes on rumor containment in PICV model. In Sect. 5.1, the simulations show the effect of change in budget and change in rumor reference interval on number of rumor-active nodes for each algorithm.

To limit the number of numerical result plots, in the remaining simulations (Sect. 5.2, Sect. 5.3 and Sect. 5.4), we have simulated the change in rumor reference interval on number of rumor-active nodes for one dataset and the change in budget on number of rumor-active nodes for another dataset. Like several papers[4,10], we have used the greedy algorithm as a benchmark for performance comparison of the proposed PWD algorithm.



(a) Dataset: soc-firm-high-tech. Number of rumor seed nodes is 5.



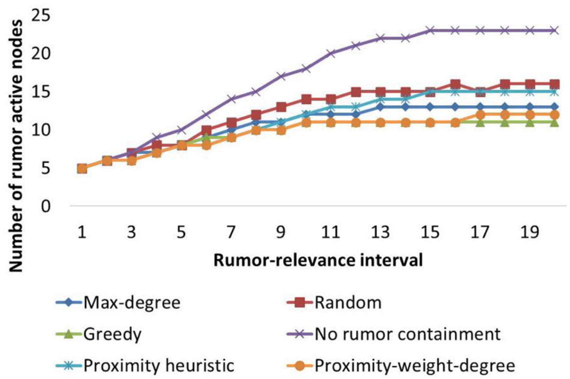
(b) Dataset: email-enron-only. Number of rumor seed nodes is 5.

Fig. 9 Variations in average number of rumor-active nodes with the budget for the PLT model with positive information diffusion for rumor containment and T= 20 time-steps

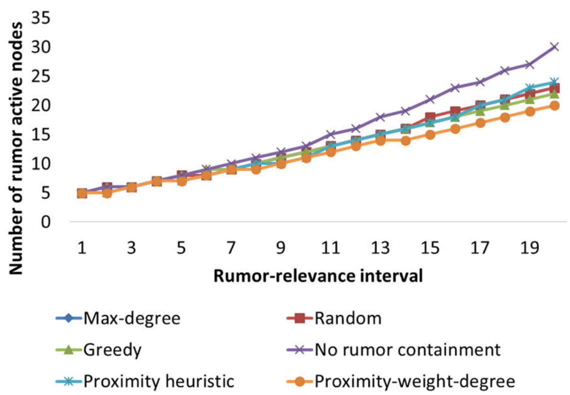
5.1 Rumor containment in PLT model by positive information diffusion

5.1.1 Effect of K on average number of rumor-active nodes

Figure 9a and 9b plots the average number of rumor-active nodes as a function of number of positive seed nodes for the soc-firm-hi-tech and the email-enron-only datasets, respectively. From Figs. 9a and b, we find that as the number of positive seed nodes increases, the average number of rumor-active nodes either remains the same or decreases for all the algorithms. This is because the more the number of positive seed nodes, the more the positive information diffusion in the network. The greedy algorithm performs the best. For soc-firm-hi-tech dataset, the performance of the proposed algorithm is close to greedy. The difference in performance of the proposed algorithm for the soc-firm-hi-tech and the



(a) Dataset: soc-firm-high-tech. Number of rumor seed nodes is 5 and the budget is 5.



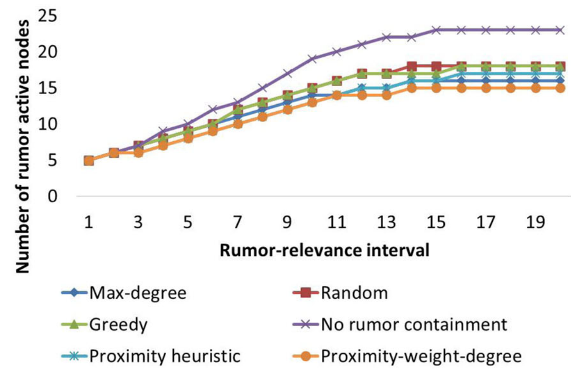
(b) Dataset: email-enron-only. Number of rumor seed nodes is 5 and the budget is 5.

Fig. 10 Variations in average number of rumor-active nodes with the rumor-relevance interval for the PLT model with positive information diffusion for rumor containment

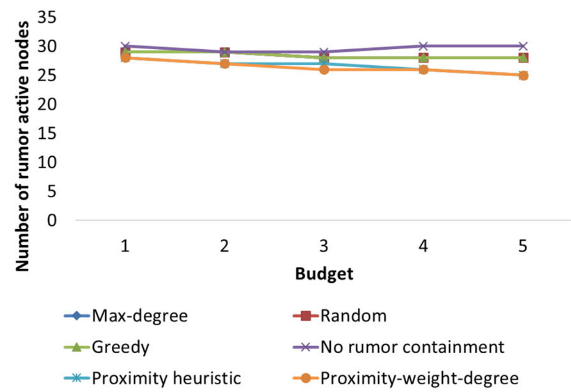
email-enron-only datasets is because the networks have different internal structures.

5.1.2 Effect of T on average number rumor-active nodes

Figure 10a and 10b plots the average number of rumor-active nodes as a function of rumor-relevance interval for the soc-firm-hi-tech and the email-enron-only datasets, respectively. We find that as steps increase, the average number of rumor-active nodes increases for all the algorithms. This is because the more the time, the more the rumor diffusion in the network. Further, at high values of T, we observe that the number of rumor-active nodes reaches a limit. This is because the rumor-active nodes have already diffused the rumor information by this time and there are no further diffusion by the rumor-active nodes. We also observe that the proximity-weight-degree algorithm performs well.



(a) Dataset: soc-firm-high-tech. Number of rumor seed nodes is 5 and the budget is 5.



(b) Dataset: email-enron only. Number of rumor seed nodes is 5 and $T = 20$ time-steps

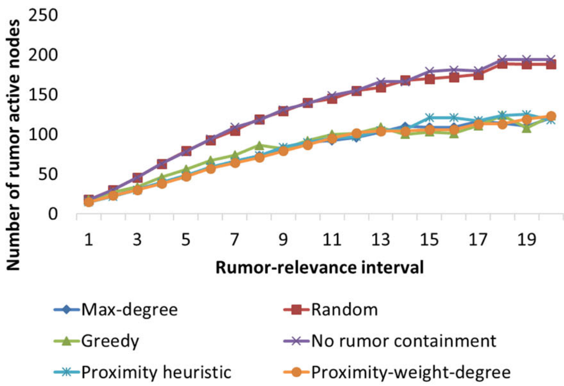
Fig. 11 Average number of rumor-active nodes for the PLT model with blocking for rumor containment

5.2 Rumor containment in PLT model by blocking

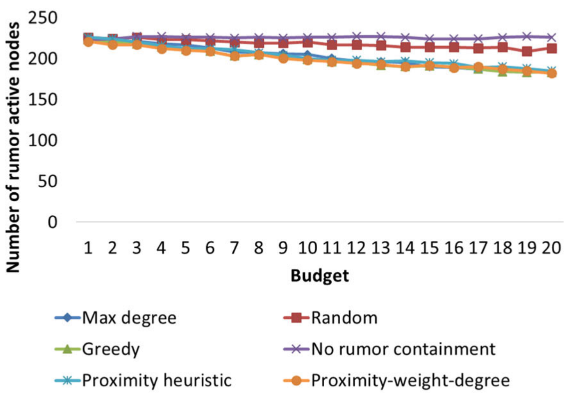
Figure 11a and 11b plots the average number of rumor-active nodes as a function of rumor relevance interval for the soc-firm-high-tech data set and the average number of rumor-active nodes as a function of number of blocked seed nodes for the email-enron-only dataset, respectively. We observe that as the budget increases, the number of rumor-active nodes decreases because more nodes in the OSN are blocked and do not spread the rumor further. We observe that as the number of time-steps increases, the number of rumor-active nodes increases as with time the rumor diffuses more in the OSN. The plots show that performance of our algorithm is the best and it outperforms the greedy algorithm. This is because greedy algorithm is not an optimal algorithm and for some datasets, our algorithm performs better than greedy.

Comparing Fig. 10a with Fig. 11a and comparing Fig. 9b with Fig. 11b, we observe that positive information diffusion performs better as compared to blocking for the same budget and the same rumor relevance interval. This is because a positively activated node not only blocks rumor but also makes

International Journal of Data Science and Analytics (2022) 13:185–198

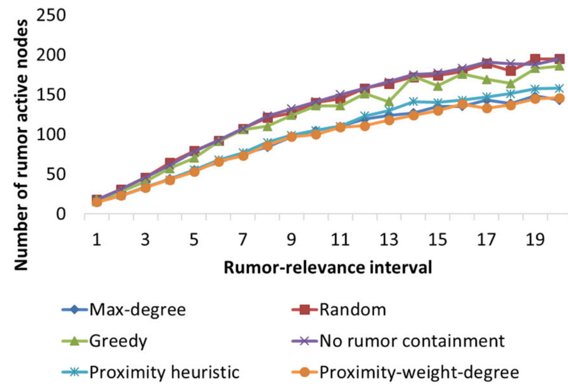


(a) Dataset: ca-GrQc. Number of rumor seed nodes is 10 and the budget is 20.

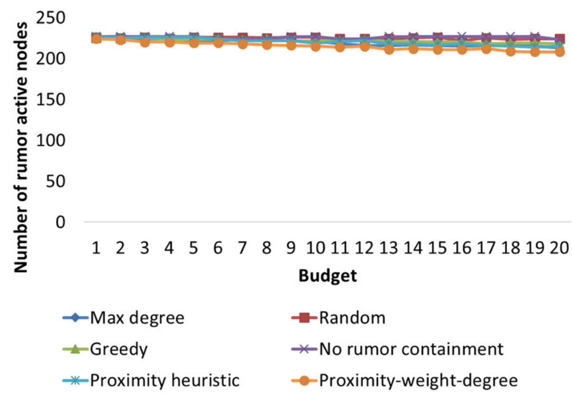


(b) Dataset: facebook. Number of rumor seed nodes is 10 and $T = 5$ time-steps.

Fig. 12 Average number of rumor-active nodes for the PICV model with positive information diffusion for rumor containment



(a) Dataset:ca-GrQc. Number of rumor seed nodes is 10 and number of blocked seed nodes is 20.



(b) Dataset: facebook. Number of rumor seed nodes is 10 and $T = 5$ time-steps.

Fig. 13 Average number of rumor-active nodes for PICV model with blocking for rumor containment

other nodes positively activated. So if we start with a few rumor containment nodes, in case of blocking, the number of nodes that block rumor remains the same, but in case of positive information diffusion, the number of nodes that block rumor (positively activated nodes) also increases. Therefore, spreading positive information is more effective than just blocking nodes.

5.3 Rumor containment in PICV model by positive information diffusion

Figure 12a and b plots the average number of rumor-active nodes as a function of rumor relevance interval for the ca-GrQc data set and the average number of rumor-active nodes as a function of number of positive seed nodes for the facebook dataset, respectively. Again, we observe that the number of rumor-active nodes decreases with the budget and increases with the rumor relevance interval, which is due to reasons given in Sect. 5.1. The plots show that the performance of the greedy is the best. We also observe that performance of the proposed algorithm is close to the greedy algorithm.

5.4 Rumor containment in PICV model by blocking

Figure 13a and 13b plots the average number of rumor-active nodes as a function of rumor relevance interval for the ca-GrQc data set and the average number of rumor-active nodes as a function of number of blocked seed nodes for the facebook dataset, respectively. We observe that the number of rumor-active nodes decreases with the budget and increases with the rumor relevance interval, which is due to reasons given in 5.2. The plots show the performance of the proposed scheme is the best. It even performs better than the Greedy algorithm for some simulations.

Figures 12a and Figure 13a are identical. Similarly, Figs. 12b and 13b are identical. This shows blocking of nodes is nearly as effective as positive cascade. However, positive information diffusion performs better than blocking for some cases.

Summary of the numerical results Our numerical results highlight the following points on rumor containment in peer-to-peer models.

- The proposed PWD algorithm for seed selection for rumor containment performs better than several low-complexity baseline algorithms such as max-degree, random, and proximity heuristic.
- The performance of the PWD algorithm is close to that of the high-complexity benchmark greedy algorithm.
- For all the algorithms, rumor containment with positive information strategy is as good as blocking strategy, or better.
- For all the algorithms, at large values of rumor relevance interval, the average number of rumor active nodes reaches a limit and does not increase further.

6 Conclusion

In this paper, we applied positive information diffusion and blocking for rumor containment in an OSN, where the rumor spreads through peer-to-peer messages. We proposed a low-complexity proximity-weight-degree algorithm for selecting the seed nodes for rumor containment. Our experimental studies, which used real-world datasets, showed that the performance of the proposed proximity-weight-degree algorithm is on par with the greedy algorithm. Further, we observed that the proposed algorithm performs better than other low-complexity algorithms like max-degree, random selection, and proximity heuristic. We also observed that among the two rumor containment schemes, blocking nodes perform nearly as good as spreading positive information in most of the simulations. In a few simulations, spreading positive information performs better than blocking nodes.

Some improvements and directions on future work are as follows. Design of a low-complexity intricate seed selection algorithm with probabilistic guarantees on lines similar to [28] is an interesting future research avenue. Rumor containment with adaptive containment seed node selection, in which the containment seed nodes are selected one by one at different time-steps based on how the rumor progresses in P2P OSN models. The rumor diffusion and containment strategies can also be investigated when a user sends peer-to-peer rumor messages to a subset of its neighbors rather than all of them.

Declarations

Conflict of Interest We have no conflicts of interest or competing interests to declare.

References

1. Tripathi, R., Rao, S.: Positive information diffusion for rumor containment in online social networks. In: Proc. COMSNETS, pp. 610–613 (Jan. 2020)
2. https://www.pewresearch.org/fact-tank/2017/11/02/more-americans-are-turning-to-multiple-social-media-sites-for-news/, (2017)
3. https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/, (2019)
4. Tong, G., Wu, W., Du, D.: Distributed rumor blocking with multiple positive cascades. IEEE Trans. Comput. Soc. Syst. **5**(2), 468–480 (2018)
5. https://foreignpolicy.com/2009/04/25/swine-flu-twitters-power-to-misinform/, (2009)
6. Zubiaga, A., Ji, H.: Tweet, but verify: epistemic study of information verification on twitter. Soc. Netw. Anal. Min. **4**(1), 163 (2014)
7. Kimura, M., Saito, K., Motoda, H.: Blocking links to minimize contamination spread in a social network. ACM Trans. Knowl. Discov. Data **3**(2), 9:1-9:23 (2009)
8. Wang, S., Zhao, X., Chen, Y., Li, Z., Zhang, K., Xia, J.: Negative influence minimizing by blocking nodes in social networks. In: Proc. Workshops 27th AAAI Conf. Artif. Intell. (2013)
9. Budak, C., Agrawal, D., El Abbadi, A.: Limiting the spread of misinformation in social networks. In: Proc, pp. 665–674. New York, NY, USA, ACM World Wide Web (2011)
10. He, X., Song, G., Chen, W., Jiang, Q.: Influence blocking maximization in social networks under the competitive linear threshold model. In: Proc. SIAM Data Mining, pp. 463–474 (2012)
11. Wang, B., Chen, G., Fu, L., Song, L., Wang, X.: Drimux: Dynamic rumor influence minimization with user experience in social networks. IEEE Trans. Knowl. Data Eng. **29**(10), 2168–2181 (2017)
12. Khalil, E., Dilkina, B., Song, L.: Cuttingedge: influence minimization in networks. In: Proc. NIPS Workshop Frontiers Netw. Analysis: Meth., Models, and Appl. NIPS (2013)
13. Doerr, B., Fouz, M., Friedrich, T.: Why rumors spread so quickly in social networks. Commun. ACM **55**(6), 70–75 (2012)
14. Liang, G., He, W., Xu, C., Chen, L., Zeng, J.: Rumor identification in microblogging systems based on users behavior. IEEE Trans. Comput. Social Syst. **2**(3), 99–108 (2015)
15. Luo, W., Tay, W.P., Leng, M.: Infection spreading and source identification: A hide and seek game. IEEE Trans. Signal Process. **64**(16), 4228–4243 (2016)
16. Guo, J., Chen, T., Wu, W.: A multi-feature diffusion model: rumor blocking in social networks. IEEE/ACM Transactions Netw. **29**(1), 386–397 (2021)
17. Song, C., Hsu, W., Lee, M.L.: Temporal influence blocking: Minimizing the effect of misinformation in social networks. In: Proc. IEEE ICDE, pp. 847–858 (Apr. 2017)
18. He, Z., Cai, Z., Yu, J., Wang, X., Sun, Y., Li, Y.: Cost-efficient strategies for restraining rumor spreading in mobile social networks. IEEE Trans. Veh. Technol. **66**(3), 2789–2800 (2017)
19. Manouchehri, M. A., Helfroush, M. S., Danyali, H.: "Temporal Rumor Blocking in Online Social Networks: A Sampling-Based Approach," in IEEE Trans. on Syst., Man, and Cybern.: Syst., Early Access (2021), https://doi.org/10.1109/TSMC.2021.3098630
20. Tong, G.A., Wu, W., Guo, L., Li, D., Liu, C., Liu, B., Du, D.-Z.: An efficient randomized algorithm for rumor blocking in online social networks. In: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, pp. 1–9 (2017)
21. Yang, L., Li, Z., Giua, A.: Containment of rumor spread in complex social networks. Information Sciences, **506**, 113–130 (2020). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519306607

22. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proc, ACM Knowl. Discovery Data Mining. New York, NY, USA, pp. 137–146 (2003)
23. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proc. ACM SIGKDD. ACM, pp. 199–208 (2009)
24. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions. Math. Programm. **14**(1), 265–294 (1978)
25. Chen, W., Lakshmanan, L.V., Castillo, C.: Information and influence propagation in social networks. Synth. Lect. Data Manag. **5**(4), 1–177 (2013)
26. Chen, W., Castillo, C., Lakshmanan, L. V. S.: Information and Influence Propagation in Social Networks, Morgan & Claypool, 2013
27. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: AAAI, (2015). [Online]. Available: http://networkrepository.com
28. Borgs, C., Brautbar, M., Chayes, J., Lucier, B.: Maximizing social influence in nearly optimal time. In: In Proc. SODA. SIAM, pp. 946–957 (2014)