Information diffusion is efficient via gossip or rumor spreading in many of the next generation networks. It is of great importance to select some seed nodes as information sources in a network so as to maximize the gossip spreading. In this paper, we deal with the issue of the selection of information sources, which are initially informed nodes (i.e., seed nodes) in a network, for pull-based gossip protocol. We prove that the gossip spreading maximization problem (GSMP) is NP-hard. We establish a temporal mapping of the gossip spreading process using virtual coupon collectors by leveraging the concept of temporal network, further prove that the gossip spreading process has the property of submodularity, and consequently propose a greedy algorithm for selecting the information sources, which yields a suboptimal solution within $(1 - 1/e)$ of the optimal value for GSMP. Experiments are carried out to study the spreading performance, illustrating the significant superiority of the greedy algorithm over heuristic and random algorithms.

1. Introduction

Information dissemination through networks is ubiquitous in the modern world [1–6], and gossip or rumor spreading is an efficient way of information diffusion: imagine that a rumor arises in a town and is epidemically like spread among the whole population [7]. There are two atomic types of gossip protocols: “pull,” an uninformed node requests an unpossessed message from a randomly chosen neighbor, and “push,” an informed node sends its possessed message to a randomly chosen neighbor. Gossip-based algorithms are simple, robust, flexible, and scalable and hence are promising for many of the next generation networks [8]. Existing applications are numerous, such as consensus and averaging problems in sensor networks [9, 10], ad hoc message routing [11], peer-to-peer (P2P) file distribution [12], and information dissemination in social networks [13].

Most existing analytical works on gossip spreading have dealt with (high probability) upper bounds of the completion time, disregarding the choice of information sources [7, 8, 12–15]. Given a static connected network graph, the completion time of a gossip protocol is the first time at which all the nodes are informed. Recently, the issue on the selection of information sources for gossip spreading has been treated in our previous work [16]. However, in [16], only the push-based gossip protocol was considered and the complexity issue of the gossip spreading maximization problem was left open.

In this paper, we consider the problem of selecting information sources for gossip spreading, focusing on the pull-based gossip protocol. The information sources are initially informed nodes (i.e., seed nodes) in a network. We ask the question: given a general network graph, the budgeted size of a seed set, and a constrained deadline, how to pick the elements in the seed set, which will be endowed with an identical message to be propagated to the rest of the network, such that the expected number of informed nodes is maximized within the deadline. The gossip spreading maximization problem (GSMP) frequently arises in many scenarios, especially when a popular content is demanded by a group of nodes. In sensor networks, one needs to decide the deployment of key sensors, capable of detecting and issuing emergence with the other hearing-and-forwarding sensors, so as to maximize the alarming area as quickly as possible. In P2P networks, one needs to decide the best choice of seeds so as to maximize the file distribution before delay tolerance. Besides, similar research topics can be found in viral marketing (a.k.a. influence maximization), detection of disease outbreak, and opportunistic cellular traffic offloading [17–19].

Our contributions are as follows. (1) We prove that GSMP is NP-hard, which means that suboptimal solutions
with performance guarantee should be exploited because polynomial-time algorithms have not yet been discovered for the class of NP-hard problems [20]. (2) We establish a temporal mapping of the pull-based gossip spreading process using virtual coupon collectors by leveraging the concept of temporal network and prove the submodularity in gossip spreading based on this temporal mapping method. Consequently, we propose a greedy algorithm for GSMP that yields a solution within \((1 - 1/e)\) of the optimal value. (3) We carry out extensive experiments to study the spreading performance, demonstrating that the greedy algorithm outperforms heuristic and random algorithms significantly.

Our work differs from previous works in several aspects. First, we deal with GSMP by selecting influential information sources, different from the analytical works on high-probability completion time (see, e.g., [8]). Second, we study GSMP in the area of gossip spreading and treat it using the coupon collecting and temporal mapping method, while GSMP in the area of gossip spreading and treat it using probability completion time (see, e.g., [8]). Second, we study GSMP in the area of gossip spreading.

In general, a directed network graph \(G = (V, E)\) consists of a set of nodes \(V\) and a set of directed edges \(E\). Each node is spoken of being either informed or uninform; and a node is informed if and only if it has possessed its desired message. For any pair of nodes \(u, v \in V\), \(u\) may send its possessed message to \(v\) if and only if the edge \((u, v) \in E\).

Time is slotted and a message can be transferred from a sender to a receiver within a time slot, which is called a round throughout this paper. No matter which content the information flow over the underlying network is carrying, we focus on only one piece of message in our model.

2.2. Problem Formulation. Consider the sample space \(\Omega\), where each sample specifies one possible realization of the gossip spreading process. Let \(X\) denote one sample in \(\Omega\), and let \(\Pr[X]\) denote its occurrence probability. We are interested in the case where the gossip spreading process runs until a constrained deadline. Given a seed set \(S\) of \(k\) nodes and a deadline of \(D\) rounds, the number of informed nodes under one sample \(X \in \Omega\) by deadline \(D\) is \(I_D(S | X)\). So the expected number of informed nodes (within \(D\) rounds) is

\[
\sigma_D(S) := E[I_D(S)] = \sum_{X \in \Omega} \Pr[X] \cdot I_D(S | X).
\]

Given a directed network graph \(G = (V, E)\), the budgeted size \(k\) of a seed set \(S\), and a constrained deadline \(D\), we wish to select the seed nodes in \(S\) such that \(\sigma_D(S)\) is maximized. This is called gossip spreading maximization problem (GSMP) and is formally given by

\[
\max_{S \subseteq V} \sigma_D(S) \quad \text{subject to } |S| \leq k.
\]

3. Complexity Analysis

GSMP belongs to the field of stochastic programming, and we show that it is NP-hard in this section.

3.1. Preliminary. First of all, we consider the decision version of GSMP.

Problem 1 (gossip spreading decision problem). Given a network graph \(G = (V, E)\), a constrained deadline \(D\), and a utility quota \(q\), we wish to determine whether there exists \(I\) of the nodes for the seed set \(S\) such that the expected number of informed nodes \(\sigma_D(S)\) is at least \(q\). Let an instance of Problem 1 be denoted by \(GSP(V, E, D, I, q)\).

We see that \(GSP(V, E, D, I, q)\) belongs to the class NP [20], since it can be validated in polynomial time given any solution of \(I\) seed nodes. In order to argue the NP-hardness of GSMP, we will show that its decision version (i.e., \(GSP(V, E, D, I, q)\)) can be reduced from the following problem.
Problem 2 (partial set cover problem). Given a ground set $U = \{u_1, u_2, \ldots, u_n\}$ and a collection of $U$'s subsets $C = \{S_1, S_2, \ldots, S_m\}$, we wish to determine whether there exist $h$ of the subsets such that the cardinality of their union is at least $p$. Let an instance of Problem 2 be denoted by SCP($U, C, h, p$). As the partial set cover problem generalizes the NP-complete set cover problem [20], it must be NP-complete.

3.2. GSMP is NP-Hard. Next we show that GSMP is NP-hard using a reduction from the NP-complete SCP($U, C, h, p$) to GSP($V, E, D, I, q$). Note that a bipartite graph is leveraged in the following proof, and similar techniques had been widely used in the complexity analysis literature, such as [17, 21, 22].

Theorem 3. The gossip spreading maximization problem (GSMP) for the pull-based gossip protocol is NP-hard.

Proof. Consider an arbitrary instance of the partial set cover problem SCP($U, C, h, p$) with $n$ elements $U = \{u_1, u_2, \ldots, u_n\}$ and $m$ subsets $C = \{S_1, S_2, \ldots, S_m\}$ of $U$, and construct a directed bipartite graph $G^* = (V^*, E^*)$ as follows. The node set $V^*$ contains $n + m$ nodes, in which a node $v_i$ ($i = 1, \ldots, m$) is corresponding to a subset $S_i$ and a node $v_j$ ($j = 1, \ldots, n$) is corresponding to an element node $u_j$. Each $v_i$ is called subset node and each $v_j$ is called element node hereafter. There is a directed edge from a subset node $v_i$ to an element node $v_j$ if $u_j \in S_i$; for example, see Figure 1. In the following, we will see that solving an arbitrary instance SCP($U, C, h, p$) of the partial set cover problem is equivalent to solving a special-case instance GSP($V^*, E^*, \infty, h, h + p$) of the gossip spreading decision problem, and we assume $h < m$ without loss of generality.

If we can find $h$ of the subsets in $C$ such that the cardinality of their union is at least $p$, then we will show that $h$ nodes can be found in $G^*$ for the seed set $S$ such that $\sigma_{oo}(S) \geq h + p$ with the deadline $D$ being infinity. For each of these $h$ selected subsets for SCP($U, C, h, p$), we use the corresponding subset node in $G^*$ as a seed node; eventually, at least $p$ element nodes can pull the desired message from their subset nodes via gossip spreading given the infinite deadline; that is, $\sigma_{oo}(S) \geq h + p$.

Conversely, if we can find $h$ of the nodes in $G^*$ for the seed set $S$ such that $\sigma_{oo}(S) \geq h + p$, then we will show that $h$ subsets can be found in $C$ such that the cardinality of their union is at least $p$. For each $v$ of these $h$ selected seed nodes for GSP($V^*, E^*, \infty, h, h + p$), if $v$ is an element node, then we replace it with a subset node as follows. If the subset node $v'$ that points to $v$ either has already been selected as a seed node or has already replaced other seed nodes, then we replace $v$ with any other available subset node; otherwise, we replace $v$ with this subset node $v'$. After all of these possible replacements, we have obtained subset nodes in $G^*$ as seed nodes, and $\sigma_{oo}(S) \geq h + p$ is clearly still satisfied. Therefore, at least $p$ elements can be covered using $h$ subsets in $C$, which are exactly corresponding to these $h$ subset nodes in $G^*$.

In total, if the gossip spreading decision problem can be solvable, then the partial set cover problem must be solvable; that is, the decision version of GSMP is at least as hard as the NP-complete partial set cover problem.

Remark 4. The above arguing method can be applied in analyzing the complexity of GSMP under the “push” model. Since GSMP is NP-hard and polynomial-time algorithms have not yet been discovered for the class of NP-hard problems [20], we should exploit suboptimal solutions with performance guarantee.

4. Submodularity and Greedy Algorithm

In this section, we establish a temporal mapping of the gossip spreading process using virtual coupon collectors by leveraging the concept of temporal network [23]. This treatment provides a tractable way to recognize the submodularity in gossip spreading and leads to a greedy algorithm which yields a solution to GSMP within $(1 - 1/e)$ of the optimal value.

4.1. Preliminary. Before the analysis, we introduce the preliminaries on the temporal network, the shortest time-respecting path, and the live diffusion path.

A temporal network embodies the information of when events occur in dynamic systems [23]. For the case of gossip spreading, the edge between any two interacting nodes is endowed with the information of contact times when these two nodes share message. For example in Figure 2, the weights on the edge from node $v_i$ to node $v_j$ indicates that $v_i$ sends data to $v_j$ in rounds $t = 6, 7, 11$. The key is to consider the causality constraints of the time sequences of nodes’ contacts [23]; for example, in Figure 2, $v_a$ cannot transmit data to $v_d$ even if there are contacts between $v_a$ and $v_d$ as well as between $v_b$ and $v_d$, since the contacts of $v_b$ and $v_d$ occur before those of $v_a$ and $v_b$.

Consider a directed temporal graph $G^T = (V, E^T, W^T)$ for the gossip spreading process over a network $G = (V, E)$, where $W^T$ is the set of weights on the set of directed edges $E^T$, indicating the time information of nodes’ contacts. According to the causality constraints of time sequences, a time-respecting path $\mathcal{P}_{uv}^T$ from $u$ to $v$ is given by

\[
\langle \mathcal{P}_{uv}^T : v_0 = u, v_1, \ldots, v_M = v | w_{v_i, v_{i+1}} \geq 0, i \leq M - 1 \rangle,
\]  

(3)

where the weight $w_{v_i, v_{i+1}}$ is the time at which $v_i$ sends data to $v_{i+1}$, and the weights of successive edges on the path $\mathcal{P}_{uv}^T$ must be strictly increasing; that is, $w_{v_i, v_{i+1}} < w_{v_{i+1}, v_{i+2}}$ for all $0 \leq i \leq M - 2$. Let $T_u$ be the informed time of user $u$, that
is, the first time at which \( u \) becomes informed; then for \( \mathcal{P}_{uv}^T \), defined in (3), its length (i.e., the distance) is defined as

\[
d(u, v) := \text{distance}(\mathcal{P}_{uv}^T) = w_{v_{M-1}, v_M} - T_u.
\] (4)

In particular, the shortest time-respecting path \( \mathcal{P}_{uv}^T \) from \( u \) to \( v \) is given by

\[
\mathcal{P}_{uv}^T = \arg \min_{\mathcal{P}_{uv}^T \in \mathcal{P}^T} w_{v_{M-1}, v_M} - T_u.
\] (5)

Given a node set \( S \), we say a node \( v \) is reachable if either \( v \in S \) or there exists a time-respecting path from one node in \( S \) to \( v \) on \( G^T \); otherwise, it is unreachable. The distance \( d(S, v) \) from \( S \) to a reachable node \( v \) on \( G^T \) is

\[
d(S, v) = \min_{u \in S} d(u, v),
\] (6)
in which \( d(u, v) \) is the length of the shortest time-respecting path from the node \( u \) to the node \( v \), and \( d(v, v) = 0 \). For an unreachable node \( v \), \( d(S, v) = \infty \).

For an ordinary path \( \langle \mathcal{P}_{uv} : v_0 = u, v_1, \ldots, v_M = v \rangle \) on the network \( G \), we say it is a live diffusion path, if \( u \in S \) becomes informed initially in round \( t = 0 \), \( v_{i+1} \) succeeds in pulling its desired message from its neighbor \( v_i \) for each \( 0 \leq i \leq M - 1 \), and the considered node \( v \) is finally informed. Note that \( v \) is reachable from \( S \) via the seed node \( u \).

4.2. A Temporal Mapping. In the following, we establish a temporal mapping of the pull-based gossip spreading process by constructing a directed temporal graph. Note that all the multiple weights on each edge in the temporal graph are absolute time since the gossip spreading process starts up initially, and the temporal mapping method used in this paper is different from the equivalent view method used in our previous work [16]. These two methods are not simply coupled, and the "pull" model brings in new ingredients to GSMP.

Consider an arbitrary node \( v \), which attempts to pull its desired message from its neighbors in each round since the beginning of round \( t = 1 \), as long as it is uninformed. Note that the pulling process of \( v \) from its neighbors \( N_v \) is exactly a coupon collecting process [24], and denote this process using \( CC(v) \). In \( CC(v) \), \( v \) has \( |N_v| \) different coupons to collect, and in each round each of these coupons is collected uniformly and independently at random with replacement. For the node \( v \), let \( Z_t(v) \) denote the stochastic process indicating the coupon collected in its \( CC(v) \) in round \( t \) (\( t \geq 1 \)). The event that a certain coupon \( u \) is collected in round \( t \) means that the corresponding neighbor \( u = Z_t(v) \) is pulled from in round \( t \) by the node \( v \). Note that in the above described \( CC(v) \), we do not care whether the message-pulling node \( v \) has already possessed the message or not.

Given the constrained deadline \( D \), for each node \( v \), we independently run \( CC(v) \) till the deadline \( D \) is reached and record all the time stamps for each collected coupon when \( v \) collects it every time. Therefore, the set \( \tau_{u,v} \) of time stamps for a neighbor \( u \) which has been contacted by \( v \) can be written as

\[
\tau_{u,v} = \{ t : Z_t(v) = u; t \geq 1 \}.
\] (7)

After all the coupon collecting processes \( \{CC(v), v \in V\} \) are completed, a directed temporal graph \( G^T = (V, E^T, W^T) \) is thus constructed; for example, see Figure 3. In \( G^T \), the set \( E^T \) of directed edges contains just the incoming edges from those nodes in \( Z_t(v) \) with \( 1 \leq t \leq D \) for each \( v \in V \), and the set \( W^T \) of edge weights is given by

\[
\forall \text{ directed edge } (u, v) \in E^T, \quad \text{weight set } w_{u,v} = \tau_{u,v}.
\] (8)

Leveraging the constructed temporal graph \( G^T \) as above, we have Theorem 5. Note that the informed time \( T_v \) of a reachable node \( v \) is the first time in which \( v \) becomes informed. In addition, the above-assumed pulling process \( CC(v) \) of a node \( v \) after it becomes informed is no longer effective; that is, the attempts of \( v \) to pull its desired message from its neighbors will no longer take place in the actual gossip spreading process.

**Theorem 5.** Given a directed network \( G = (V, E) \), an arbitrary seed set \( S \) of nodes, and a constrained deadline \( D \), the expectation of the informed time of each reachable node is equal to the expectation of the length of the shortest time-respecting path (i.e., the distance) from \( S \) to the considered node on \( G^T = (V, E^T, W^T) \).

**Proof.** For each node \( v \in V \), consider its \( CC(v) \). From the memoryless property of the pull-based gossip spreading process, the \( CC(v) \) can be started up from the very beginning of the spreading process till a certain time (i.e., the deadline \( D \)) even after \( v \) becomes informed. The attempts of \( v \) to pull its desired message from its neighbors are no longer effective after its informed time \( t = T_v \). Consequently, for each node \( v \), we can let the \( CC(v) \) be run at the very beginning and independently of the coupon collecting processes of all the other nodes.

With all the coupon collecting processes \( \{CC(v), v \in V\} \) run till the deadline \( D \) is reached, their results are then recorded and can be later used for revealing the (absolute) time stamps of the events that \( v \) succeeds in pulling its desired message from its neighbors for the first time within \( D \) rounds. Therefore, a temporal graph \( G^T = (V, E^T, W^T) \) can be constructed from \( \{CC(v), v \in V\} \), containing all the information about one sample realization of the spreading process over the network \( G = (V, E) \) within \( D \) rounds.

Specially, given an arbitrary seed set \( S \) of nodes, the informed time \( T_v \) for each reachable node \( v \in V \) in which
it becomes informed for the first time is equal to the length of the shortest time-respecting path from $S$ to $v$ on $G^T$, and thus their expectations are also equal by taking expectations over all possible realizations of the gossip spreading process within $D$ rounds.

\textbf{Remark 6.} For any sample realization of the gossip spreading process, each of the resulting live diffusion paths from $S$ to all the other reachable nodes on $G$ is equivalent to the shortest time-respecting path from $S$ to the considered node on $G^T$.

4.3. \textit{Submodularity in Gossip Spreading.} The following arguments lead to a greedy algorithm that yields a solution within $(1 - 1/e)$ of the optimal value for GSMP. Given a finite ground set $U = \{u_1, u_2, \ldots, u_n\}$ of $n$ elements and an arbitrary function $f(\cdot) : U \rightarrow R$, $f(\cdot)$ maps subsets of $U$ to real numbers. Formally, $f(\cdot)$ is called \textit{submodular} function if satisfying

$$f(A_1 \cup \{v\}) - f(A_1) \geq f(A_2 \cup \{v\}) - f(A_2),$$

for all pairs of subsets $A_1 \subseteq A_2$ and all elements $v \in U \setminus A_2$ [25]. The quantity $f(A \cup \{v\})$ is called the \textit{marginal increase} by adding a new element $v$ into a given subset $A$. Besides, $f(\cdot)$ is called \textit{monotone} function if satisfying

$$f(A \cup \{v\}) - f(A) \geq 0,$$

for all subsets $A \subseteq U$ and all elements $v \in U \setminus A$.

Leveraging the temporal mapping of the gossip spreading process established in Theorem 5, we have Theorem 7. Note that $\sigma_D(\cdot)$ is called \textit{gossip spreading function}, and $\sigma_D(A)$ is defined in (1) for any given seed set $A \subseteq U$ in the network $G = (V, E)$.

\textbf{Theorem 7.} Given a directed network $G = (V, E)$ and a constrained deadline $D$, the gossip spreading function $\sigma_D(\cdot)$ is submodular for the pull-based gossip protocol.

\textbf{Proof.} Recall Theorem 5 and consider the sample space $\Omega$, where each sample specifies one possible realization of \{CC($v$), $v \in V$\}. Conditioned upon $X \in \Omega$, define $I_D(A \mid X)$ as the number of informed nodes within $D$ rounds using $A$ as the seed set. Let $R(s, X)$ denote the set of nodes that are reachable from a node $s$ on $G^T = (V, E^T, W^T)$ with the length of the shortest time-respecting path no larger than $D$. Therefore, $I_D(A \mid X)$ is equal to the cardinality of the union $\bigcup_{s \in A} R(s, X)$; that is,

$$I_D(A \mid X) = \bigcup_{s \in A} R(s, X).$$

(11)

We now prove that the function $I_D(\cdot \mid X)$ is submodular for each sample $X$, similar to [16, 17]. Let $A_1$ and $A_2$ denote two seed sets with $A_1 \subseteq A_2$. For a node $v$, consider the following quantity:

$$I_D(A_1 \cup \{v\} \mid X) - I_D(A_1 \mid X) = |R(v, X) \setminus \bigcup_{s \in A_1} R(s, X)|,$$

(12)

which is the number of elements in $R(v, X)$ that are not already in the union $\bigcup_{s \in A_1} R(s, X)$. Therefore, we have

$$I_D(A_1 \cup \{v\} \mid X) - I_D(A_1 \mid X)$$

$$= |R(v, X) \setminus \bigcup_{s \in A_1} R(s, X)|$$

$$\geq |R(v, X) \setminus \bigcup_{s \in A_2} R(s, X)|$$

$$= I_D(A_2 \cup \{v\} \mid X) - I_D(A_2 \mid X).$$

According to the defining property of submodularity in (9), we see that $I_D(\cdot \mid X)$ is submodular from (13). To complete the proof, we have

$$\sigma_D(A) = \sum_{X \in \Omega} \Pr[X] \cdot I_D(A \mid X),$$

(14)

which means that within $D$ rounds the expected number of informed nodes is just the weighted average over all the sample realizations in $\Omega$ of the gossip spreading process. Since a nonnegative linear combination of submodular functions is still submodular [25], $\sigma_D(\cdot)$ is submodular.

\textbf{Remark 8.} Submodularity is a widely applied mathematical tool in tackling a class of nonconvex combinatorial optimization problems, such as social influence maximization, maximum facility location, sensor placement, and optimization design of cellular networks [17, 26–28].

4.4. \textit{Greedy Algorithm with Performance Guarantee.} We invoke the following result from [25]. Note that the greedy algorithm, presented in “Algorithm 1,” selects each new element with the largest marginal increase in the gossip spreading function $\sigma_D(\cdot)$ till the seed set $S$ is filled in with $k$ nodes; and in “Algorithm 1,” $\lfloor$Gossip$(A)\rfloor$ is the number of finally informed users when the gossip spreading process runs till the delay-tolerant deadline $D$ using the seed set $A$, and the Monte Carlo method is leveraged to evaluate the average value for $R$ repeating times. The complexity
of the greedy algorithm is $O(kR|V|)$. In addition, the lazy evaluation method in [18] can be leveraged to accelerate the greedy algorithm.

**Theorem 9** (see [25]). For a nonnegative monotone submodular function $f(\cdot)$, let $A$ be the solution of the greedy algorithm and let $A^*$ be any arbitrary solution; then

$$f(A) \geq \left(1 - \frac{1}{e}\right) \cdot f(A^*).$$

**Remark 10.** Since $\sigma_D(\cdot)$ is submodular and clearly nonnegative monotone as well, the greedy algorithm yields a solution within $(1 - 1/e)$ of the optimal value for GSMP. Note that this $(1 - 1/e)$-factor is the best reported one for the class of submodular maximization problems and the infeasibility to improve this factor is argued in [29].

### 5. Experiments

In this section, we carry out simulation experiments to study the spreading performance of the greedy algorithm. For comparison, we also implement two heuristic algorithms and a random algorithm.

#### 5.1. Setup

**5.1.1. Random Geometric Network with Hotspots.** We leverage the widely used random geometric graph to generate a network graph with hotspots. There are $n_1 = 350$ nodes uniformly distributed on the $[0, 1]^2$ square; moreover, additional $n_2 = 40$, $n_3 = 50$, and $n_4 = 60$ nodes are densely clustered and uniformly distributed in three separate hotspots on this square.

In total, there are $n = 500$ nodes in the network; and any pair of nodes is connected with two edges in both directions if their Euclidean distance is no larger than the connectivity radius $r_n = \sqrt{(\log n)/n}$. Note that this $r_n$ guarantees that a random geometric graph of $n$ nodes is connected with high probability [30]. Besides, each hotspot is restricted in a rectangular region within the offset ranges $[-r_n, +r_n]^2$ from its hotspot center. In Figure 4, the network layout is illustrated.

**5.1.2. Four Typical Social Networks.** We further implement our approach in four typical social networks to evaluate the spreading performance. The small-world network [31], the scale-free network [32], the scientific-collaboration network [33], and the autonomous-system network [34] and the main network statistics are presented in Table I.

**5.2. Heuristic and Random Algorithms.** The degree centrality and distance centrality-based heuristic algorithms are widely used in social network analysis [35]; and we can select the seed nodes from the network, using their degree centralities and distance centralities as the decision criteria.

The degree centrality $\deg_u$ of each node $u$ is equal to the number of its neighbors (connected by outgoing edges). The degree-centrality algorithm selects each new element with the largest degree centrality till the seed set $S$ is filled in with $k$ nodes.

In particular, we assume the network is connected for the distance-centrality algorithm. The distance centrality $\text{dist}_u$ of each node $u$ is equal to the summarization of the distances from $u$ to all the other nodes in the network. The distance from $u$ to each of the rest nodes is measured by the number of hops on the shortest path from $u$ to that node. The distance-centrality algorithm selects each new element with the smallest distance centrality till the seed set $S$ is filled in with $k$ nodes. If the network is not connected, we can use the summarization of these distances’ reciprocals instead of these distances themselves to measure the distance centrality.

**Algorithm 1: Greedy** $(G; k; D)$.

```plaintext
Initialize $S \leftarrow \emptyset$ and $R$
for $i = 1 \rightarrow k$ do
    for each node $v \in V \setminus S$ do
        $\eta_v \leftarrow 0$
    for $j = 1 \rightarrow R$ do
        $\eta_v \leftarrow \eta_v + |\text{Gossip}(S \cup \{v\})|
    end for
    $\eta_i \leftarrow \eta_i/R$
end for
$S \leftarrow S \cup \{\arg\max_v_{v \in V \setminus S}(\eta_v)\}$
end for
Output $S$
```

**Table 1: Node and edge numbers of four social networks.**

| Network                  | Node number | Edge number |
|--------------------------|-------------|-------------|
| Small-world network      | 5000        | 90000       |
| Scale-free network       | 5000        | 49970       |
| Scientific-collaboration network | 13861 | 89238 |
| Autonomous-system network | 11157   | 61886       |

**Figure 4:** Network layout of a random geometric network with hotspots on the $[0, 1]^2$ square.
The random algorithm is the baseline, selecting $k$ distinct seed nodes uniformly at random from all the network nodes.

5.3. Simulation Results

5.3.1. Spreading Performance in Random Geometric Network with Hotspots. We evaluate the spreading performance of the greedy algorithm using different deadlines and seed set sizes, and the results are presented in Figure 5. It is shown that the informed population (i.e., the expected number of informed nodes) grows larger as more seed nodes are selected. In addition, if the network nodes are willing to tolerate longer deadline, more of them can successfully receive their desired information.

Next, we present the spreading performance of different seed-selecting algorithms in Figure 6. When the seed set size is 10, the greedy algorithm outperforms the degree-centrality algorithm by 76%, the distance-centrality algorithm by 426%, and the random algorithm by 37%.

The greedy algorithm that leverages the dynamics of the gossip spreading process in the network performs much better than those centrality-based heuristic algorithms that rely only on the network's structural properties. In actual fact, many of the most central nodes (e.g., with high degree centralities or low distance centralities) are clustered, and thus it is unnecessary to select all of them. In Figure 7, this clustering effect of both centrality-based algorithms is illustrated.

Besides, we see that the random algorithm outperforms both these heuristic algorithms. The reason lies in the fact that the underlying network is generated from the random geometric graph and the random algorithm may often select some seed nodes which have high power to cause a large gossip spreading in different regions over the $[0,1]^2$ square.

5.3.2. Spreading Performance in Four Social Networks. We further evaluate the spreading performance of the greedy algorithm in Figures 8, 9, 10, and 11, with the random algorithm as a baseline for comparison. The deadline is 10 in Figures 8 and 10 and is 5 in Figures 9 and 11. Note that the spreading performances of different deadlines and algorithms in these four social networks have similar trends as those in the random geometric network with hotspots and hence omitted.

From Figures 8, 9, 10, and 11, we see that a large portion of nodes in each network are informed when the seed set size is around 20 for the greedy algorithm. Besides, we see that the greedy algorithm significantly outperforms the random algorithm, especially in the scale-free network, the scientific-collaboration network, and the autonomous-system network.
6. Conclusions

In this paper, we have investigated the problem on the selection of information sources for pull-based gossip spreading. We have proved the NP-hardness of the gossip spreading maximization problem and proposed a suboptimal solution (i.e., the greedy algorithm) for this problem to select seed nodes as information sources. A temporal mapping of the dynamic gossip spreading process has been established via virtual coupon collectors by leveraging the concept of temporal network to analyze this problem. In addition, the temporal mapping method helps to bridge the connection from graph theoretic problems to submodularity and further lead to the greedy algorithm with performance guarantee. In the future, it is interesting to leverage the methods developed in the treatment of GSMP to deal with the problems on gossip spreading when implementing gossip-based algorithms in real-world networks.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work has been supported by National Basic Research Program of China (973 Program) through Grant 2012CB316004,
SRFPD and RGC ERG Joint Research Scheme through Grant 20133402140001, National Natural Science Foundation of China through Grant 61379003, and the 100 Talents Program of Chinese Academy of Sciences.

References

[1] S. Selvakkeneddy and S. Sinnappan, "An adaptive data dissemination strategy for wireless sensor networks," International Journal of Distributed Sensor Networks, vol. 3, no. 1, pp. 23–40, 2007.

[2] C. Boldrini, M. Conti, and A. Passarella, "Contentplace: social-aware data dissemination in opportunistic networks," in Proceedings of the 11th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems (MSWIM '08), pp. 203–210, October 2008.

[3] W. Gao and G. Cao, "User-centric data dissemination in disruption tolerant networks," in Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM '11), pp. 3109–3127, April 2011.

[4] Z. Li, G. Xie, K. Hwang, and Z. Li, "Churn-resilient protocol for massive data dissemination in P2P networks," IEEE Transactions on Parallel and Distributed Systems, vol. 22, no. 8, pp. 1342–1349, 2011.

[5] H. Zhao and J. Zhu, "Efficient data dissemination in urban VANETs: parked vehicles are natural infrastructures," International Journal of Distributed Sensor Networks, vol. 2012, Article ID 151795, 11 pages, 2012.

[6] W. Dong, Y. Yang, and W. Zhang, "Gossiping with message splitting on structured networks," International Journal of Distributed Sensor Networks, vol. 2015, Article ID 504581, 8 pages, 2015.

[7] R. Karp, C. Schindelhauer, S. Shenker, and B. Vocking, "Randomized rumor spreading," in Proceedings of the 41st Annual Symposium on Foundations of Computer Science, pp. 565–574, Redondo Beach, Calif, USA, November 2000.

[8] D. Shah, Gossip Algorithms, Now Publishers, 2009.

[9] J. Tang, S. Dai, J. Li, and S. Li, "Gossip-based scalable diffusion for wireless sensor networks," International Journal of Communication Systems, vol. 24, no. 11, pp. 1418–1430, 2011.

[10] Z. Huang and Q. Huang, "To reach consensus using uninorm aggregation operator: a gossip-based protocol," International Journal of Intelligent Systems, vol. 27, no. 4, pp. 375–395, 2012.

[11] A. Vahdat and D. Becker, "Epidemic routing for partially connected Ad hoc networks," Tech. Rep. CS-200006, Duke University, 2000.

[12] S. Sanghavi, B. Hajek, and L. Massoulie, "Gossiping with multiple messages," IEEE Transactions on Information Theory, vol. 53, no. 12, pp. 4640–4654, 2007.

[13] F. Chierichetti, S. Lattanzi, and A. Panconesi, "Rumor spreading in social networks," Theoretical Computer Science, vol. 412, no. 24, pp. 2602–2610, 2011.

[14] T. Sauerwald, "On mixing and edge expansion properties in randomized broadcasting," Algorithmica, vol. 56, no. 1, pp. 51–88, 2010.

[15] G. Giakkoupis and T. Sauerwald, "Rumor spreading and vertex expansion," in Proceedings of 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '12), pp. 1623–1641, 2012.

[16] W. Dong, W. Zhang, and G. Wei, "Extracting influential information sources for gossiping," in Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton '12), pp. 1438–1444, Monticello, Ill, USA, October 2012.

[17] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03), pp. 137–146, Washington, DC, USA, August 2003.

[18] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Vanbriens, and N. Glance, "Cost-effective outbreak detection in networks," in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07), pp. 420–429, August 2007.

[19] B. Han, P. Hui, V. S. A. Kumar, M. V. Marathe, G. Pei, and A. Srinivasan, "Cellular traffic offloading through opportunistic communications: a case study," in Proceedings of the 5th ACM Workshop on Challenged Networks (CHANTS '10), pp. 31–38, ACM, September 2010.

[20] R. M. Karp, Reducibility Among Combinatorial Problems, Springer, Berlin, Germany, 1972.

[21] G. Wilfong and P. Winkler, "Ring routing and wavelength translation," in Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '08), pp. 333–341, 1998.

[22] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: wireless video content delivery through distributed caching helpers," in Proceedings of the 31st IEEE International Conference on Computer Communications (INFOCOM '12), pp. 1107–1115, 2012.

[23] P. Holme and J. Saramäki, "Temporal networks," Physics Reports, vol. 519, no. 3, pp. 97–125, 2012.

[24] M. Mitzenmacher and E. Upfal, Probability and Computing: Randomized Algorithms and Probabilistic Analysis, Cambridge University Press, Cambridge, UK, 2005.

[25] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," Mathematical Programming, vol. 14, no. 1, pp. 265–294, 1978.

[26] Z. Drezner and H. W. Hamacher, Facility Location: Applications and Theory, Springer, 2004.

[27] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in gaussian processes: theory, efficient algorithms and empirical studies," Journal of Machine Learning Research, vol. 9, no. 1, pp. 235–284, 2008.

[28] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," IEEE Journal on Selected Areas in Communications, vol. 29, no. 8, pp. 1525–1536, 2011.

[29] U. Feige, "A threshold of LN n for approximating set cover," Journal of the ACM, vol. 45, no. 4, pp. 634–652, 1998.

[30] P. Gupta and P. R. Kumar, "The capacity of wireless networks," IEEE Transactions on Information Theory, vol. 46, no. 2, pp. 388–404, 2000.

[31] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature, vol. 393, no. 6684, pp. 440–442, 1998.

[32] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," Science, vol. 286, no. 5439, pp. 509–512, 1999.

[33] M. E. Newman, "The structure of scientific collaboration networks," Proceedings of the National Academy of Sciences of the United States of America, vol. 98, no. 2, pp. 404–409, 2001.

[34] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in Proceedings of the 11th ACM SIGKDD International
Conference on Knowledge Discovery and Data Mining (KDD '05), pp. 177–187, Chicago, Ill, USA, August 2005.

[35] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, UK, 1994.