Extracting Influential Information Sources For Gossiping

Wenxiang Dong, Wenyi Zhang, and Guo Wei
Department of Electronic Engineering and Information Science
University of Science and Technology of China, Hefei, 230027, China
javin@mail.ustc.edu.cn, {wenyizha, wei}@ustc.edu.cn

Abstract—Gossiping or rumor spreading is an effective way of information dissemination, and gossiping-based schemes are promising solutions for many of the next generation networks. Despite the abundant existing works on (high-probability) completion time of gossiping processes, it remains an interesting, important, and largely unaddressed problem how to select the initial information sources for gossiping, so as to maximize the expected spreading or minimize the expected completion time. In this paper, we deal with problems related to selecting influential information sources, which are initially informed nodes in a network for gossiping. The considered problems include the gossip spreading maximization problem (GSMP) and the completion time minimization problem (CTMP). Both problems are conjectured to be at least NP-hard, whereas a greedy algorithm for GSMP is proposed to guarantee an approximation factor of \(\frac{1}{e}\), due to the inherent submodularity property of GSMP. Additionally, based on an all-ordered-pair matrix, three heuristics are proposed for solving CTMP. Throughout the analysis and development, a key observation is that there exists an equivalent view that relates the expected gossiping time to the expected shortest path length in a random weighted directed graph induced by virtual coupon collectors run at each node in the network graph.

I. INTRODUCTION

Gossiping or rumor spreading is a way of information dissemination, and the classical gossiping process is described by the random phone call model [1]: in a communication network some nodes get informed with a rumor, then in each round each informed node selects a partner from its neighbors at random, and calls to tell the selected neighbor the rumor. The goal is to spread the rumor to the whole network, and all nodes will know the rumor if the network is connected.

Gossiping-based algorithms are promising for many of the next generation networks [2], because of their simplicity, robustness, flexibility and scalability. Potential applications of gossiping-based algorithms are enormous, such as function computing on sensor networks [3], message routing on ad-hoc networks [4], file distributing on peer-to-peer networks [5], and information diffusion on social networks [6].

Existing theoretical works on gossiping have mainly dealt with (high probability) upper bounds of the completion time, regardless of the choice of information sources. With a network abstracted as a graph and assumed to be static and connected, the completion time of gossiping is defined as the first time when all the nodes are informed. Several important classes of graphs have been analyzed, including complete graph [1][3], random geometric graph [7], and graphs with vertex expansion [8].

In this paper, we deal with several problems of choosing influential information sources, i.e., nodes in the initial target set, for gossiping:

1) Completion Time Estimation Problem (CTEP): given a connected undirected graph and an initial target set of nodes with an identical message to be propagated to all the other nodes, what is the expected completion time?
2) Gossip Spreading Maximization Problem (GSMP): given a connected undirected graph, the cardinality of an initial target set and a delay constraint, how to choose the nodes in the initial target set, which will be provided with an identical message to be propagated to all the other nodes, such that the expected number of informed nodes is maximized within the delay constraint?
3) Completion Time Minimization Problem (CTMP): given a connected undirected graph and the cardinality of an initial target set, how to choose the nodes in the initial target set, which will be provided with an identical message to be propagated to all the other nodes, such that the expected completion time is minimized?

The above three problems and their variants frequently arise in many scenarios, when a common message is demanded by a group of nodes in a two-stage communication procedure, where the message is distributed from an information center to a subset of nodes in the first stage and then is propagated to the whole network via gossiping in the second stage.

- In sensor networks, given that an emergence is detected by a set of sensors, one wants to estimate the expected time needed to alarm the whole network. Sometimes, 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 minimize the expected alarming time.
- In ad-hoc networks, given that a message is possessed by a set of routers, one wants to estimate the expected time needed for all agents to get the message. Sometimes, one needs to decide the smart choice of initially informed routers so as to maximize the expected number of informed nodes before a message is updated.
- In P2P networks, given a set of peers with a file, one wants to estimate the expected time needed for all peers to download the file. Sometimes, one needs to decide the best set of seed peers so as to minimize the expected downloading time.
- In social networks, given a set of infected people, one wants to estimate the trend of the spread of a disease. Sometimes,
before the wide spread of a disease one needs to control the
most infectious people.

CTEP is a basic step for GSMP and CTMP, and our main
attention is focused on the later two optimization problems.
To our best knowledge, there has not been explicit prior study
on choosing influential information sources for gossiping,
and our findings are as follows.
1) For push-based gossiping process, we establish an equi-
valent view of it via the coupon collector’s problem, so that
the solution of CTEP can be equivalently evaluated as the
expected shortest path length in a random weighted directed
graph induced by virtual coupon collectors run at each node
in the network graph.
2) GSMP under each sample instance can be reduced to
the maximum coverage problem in graph theory, and we
conjecture that GSMP is at least NP-hard in light of the
partial set cover problem. Additionally, through recognizing
the submodularity of the delay constrained gossip spreading
function, defined as the expected number of informed nodes
within a delay constraint, we develop a greedy algorithm
for GSMP, which guarantees an approximation factor of
\((1 - 1/e)\).
3) CTMP under each sample instance can be reduced to
the \(k\)-center problem in network location theory, and we
conjecture that CTMP is at least NP-hard in light of the
generalized dominating set problem. Additionally, motivated
by existing works on the \(k\)-center problem, we suggest three
heuristics to choose the initial target set for CTMP.

Our work differs from previous works in several aspects.
First, we deal with problems of choosing influential infor-
mation sources and of optimizing the expected performance,
instead of the high-probability performance guarantee as con-
considered in most existing works. Second, our work on GSMP
is related to the classical gossiping protocol, while the social
influence maximization problem (see, e.g., [9]) is related to
the social influence diffusion model, where an active node
can inform as many neighbors as possible in one time slot
and no deadline is imposed. Third, our work on CTMP is
about selecting the initial target set, while in existing works
that treat similar problems (see, e.g., [10][11]), both the
problem formulations and the corresponding solutions are
rather different from ours.

The remaining part of the paper is organized as follows.
Section II describes the system model, gossiping protocol,
and problems formulation. In Section III, we first establish
the equivalent view of gossiping, then discuss the complexity
of GSMP and CTMP, and finally show the submodularity of
the delay constrained gossip spreading function. In Section
IV, we propose algorithms for solving the problems, and
present results of numerical experiments. Section V con-
cludes the paper.

II. FRAMEWORK AND PROBLEM FORMULATION

In general, an undirected network \(G = (V, E)\) is defined
by a set of nodes \(V = \{1, \ldots, |V|\}\) and a set of undirected
edges \(E\). There are \(n = |V|\) nodes and \(m = |E|\) edges in
the network. Any pair of two nodes can communicate directly
with each other if and only if they are connected by an edge
in \(E\). The network is assumed connected and static, without
nodes joining and leaving.

There have been three traditional types of gossiping [5]:
single source with single message, single source with mul-
tiple messages, and all-to-all communication. The type of
gossiping in the current paper may be called multiple sources
with single message: initially a target set \(S_0\) of nodes with
cardinality \(k\) are endowed with an identical message, and all
the other nodes wish to receive a copy of that message. Each
node in \(S_0\) can push the message to a partner in each round,
and any initially uninformed node can also push the message
to its partner after obtaining the message.

In this paper, the main attention is paid to the completion
time \(T(S_0)\) and the informed population \(I_t(S_0)\), which are
the first time when the last node in graph \(G\) gets informed,
and the number of informed nodes at a certain time \(t\),
respectively. In the following, we describe the gossiping
protocol, and formulate the problems formally.

A. Gossiping Protocol

Time is slotted and a message can be transferred from a
sender to a receiver within a time slot, which is also called
round throughout the paper. In each round, nodes contact
their neighbors in the following manner: each node chooses
a partner uniformly at random from the set of the node’s
all neighbors, independently of any state, history or other
nodes’ choices. Once a partner is chosen, the informed node
“pushes” its possessed message to the chosen partner. Pull-
based and hybrid gossiping protocols will be considered in
future works.

In any given round, a node can push the message to only
one partner, and a uninformed node obtains the message as
long as it is pushed to by at least a neighbor. If the chosen
partner already has possessed the pushed message, the time
slot is wasted. Besides, the system is supposed to operate
in a progressive fashion, so that nodes that are informed
in some round will stay informed throughout the remaining
gossiping process. All communications are assumed to be
error-free, and the issues of protocol overhead and incentives
are ignored.

B. Problem Formulation

Consider the probability space \((\Omega, \mathcal{F}, \Pr)\), where each
sample specifies one possible outcome for a full run of
gossiping process. Let \(X\) denote one sample in \(\Omega\), and \(\Pr[X]\)
denote the probability that \(X\) occurs. Given an initial target
set \(S_0\) of nodes and a delay constraint \(D\), the number of
informed nodes under the occurrence of \(X \in \Omega\) is \(I_D(S_0|X)\).
So the expected number of informed nodes is
\[
\sigma_D(S_0) := \mathbb{E}[I_D(S_0)] = \sum_{X \in \Omega} \Pr[X] \cdot I_D(S_0|X).
\] (1)

Similarly, given an initial target set \(S_0\) of nodes, the com-
pletion time under \(X \in \Omega\) is \(T(S_0|X)\). So the expected
completion time is
\[
\tau(S_0) := \mathbb{E}[T(S_0)] = \sum_{X \in \Omega} \Pr[X] \cdot T(S_0|X).
\] (2)
(1) Problem 1: CTEP(G, S_0)
CTEP(G, S_0) is described as: given a graph G = (V, E) and an initial target set S_0, evaluate the expected completion time \( \tau(S_0) \). It can be formulated as:

\[
\text{find } \tau(S_0), \text{ given } S_0. \tag{3}
\]

(2) Problem 2: GSMP(G, k, D)
GSMP(G, k, D) is described as: given a graph G = (V, E), the cardinality k of the initial target set S_0 and a delay constraint D, determine S_0 so that the expected number of informed nodes \( \sigma_D(S_0) \) is maximized within the delay constraint D. It can be formulated as:

\[
\max_{S_0 \subseteq V} \sigma_D(S_0) \text{ s.t. } |S_0| \leq k. \tag{4}
\]

(3) Problem 3: CTMP(G, k)
CTMP(G, k) is described as: given a graph G = (V, E) and the cardinality k of the initial target set S_0, determine S_0 so that the expected completion time \( \tau(S_0) \) is minimized. It can be formulated as:

\[
\min_{S_0 \subseteq V} \tau(S_0) \text{ s.t. } |S_0| \leq k. \tag{5}
\]

III. ANALYSES

Our subsequent arguments deal with the gossiping process by formulating an equivalent view, which provides an alternate and perhaps more transparent way to characterize the informed times for the nodes. Following that, we discuss the complexity issue of GSMP and CTMP, and establish the submodularity property in GSMP.

A. Preliminary

Herein let us introduce the shortest path from a set of nodes to a node and the gossiping path of a gossiping process.

(1) Shortest path from a set of nodes to a node
The term “asymmetrically weighted connected graph” is used to denote a class of graphs \( G^* = (V, E, W) \), which are connected and in which each pair of two neighboring nodes has edges in both directions with generally different weights. We consider the distance \( d(S_0,v) \) from a given set \( S_0 \) of nodes to a given node \( v \) in \( G^* = (V, E, W) \), and it is defined by

\[
d(S_0,v) = \min_{u \in S_0} d(u,v), \tag{6}
\]

where \( d(u,v) \) is the length of the shortest path from node \( u \) to \( v \) (assuming \( d(v,v) = 0 \) for each \( v \in V \)).

(2) Gossiping path
A gossiping path from node \( v \) to node \( u \) is written as \( < P_{vu} : v_0 = v, v_1, \ldots, v_N = u > \), where \( v \in S_0 \) is an initially informed node, and node \( v_i \) pushes the message to its neighbor \( v_{i+1} \) for each \( i \in \{0,1,\ldots,N-1\} \) and the finally informed node is \( u \).

B. An Equivalent View
Consider a round when node \( v \) has just become informed in a gossiping process, and it then attempts to push its possessed message to a neighbor in each of the following rounds. We note that the pushing process from node \( v \) to all of its neighbors \( N(v) \) (including the node which pushed the message to \( v \)) is exactly a coupon collection process (see, e.g., [12]), and we term it CC\((v)\). In CC\((v)\), \( v \) has \( r_v = |N(v)| \) different coupons to collect, and in each round a coupon is collected uniformly and independently at random with replacement. In CC\((v)\), let \( Z_t(w) \) denote the collected coupon in round \( t \). The event that a certain coupon \( z \) is collected in round \( t \) means that the neighbor node \( z \in N(v) \) is pushed to in round \( t \) by node \( v \). Note that in the above described pushing process we do not care whether a pushed node has already possessed the message.

For each node \( v \), we independently run CC\((v)\) and record the time stamp for each coupon when \( v \) collects that coupon for the first time. Therefore, the time stamp \( w_{vu} \) for a neighbor \( u \in N(v) \) is pushed to by \( v \) can be written as:

\[
w_{vu} = \min\{t \geq 1 : Z_t(w) = u\}. \tag{7}
\]

After all the coupon collection processes \( \{CC\((v), v \in V\) \) are completed, an asymmetrically weighted connected graph \( G^* = (V, E, W) \) is thus constructed, in which the set of edge weights \( W \) is given by

\[
\forall \text{ directed edge } (v, u) \in E, \text{ edge weight } w(v,u) = w_{vu}. \tag{8}
\]

Proposition 3.1: Given a connected undirected graph \( G = (V, E) \) and an initial target set \( S_0 \) of nodes, the expectation of the informed time of each node is equal to the expectation of the shortest path length (i.e., distance) from \( S_0 \) to the considered node on graph \( G^* = (V, E, W) \), which is constructed as above.

Proof: For each node \( v \in V \), consider its CC\((v)\). From the memoryless property of the gossiping process, it clearly does not matter whether the CC\((v)\) is started up at the moment when node \( v \) becomes informed, or at the very beginning of the whole gossiping process. Going forward with this argument, we can thus let, for each node \( v \), the CC\((v)\) be run at the very beginning and independently of all the other nodes’ coupon collection processes.

With all the coupon collection processes \( \{CC\((v), v \in V\) \) run, their results are then recorded so can be later used for revealing the time stamps of the events that node \( v \) pushes to each of its neighbors for the first time after becoming informed. Therefore, \( G^* = (V, E, W) \) is constructed which contains all the information about the gossiping process. Specifically, given an initial target set \( S_0 \), the informed time \( T_v \) for each node \( v \in V \) when it becomes informed for the first time is equal to the length of the shortest path from \( S_0 \) to \( v \) on graph \( G^* \), and thus their expectations are also equal, by taking expectations over \( G^* \).

Regarding the quantifiability of the equivalent view via computing the shortest paths of the induced asymmetrically weighted connected graph, we have the following result.
Proposition 3.2: Given a connected undirected graph $G = (V, E)$ and the maximum degree $\Delta$ of $G$, all the coupon collection processes $\{CC(v), v \in V\}$, when computed in parallel, are completed within $(2\Delta \log n)$ steps with high probability at least $(1 - n^{-1})$ so as to quantify the informed time $T_n$ for each node $v$ in $G$.

Proof: In $CC(v)$, for each node $v$ with neighbor set $N(v)$, $r_v = |N(v)| \leq \Delta$ different coupons need to be collected by $v$. The probability that all the $r_v$ coupons are not collected after $(2r_v \log n)$ steps is at most $n^{-2}$ [12]. Taking the union bound over all the nodes in $G$, the probability that all the coupon collection processes $\{CC(v), v \in V\}$ are completed within $(2\Delta \log n)$ steps is at least $(1 - n^{-1})$.

C. Complexity Issues

As a general remark, the stochastic version of an optimization problem is typically harder than the deterministic version; see, e.g., [13].

GSMP and CTMP under an arbitrary sample instance can be reduced to the maximum coverage problem [14] and the $k$-center problem [15], respectively. We thus make the following conjecture regarding the complexity of GSMP and CTMP.

Conjecture 3.3: Both GSMP and CTMP are at least NP-hard.

1) Discussion on GSMP

The decision problem for GSMP is to decide if there is a set $A$ of $k$ nodes such that $\sigma_D(A) \geq p$, where $\sigma_D(A)$ is defined in (1). For each instance of the gossiping process $X$, i.e., $G^* = (V, E, W)$ equivalently, the decision problem can be cast for $I_D(A,X)$ instead of the expectation $\sigma_D(A)$, and it can be reduced into an instance of the NP-complete $p$ partial set cover problem ($p$-PSC) [16] over $G^*$: given the ground set $V$ and the collection of subsets $\{R(v,X): v \in V\}$ in which $R(v,X)$ denotes the set of nodes that can be reached from $v \in V$ on $G^*$ within deadline $D$, decide if there exist $k$ of the subsets such that their union has cardinality at least $p$.

2) Discussion on CTMP

The decision problem for CTMP is to decide if there is a set $A$ of $k$ nodes such that $\tau(A) \leq r$, where $\tau(A)$ is defined in (2). For each instance of the gossiping process $X$, i.e., $G^* = (V, E, W)$ equivalently, the decision problem can be cast for $T(A,X)$ instead of the expectation $\tau(A)$, and it can be reduced into an instance of the NP-complete dominating set problem of radius $r$ ($r$-DSD) [17] over $G^*$: decide if there exists a set $A$ of $k$ nodes such that the maximum of the shortest path lengths from $A$ to the remaining nodes on $G^*$ is no greater than $r$.

D. Submodularity in GSMP

The following results in this subsection will be used for guaranteeing that a greedy algorithm can provide an approximation factor of $(1 - 1/e)$ for GSMP. In general, we are given a finite ground set $U = \{u_1, u_2, \cdots, u_n\}$ and an arbitrary function $f(\cdot)$ which maps subsets of $U$ to non-negative real numbers. Formally, $f(\cdot)$ is called a submodular function if it satisfies

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

for all elements $v \in U$ and all pairs of subsets $A_1 \subseteq A_2$ [18]. The difference $f(A \cup \{v\}) - f(A)$ is referred to as the “marginal increase” when adding a new element $v$ to a given subset $A$.

Proposition 3.4: Given a connected undirected graph $G = (V, E)$ and a delay constraint $D$, the delay constrained gossip spreading function $\sigma_D(\cdot) := E[I_D(\cdot)]$ is submodular, where $E[I_D(\cdot)]$ is the expected number of informed nodes within the delay constraint $D$.

Proof: Recall Proposition 3.1 and consider the probability space $(\Omega, \mathcal{F}, \Pr)$, where each sample specifies one possible outcome for a full run of $\{CC(v), v \in V\}$. Conditioned upon $X$, define $I_D(A|X)$ as the number of informed nodes within $D$ rounds when $A$ is the initial target set. Let $R(v,X)$ denote the set of nodes that can be reached from $v \in V$ on graph $G^* = (V, E, W)$, with the shortest path length no larger than $D$. Hence $I_D(A|X)$ is simply the cardinality of the union $\bigcup_{v \in A} R(v,X)$.

We now prove that the function $I_D(\cdot|X)$ is submodular for each outcome $X$, following a similar way as that in [9]. Let $A_1$ and $A_2$ denote two sets of nodes with $A_1 \subseteq A_2$, and for a node $u \in U$ consider the marginal increase $I_D(A_1 \cup \{u\}) - I_D(A_1)$, which is the number of elements in $R(u,X)$ that are not already in the union $\bigcup_{v \in A_1} R(v,X)$; clearly this quantity is at least as large as the number of elements in $R(u,X)$ that are not already in the union $\bigcup_{v \in A_2} R(v,X)$. That is,

$$I_D(A_1 \cup \{u\}) - I_D(A_1) \geq I_D(A_2 \cup \{u\}) - I_D(A_2),$$

which is exactly the defining property of submodularity in (9). To complete the proof, we have

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

which means that within $D$ rounds the expected number of informed nodes is just the weighted average of the conditioned numbers of informed nodes, over all possible outcomes. Since a non-negative linear combination of submodular functions is still submodular [18], $\sigma_D(\cdot)$ is submodular.

We invoke the following result from [18].

Proposition 3.5: For a non-negative, monotone submodular function $f(\cdot)$, let $H$ be the solution of the greedy algorithm by choosing a new element with the largest marginal increase in the function and $H^*$ be the optimal solution. Then $f(H) \geq (1 - 1/e) f(H^*)$, i.e., the greedy algorithm provides an $(1 - 1/e)$-approximation.

IV. EXPERIMENTS

After analyzing the problems, it is important to develop effective algorithms for solving them. In this section, we propose a greedy algorithm for GSMP, and three heuristic algorithms for CTMP. During the experiments, Monte Carlo simulation is used in order to estimate the expectation of performance.
TABLE I
TRACES OF NETWORK TOPOLOGY.

<table>
<thead>
<tr>
<th>traces</th>
<th>nodes number</th>
<th>graph classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>wdm32</td>
<td>32</td>
<td>general graph</td>
</tr>
<tr>
<td>wdm47</td>
<td>47</td>
<td>general graph</td>
</tr>
<tr>
<td>grid100</td>
<td>100 (10 × 10)</td>
<td>grid graph</td>
</tr>
<tr>
<td>grid400</td>
<td>400 (20 × 20)</td>
<td>grid graph</td>
</tr>
<tr>
<td>rgg100</td>
<td>100</td>
<td>random geometric graph</td>
</tr>
<tr>
<td>rgg350</td>
<td>350</td>
<td>random geometric graph</td>
</tr>
</tbody>
</table>

A. CTEP: Matrix of Informed Times

Motivated by the all-ordered-pair shortest path problem in graph theory, we introduce a matrix \( T \) of all-ordered-pair informed times for gossiping. Given \( G = (V, E) \), the \( n \times n \) \((n = |V|)\) matrix \( T \) of all ordered pairs is defined by

\[
T(u, v) := E[T_{uv}],
\]

where \( E[T_{uv}] \) is the expected informed time when node \( v \) becomes informed for the first time given that the initially informed node is only \( u \).

Based on \( T \), we can validate the equivalent view of gossiping in Proposition 3.1. On the one hand, we run the gossiping protocol for each \( v \in V \) to estimate \( T \). On the other hand, we run all the coupon collection processes \( \{CC(v), v \in G\} \) and Dijkstra's algorithm to compute the all-ordered-pair shortest paths on \( G^* = (V, E, W) \) to obtain another estimate of \( T \).

Let \( T_1 \) and \( T_2 \) denote the experiment results when running the above two experiments. To measure their consistency, both absolute error and relative error are considered, where the absolute error matrix \( E_1 \) is \( E_1 = T_1 - T_2 \), and the relative error matrix \( E_2 \) is \( E_2 = (T_1 - T_2)/T_2 \).

Four aggregate statistics: error \( err_i \), absolute error \( abs_i \), variance \( var_i \) and standard variance \( std_i \) are respectively defined by, for each error matrix \( E_i \) \((i = 1, 2)\):

\[
\begin{align*}
err_i &= \sum_{u \in V} \sum_{v \in V} E_i(u, v)/ (n(n-1)), \\
abs_i &= \sum_{u \in V} \sum_{v \in V} |E_i(u, v)|/ (n(n-1)), \\
var_i &= \sum_{u \in V} \sum_{v \in V} (E_i(u, v))^2/ (n(n-1)), \\
std_i &= \sqrt{var_i}.
\end{align*}
\]

(13)

Our experiments are carried out on six traces, as listed in Table I. Among the traces, wdm32 and wdm47 are of general topology [19], grid100 and grid400 are grid graphs on \([0,1]^2\) square, and rgg100 and rgg350 are random geometric graphs \(G(n, r_n)\) on \([0,1]^2\) square, where \( n \) is the number of nodes and \( r_n = \sqrt{2 \log(n)/n} \) is the connectivity radius [20]. As validation of the equivalent view of gossiping in Proposition 3.1, the experiment results are presented in Table II. It can be seen that all the error statistics are sufficiently small and hence the equivalent view of gossiping is numerically dependable.

B. GSMP: Greedy Algorithm

In Proposition 3.7, the argument on the submodularity of the delay constrained gossip spreading function \( \sigma_D(\cdot) \) in GSMP is presented. Besides, it is clear that \( \sigma_D(\cdot) \) is non-negative and monotone. As a result, the greedy algorithm by always choosing a new node with the largest marginal increase in \( \sigma_D(\cdot) \) for GSMP provides an approximation factor of \((1−1/e)\), and the developed algorithm is referred to as “greedy”.

Our experiments are carried out on three traces: wdm47, grid400 and rgg350 (the delay constraints are 5, 10 and 8 rounds, respectively) in Table I. Throughout the experiments, we take the uniformly random selection of the \( k \) initially informed nodes as the baseline, referred to as “random”.

TABLE II
VALIDATION OF THE EQUIVALENT VIEW OF GossipING.

<table>
<thead>
<tr>
<th>traces</th>
<th>wdm32</th>
<th>wdm47</th>
<th>grid100</th>
<th>grid400</th>
<th>rgg100</th>
<th>rgg350</th>
</tr>
</thead>
<tbody>
<tr>
<td>(err_1)</td>
<td>-0.0215</td>
<td>-0.0042</td>
<td>0.0071</td>
<td>0.0011</td>
<td>0.0053</td>
<td>0.0040</td>
</tr>
<tr>
<td>(err_2)</td>
<td>-0.0024</td>
<td>-0.0002</td>
<td>0.0004</td>
<td>0.0001</td>
<td>0.0006</td>
<td>0.0003</td>
</tr>
<tr>
<td>(abs_1)</td>
<td>0.0795</td>
<td>0.0751</td>
<td>0.0801</td>
<td>0.0928</td>
<td>0.0587</td>
<td>0.0631</td>
</tr>
<tr>
<td>(abs_2)</td>
<td>0.0108</td>
<td>0.0087</td>
<td>0.0066</td>
<td>0.0042</td>
<td>0.0074</td>
<td>0.0057</td>
</tr>
<tr>
<td>(var_1)</td>
<td>0.0106</td>
<td>0.0092</td>
<td>0.0101</td>
<td>0.0138</td>
<td>0.0054</td>
<td>0.0063</td>
</tr>
<tr>
<td>(var_2)</td>
<td>0.0002</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>(std_1)</td>
<td>0.1029</td>
<td>0.0958</td>
<td>0.1007</td>
<td>0.1175</td>
<td>0.0737</td>
<td>0.0795</td>
</tr>
<tr>
<td>(std_2)</td>
<td>0.0144</td>
<td>0.0115</td>
<td>0.0090</td>
<td>0.0059</td>
<td>0.0096</td>
<td>0.0075</td>
</tr>
</tbody>
</table>
In Figures 1-3 it is shown that “greedy” outperforms the “random” baseline significantly, especially for the trace of wdm47 and for moderate target set sizes.

C. CTMP: Three Heuristic Algorithms

When dealing with CTMP, unlike that in GSMP, it is more difficult to provide a performance guarantee with a definite approximation factor. Based on $T$, we propose to make use of the two algorithms in [15], so as to develop two heuristic algorithms for CTMP. For each element in $T$ (considered as a cover radius $r$), we make use of Algorithm Recursive Cover($V, k/2, r$) in [15] to generate a feasible set of $k$ center nodes as a candidate of $S_0$. Among all the candidate sets, we choose the one with the minimum average completion time as the solution $S_0$. This heuristic algorithm is referred to as “recv-greedy”. Instead of using Algorithm Recursive Cover, we can make use of a further variant of it, Algorithm Approximate p-Center($V, k, r$) in [15], to generate a feasible set of $k$ center nodes as a candidate of $S_0$. Again, we choose among the candidate sets the one with the minimum average completion time as the solution $S_0$. This heuristic algorithm is referred to as “ccv-greedy”.

In addition, for graphs with regular structure such as grid graphs and random geometric graphs, both of which are widely considered in wireless networks, we also propose to use a third heuristic algorithm. We divide the $[0, 1]^2$ square into $k$ equal parts and choose the node nearest to the geometric center in each part as a center. The idea exploits the symmetry of the graph, and the heuristic algorithm is referred to as “central”.

Our experiments are carried out on three traces: wdm32, grid100 and rgg100 in Table I. Throughout the experiments, we also use as baseline the “random” selection of initially informed nodes. It is shown in Figure 4 that both “recv-greedy” and “ccv-greedy” heuristics outperform the “random” baseline significantly, and the two heuristics take turns to behave better. Besides, it is shown in Figures 5 and 6 that the “central” heuristic outperforms the “random” baseline noticeably when the target set size is small, but indeed performs poorly when the target set size gets large. The reason lies in the fact that when the area is divided into many equal parts, nodes are no longer uniformly distributed within each part, and thus the fixed choice of geometric centers degrades the performance. Still, it is shown in Figures 5 and 6 that “recv-greedy” outperforms both “central” and the “random” baseline.

In summary, both “recv-greedy” and “ccv-greedy” heuristics are good solutions for CTMP. In addition, “central” is a good solution for grid graphs and random geometric graphs, when the target set size is relatively small. It still remains a
unaddressed work to find CTMP algorithms with guaranteed performance approximations.

V. CONCLUSION

In this paper, we deal with the problems of choosing the influential information sources, i.e., nodes in the initially informed target set, for gossiping. The three problems, CTEP, GSMP, and CTMP, frequently arise in many application scenarios, when a common message is demanded by a group of nodes in a two-stage communication procedure. It is interesting to further investigate the equivalent view via coupon collection processes and exploit the revealed connection with graph-theoretic problems to develop effective gossiping-based algorithms.

ACKNOWLEDGMENT

This work was supported by 100 Talents Program of Chinese Academy of Sciences, and National Basic Research Program of China (973 Program) through 2012CB316004.

REFERENCES