Necessary and Sufficient Budgets in Information Source Finding with Querying: Adaptivity Gap

Jaeyoung Choi* and Yung Yi†

***Abstract*—In this paper, we study a problem of detecting the source of diffused information by querying individuals, given a sample snapshot of the information diffusion graph, where two queries are asked: *(i)* whether the respondent is the source or not, and *(ii)* if not, which neighbor spreads the information to the respondent. We consider the case when respondents may not always be truthful and some cost is taken for each query. Our goal is to quantify the necessary and sufficient budgets to achieve the detection probability $1-\delta$ for any given $0<\delta<1$. To this end, we study two types of algorithms: adaptive and non-adaptive ones, each of which corresponds to whether we adaptively select the next respondents based on the answers of the previous respondents or not. We first provide the information theoretic lower bounds for the necessary budgets in both algorithm types. In terms of the sufficient budgets, we propose two practical estimation algorithms, each of non-adaptive and adaptive types, and for each algorithm, we quantitatively analyze the budget which ensures $1-\delta$ detection accuracy. This theoretical analysis not only quantifies the budgets needed by practical estimation algorithms achieving a given target detection accuracy in finding the diffusion source, but also enables us to quantitatively characterize the amount of extra budget required in non-adaptive type of estimation, refereed to as *adaptivity gap*. We validate our theoretical findings over synthetic and real-world social network topologies.**

I. Introduction

Information spread in networks is universal to model many real-world phenomena such as propagation of infectious diseases, diffusion of a new technology, computer virus/spam infection in the Internet, and tweeting and retweeting of popular topics. The problem of finding the information source is to identify the true source of information spread. This is clearly of practical importance, because harmful diffusion can be mitigated or even blocked, *e.g.*, by vaccinating humans or installing security updates [1]. Recently, extensive research attentions for this problem have been paid for various network topologies and diffusion models [1]–[8], whose major interests lie in constructing an efficient estimator and providing theoretical analysis on its detection performance.

Prior work directly or indirectly conclude that this information source finding turns out to be a challenging task unless sufficient side information or multiple diffusion snapshots are provided. There have been several research efforts which use multiple snapshots [9] or a side information about a restricted superset the true source belongs [10], thereby the detection performance is significantly improved. Another type of side information is the one obtained from *querying*, *i.e.*, asking questions to a subset of infected nodes and gathering more hints about who would be the true information source [11]. The focus of this paper is also on querying-based approach (we will shortly present the difference of this paper from [11] at the end of this section).

In this paper, we consider an *identity with direction* (id/dir in short) question as follows. First, a querier asks an identity question of whether the respondent (throughout the paper, we call 'respondent' by the node who is asked a question from the querier) is the source or not, and if "no", the respondent is subsequently asked the direction question of which neighbor spreads the information to the respondent. Respondents may be untruthful with some probability so that the multiple questions to the same respondent are allowed to filter the untruthful answers, and the total number of questions can be asked within a given budget. We consider two types of querying schemes: *(a) Non-Adaptive (***NA***)* and *(b) ADaptive (***AD***)*. In **NA**-querying, a candidate respondent set is first chosen, and the id/dir queries are asked in a batch manner. In **AD**-querying, we start with some initial respondent, iteratively ask a series of id/dir questions to the current respondent, and adaptively determine the next respondent using the (possibly untruthful) answers from the previous respondent, where this iterative querying process lasts until the entire budget is used up.

We summarize our main contributions of this paper. First, we obtain the necessary budgets for both querying schemes to achieve the $(1-\delta)$ detection probability for any given $0<\delta<1$. To this end, we establish information theoretical lower bounds from the given diffusion snapshot and the answer samples from querying. Our results show that it is necessary to use the budget $\Omega\left(\frac{(1/\delta)^{1/2}}{\log(\log(1/\delta))}\right)$ for the **NA**-querying, whereas $\Omega\left(\frac{\log^{1/2}(1/\delta)}{\log(\log(1/\delta))}\right)$ for the **AD**-querying, respectively. Second, to obtain the sufficient amount of budget for $(1-\delta)$ detection performance, we consider two estimation algorithms, each for both querying schemes, based on a simple majority voting to handle the untruthful answer samples. We analyze simple, yet powerful estimation algorithms and characterize their detection probabilities for given parameters. Our results show that it suffices to use $O\left(\frac{(1/\delta)}{\log(\log(1/\delta))}\right)$ for the **NA**-querying, whereas $O\left(\frac{\log^{2}(1/\delta)}{\log(\log(1/\delta))}\right)$ is sufficient for the **AD**-querying, respectively. The gap between necessary and sufficient budgets in both querying schemes is due to our consideration of simple, yet practical estimation algorithms

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT(No. 2016R1A2A2A05921755)

*Department of Automotive Engineering, Honam University, Republic of Korea. (e-mails: jychoi@honam.ac.kr) †Department of Electrical Engineering, KAIST, Republic of Korea. (e-mails: yiyung@kaist.ac.kr)

based on majority voting, caused by the fact that the classical ML-based estimation is computationally prohibitive and even its analytical challenge is significant. Our quantification of necessary and sufficient budgets enables us to obtain the lower and upper bounds of the *adaptive gap*, *i.e.*, the gain of adaptive querying scheme compared to non-adaptive one. Finally, we validate our findings via extensive simulations over popular random graphs (*Erdös-Rényi* and scale-free graphs) and a real-world Facebook graph.

We end this section by presenting the difference of this paper from our preliminary work [11]. In [11], (i) only identity question in the non-adaptive case is considered and (ii) untruthfulness for the answers of identity questions in the adaptive case is not modeled. In this paper, we generalize and complete the model in terms of query types and schemes, which add non-negligible analytical challenges, and we establish information-theoretic lower bounds for the necessary amount of budget, which is the key step to quantifying the adaptivity gap.

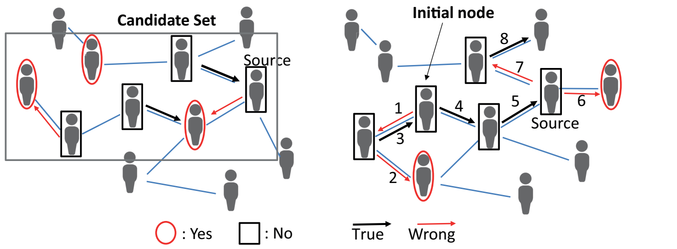
II. Model Preliminaries

A. Diffusion Model and MLE

We consider an undirected graph $G = (V, E)$, where V is a countably infinite set of nodes and E is the set of edges of the form (i, j) for $i, j \in V$. Each node represents an individual in human social networks or a computer host in the Internet, and each edge corresponds to a social relationship between two individuals or a physical connection between two Internet hosts. As an information spreading model, we consider a *Susceptible-Infected* (SI) model under exponential distribution with rate of λ_{ij} for the edge (i, j), and all nodes are initialized to be susceptible except the information source. Once a node i has an information, it is able to spread the information to another node j if and only if there is an edge between them. We denote by $v_1 \in V$ the information source, which acts as a node that initiates diffusion and denote by $V_N \subset V$, N infected nodes under the observed snapshot $G_N \subset G$. In this paper, we consider the case when G is a regular tree, the diffusion rate λ_{ij} is homogeneous with unit rate, *i.e.*, $\lambda_{ij} = \lambda = 1$, and N is large, as done in many prior work [2], [3], [9], [10], [12]. We assume that there is no prior distribution about the source, *i.e.*, the uniform distribution. As a useful prior result, under the SI-diffusion with homogeneous rate over regular tree, the authors [2] first show that the source chosen by the Maximum Likelihood Estimator (MLE) becomes the node with a highest graph-theoretic score metric, called *rumor centrality*. Formally, the estimator chooses v_{RC} as the rumor source defined as $v_{RC} = \arg\max_{v \in V_N} \mathbb{P}(G_N | v = v_1)$ where v_{RC} is called *rumor center* (RC).

B. Querying Model and Algorithm Classes

Querying with untruthful answers. Using the diffusion snapshot of the information, a detector performs querying which refers to a process of asking some questions. We assume that a fixed budget K is given to the detector (or the querier) and a unit budget has worth of asking one pair of id/dir question,



(a) Non-adaptive (NA)-querying. (b) Adaptive (AD)-querying.

Fig. 1. Examples of two querying types with untruthful answers ($r = 1$). In (a), the querier selects a candidate set (a large square) and asks just one id/direction question in a batch manner under the untruthful answers. In (b), starting from the initial node, the querier first asks one id/direction question and adaptively tracks the true source with the untruthful answers. (In (b), True is the direction of true parent and Wrong is the wrong direction.)

i.e., "Are you the source?" first and if the respondent answers "yes" then it is done. Otherwise, the detector subsequently asks a direction question as "Which neighbor spreads the information to you?". In answering a query, we consider that each respondent v is only partially truthful in answering id and dir questions, with probabilities of being truthful, p_v and q_v, respectively. To handle untruthful answers, the querier may ask to a respondent v the question multiple times, in which v's truthfulness is assumed to be independent. We also assume that homogeneous truthfulness across individuals, *i.e.*, $p_v = p$ and $q_v = q$ for all $v \in V_N$, and $p > 1/2, q > 1/d$ meaning that all answers are more biased to the truth. In terms of querying schemes, we consider the following two types, *non-adaptive* and *adaptive,* for each of which we restrict ourselves to a certain class of querying mechanisms:

NA-querying. In this querying, we first choose a subset of infected nodes *in a batch* as a candidate set which is believed to contain the true source, then ask (multiple) id/dir question to each respondent inside the candidate set, and finally run an estimation algorithm based on the answers from all the respondents. We consider the following class of **NA**-querying mechanisms, denoted by $\mathcal{NA}(r, K)$, in this paper:

Definition 1: (Class $\mathcal{NA}(r, K)$) In this class of **NA**-querying schemes with the parameter r and a given budget K, the querier first chooses the candidate set of $\lfloor K/r \rfloor$ infected nodes according to the following selection rule: We initially select the node RC and add other infected nodes in the increasing sequence in terms of the *hop-distance* from the RC. Then, the querier asks the id/dir question r times to each node in the selected candidate set.

AD-querying. A querier first chooses an initial node to ask the id/dir question, possibly multiple times, and the querier adaptively determines the next respondent using the answers from the previous queriee, which is repeated until the entire budget is exhausted. We consider the following class of AD-querying mechanisms, denoted by $\mathcal{AD}(r, K)$, in this paper:

Definition 2: (Class $\mathcal{AD}(r, K)$) In this class of **AD**-querying schemes with the parameter r and a given budget K, the querier first chooses the RC as a starting node, and performs the repeated procedure mentioned earlier, but in choosing the next respondent, we only consider one of the neighbors of the previous node, where each chosen respondent is asked the id/dir question r times. If the querier can not obtain any information about the direction (due to all "yes" answers for id questions), it chooses one of the neighbors as the next respondent uniformly at random.

In **NA**-querying, Fig. 1(a) illustrates a candidate set of nodes inside a square, id/dir querying is performed in a batch with $r = 1$. This hop-based candidate set selection has also been considered in [11], [12], revealing that it is a good approximation for the optimal one. In **AD**-querying, Fig. 1(b) shows an example scenario that starting from the initial node, a sequence of nodes answer the queries truthfully or untruthfully for $r = 1$.

III. Main Results

We now present our main results which state the necessary and sufficient budgets to achieve $1 - \delta$ detection accuracy for both querying types defined in the class of querying schemes $\mathcal{NA}(r, K)$ and $\mathcal{AD}(r, K)$, respectively. Due to space limitation, we present the proof of all theorems in our technical report [13].

For presentational convenience, we define a Bernoulli random variable X that represents a querier's answer for an id question, such that X is one with probability (w.p.) p and X is zero w.p. $1 - p$. Similarly, we define a querier's random answer Y for a dir question, such that Y is one w.p. q and Y is i w.p. $(1 - q)/(d - 1)$, for $i = 2, \ldots, d$. To abuse the notation, we use $H(p)$ and $H(q)$ to refer to the entropies of X and Y, respectively. Throughout this paper, we also use the standard notation $H(\cdot)$ to denote the entropy of a given random variable or vector.

A. NA-Querying: Necessary and Sufficient Budgets

(1) Necessary budget. We present an information theoretic lower bound of the budget for the target detection probability $1 - \delta$ inside the class of $\mathcal{NA}(r, K)$. We let $\mathcal{T}(r) = [T_1, T_2, \ldots, T_{\lfloor K/r \rfloor}]$ be the random vector where each T_i is the random variable of infection time of the i-th node in the candidate set. Then, by appropriately choosing r, we have the following theorem.

Theorem 1: Under d-regular tree G, as $N \to \infty$, for any $0 < \delta < 1$, there exists a constant $C = C(d)$, such that if

$$K \le \frac{C \cdot H(\mathcal{T}(r^\star))(2/\delta)^{1/2}}{f_{LN}(p, q) \log(\log(2/\delta))}, \tag{1}$$

where

$$f_{LN}(p, q) = (1 - H(p)) + p(1 - p)(\log_2 d - H(q)),$$
$$r^\star = \left\lfloor 1 + \frac{4(1 - p)\{7H(p) + 2H(q)\} \log K}{3e \log(d - 1)} \right\rfloor, \tag{2}$$

then no algorithm in the class $\mathcal{NA}(r, K)$ can achieve the detection probability $1 - \delta$.

Note that $H(\mathcal{T}(r))$ can be expressed as a function of the diffusion rate λ, see [14]. The implications of Theorem 1 are in order. First, if the entropy $H(\mathcal{T}(r^\star))$ of the infection time is large, the necessary amount of budget increases due to large uncertainty in figuring out a predecessor in the diffusion snapshot. Second, larger entropy for the answers of id/dir questions requires more budget to achieve the target detection accuracy. Also, when p goes to $1/2$ and q goes to $1/d$, *i.e.*, no information from the querying, results in diverging the required budget (because f_{LN} goes to zero). Finally, if respondents are truthful in answering for the id question (*i.e.*, $p = 1$), the direction answers does not effect the amount of necessary budget.

Algorithm 1: MVNA(r)

```
Input: Diffusion snapshot $G_N$, budget $K$, degree $d$,
       truthfulness probabilities $p > 1/2$, $q > 1/d$.
Output: Estimator $\hat{v}$
1  $C_r = S_I = S_D = \emptyset$;
2  Choose the candidate set $C_r$ as in Definition 1 and ask the
     id/dir questions $r$ times to each node in $C_r$;
3  for each $v \in C_r$ do
4      Step1: Count the number of 'yes'es for the identity
         question, stored at $\mu(v)$, and if $\mu(v)/r \ge 1/2$ then add $v$
         to $S_I$;
5      Step2: For each of $v$'s neighbors, count the number of
         designations for the dir question, choose the $v$'s neighbor,
         say $w$, with the largest count (under the rule of random
         tie breaking) as $v$'s 'predecessor', and save a directed
         edge, called predecessor edge, $w \to v$ ;
6  Make a graph $G_{\text{pre}}$with all the predecessor edges and for each
     $v \in C_r$, set $E(v) \leftarrow$ the number of all the descendants of $v$
7  $S_D \leftarrow \arg\max_{v \in C_r} |E(v)|$;
8  if $S_I \cap S_D = \emptyset$ then
9      If $p = 1$, set $\hat{v} \leftarrow \arg\max_{v \in S_I} \mathbb{P}(G_N | v = v_1)$ otherwise,
         set $\hat{v} \leftarrow \arg\max_{v \in S_I \cup S_D} \mathbb{P}(G_N | v = v_1)$;
10 else
11     $\hat{v} \leftarrow \arg\max_{v \in S_I \cap S_D} \mathbb{P}(G_N | v = v_1)$;
12 Return $\hat{v}$;
```

(2) Sufficient budget. To compute a sufficient budget, a natural choice would be to use the MLE (Maximum Likelihood Estimator), which, however, turns out to be computationally intractable for large N due to too much randomness of the diffusion snapshot and query answers. Hence, we consider a simple estimation algorithm named **MVNA**(r) that is based on majority voting for both the id and dir questions. To briefly explain how the algorithm behaves, we first select the candidate set C_r of size $\lfloor K/r \rfloor$ that has the least hop-distance from the RC, then we ask r times of id/dir questions to each node in the candidate set (Line 1). Then, we filter out the nodes that are more likely to be the source and save them in S_I (Line 4) and using the results of the dir questions, compute $E(v)$ that correspond to how many nodes in C_r hints that v is likely to be the source node (Lines 5 and 6). Finally, we choose a node with maximal likelihood in $S_I \cap S_D$ and if $S_I \cap S_D = \emptyset$, we simply perform the same task for $S_I \cup S_D$. It is easy to see that the time complexity is $O(\max\{N, K^2/r\})$.

Now, Theorem 2 quantifies the amount of querying budget that is sufficient to obtain arbitrary detection probability by appropriately choosing the number of questions to be asked.

Theorem 2: For any $0 < \delta < 1$, the detection probability under d-regular tree G is at least $1 - \delta$, as $N \to \infty$, if

$$K \geq \frac{12d/(d-2)(2/\delta)}{f_N(p,q)\log(\log(2/\delta))}, \qquad (3)$$

where $f_N(p,q) = 3(p-1/2)^2 + \frac{(d-1)p(1-p)}{3d}(q-1/d)^2$ under **MVNA**$(r^\star)$, where

$$r^\star = \left\lfloor 1 + \frac{2(1-p)\{1+(1-q)^2\}\log K}{e\log(d-1)} \right\rfloor .$$

We briefly discuss the implications of the above theorem. First, we see that $(1/\delta)^{1/2}$ times more budget is required that the necessary one, which is because we consider a simple, approximate estimation algorithm. Second, the dir question does not effect the sufficient budget K if $p = 1$ *i.e.*, no untruthfulness for the id question as in Theorem 1. However, if $p < 1$, the information from the answers for the dir questions reduces the sufficient amount of budget, because f_N increases in the denominator of (3). Finally, when p goes to $1/2$ and q goes to $1/d$, the required budget diverges due to the lack of information from the querying.

B. AD-Querying: Necessary and Sufficient Budgets

(1) Necessary budget. Next, we present an information theoretic lower bound of the budget for the target detection probability $1 - \delta$ for the algorithms in the class $\mathcal{AD}(r, K)$ in Theorem 3 by choosing r, appropriately.

Theorem 3: Under d-regular tree G, as $N \to \infty$, for any $0 < \delta < 1$, there exists a constant $C = C(d)$, such that if

$$K \leq \frac{C \cdot H(\mathcal{T}(r^\star))(\log(7/\delta))^{\alpha/2}}{f_{LA}(p,q)\log(\log(7/\delta))}, \qquad (4)$$

for $\alpha = 2$ if $p < 1$ and $\alpha = 1$ if $p = 1$ where

$$\begin{aligned} f_{LA}(p,q) &= (1 - H(p)) + p(\log_2 d - H(q)), \\ r^\star &= \left\lfloor 1 + \frac{7dp\{3H(p) + 2dH(q)\}\log\log K}{2(d-1)} \right\rfloor, \end{aligned} \qquad (5)$$

then no algorithm in the class $\mathcal{AD}(r, K)$ can achieve the detection probability $1 - \delta$.

We describe the implications of Theorem 3 as follows. First, when p goes to $1/2$ and q goes to $1/d$, *i.e.*, no information from the querying causes diverging the required budget (because f_{LA} becomes zero). Second, the positive untruthfulness for the id question ($p < 1$) requires $\log^{1/2}(1/\delta)$ times more budget than that under the perfect truthfulness ($p = 1$). This is because more sampling is necessary to learn the source from the answers of the id questions when $p < 1$, whereas no such learning is required for finding the source when $p = 1$. Third, large truthfulness (*i.e.*, large p) gives more chances to get the direction answers which decreases the amount of budget. Finally, we see that the order is reduced from $1/\delta$ to $\log(1/\delta)$, compared to that in Theorem 1.

(2) Sufficient budget. In **AD**-querying, due to the similar computational issue to **NA**-querying in using the MLE, we also consider a simple estimation algorithm to obtain a sufficient budget named by **MVAD**(r), which is again based on majority voting for both the id and dir questions. In this algorithm, we choose the RC as the initial node and perform different querying procedures for the following two cases: (i) $p = 1$ and (ii) $p < 1$. First, when $p = 1$, since there is no untruthfulness of the answers of the id questions, we check whether the current respondent s is the source or not. If yes, then the algorithm is terminated and it outputs the node s as a result (Line 5). If not, it asks of s the dir question r times and chooses one predecessor by majority voting with random tie breaking (Line 8). Then, for the chosen respondent, we perform the same procedure until we meet the source or the budget is exhausted. Second, when $p < 1$, we first add one in $\eta(s)$ which is the count that the node s is taken as the respondent. Next, due to untruthfulness, we count the number of "yes" answers for the id question and apply majority voting to filter out the nodes that are highly likely to be the source and save them in S_I (Line 7). For the negative answers for id questions, we count the designations of neighbors and apply majority voting to choose the next respondent. Then, we perform the same procedure to the chosen node and repeat this until the budget is exhausted. To filter out more probable source node from the direction answers, we compare the number that is taken as the respondent by designation from the neighbors in $\eta(v)$, and we choose the node which has the maximal count of it and save them into S_D (Line 10). Finally, we select a node with maximal likelihood in $S_I \cap S_D$ or $S_I \cup S_D$ (Lines 11-14). We easily see that the time complexity of this algorithm is $O(\max\{N, K\})$. Now, Theorem 4 quantifies the sufficient amount of budget to obtain arbitrary detection probability by appropriately choosing the number of questions to be asked.

Algorithm 2: MVAD(r)

```
   Input: Diffusion snapshot $G_N$, querying budget $K$, degree $d$,
          truthful probabilities $p > 1/2$, $q > 1/d$
   Output: Estimated rumor source $\hat{v}$
1  $S_I = S_D = \emptyset$ and $\eta(v) = 0$ for all $v \in V_N$;
2  Set the initial node $s$ by RC;
3  while $K \geq r$ do
4      if $p = 1$ then
5          If $s = v_1$, return $\hat{v} = s$ otherwise, go to step 2;
6      else
7          Step1: Set $\eta(s) \leftarrow \eta(s) + 1$ which describes that the
             node $s$ is taken as a respondent and count the number
             of "yes"es for the identity question, stored at $\mu(s)$,
             and if $\mu(v)/r \geq 1/2$ then add $v$ to $S_I$;
8      Step2: Count the number of "designations" for the
         direction question among $s$'s neighbors, and choose the
         largest counted node as the predecessor with a random tie
         breaking;
9      Set such chosen node by $s$ and $K \leftarrow K - r$;
10 $S_D \leftarrow \arg\max_{v \in V_N} \eta(v)$;
11 if $S_I \cap S_D = \emptyset$ then
12     $\hat{v} \leftarrow \arg\max_{v \in S_I \cup S_D} \mathbb{P}(G_N | v = v_1)$;
13 else
14     $\hat{v} \leftarrow \arg\max_{v \in S_I \cap S_D} \mathbb{P}(G_N | v = v_1)$;
15 Return $\hat{v} = s$;
```

Theorem 4: For any $0 < \delta < 1$, the detection probability under d-regular tree G is at least $1 - \delta$, as $N \to \infty$, if

$$K \geq \frac{2(2d-3)/d(\log(7/\delta))^{\alpha}}{f_A(p,q)\log(\log(7/\delta))}, \tag{6}$$

where $f_A(p,q) = \frac{2d}{d-1}(p-1/2)^2 + \frac{d-1}{d-2}(q-1/d)^3$ and $\alpha = 2$ if $p < 1$ and $\alpha = 1$ if $p = 1$ under **MVAD**$(r^\star)$, where

$$r^\star = \left\lfloor 1 + \frac{7d^2\{2(1-p)^3 + (1-q)^2\}\log\log K}{3(d-1)} \right\rfloor.$$

The gap between necessary and sufficient budgets is $\log(1/\delta)$ when $p < 1$, and $\log^{1/2}(1/\delta)$, when $p = 1$. Note that we have $\log(1/\delta)$ factor reduction from what is sufficient under **MVNA**$(r^\star)$ in the non-adaptive case. Further, as expected, we see that the sufficient budget arbitrarily grows as p goes to $1/2$ and q goes to $1/d$, respectively.

C. Adaptivity Gap: Lower and Upper Bounds

Using our analytical results stated in Theorems 1-4, we now establish the quantified adaptivity gap defined as follows:

Definition 3: (Adaptivity Gap) Let $K_{na}(\delta)$ and $K_{ad}(\delta)$ be the amount of budget needed to obtain $(1-\delta)$ detection probability for $0 < \delta < 1$ by the optimal algorithms in the classes $\mathcal{NA}(r,K)$ and $\mathcal{AD}(r,K)$, respectively. Then, the adaptivity gap, AG(δ) is defined as $K_{na}(\delta)/K_{ad}(\delta)$.

Theorem 5: For a given $0 < \delta < 1$, there exist a constant r and two other constants $U_1 = U_1(r,p,q)$ and $U_2 = U_2(r,p,q)$, where the constant r corresponds to the number of repeated id/dir questions for each respondent in both classes $\mathcal{NA}(r,K)$ and $\mathcal{AD}(r,K)$, such that

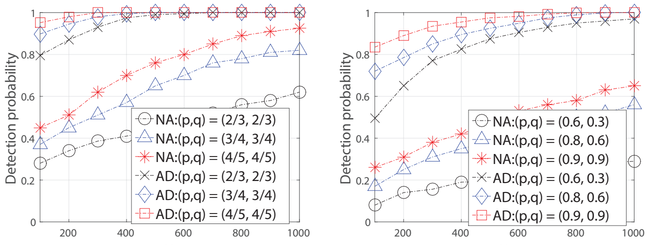
$$\frac{U_1 \cdot (1/\delta)^{1/2}}{\log^{\alpha}(1/\delta)} \leq AG(\delta) \leq \frac{U_2 \cdot (1/\delta)}{\log^{\alpha/2}(1/\delta)}, \tag{7}$$

where $\alpha = 2$ if $p < 1$, and $\alpha = 1$ if $p = 1$.

In Theorem 5, we see that for a given target detection probability $1 - \delta$, the required amount of querying budget by adaptive querying asymptotically decreases from $(1/\delta)$ to $\log(1/\delta)$, This implies that there is a significant gain of querying in the adaptive manner. Further, the difference of upper and lower bounds of AG(δ) is expressed by square root in our algorithm classes, when we use **MVNA**$(r^\star)$ and **MVAD**$(r^\star)$ for sufficient budgets, respectively.

IV. Simulation Results

In the simulation, we consider two graph topologies: regular trees, and a Facebook graph. We propagate an information from a randomly chosen node to 400 infected nodes at maximum, and plot the detection probability from 200 iterations, see [13] for more details on the simulation setup. We obtain the detection probabilities with varying budgets K under different parameters (p,q). In the regular tree, we use **MVNA**$(r^\star)$ and **MVAD**$(r^\star)$ for both querying schemes with $d = 3$ and Fig. 2(a) shows that there is a significant adaptive gain for various parameters (p,q), validating our theoretical results. Different from the regular tree, there exist loops in a general graph such as Facebook network. It is known that computing the MLE in such a general loopy graph is #P-complete [3]. Hence, as a heuristic, we use a Breath First Search (BFS) to the graph and use the BFS estimator defined in [13] as the initial center node of candidate set in **NA**-querying and the initial node in **AD**-querying, respectively. Further, in **NA**-querying, we count the number of descendants of each node in the candidate set on the BFS tree due to the loop in the general graph. Fig. 2(b) shows the detection probabilities with varying K for **NA**-querying and **AD**-querying with different parameters (p,q) and we observe similar trends to those in the regular tree. We see that the **AD**-querying is powerful for finding the source because it uses the sampled data more efficiently in an interactive manner.



(a) Regular tree ($d = 3$). (b) Facebook network.

Fig. 2. Detection probabilities with varying K for regular tree (a) and Facebook network (b), respectively.

V. Conclusion

In this paper, we considered querying for the information source inference problem in both non-adaptive and adaptive setting. We obtained the answer for the fundamental question of how much benefit adaptiveness in querying provides in finding the source with analytical characterization in presence of individuals' untruthfulness.

References

[1] K. Zhu and L. Ying, "Information Source Detection in Network: Possiblity and Impossibility Results," in *Proc. IEEE INFOCOM*, 2016.

[2] D. Shah and T. Zaman, "Detecting Sources of Computer Viruses in Networks: Theory and Experiment," in *Proc. ACM SIGMETRICS*, 2010.

[3] ——, "Rumor Centrality: A Universal Source Estimator," in *Proc. ACM SIGMETRICS*, 2012.

[4] K. Zhu and L. Ying, "Information Source Detection in the SIR Model: A Sample Path Based Approach," in *Proc. Information Theory and Applications Workshop (ITA)*. IEEE, 2013.

[5] W. Luo and W.-P. Tay, "Finding an infection source under the SIS model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[6] S. Bubeck, L. Devroye, and G. Lugosi, "Finding Adam in random growing trees," in *arXiv:1411.3317*, 2014.

[7] B. Chang, F. Zhu, E. Chen, and Q. Liu, "Information Source Detection via Maximum A Posteriori Estimation," in *Proc. IEEE ICDM*, 2015.

[8] M. Farajtabar, M. Gomez-Rodriguez, N. Du, M. Zamani, H. Zha, and L. Song, "Back to the Past: Source Identification in Diffusion Networks from Partially Observed Cascades," in *Proc. AISTATS*, 2015.

[9] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rumor source detection with multiple observations: fundamental limits and algorithms," in *Proc. ACM SIGMETRICS*, 2014.

[10] W. Dong, W. Zhang, and C. W. Tan, "Rooting Out the Rumor Culprit from Suspects," in *Proc. IEEE ISIT*. IEEE, 2013.

[11] J. Choi, S. Moon, J. Woo, K. Son, J. Shin, and Y. Yi, "Rumor Source Detection under Querying with Untruthful Answers," in *Proc. IEEE INFOCOM*, 2017.

[12] J. Khim and P.-L. Loh, "Confidence Sets for Source of a Diffusion in Regular Trees," in *arXiv:1510.05461*, 2015.

[13] J. Choi and Y. Yi "Necessary and Sufficient Budgets in Information Source Finding with Querying: Adaptivity Gap" ArXiv: 1805.03532, 2018.

[14] P. Netrapalli and S. Sangavi, "Learning the Graph of Epidemic Cascades," in *Proc. ACM SIGMETRICS*, 2012.