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 theoretical 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, referred 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 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 $\Omega\left(\frac{\log^{2}(1/\delta)}{\log(\log(1/\delta))}\right)$ 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 based on majority voting, caused by the fact that the classical

*†‡ 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)
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 (Erdős-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.

Related work. The research on rumor source detection has recently received significant attentions. The first theoretical approach was done by Shah and Zaman [2], [3], [12] and they introduced the metric called rumor centrality, which is a simple topology-dependent metric. They proved that the rumor centrality describes the likelihood function when the underlying network is a regular tree and the diffusion follows the SI (Susceptible-Infected) model, which is extended to a random graph network in [13]. Zhu and Ying [4] solved the rumor source detection problem under the SIR (Susceptible-Infected-Removed) model and took a sample path approach to solve the problem, where a notion of Jordan center was introduced, being extended to the case of sparse observations [14]. Recently, there has been some approaches for the general graphs in [1], [7] to find the information source of epidemic. All the detection mechanisms so far correspond to point estimators, whose detection performance tends to be low. There was several attempts to boost up the detection probability. Wang et al. [9] showed that observing multiple different epidemic instances can significantly increase the detection probability. Dong et al. [10] assumed that there exist a restricted set of source candidates, where they showed the increased detection probability based on the MAPE (maximum a posteriori estimator). Choi et al. [15], [16] showed that the anti-rumor spreading under some distance distribution of rumor and anti-rumor sources helps finding the rumor source by using the MAPE. The authors in [6], [17], [18] introduced the notion of set estimation and provide the analytical results on the detection performance. Choi et al. [11], [18] first considered the querying approach under the untruthful answer for this problem. They obtain the sufficient number of budget to achieve the target detection probability for the proposed MV-based algorithms. However, our work is done in a much more generalized and practical setup in the sense that we consider the case when users may be untruthful for both identity and direction questions, and also two types of practical querying scenarios are studied. There have been studies to analyze the adaptiveness gain in other literatures. Oh et al. [19] considered the adaptiveness gain in the crowdsourcing problem in the sense that how tasks are assigned to the workers and they showed that there is significant adaptiveness gain compared to the non-adaptive allocation scheme. As an information maximization problem, the authors in [20], [21] considered the adaptiveness for choosing the seed nodes after observing a partial snapshot and obtained the adaptiveness gain for maximizing the information in social network. Our work first considers a similar adaptiveness gain on the information source detection problem using querying under the untruthful answers.

II. Model Preliminaries

A. Diffusion Model and MLE

Diffusion Model 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 \( \lambda_{ij} \) 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 \( \lambda_{ij} \) 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], [17]. We assume that there is no prior distribution about the source, i.e., the uniform distribution.

Maximum Likelihood Estimator. As a preliminary, we explain the notion of rumor centrality, which is a graph-theoretic score metric and is originally used in detecting the rumor source in absence of querying and users’ untruthfulness, see [2]. This notion is also importantly used in our framework as a sub-component of the algorithms for both NA-querying and AD-querying. In regular tree graphs, Shah and Zaman [2] showed that the source chosen by the MLE becomes the node with highest rumor centrality. Formally, the estimator chooses \( v_{RC} \) as the rumor source defined as

\[
v_{RC} = \arg \max_{v \in V_N} \Pr(G_N | v = v_1) = \arg \max_{v \in V_N} R(v, G_N),
\]

where \( v_{RC} \) is called rumor center and \( R(v, G_N) \) is the rumor centrality of a node \( v \) in \( V_N \). The rumor centrality of a particular node is calculated only by understanding the graphical structure of the rumor spreading snapshot, i.e.,

\[\text{We use the terms “information” and “rumor” interchangeably.}\]
respondents. We consider the following class of adaptive, schemes, we consider the following two types,

- **Non-adaptive (NA)-querying.** 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.)

$$R(v, G_N) = N! \prod_{u \in V_N} (1/T_u^v)$$ where $T_u^v$ denotes the number of nodes in the subtree rooted at node $u$, assuming $v$ is the root of tree $G_N$ (see [2] for details).

### 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, 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 $\mathcal{N}_A(r, K)$, in this paper:

**Definition 1:** (Class $\mathcal{N}_A(r, K)$) In this class of NA-querying schemes with the parameter $r$ and a given budget $K$, the querier first chooses the candidate set of $\lceil K/r \rceil$ 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 respondent, which is repeated until the entire budget is exhausted. We consider the following class of AD-querying mechanisms, denoted by $\mathcal{A}(r, K)$, in this paper:

**Definition 2:** (Class $\mathcal{A}(r, K)$) 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 respondent 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], [17], 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 $r = 1$.

**Goal.** Our goal is to quantify that how much adaptivity on the querying scheme enriches the detection probability of the information source under the individual’s untruthfulness. As a performance measure, we consider the required number of budget $K$ to achieve the detection probability at least $1 - \delta$ for a given $0 < \delta < 1$. In the next section, we will provide the necessary and sufficient number of budgets to guarantee the detection probability under the algorithm classes by information theoretic techniques and designing Majority Voting (MV)-based algorithms, respectively. Finally, we define the adaptivity gap and quantify the lower and upper bound of this gap using the obtained results, respectively.

### 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 $\mathcal{N}_A(r, K)$ and $\mathcal{A}(r, K)$, respectively. Due to space limitation, we present the proof of all theorems in our technical report [22].

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 − δ) inside the class of \( N \mathcal{A}(r, K) \). We let \( T(r) = [T_1, T_2, ..., T_{[K/r]}] \) 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

\[
K \leq \frac{C \cdot H(T(r^*)) (2/\delta)^{1/2}}{f_{LN}(p, q) \log(\log(2/\delta))},
\]

where

\[
f_{LN}(p, q) = (1 - H(p)) + p(1 - p)(\log_2 d - H(q)),
\]

\[
r^* = \left[ 1 + \frac{4(1 - p)(7H(p) + 2H(q))}{3e \log(d - 1)} \right] \log K,
\]

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

Note that \( H(T(r)) \) can be expressed as a function of the diffusion rate \( \lambda \), see \([23]\). The implications of Theorem 1 are in order. First, if the entropy \( H(T(r^*)) \) 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 affect the amount of necessary budget.

**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. To see this, we first describe the data of querying answer as follows. Let \( A_r(p, q) := \{X(p), Y(q)\} \) be the answer vector where \( X(p) := [x_1, x_2, ..., x_{K/r}] \) with \( 0 \leq x_i \leq r \) representing the number of answers “yes” and \( Y(q) = \{Y_1(q), ..., Y_{K/r}(q)\} \) where \( Y_i(q) = [y_1, y_2, ..., y_d] \) is the answers of designations of the \( i \)-th node with \( 0 \leq y_j \leq r \) representing the answer “designations” for \( j \) node for \( 1 \leq j \leq d \). Then, the MLE of querying as an optimal estimation algorithm is to solve the following problem:

**OPT-NA:**

\[
\max_{1 \leq r \leq K} \max_{C_r} \max_{v \in C_r} \mathbb{P}[G_N, A_r(p, q) | v = v_1], \tag{4}
\]

where the inner-most max corresponds to the MLE given the diffusion snapshot \( G_N \) and the query answer sample \( A_r(p, q) \).

**Challenges.** We now explain the technical challenges in solving OPT-NA as similar in \([11]\). To that end, let us consider the following sub-optimization in OPT-NA for a fixed \( 1 \leq r \leq K \):

**SUB-OPT-NA:**

\[
\max_{C_r} \max_{v \in C_r} \mathbb{P}[G_N, A_r(p, q) | v = v_1]. \tag{5}
\]

Note that construct \( C_r^* \) by including the \( K/r \) nodes in the decreasing order of their likelihood i.e., rumor centrality. Then, \( C_r^* \) is the solution of SUB-OPT-NA as in \([11]\). Despite our knowledge of the solution of SUB-OPT-NA, solving OPT-NA requires an analytical form of the objective value of SUB-OPT-NA for \( C_r^* \) to find the optimal repetition count, say \( r^* \). However, analytically computing the detection probability for a given general snapshot is highly challenging due to the following reasons. We first note that

\[
\max_{v \in C_r^*} \mathbb{P}[G_N, A_r(p, q) | v = v_1] = \mathbb{P}[v_1 \in C_r^*] \times \max_{v_1 \in C_r^*} \mathbb{P}[G_N, A_r(p, q) | v = v_1, v_1 \in C_r^*]. \tag{6}
\]

First, the term \((a)\) is difficult to analyze, because only the MLE of snapshot allows graphical and thus analytical characterization as discussed in \([2]\), but other nodes with high rumor centrality is difficult to handle due to the randomness of the diffusion snapshot. Second, in \((b)\), we observe that using the independence between \( G_N \) and \( A_r \), by letting the event
that is sufficient to obtain arbitrary detection probability by appropriately choosing the number of questions to be asked. Under weighted has the maximum weighted rumor centrality which is hard to obtain a characterization due to the randomness of the answer for querying, thus resulting in the challenge of computing $r$ that maximizes the detection probability in OPT-NA.

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 $\lceil K/r \rceil$ 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 $O(\max\{N, K^2/r\})$.

**Rationale.** The rationale of MVNA$(r)$ from the perspective of how we handle the analytical challenges by an approximate manner is described as follows.

- **Construction of the filtered sets $S_I$ and $S_D$ from querying:** First, for the identity questions, consider the answer sample of node $v$ for $r$ questions, $x_v \ (1 \leq x_v \leq r)$, where one can easily check that for $x_v \geq r/2$ then the weight $\mathbb{P}[X(p)\mid v = v_1]$ becomes larger than that for $x_v < r/2$ due to $p > 1/2$. We use an approximated version of the weight from the answer samples by setting $\mathbb{P}[X(p)\mid v = v_1] = 1$ if $x_v \geq r/2$, and $\mathbb{P}[X(p)\mid v = v_1] = 0$ if $x_v < r/2$. For the direction questions, we see that $\mathbb{P}[Y(q)\mid v = v_1] = 1$ for the maximum consistent edge node and $\mathbb{P}[Y(q)\mid v = v_1] = 0$, otherwise. Hence, this is two step $\{0, 1\}$-weighted algorithm instead of using MLE of answer data.

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

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

where $f_N(p, q) = 3(p-1/2)^2 + \frac{(d-1)p(1-p)}{4d}(q-1/d)^2$ under MVNA$(r^*)$, where

$$r^* = \left[1 + \frac{2(1-p)\{1 + (1-q)^2\} \log K}{e \log(d-1)}\right].$$

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 $\delta$. 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 $AD(r, K)$ 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

$$K \geq C \cdot \frac{H(T(r^*))(\log(7/\delta))^{n/2}}{f_{LA}(p, q) \log(\log(7/\delta))},$$

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

$$f_{LA}(p, q) = \left(1 - H(p)\right) + p(\log_2 d - H(q)),$$

$$r^* = \left[1 + \frac{7dp(3H(p) + 2dH(q)) \log \log K}{2(d-1)}\right],$$

then no algorithm in the class $AD(r, K)$ 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 $\log(1/\delta) \times (1/\delta)$ 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, it will be given a sample of the answers to the node which we denote by a vector $Z_{r,i} := (X_i(p), Y_i(q))$ where $X_i(p)$ is the number of “yes” for the identity question and $Y_i(q) = [y_1, y_2, \ldots, y_d]$ be the answer vector for the respondent $i$, where $0 \leq y_j \leq r$ that represents the number of “designations” to $j$-th neighbor $(1 \leq j \leq d)$ of the queried node $i$. Let $\mathcal{P}(v_I)$ be a set of all policies, each of which provides a rule of choosing a next respondent at each querying step, when the initial respondent is $v_I$. We denote $W(P) = \{w_I, \ldots, w_{K/r}\}$ as set of selected queried nodes under the next node selection policy $P \in \mathcal{P}(v_I)$ with the $i$-th respondent $w_i$ for $1 \leq i \leq K/r$ and denote $A_r(P) := [Z_{r,1}, \ldots, Z_{r,K/r}]$ as the answer vector for all
natural to consider an algorithm based on MLE over all the actions at each node; pretty challenging to find an optimal policy \( P \) because \( P \)'s action at each \( i \)-th respondent can be considered as a mapping \( \mathcal{F}_i \) that uses the entire history of the respondents and their answers as discussed in [11]. Hence, as an approximation, it is natural consider that the next query is determined only by the information at the moment. Then, we have the inner part of (12) by:

\[
\mathbb{P}[G_N, A_r(P)|v = v_1, v_I] = \mathbb{P}[G_N|v = v_1, v_I] \times \mathbb{P}[A_r(P)|v = v_1, v_I] = \mathbb{P}[G_N|v = v_1] \times \mathbb{P}[Z_r, K/r, \ldots , Z_r, 1|v = v_1, v_I] \\
\times \ldots \times \mathbb{P}[Z_r, K/r|v = v_1, Z_r, K/r-1].
\]

Even under this approximation, this is also not easy to analyze because for a fixed node \( v \), the probability that it is a true parent requires to compute the probability that the true source is located in \( v \)'s subtree which does not contain \( w_i \) and there are \( O(K(r-1)) \) different answers for the direction questions. Thus, we propose a heuristic algorithm that is designed to produce an approximate solution of \( \text{OPT-AD} \). The key of our approximate algorithm is to choose the policy that allows us to analytically compute the detection probability for a given \( r \) so as to compute \( r \) easily, yet its performance is close to that of \( \text{OPT-AD} \).

Hence, in \( \text{AD- querying} \), we also consider a simple estimation algorithm to obtain a sufficient budget named by \( \text{MVAD}(r) \), 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 \( O(\max\{N, K\}) \).

Rationale. We now provide the rationale of AD-Q\((r)\) from the perspective of how we handle the analytical challenges in [4], so as to solve OPT-AD in an approximate manner.

- Construction of the filtered set \( S_I \) and \( S_D \) from querying: 
  First, for the identity questions, the intuition is similar to the NA-Q\((r)\) because it use simple MV-based rule for filtration. However, for the direction questions with the next queried node selection, we see that large designation node has large value of \( P[A_r(P)]|v = v_1, v_l] \) because it gives high value of

\[
P[Z_{r,i}]|v = v_1, Z_{r,i-1} = P[X_i|v = v_1, Z_{r,i-1}] \times P[Y_i|v = v_1, Z_{r,i-1}] 
\]

for the node \( i \) in [3].

In selecting a parent node of the target respondent, instead of the exact calculation of MLE, a simple majority voting is used by selecting the node with the highest number of designations, motivated by the fact that when \( q > 1/d \), such designation sample can provide a good clue of who is the true parent.

Now, Theorem 4 quantifies the sufficient amount of budget to obtain arbitrary detection probability by appropriately choosing the number of questions to be asked.

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

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

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

\[
r^* = \left[ 1 + \frac{7^2d(2(1-p)^3 + (1-q)^2)}{3(d-1)} \log \log K \right].
\]

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\((r^*)\) 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 3-4, 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 \( \mathcal{N}_A(r, K) \) and \( \mathcal{A}(r, K) \), respectively. Then, the adaptivity gap, \( \text{AG}(\delta) \) 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 \( \mathcal{N}_A(r, K) \) and \( \mathcal{A}(r, K) \), such that

\[
\frac{U_1 \cdot (1/\delta)^{1/2}}{\log^{\alpha}(1/\delta)} \leq \text{AG}(\delta) \leq \frac{U_2 \cdot (1/\delta)}{\log^{\alpha/2}(1/\delta)},
\]

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 \( \text{AG}(\delta) \) is expressed by square root in our algorithm classes, when we use MVNA\((r^*)\) and MVAD\((r^*)\) for sufficient budgets, respectively.

IV. SIMULATION RESULTS

In this section, we will provide simulation results of our two proposed algorithms over three types of graph topologies: (i) regular trees, (ii) two random graphs, and (ii) a Facebook graph. We propagate a rumor from a randomly chosen source up to 400 infected nodes, and plot the detection probability from 200 iterations.

Regular trees. We first obtain the numerical and simulation result for the obtained best \( r^* \) in each theorems of both querying schemes as in Fig 3(a). In this result, we plot the obtained \( r^* \) without the flooring and run the simulations 100 iterations to obtain the mean under the \( d = 3 \) for both querying schemes. We see that if we use the parameters \( p = q = 2/3 \) then the best \( r^* \) are around \( 4 - 6 \) but if the parameters are decreased by \( p = q = 4/9 \) then the algorithms use higher value of \( r^* \). Especially, we check that the chosen \( r^* \) is about 10. for the adaptive querying scheme due to increasing of the untruthfulness for direction answers. Next, we use obtain the detection probability for both querying schemes with \( d = 3 \) that is the ratio of the number of correct detections and iterations as varying two important parameters \( p \) and \( q \), respectively. First, we see that NA-Q needs more number of budget to achieve the same target detection probability under the same parameter \( (p, q) \) in Fig 3(b) and (c) as we expected. Further, we also find that the required budget is smaller than that of SB-Q in (11) where only the identity question is used. Next, we obtain the detection probabilities when one parameter is fixed (\( q \) for the NA querying and \( p \) for the AD querying) under the given budget \( K = 200 \) for the NA-querying and \( K = 100 \) for the AD-querying as in Fig 3(d) and (e), respectively. We check that if one parameter goes to one, then the detection probability also goes to one regardless for the other parameter. This is because the truthfulness for the identity question is enough to find the source in the NA-querying if it is sufficiently large and the truthful direction query can enlarge the detection for the AD-querying, respectively.

Random and real world graphs. We consider Erdös-Rényi (ER) and scale-free (SF) graphs. In the ER graph, we choose
where the answer for the direction question can be a consistent edge. To NA-querying, due to the loop in the general graph, the wrong corresponds to a social relationship (called FriendList) and consisting of 4039 nodes and 88234 edges where each edge Facebook ego network in [24] which is a undirected graph a Facebook topology as depicted in Fig. 3 (h). We use the sense of using the queries. Finally, we show the results for gives the chance to find the source more efficiently in the NA-querying scheme. We also check that the adaptivity to the case in [11] where does not use the direction information to those in regular trees. We check that how much direction edge for all the nodes in the candidate set. Figs. 3 (f) and 3 (g) shows the diameter is 8 hops. We perform the same algorithm used in NA-querying scheme. 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.

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: Possibility 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. Zaman, 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.
been infected in different moments so that the estimated node is exactly the source in the following lemma.

\[ \lim_{l \to \infty} P[d(v_1, v_k) = l] = \frac{G(k - 2, k - l - 1)}{k^{k-1}} \prod_{j=1}^{l-1} (2 + j(d - 2)), \]

where \( v_k \) is the \( k \)-th infected node. The summation of series starts from \( k = i + 1 \), since the node infected after \((l+1)\)-th order cannot be placed within distance \( l \) from \( v_1 \). We now obtain \( P[d(v_1, v_k) = l] \) as follows.

**Proposition 1:** \( P[d(v_1, v_k) = l] \) is given by

\[ P[d(v_1, v_k) = l] = \frac{G(k - 2, k - l - 1)}{k^{k-1}} \prod_{j=1}^{l-1} (2 + j(d - 2)), \]

where \( G(a, b) \triangleq \sum_{i \in \{1, \cdots, a\}, |I| = b} \prod_{j \in I} \{1 + l(d - 2)\} \), where \( I \) is any subset of \( \{1, \cdots, a\} \) with cardinality \( |I| = b \).

We will provide the proof of Proposition I in later. From this result, we have

\[ P[d(v_1, v_k) = l] = \frac{G(k - 2, k - l - 1)}{k^{k-1}} \prod_{j=1}^{l-1} (2 + j(d - 2)), \]

where \( (a) \) follows from the fact that the cardinality of summation is same for the number of possible ways of \( k - l - 1 \) among \( k - 2 \). It is shown that \[ P[v_{RC} = v_k] \leq \frac{1}{k}. \]

by using these lower bound, we have

\[ P[d(v_1, v_{RC}) = l] = \sum_{k=1+1}^{\infty} \Pr(v_{RC} = v_k) \Pr(d(v_1, v_k) = l) \]

\[ \leq \sum_{k=1+1}^{\infty} \frac{(k - l - 1)^l d^l}{k^{k-1}} \prod_{j=1}^{l-1} (2 + j(d - 2)), \]

Then, we obtain the probability \( P[v_1 \in V_1] = \sum_{k=1}^{l-1} P[d(v_1, v_{RC}) = l] \leq 1 - c \cdot e^{-\log l} \) where \( c = 4d/3(d - 2) \) and this completes the proof of Lemma I.

We will closely look at the case of each \( l \), to derive the probability that the rumor center \( v_{RC} \) is exactly \( l \)-hop distant from the rumor source \( v_1 \). Let \( \delta_1 \) be the error for the \( P[v_1 \notin V_i] \) then it is lower bounded by \( \delta_1 \geq c \cdot e^{-\log l} \).

To obtain the second term in [15], we use the information theoretical techniques for the direct graph in [23] with partial observation because, if the rumor spread from the source we can obtain a direct tree where all direction of edges are outgoing from the source. By asking the id/dir question, we can infer the revered direct tree where direction of all edges are forward to the source and in our querying model, there may
be no direction answer which gives partial observation. From the assumption of independent answers of queries, we see that the snapshot from one querying process is equivalent to the snapshot of diffusion from the source under the Independent Cascade (IC) diffusion model. By using these fact and result of graph learning from the epidemic cascades in [23], we have the following lemma.

**Lemma 2:** For any graph estimator to have a probability of error of $\delta_2$, it needs $r$ queries to the candidate set $V_i$ with $|V_i| = n$ that satisfies

$$r \geq \frac{(1 - \delta_2) H(T)(n - 1) \log \frac{2}{\delta}}{n((1 - H(p)) + p(1-p)(\log_2 d - H(q)))},$$

(21)

where $H(T)$ is the entropy of infection time vector and $H(p) = p \log p + (1-p)(\log(1-p))$ and $H(q) = q \log q + (1-q)(\log(1-q))$, respectively. The inequality (a) follows from the sub-additivity of entropy and (b) is from the fact that two random variable $X_i$ and $Y_i$ are independent. Then, we have

$$\Pr[G = \hat{G}] \leq \frac{r \cdot d((1 - H(p)) + p(1-p)(\log_2 d - H(q)))}{H(T)(n - 1)d \log(n/2)}.$$  

By setting the right term less than $1 - \delta_2$, we obtain the result and this completes the proof of Lemma 2.

From the disjoint of two error event and by setting $\delta_1 = \delta_2 = \delta/2$ with $l = \log \left(\frac{K(d-2)}{3\delta} + 2\right)/\log(d-1)$, we have

$$\Pr[\hat{v} \neq v_1] \geq \frac{e^{-h_1(T, p, q) \log K \log(K) + \frac{c}{4} e^{-2h_1(T, p, q) \log K \log(K) + \frac{l}{3e \log(d-1)}}}}{C_d e^{-h_1(T, p, q) \log K \log(K)}},$$

(22)

where $C_d = (c + 3)/4$ and $h_1(T, p, q) = H(T)^{-1}(1 - H(p)) + p(1-p)(\log_2 d - H(q))$. If we set $\delta \leq C_d e^{-h_1(T, p, q) \log K \log(K)}$, we find the value of $K$ such that its assignment to 22 produces the error probability $\delta$, and we finally obtain the desired lower-bound of $K$ as in Theorem [2].

**B. Proof of Theorem 2**

We first provide the lower bound of detection probability of $\hat{v}$ for a given $K$ and $r$ in the following lemma.

**Lemma 3:** For $d$-regular trees ($d \geq 3$), a given budget $K$, our estimator $\hat{v}$ from $\text{MVNA}(r)$ has the following lower-bound of the detection probability:

$$\Pr[\hat{v} = v_1] \geq 1 - \frac{c}{r + p + q + \frac{d}{2}} \cdot \exp \left(\frac{-h_d(K, r) w_d(p, q)}{2}\right),$$

(23)

where $c = 7(d + 1)/d$ and $w_d(p, q) = \frac{1}{2} (4(p - 1/2)^2 + (d/(d - 1))^3(q - 1/d)^3)$. The term $h_d(K, r)$ is given by

$$h_d(K, r) := \frac{\log \left(\frac{K}{r}\right)}{\log(d - 1)} \log \left(\frac{\log \left(\frac{K}{r}\right)}{\log(d - 1)}\right).$$
Proof: Under the MVNA(r), the detection probability is expressed as the product of the three terms:

\[ P[\hat{v} = v_1] = P[v_1 \in V] \times P[\hat{v} = v_1 | v_1 \in V] \times P[v_1 = v_{LRC} | v_1 \in V], \]

(24)

where \( \hat{V} := S_J \cap S_D \) if it is not empty or \( \hat{V} := S_J \cup S_D \), otherwise. This is the filtered candidate set in MVNA(r) and \( v_{LRC} \) is the node in \( \hat{V} \) that has the highest rumor centrality \( i.e., \) likelihood, where LRC means the local rumor center. We will drive the lower bounds of the first, second, and the third terms of RHS of (24). The first term of RHS of (24) is bounded by

\[ P[v_1 \in V] \geq 1 - c \cdot e^{-(l/2) \log l}, \]

(25)

where the constant \( c = 7(d + 1)/d \) from Corollary 2 of [17].

Let \( S_N \) be the set of revealed nodes itself as the rumor source and let \( S_J \) be the set of nodes which minimizing the errors. If the true source is in \( V_i \), then the probability that it is most indicated node for a given budget \( K \) with the repetition count \( r \) and truth probability \( p \) is given by

\[ P[v_1 \in LRC | v_1 \in \hat{V}] = \frac{\max \{ P[v_1 > v, G_{N}] | K, p, q \}}{P[\hat{v} = v_1 | v_1 \in V]}. \]

(26)

To obtain this, we consider that if \( p > 1/2 \), the probability \( v_1 \in S_J \) by the majority voting, because the selected node can be designation again in the algorithm. We let total number of queries by \( r \geq 1 \), we let \( W = \sum_{i=1}^{r} X_i(v_1) \) for the source node \( v_1 \), then the probability that true source is in the filtration set \( S_J \) is given by

\[ P[W \geq r/2] = \sum_{j=0}^{r/2} \binom{r}{j} (1 - p)^j p^{r-j}. \]

Then, from this relation, we have the following lemmas whose proofs are all provided in [11]:

Lemma 4: When \( p > 1/2 \),

\[ P[v_1 \in S_J | v_1 \in V] \geq p + (1 - p)(1 - e^{-p^2 \log r}). \]

This result shows that the filtration step for the identity questions to include the source in \( S_J \). Next, we will provide the probability that the source is in the filtration set for the direction question \( S_D \) as follows. Next, we will obtain the filtered direction answer as follows. First, consider the total number of direction queries \( N_d \) is a random variable with the distribution is Binomial:

\[ P(N_d = k) = \binom{d}{k} p^k (1 - p)^{d-k}, \]

where \( k \) is less than the total repetition count \( r \). Using this fact, we obtain the following result that the source is also in the filtration set \( S_D \) for the direction question.

Lemma 5: When \( p > 1/2 \) and \( q > 1/d \),

\[ P[v_1 \in S_D | v_1 \in V] \geq 1 - e^{-(d-1)(q-1)/d^2}. \]

Proof: We first consider the following result which is the lower bound of probability of determine the true parent among the direction answers. To see this, we let \( Y_{i}^1(v) \) be the random variable which takes \( +1 \) for the \( i \)-th query when the true parent node is designated by the respondent \( v \) with probability \( q \) and let \( Y_{i}^2(v) \) be the random variable which takes \( +1 \) for the \( i \)-th query when one of other neighbor nodes \( 2 \leq j \leq d \) is designated by \( v \) with probability \( (1-q)/(d-1) \). Let \( Z_j(v) := \sum_{i=1}^{k} Y_{i}^2(v) \) be the total number of designations by the node \( v \) for the \( j \)-th neighbor \( (1 \leq j \leq d) \). Then, we need to find \( P[Z_1(v) \geq Z_j(v), \forall j] \) which is the probability that the true parent is the node with maximum designations by respondent \( v \). Then, from Hoeffding bound, we obtain the following lemma.

Lemma 6: Suppose there are \( k \leq r \) answers for the direction questions with \( q > 1/d \) then

\[ P[Z_1(v) > Z_j(v), \forall j] \geq 1 - e^{-(k-1)(q-1)/d^2}. \]

Proof: For a given \( r \), we have

\[ P[Z_1(v) \geq Z_j(v), \forall j] = \sum_{i=1}^{k} Y_{i}^1(v) \geq \sum_{i=1}^{k} Y_{i}^2(v), \forall j \neq i \]

\[ \geq \max \left \{ \prod_{i=1}^{k} Y_{i}^1(v) \geq \mu_1 + \varepsilon_1 \right \} \]

(a)

\[ \geq 1 - \sum_{j=2}^{d} \sum_{i=1}^{k} Y_{i}^2(v) \geq \mu_j + \varepsilon_j \]

(b)

\[ \geq 1 - \frac{(1 - e^{-2\mu_1})}{s} \geq 1 - e^{\frac{2(1-q)/(d-1)}{3d(1-q)^2}}, \]

where \( \mu_1 = \mathbb{E}[Z_j(v) = \sum_{i=1}^{k} Y_{i}^1(v)] \) \( kq \) and \( \mu_j = \mathbb{E}[Z_j(v) = \sum_{i=1}^{k} Y_{i}^j(v)] = k(1-q)/(d-1) \). The inequality \( (a) \) comes from the fact that \( \mu_1 \geq \mu_j \) for \( 2 \leq j \leq d \) and the union bound of probability. From Chernoff-Hoeffding bound of \( Y_{i}^j(v) \), we obtain the inequality \( (b) \) by using \( \varepsilon_j = \varepsilon \mu_j \).

If we set \( \varepsilon = q^{1/2} \) then we obtain \( \hat{q} \geq 1 - e^{\frac{(1-q)/(d-1)^2}{3d(1-q)^2}} \)

which completes the proof of Lemma 6.

Now, we let \( \text{Err}(v) \) be the set of inconsistency edges in the snapshot of direction question from a node \( v \), and \( |\text{Err}(v)| \) be the total number of such edges. Then, we have the following lemma.

Lemma 7: Suppose \( v_1 \in V_l \) and \( |\text{Err}(v_1)| \) is \( s \) for \( 0 \leq s \leq |V_l| \) with \( p \). Then

\[ P[v_1 \in S_D | |\text{Err}(v_1)| = s] \geq \frac{|V_l| - s}{|V_l|}, \]

(27)

where \( |V_l| = \frac{d(d-1)^{l-2}}{d-2} \).
where \((a)\) is from the fact that for a fixed node \(v \in V_i\), the probability of selecting the edges between of them is \(d(v)/|V_i|\) with the union bound and the second term is the probability that some of them are changed which is a binomial distribution. The inequality \((b)\) by letting \(2l < |V_i|\). Hence, we have the result of (27) and this completes the proof of Lemma 8.

Using above two lemmas and from the fact that there are \(np\) nodes that are not determined as the source in expectation by the Binomial distribution, we finally obtain the result of Lemma 5 and this completes the proof.

We can use this result directly in our case because the identity question is queried by the number of \(r\) that is the total query to a node. The number of direction questions may be different. By considering the two results in the above, we have the following lemma.

Lemma 8: For given repetition count \(r\),

\[
P(v_1 \in S_I \cap S_D|v_1 \in V_I) \geq 1 - 2e^{-f(p,q)2r \log r} \tag{29}
\]

where \(f(p,q) = 3(p-1/2)^2 + \frac{d}{2}p(1-p)(q - 1/d)^2\).

Proof: From that fact that the events of \(v_1 \in S_I \) and \(v_1 \in S_D\) are independent for given \(K,p,q\) since first one is the results of source itself and the second one is the result from the other nodes. Then, we have

\[
P[v_1 \in S_I \cap S_D|v_1 \in V_I] = P[v_1 \in S_I|v_1 \in V_I]P(v_1 \in S_D|v_1 \in V_I)
\]

\[
\geq \left( p + (1-p)(1- e^{-p^2 \log r}) \right) \left( 1 - e^{-\frac{r(d-1)(q-1/d)^2}{2a}} \right)
\]

\[
\geq \left( 1 - e^{-3(p-1/2)^2 \log r} \right) \left( 1 - e^{-\frac{r(d-1)(q-1/d)^2}{2a}} \right)
\]

\[
\geq 1 - 2e^{-f(p,q)2r \log r},
\]

where \((a)\) is from the fact that \(p > 1/2\) and \((b)\) can be obtained by simple algebra with \(f_1(p,q) = c_1(p-1/2)^2 + c_2p(1-p)(q - 1/d)^2\) for some constants \(c_1 = 3\) and \(c_2 = (d-1)/3d\). This completes the proof of Lemma 8.

Next, we consider the following lemma which indicates the lower bound of detection probability among the final candidate set.

Lemma 9: When \(d \geq 3\), \(p > 1/2\) and \(q > 1/d\),

\[
P(v_1 = v_{LRC}|v_1 \in S_I \cap S_D) \geq 1 - e^{-f(p,q)r \log r},
\]

Proof: To prove this, we let \(X(n,\hat{p},\hat{q})\) be the random variable that indicates the number of nodes which reveals itself as the rumor source among \(n\) nodes. Then, from [10], we have that if \(X(n,\hat{p},\hat{q}) = k\) then

\[
\varphi_d(K,r) = 1 - \frac{2(k-1)}{k} \left( 1 - I_{1/2} \left( \frac{1}{d-2}, \frac{d-1}{d} \right) \right),
\]

where \(I_{\alpha}(a,b)\) is a regularized incomplete Beta function: \(I_{\alpha}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^\alpha t^{a-1}(1-t)^{b-1}dt\),

where \(\Gamma(\cdot)\) is the standard Gamma function. Clearly, the right term is a decreasing function with respect to \(k\) and it has the minimum at \(k = 1\) and maximum at \(k = 1\) as one. From the previous result, we know that if the repetition count \(r\) increase then \(\hat{p}\) goes to one. Based on this, we will show that if \(p > 1/2\) and \(r \geq 1\) then \(P[X(n,\hat{p}) > 1] \leq e^{-p^2 \log r}\).

From (30), we can check that \(\hat{p} \geq p + (1-p)(1- e^{-p^2 \log r}) < 1 - \frac{1}{n}\), which implies \(\mu = (n-1)(1-\hat{p}) < 1\). Then, by using this and (30), we obtain \(P[X \geq 1] \leq e^{-p^2 r \log r}\). Finally, from these results, the probability \(\varphi_d(n,\hat{p})\) is given by

\[
\varphi_d(K,r) = \sum_{k=1}^{n} \mathbb{P}[\hat{v} = v_1|X(n,\hat{p}) = k] \mathbb{P}[X(n,\hat{p}) = k]
\]

\[
= \mathbb{P}[\hat{v} = v_1|X(n,\hat{p}) = k] \mathbb{P}[X(n,\hat{p}) = k]
\]

\[
+ \sum_{k=2}^{n} \mathbb{P}[\hat{v} = v_1|X(n,\hat{p}) = k] \mathbb{P}[X(n,\hat{p}) = k]
\]

\[
\geq 1 - e^{-f(p,q)r \log r},
\]

which completes the proof of Lemma 9.
Merging these lower-bound with the lower-bound in (23), where we plug in $l = \frac{\log(K/d-2)+2}{\log(d-1)}$, we finally get the lower bound of detection probability of $\text{MVNA}(r)$ for a given repetition count $r$ and this completes the proof of Lemma 3.

The second term of RHS of (23) is the probability that the source is in the candidate set for given $K$ and $r$. Hence, one can see that for a fixed $K$, large $r$ leads to the decreasing detection probability due to the smaller candidate set. However, increasing $r$ positively affects the first term of RHS of (23), so that there is a trade off in selecting a proper $r$. By derivation of the result with respect to $r$, we first obtain $r^*$ which maximizes the detection probability by

$$r^* = \left[1 + \frac{2(1-p)(1+q^2)}{e \log(d-1)} \log K \right]$$

in $\text{MVNA}(r^*)$ and put this into the error probability $P[\hat{v} \neq v_1]$ such as

$$P[\hat{v} \neq v_1] \leq e^{-f(p,q)r \log r} + 2e^{-f(p,q)2r \log r} + c \cdot e^{-\frac{1}{2} \log l},$$

where the constant $c$ is the same as that in (25). Now, we first put $l = \frac{\log(K/d-2)+2}{\log(d-1)}$ into (32) and obtained the upper-bound of (32), expressed as a function of $r$, for a given $p$ and $q$ and the constant $c$. Then, we take $r^*$ and put it to the obtained upper-bound which is expressed as a function of $K$, as follows:

$$P[\hat{v} \neq v_1] \leq 3e^{-f(p,q)K \log K + \log K} + e^{-\frac{1}{2} \log K \log K} \leq c_1e^{-f(p,q)K \log K} + e^{-\frac{1}{2} \log K \log K},$$

where $c_1 = c + 3$. If we set $\delta \geq c_1e^{-f(p,q)K \log K + \log K}$, we find the value of $K$ such that its assignment to (33) produces the error probability $\delta$, and we get the desired lower-bound of $K$ as in the theorem statement. This completes the proof.

C. Proof of Theorem 4

We will show the lower bound for given $K$ and $r$ of the case $p < 1$.

For a given $r$, let $V_s$ be the set of all infected nodes from the rumor source within distance $K/r$ then we see that the querying dynamic still becomes a directed tree construction rooted by the source $v^*$. Then, the detection probability is expressed as the product of the two terms:

$$P[\hat{v} = v_1] = P[v_1 \in V_s] \times P[\hat{v} = v_1 | v_1 \in V_s],$$

where the first one is the probability that the distance between source and rumor center is less than $K/r$ and the second term is the probability that the estimated node is exactly the source in the candidate set for any learning algorithm under the algorithm class $C(n,b,r)$. First, from Lemma 4 we have that the probability of first term in (34) is upper bounded by

$$1 - ce^{-(K/r) \log K}$$

where $c = \frac{4d}{3(d-2)}$ for a given budget $K$ and repetition count $r$. We see that the querying dynamic still becomes a directed tree construction rooted by the source $v^*$. However, different to the NA-querying, the querying process gives direction data of a subgraph of the original direct tree because the querier chooses a node, interactively. For a given $r$, let $Z_{r,i}$ be the answer data of querying for a selected queried node $i$ where $1 \leq i \leq K/r$. Then, from the assumption of the algorithm class $C(nb,r)$, the joint entropy for the random answers with the infection time random vector $T$, $(T, Z_{r,1}, \ldots, Z_{r,K/r})$ is given by

$$H(T, Z_{r,1}, \ldots, Z_{r,K/r}) = \sum_{i=1}^{K/r} H(T, Z_{r,i} | Z_{r,i-1}, \ldots, Z_{r,1})$$

$$= \sum_{i=1}^{K/r} \sum_{j=i}^{K/r} H(T, Z_{r,i} | Z_{r,j})$$

where $(a)$ is from the fact that all data $Z_{r,i}$ are independent. Let $G^*$ be the true directed graph and let $\hat{G}$ be an estimated directed tree from the sequential answers of adaptive querying $(Z_{r,1}, \ldots, Z_{r,K/r})$. Then, we see that this defines a Markov chain

$$G^* \rightarrow (T, Z_{r,1}, \ldots, Z_{r,K/r}) \rightarrow \hat{G},$$

from the defined algorithm class $C(nb,r)$. By property of the mutual information, we have

$$I(G^*; T, Z_{r,1}, \ldots, Z_{r,K/r})$$

$$\leq H(T, Z_{r,1}, \ldots, Z_{r,K/r})$$

$$= \sum_{i=1}^{K/r} H(T, Z_{r,i})$$

$$\geq \frac{(K/r)H(T, Z_{r,1})}{\log(K/2r)}$$

where $(a)$ follows from the fact that the answers $Z_{r,i}$ are mutually exclusive and $(b)$ is from the fact that $H(T, Z_{r,1}) = (1 - H(p)) + p(\log_2 d - H(q))$ since the number of direction answers follows binomial distribution. Let $G_{K/r}$ be the set of possible directed tree in $V_s$ then we have $|G_{K/r}| \leq (K/r) \log(K/2r)$. Using the Fano’s inequality on the Markov chain $G^* \rightarrow (Z_{r,1}, \ldots, Z_{r,K/r}) \rightarrow \hat{G}$, we obtain

$$P[G \neq G^*] = P[\hat{v} = v_1 | v^* \in V_s]$$

$$\geq 1 - \frac{I(G^*; Z_{r,1}, \ldots, Z_{r,K/r}) + h(p,q)}{H(T) \log |G_{K/r}|}$$

$$\geq 1 - \frac{K(h(p,q) + h(p,q))}{K r H(T) \log(K/2r)}$$

By solving $r$, we have if $r \geq (1 - \delta)H(T)(\log K - 1)/h(p,q) \log_2 d$ then, $P[\hat{v} \neq v_1 | v^* \in V_s] \geq \delta$. From the disjoint of two error event and by setting $\delta_1 = \delta_2 = \delta/2$ for each error, we have

$$P[\hat{v} \neq v_1] \geq c \cdot e^{-(K/r) \log(K)}$$

$$+ 1 - \frac{K(h(p,q) + h(p,q))}{K r H(T) \log(K/2r)} \geq \delta.$$
From the fact that $\lambda = 1$ in our setting and Lemma 2 in [23], we obtain $H(T) \leq K/r$ and by differentiation of above lower bound with respect to $r$, we obtain $r^* = 1 + 2d_p H(p) + 2d_q H(q) \log \log K$ where the derivation is given in [22]. Since if we use the $r^*$, it gives the upper bound of detection probability hence, we put it to the obtained upper-bound which is expressed as a function of $K$, as follows:

$$
\mathbb{P}[\hat{v} \neq v_1] \geq \frac{1}{3} e^{-h_2(T,p,q) \log \log K} + \frac{c}{4} e^{-7h_2(T,p,q) \log \log K} \geq C_d e^{-h_2(T,p,q) \log \log K}, \tag{39}
$$

where $C_d = 2(c + 3)/7$ and $h_2(T,p,q) = H(T)^{-1}(1 - H(p) + (1 - p)(\log_2 d - H(q)))$. If we set $\delta \leq C_d e^{-h_2(T,p,q) \log \log K}$, we find the value of $K$ such that its assignment to (39) produces the error probability $\delta$, and we get the desired lower-bound of $K$ as in the theorem statement. Then, we finally obtain the result and this completes the proof of Theorem 3.

D. Proof of Theorem 2

We will show the lower bound of the detection probability for given $K$ and $r$ of the case $p < 1/3$ in Lemma 10.

Lemma 10: For $d$-regular trees ($d \geq 3$), a given budget $K$, our estimator $\hat{v}$ from MVAD($r$) has the detection probability lower-bounded by:

$$
\mathbb{P}[\hat{v} = v_1] \geq 1 - c(g_d(r,q))^3 \cdot \exp \left( - \left( \frac{p - 1}{2} \right)^{\frac{K}{r}} \log \left( \frac{K}{r} \right) \right), \tag{40}
$$

where $g_d(r,q) := 1 - e^{-\frac{r(d-1)(q-1)-d^2}{d^4}}$ and $c = (5d + 1)/d$.

Proof: For the MVAD($r$), for a given $r$, we introduce the notation $V_s$, where the set of all queried nodes of the algorithm. From the initial queried node, we need the probability that the source is in the set of queried node by some policy $P \in \mathcal{P}(v_T)$. Then, the detection probability is also expressed by the product of the three terms:

$$
\mathbb{P}[\hat{v} = v_1] = \mathbb{P}[v_1 \in V_s] \times \mathbb{P}[\hat{v} = v_1 | v_1 \in V_s] = \mathbb{P}[v_1 \in V_s] \times \mathbb{P}[v_1 \in \hat{V} | v_1 \in V_s] \times \mathbb{P}[v_1 \in v_{RC} | v_1 \in \hat{V}], \tag{41}
$$

where $V_s = \{v|d(v_{RC}, v) \leq K/r\}$ because the number of budget is $K$ and $V = S_I \cap S_D$ if it is not empty or $V = S_I \cup S_D$, otherwise. From the result in Corollary 2 of [17], we have $\mathbb{P}[E_1] \leq c \cdot e^{-\frac{K}{r}) \log \log K}/r$ since we use additional direction query with identity question. For the second part of probability in (41), we obtain the following lemma.

Lemma 11: When $p > 1/2$,

$$
\mathbb{P}[v_1 \in S_I | v_1 \in V_s] \geq \left( p + (1 - p)(1 - e^{-p^2 \log r}) \right) \left( 1 - ce^{-\frac{K}{r}(q-1)/d} \right),
$$

$^4$The result for $p = 1$ is given in [11] hence, we omit it here.

Proof: Let $Q_K(v)$ be the number of queries to a node $v \in V$ when there are $K$ queries then we have

$$
\mathbb{P}[Q(v_1) \geq 1] = \sum_{i=1}^{\infty} \mathbb{P}[Q(v_1) \geq 1 | d(v_1, v_{RC}) = i] \mathbb{P}[d(v_1, v_{RC}) = i].
$$

where $\mathbb{P}[d(v_1, v_{RC}) = i]$ is the probability that the distance from the rumor center to rumor source is $i$ and this probability become smaller if the distance between rumor source and rumor center is larger. From this, we have the following result for the lower bound of the probability of distance between the rumor center and source.

Lemma 12: For $d$-regular trees,

$$
\mathbb{P}[d(v_1, v_{RC}) = i] \geq \left( \frac{d - 1}{d} \right)^i e^{-(i+1)}. \tag{42}
$$

We will closely look at the case of each $i$, to derive the probability that the rumor center $v_{RC}$ is exactly $i$-hop distant from the rumor source $v_1$.

Proof: In this section, we will provide the proof of Lemma 12. To obtain the results. To do this, we decompose the event $\{d(v_1, v_{RC}) = i\}$ into sub-events where $v_{RC}$ has been infected in different moment so that

$$
\mathbb{P}[d(v_1, v_{RC}) = i] = \sum_{k=i+1}^{\infty} \mathbb{P}[v_{RC} = v_k] \mathbb{P}[d(v_1, v_k) = i].
$$

The summation of series starts from $k = i + 1$, since the node infected after $(i+1)$-th order cannot be placed within distance $i$ from $v_1$. We now obtain $\mathbb{P}[d(v_1, v_k) = i]$ as follows.

From the results of Proposition 1, we have

$$
\mathbb{P}[d(v_1, v_k) = i] = \frac{G(k-2, k-i-1)}{\Pi_{j=1}^{k-1}(2 + j(d - 2))} \frac{d(d-1)^{i-1}}{(k-1)d} \frac{(d - 1)^{k-2}}{(d - 1)}
$$

$$
\geq \sum_{x \in \{k+1, \ldots, k-2\} | x = k-i-1} \frac{(a)}{(k-2)} \frac{(d-1)^{i-1}}{(k-1)} \frac{(d-1)^{k-2}}{(d-1)}
$$

where (a) follows from the fact that the cardinality of summation is same for the number of possible ways of $k-i-1$ among $k-2$. It is not hard to show the following inequality that

$$
\mathbb{P}[v_{RC} = v_k] = \mathcal{I}_{1/2} \left( \left( k - 1 + \frac{1}{d-2}, \frac{1}{d-2} \right) + (d-1) \left( (k-2) + \frac{1}{d-2} \right) \right) \geq \left( \frac{1}{2} \right)^{k-i+1}.
$$

By using these lower bound, we have

$$
\mathbb{P}[d(v_1, v_{RC}) = i] = \sum_{k=i+1}^{\infty} \mathbb{P}(v_{RC} = v_k) \mathbb{P}(d(v_1, v_k) = i).
$$
Lemma 12. choose one of neighbor nodes uniformly at random and the probability is the same that after additional querying follows the answer. Then this is the rumor source. To do that, we consider the case that node as its parent and how many times a node reveals itself so that the Markov chain has the absorbing state, in this case, the state is the rumor center such that \( X = \infty \). This completes the proof of Lemma 12.

Next, we construct the following Markov chain. Let \( \hat{p} := \mathbb{P}[W = v] \) for the identity questions i.e., there is no “no” answers for the identity questions so that the algorithm should chooses one of neighbor nodes uniformly at random and let \( \hat{q} := \mathbb{P}[Z_1(v) > Z_0(v), \forall j] \) for the direction question, respectively. Different to the case for \( p = 1 \) which the node reveals itself as the rumor source or not with probability one so that the Markov chain has the absorbing state, in this case, there is no such a state. To handle this issue, we use the information that how many times the neighbors indicate a node as its parent and how many times a node reveals itself as the rumor source. To do that, we consider the case that there is a token\(^5\) from the initial state and it move to the next state after additional querying follows the answer. Then this probability is the same that after \( K/r \)-step of Markov chain, and we expect that the rumor source \( v_1 \) will have the largest chance to keeping this token due to the assumption of biased answer. Let \( X_n \) be the state (node) which keep this token at time \( n \) where the state is consist of all node in \( V_N \). The initial state is the rumor center such that \( X_0 = 0 \) where 0 indicates the rumor center. Then there are \( (d(d-1)K/r - 2)/(d-2) \) states and we can index all the state properly. Let \( p_{i,j}^0 \) be the \( n \) step transition probability from the state \( k \) to the state \( j \). To obtain these probabilities, we first label an index ordering by counter-clockwise from the rumor center \( X_0 = 0 \). Then, we have \( P(X_{n+1} = k | X_n = k) = 0 \) for all \( k \) and \( n \), respectively. Furthermore, \( P(X_{n+1} = j | X_n = k) = 0 \) for all \( d(k,j) > 1 \) since the token is moved one-hop at one-step (r querying). Then, the transition probability for the node \( k \) which is not a leaf node in \( V_k \) is as follows.

\[
p_{k,j} = \begin{cases} \frac{\hat{p}}{d} + (1 - \hat{p})\frac{1-\hat{q}}{d-1} & \text{if } j \notin nb(k, v_1) \\ \frac{\hat{q}}{d} + (1 - \hat{p})\hat{q} & \text{if } j \in nb(k, v_1), \end{cases}
\]

where \( nb(k, v_1) \) is the set of neighbors of the node \( k \) on the path between the node \( k \) and \( v_1 \). From the simple Markov property of querying scheme, if we assume that the source node is an absorbing state then we obtain for a given budget \( K/r \geq l \),

\[
\mathbb{P}[Q(v_1) \geq 1 | d(v_1, v_{RC}) = i] \geq 1 - \mathbb{P}[Q(v_1) = 0 | d(v_1, v_{RC}) = i] = 1 - \mathbb{P}[\sum_{n=0}^{K} I_n(v_1) = 0 | d(v_1, v_{RC}) = i] = 1 - \mathbb{P}[I_n(v_1) = 0, \forall i \leq n \leq K/r | d(v_1, v_{RC}) = i] = 1 - \prod_{n=i}^{K/r} (1 - p_{0,v}) = \prod_{n=i}^{K/r} (1 - p_{0,v}) - (1 - p_{0,v})^{K/r - n} \geq 1 - (1 - p_{0,v})^{K/r} - (1 - p_{0,v})^{K/r - n} \geq 1 - e^{-K/r} \cdot p_{0,v}^{K/r} - e^{-K/r} \cdot p_{0,v}^{K/r - n},
\]

(43)

where \( (a) \) follows from the fact that \( p_{0,v}^{K/r} \geq 0 \) for all \( i \leq n \leq K/r \) and \( (b) \) is from the relation of \( (1 - p)^{K/r} = e^{-K/r \log(1 - p)} \leq e^{-pK/r} \) where we use the inequality \( \log(1 - x) \leq -x \) for \( 0 \leq x \leq 1 \). Note that the transition probability is the case of \( d(v_1, v_{RC}) = i \). From Lemma, we have

\[
\mathbb{P}[Q(v_1) \geq 1 | d(v_1, v_{RC}) = i] \geq \sum_{i=1}^{\infty} \left( 1 - e^{-(K/r-i)p_{0,v}^{K/r}} \right) P(d(v_1, v_{RC}) = i) \geq \sum_{i=1}^{\infty} \left( 1 - e^{-(K/r-i)p_{0,v}^{K/r}} \right) \left( \frac{d-1}{d} \right)^i e^{-(i+1)} \geq 1 - e^{-K/r} \cdot p_{0,v}^{K/r} \geq 1 - e^{-K/r} \cdot p_{0,v}^{K/r - n},
\]

where the last inequality is follows from the fact that

Next, we have the following result.

Lemma 13: When \( p > 1/2 \) and \( q > 1/d \),

\[
\mathbb{P}[v_1 \in S_D | v_1 \in V_s] \geq 1 - e^{-Kp(d-1)(q-1)/2d^2}.
\]

Proof: Due to the Markov property of counting the designations from neighbors, we use following known result\(^2\) as a lemma.

Lemma 14: (\(^2\)) Let \( \mathcal{X} \) denote a Markov chain with state space \( E \). Then, the total number \( N_j \) of visits to a state \( j \in E \) from starting state \( i \) under \( K \) step is given by

\[
\mathbb{P}[N_j = m | X_0 = j] = f_{m-1}^{m-1} (1 - f_{m-1}),
\]

(44)

for \( i = j \) and \( 0 \leq m \leq K/2 \). For \( i \neq j \),

\[
\mathbb{P}[N_j = m | X_0 = i] = \begin{cases} 1 - f_{m} & \text{if } m = 0 \\ f_{m-1}^{m-1} (1 - f_{m-1}) & \text{if } 1 \leq m \leq K/2, \end{cases}
\]

where \( f_{m} := \mathbb{P}[\tau_j \leq K | X_0 = i] \). Here, \( \tau_j \) is the first hitting time from starting state \( i \).

Hence, we see that the term \( f_{m} \) is same for the probability \( \mathbb{P}[Q(v_j) \geq 1 | d(v_1, v_{RC}) = i] \) for given budget \( K \). In our case, the initial state is \( X_0 = 0 \) which is the rumor center then from
similar steps in [43], we have \( f_{0,v} \geq 1 - e^{-(K - j)p_0^K} \), for node \( v \) with \( d(v, v_{RC}) = j \). Let \( N(v) := \sum_{n=0}^{K} I_n(v) \) then for all node \( v \in V_i \),
\[
P[v_1 \in S_D | d(v_1, v_{RC}) = i] = P[N(v_1) > N(v), \forall v \in V_i | d(v_1, v_{RC}) = i]
\]
\[
\geq 1 - \sum_{v \in V_i} \left[ \frac{K/r}{m} \sum_{m=1}^{K/r} |P[N(v) \geq m | N(v_1) = m]|P[N(v_1) = m] \right]
\]
\[
= 1 - n_1 \sum_{m=1}^{K/r} \left( \frac{K/r}{m} \sum_{j=0}^{K/r} |P[N(v) = j]| \right) P[N(v_1) = m]
\]
\[
\geq 1 - n_1 \sum_{m=1}^{K/r} e^{-mK} P[N(v_1) = m] \geq 1 - ie^{-K^3(d-1)(q-1)/4d^3} \tag{45}
\]
where (a) is from the union bound of the joint probability and the inequality (b) is from the fact that \( \sum_{j=0}^{K/r} |P[N(v) = j]| \leq e^{-mK} \) where \( n_1 \) is the number of queried nodes in \( V_i \).

The last inequality (c) is from the geometric sum of \( e^{-mK} \) with mean of binomial distribution \( Kp \). The probability \( f_{ij} \) for a node \( v \) can be obtained by considering (1) distance from \( v_{RC} \) \( d(v, v_{RC}) = i \) and (2) indication whether it is on the path between rumor center and source because the transition probability will be different. Then, using this result, we have
\[
P[v_1 \in S_D | v_1 \in V_s]
\]
\[
= \sum_{i=0}^{l} P[v_1 \in S_f | d(v_1, v_{RC}) = i] P[d(v_1, v_{RC}) = i]
\]
\[
\geq \sum_{i=0}^{l} \left( 1 - ie^{-Kp(d-1)(q-1)/3d^3} \right) \frac{d-1}{d} e^{-(i+1)}
\]
\[
\geq \sum_{i=0}^{l} \left( 1 - ie^{-Kp(d-1)(q-1)/3d^3} \right) \frac{d-1}{ed} e^{i} \frac{1}{e}
\]
\[
\geq \sum_{i=0}^{l} \left( 1 - ie^{-Kp(d-1)(q-1)/3d^3} \right) \frac{1}{3} \tag{46}
\]
\[
\geq 1 - e^{-Kp(d-1)(q-1)/3d^3} \tag{47}
\]
where (a) is from the fact that \( \left( \frac{3d-1}{ed} \right)^i > 0 \) for all \( 0 \leq i \leq l \) and (b) comes from some algebra using the geometric sum for the \( i \). Hence, this completes the proof of Lemma 15. ■

Similar to the previous one, to obtain the detection probability, we need to find the probability \( P(v_1 \in S_N \cap S_f | v_1 \in V_s) \). From this, we consider \( r \) repetition count for identity question and \( r - X_r(v) \) for direction question where \( X_r(v) \) be the number of yes answers of the queried node \( v \) with probability \( p_r \). Hence, we see that the number of repetition count for the direction questions is also a random variable which follows a

\[
\text{Lemma 15: Suppose } v_1 \in V_1 \text{ then we have }
\]
\[
P(v_1 \in S_f \cap S_D | v_1 \in V_s) \geq 1 - e^{-3g(p,q)^2(K/r)} \log r, \tag{46}
\]
where \( g(p,q) = \frac{2d}{3d-1} \left( p - \frac{1}{2} \right)^2 + \frac{q}{1 - d} (q - 1/d)^2 \).

\textbf{Proof:} Since the events \( v_1 \in S_f \) and \( v_1 \in S_D \) are independent for a given \( v_1 \in V_s \), by using Lemma 15 and [13] and some algebra, we have
\[
P[v_1 \in S_f \cap S_D | v_1 \in V_s]
\]
\[
\geq \left( p + (1 - p)(1 - e^{-p^2 \log r}) \right) \left( 1 - ce^{-Kp(q-1)/d} \right)
\]
\[
\cdot \left( 1 - e^{-K(d-1)(q-1)/d^3} \right)
\]
\[
\geq 1 - e^{-3g(p,q)^2(K/r)} \log r,
\]
where \( g(p,q) = c_1 \left( p - \frac{1}{2} \right)^2 + c_2 (q - 1/d)^2 \) for some constants \( c_1 \) and \( c_2 \) which are only depends on the degree \( d \). This completes the proof of Lemma 15. ■

Next, we consider the following lemma which indicates the lower bound of detection probability among the final candidate set.

\textbf{Lemma 16:} When \( d \geq 3, p > 1/2 \) and \( q > 1/d \),
\[
P[v_1 = v_{LRC} | v_1 \in S_f \cap S_D] \geq 1 - e^{-g(p,q)r} \log r.
\]

The proof technique is similar to the Lemma 27, so we omit it. Using the obtained lemmas 5-8, we finally get the lower bound of detection probability of MVAD(r) for a given repetition count \( r \) and this completes the proof of Lemma 10. ■

The term \( g_r(r, q) \) in (40) is the probability that the respondent reveals the true parent for given \( r \) and \( q \). Hence, one can see that for a fixed \( K \), large \( r \) leads to the increasing this probability due to the improvement for the quality of the direction answer. However, increasing \( r \) negatively affects the term \( K/(r+1) \) in (40), so that there is a trade off in selecting a proper \( r \). The detailed proof will be provided in our technical paper [22]. By considering the error probabilities, we obtain
\[
P[\hat{v} \neq v_1] \leq c \cdot e^{-K \log(K/r)} + e^{-3g(p,q)^2(K/r) \log r}
\]
\[
+ e^{-g(p,q)r} \log r \tag{47}
\]
\[
\leq (c + 1) e^{-g(p,q)^2(K/r) \log K},
\]
where \( c_1 = c + 1 \) and the inequality (a) is from the fact that \( g(p,q) < 1 \). By derivation of the result with respect to \( r \), we first obtain \( r^* \) which maximizes the detection probability by
\[
r^* = \left[ 1 + \frac{\delta (2(1-p)^2 + (1-q)^2) \log \log K}{3d-1} \right] \text{ in MVAD}(r^*)
\]
and put this into the error probability \( P[\hat{v} \neq v_1] \), we have
\[
P[\hat{v} \neq v_1] \leq (c + 1) e^{-2g(p,q)^2(K/(r^*)) \log(K/(r^*))}
\]
\[
\leq (c + 1) e^{-g(p,q)K^2 \log(K/(r^*))}
\]
\[
\leq (c + 1) e^{-g(p,q)K \log (K)}, \tag{48}
\]
where the inequality (a) is from the fact that \( K \) \( \log(K/(r^*)) \) \( \log K \) (b) comes from the obtained result of \( r^* \). Let \( \delta \geq 1 \).
\((c + 1)e^{-f_3(p,q)K \log(\log K)}\), then, we obtain the value of \(K\) which produces the error probability \(\delta\) in later and we obtain the desired lower-bound of \(K\) as in the theorem statement. This completes the proof.

E. Proof of Proposition 7

For simple understanding, we first see the case for single-hop distance which is more simple than that of general case. Consider \(d(v_1, v_{RC}) \leq 1\), that the rumor source is within the one-hop distant from the rumor center and for this case, the required number of budget is \(K = 1\). To obtain the detection probability, we consider the following lemma.

Proposition 2: The detection probability when \(K = 1\) is given by

\[
P[i = v_1] = P[v_1 = v_{RC}] + q \cdot P[d(v_1, v_{RC}) = 1],
\]

where

\[
P[d(v_1, v_{RC}) = 1] = \sum_{k=2}^{\infty} P[d(v_1, v_k) = 1] P[v_k = v_{RC}],
\]

and

\[
P[d(v_1, v_k) = 1] = \prod_{i=1}^{k-2} \left(1 + i(d - 2)\right) \cdot d.
\]

Next, from the single-hop distant case, we extend the result to multi-hop Extant case as follows. From [3], the first term \(P[v_1 = v_{RC}]\) can be directly obtained so it remains to find the second term \(P[d(v_1, v_{RC}) = 1]\). To obtain this, we will decompose the event \(d(v_1, v_{RC}) = 1\) by disjoint events, considering subcases where \(v_{RC} = v_k\) for different \(k\). Because the rumor center becomes a single node in case of \(K = 1\), the event \(\{v_{RC} = v_k\}\) can be decomposed into disjoint sub-events. First, let’s consider the case when \(v_{RC} = v_2\). Then, it is clear that \(d(v_1, v_2) = 1\) because \(v_1\) and \(v_2\) are always adjacent. In case of \(v_{RC} = v_k\) for \(k \geq 3\), \(v_k\) is not always adjacent to \(v_1\), and therefore we express \(P[d(v_1, v_{RC}) = 1]\) as

\[
P[d(v_1, v_{RC}) = 1] = \sum_{k=2}^{\infty} P[v_{RC} = v_k, d(v_1, v_k) = 1] = \sum_{k=2}^{\infty} \left( P[v_{RC} = v_k] P[d(v_1, v_k) = 1] \right),
\]

where the equality \((a)\) holds because for all \(k \geq 2\), the events \(\{v_{RC} = v_k\}\) and \(\{d(v_1, v_k) = 1\}\) are independent. This can be explained from the fact that \(\{d(v_1, v_k) = 1\}\) is caused from the first \(k\) infections, and \(\{v_{RC} = v_k\}\) is occurred from beyond \(k\)th infection, which makes there are no causal relationships between two events.\(^7\) From [3], the probability \(P[v_{RC} = v_k]\)

Therefore, we only need to characterize \(P[d(v_1, v_k) = 1]\). To obtain this, we need to count two kinds of permutations, (i) the total number of possible topological realizations of \(\{v_1, \ldots, v_k\}\), (ii) the number of realizations of \(\{v_1, \ldots, v_k\}\) satisfying \(d(v_1, v_k) = 1\). Fig. 4(a) describes the evaluation of the probability.

(i) The total number of possible realizations is given by

\[
P[d(v_1, v_k) = 1] = \prod_{i=1}^{k-2} \left(1 + i(d - 2)\right) \cdot d.
\]

(ii) The total number of realizations where \(d(v_1, v_k) = 1\) can be counted as follows. First we fix \(v_1\) and \(v_k\) as one-hop neighbor, then we place remaining nodes \(\{v_2, \ldots, v_{k-1}\}\) around \((d - 1)\) subtrees around \(v_1\). This makes \((d - 1)(d - 1 + d - 2)\cdots(d - 1 + (k - 3)(d - 2)) = \prod_{i=1}^{k-2} (1 + i(d - 2)).\) Then we multiply this number by \(d\) since there are \(d\) possible ways to place \(v_1\) and \(v_k\) adjacent.

Since all permutation have the equal probability due to the exponential distribution under regular tree, we obtain the probability \(Pr(d(v_1, v_k) = 1)\) as

\[
P[d(v_1, v_k) = 1] = \prod_{i=1}^{k-2} \left(1 + i(d - 2)\right) \cdot d.
\]

This completes the proof of Proposition 2.

In this subsection, we will provide the derivation of (20). Different to the case \(i = 1\), the probability \(P[d(v_1, v_k) = i]\) has more complex and general form. To find this, we first define \(G(a, b)\) for given \(a \in Z_+\) and \(b \in Z\) such that

\[
G(a, b) \triangleq \sum_{\mathcal{I} \subseteq \{1, \cdots, a\}, |\mathcal{I}| = b} \left( \prod_{i \in \mathcal{I}} x_i \right),
\]

where \(\mathcal{I}\) is any subset of \(\{1, \cdots, a\}\) with cardinality \(|\mathcal{I}| = b\) and \(x_i = 1 + i(d - 2)\). For example, \(G(a, 0) = 1\) and \(G(3, 2) = x_1 x_2 + x_2 x_3 + x_3 x_1\). Note that (20) can be described as the case when \(a = 1\), since \(G(k - 2, k - 2) = x_1 x_2 \cdots x_{k-2} = \prod_{i=1}^{k-2} (1 + i(d - 2)).\) For the more precisely understanding of \(G(k - 2, k - i - 1)\) in (20), consider the illustrative following example when \(k = 6, i = 3\). Note that all we have to do is just counting the number of possible topologies satisfying \(d(v_1, v_6) = 3\). Fig. 4(b), 4(d) show the example of deriving \(Pr(d(v_1, v_6) = 3)\). Fig. 4(b) shows one possible realization partially satisfies the conditions. Here in nodes \(a\) and \(b\), nodes among \(\{v_2, v_3, v_4, v_5\}\) can be placed, making \(\binom{k-i-1}{2} = \binom{4}{2}\) possible node placements. After dividing into sub-cases, we consider the number of possible topologies one by one.

- Case 1: In Fig. 4(c), the remaining unplaced nodes \(\{v_4, v_5\}\) can be placed in any subtrees, since they are all infected after both \(v_2\) and \(v_3\) are infected. The node \(v_4\) can be placed among \((d - 1) + 2(d - 2))\) subtrees,
and this placement makes \((d - 2)\) additional subtrees. So the node \(v_5\) can be placed among \(((d - 1) + 3(d - 2))\) subtrees. This makes the number of possible topologies \((1 + 3(d - 2))(1 + 4(d - 2)) = x_3 x_4\) in Fig. 4(c).

- **Case II:** We can apply the same argument in the subcase stated in Fig. 4(d). In this case, among remaining nodes \(\{v_2, v_3\}\), the node \(v_2\) can be placed in only first subtree since \(v_3\) is “blocking” the infection path and \(v_2\) cannot be located beyond \(v_3\). The node \(v_5\) can be placed in any subtrees. This makes the number of possible topologies \((d - 1)(1 + 4(d - 2)) = x_1 x_4\) by considering similar combinatorics as in Fig. 4(c).

From above subcases, we can infer that there are \(x_1 x_2 + x_1 x_3 + x_1 x_4 + x_2 x_3 + x_2 x_4 + x_3 x_4 = G(4, 2)\) possible realizations when the location of \(v_1\) and \(v_0\) are fixed. Fixing the location of \(v_1\), the node \(v_0\) can be placed among \((d - 1)(d - 1) = d(d - 1)^2\) possible positions. This makes the overall \(G(4, 2) \cdot d(d - 1)^2\) possible topologies. In general, one can easily obtain

\[
\text{# of possible topologies} = G(k - 2, k - i - 1) \cdot d(d - 1)^{i-1}.
\]

The denominator of (20) is same as the that of single-hop case, which refers the number of possible topological realizations of \(k\) infections. This conclude the proof of Proposition 1.

**F. Deriving \(r^*\) in (23)**

First, for given \(K, r, d, w\), we define

\[
s_d(K, r, p, q) := c \cdot e^{-\frac{\log((K-1)(1-H(p)) + p(1-p)(1-H(q)))}{\log(r)}} - \frac{\log(K-1)(1-H(p)) + p(1-p)(1-H(q))}{\log(r)}
\]

where \(c = 4d/(d-1)\). To obtain \(\frac{\partial s_d(K, r, p, q)}{\partial r} = 0\), first consider that

\[
\frac{\partial s_d(K, r, p, q)}{\partial r} = c(K/r)^{-\frac{\log((K-1)(1-H(p)) + p(1-p)(1-H(q)))}{\log(r)}} (\log((K-1)(1-H(p)) + p(1-p)(1-H(q))))
\]

and we also have that

\[
\frac{\partial s_d^2(K, r, p, q)}{\partial r} = \frac{h(p, q)K^2 \log(K/2r) + K - 1}{(K/r)^2 \log(K/2r)}
\]

where \(h(p, q) = ((1-H(p)) + p(1-p)(1-H(q)))\). By setting \(h(p, q) = 0\) and after some algebra, we obtain

\[
\log r \approx \log \left(1 + \frac{4(1-p)(7H(p) + H(q))}{3e \log(d - 1)}\right),
\]

which implies that \(r^* = \Theta \left(\frac{(4(1-p)(7H(p) + 3e \log(d - 1))) \log K}{3d \log(d - 1)}\right)\) and this completes the proof.

**G. Deriving \(r^*\) in (23)**

First, for given \(K, r, p, q, d\), we define

\[
f_d(K, r, p, q) := c \left(\frac{r + p + q}{r + 2}\right)^3 \cdot \left(\frac{\log((K-1)(1-H(p)) + p(1-p)(1-H(q)))}{\log(r)}\right) - \frac{\log((K-1)(1-H(p)) + p(1-p)(1-H(q)))}{\log(r)}
\]

where \(c = 7(d+1)/d\) and \(w_d(p, q) = \frac{1}{2}(4(p-1/2)^2 + d/(d-1)^3(q - 1/d)^t)\). The term \(h_d(K, r)\) is given by

\[
h_d(K, r) := \frac{\log(K/r)}{\log(d - 1)} \cdot \frac{\log((K-1)(1-H(p)) + p(1-p)(1-H(q)))}{\log(r)}
\]

Then one can check that this is a differential unimodular function with respect to \(r\). Hence, by taking derivative with respect to \(r\) to obtain the optimal choice of \(r\), we have

\[
\frac{\partial f_d(K, r, p, q)}{\partial r} = \frac{\partial}{\partial r} \left(\frac{r + p + q}{r + 2}\right)^3 \cdot \left(\frac{\log((K-1)(1-H(p)) + p(1-p)(1-H(q)))}{\log(r)}\right)
\]

To finish this, we first consider that \(\frac{\partial}{\partial r} \left(\frac{r + p + q}{r + 2}\right)^3 = \frac{3(r + p + q)}{(r + 2)^2}\), and

\[
\frac{\partial}{\partial r} (c \cdot \exp(-h_d(K, r)w_d(p, q)/2)) = \frac{c \cdot \exp(-h_d(K, r)w_d(p, q)/2)}{2r \log(d - 1)}
\]

where \(Q_d(K, r) = \log(K/r) / \log(d - 1)\). From the relation \(h_d(K, r) = Q_d(K, r) \log(Q_d(K, r))\) and \(\frac{\partial}{\partial r} (c \cdot \exp(-h_d(K, r)w_d(p, q)/2)) = 0\), we have

\[
3(p + q - 2)(p + q + r)^2 = \frac{c(K/r)^{-\frac{\log((K-1)(1-H(p)) + p(1-p)(1-H(q)))}{\log(r)}} (\log((K-1)(1-H(p)) + p(1-p)(1-H(q))))}{2r \log(d - 1)}
\]

where \(Q_d(K, r) = \log(K/r) / \log(d - 1)\). From the relation \(h_d(K, r) = Q_d(K, r) \log(Q_d(K, r))\) and \(\frac{\partial}{\partial r} (c \cdot \exp(-h_d(K, r)w_d(p, q)/2)) = 0\), we have

\[
3(p + q - 2)(p + q + r)^2 = \frac{c(K/r)^{-\frac{\log((K-1)(1-H(p)) + p(1-p)(1-H(q)))}{\log(r)}} (\log((K-1)(1-H(p)) + p(1-p)(1-H(q))))}{2r \log(d - 1)}
\]
By taking logarithm to both sides and after some algebra, 
\[
\log(1-p) - 2\log(r+1) \\
\simeq -\log Q_d(K, r)w_d(p,q)\frac{\log(K/r)}{\log(d-1)} + \log\log Q_d(K, r) \\
+ \log c - \log r - \log\log(d-1) \\
\simeq -\log\log K - e\log r - 1 + \log(1-q)^2 - \log\log(d-1)^2,
\]
where we use \(Q_d(K, r) = \log(K/r)/\log(d-1)\). Then we have
\[
\log r \simeq \log \left(1 + \frac{2(1-p)(1+q^2)\log K}{e\log(d-1)} \right),
\]
which implies that \(r^* = \Theta \left(\frac{2(1-p)(1+q^2)\log K}{e\log(d-1)} \right)\) and this completes the proof.

**H. Deriving \(r^*\) in (38)**

For given \(K, r\) and \(d\), let
\[
t_d(K, r, p, q) := c \cdot e^{-(K/r)\log(K/r)} - \frac{K\log(K/r)}{r}\frac{K\log(K/r)}{r}(\log(K/r) - 1),
\]
where \(h(p, q) = (1 - H(p)) + (1 - p)(\log_2 d - H(q))\) and \(c = 4d/3(d-2)\). Then one can check that this is a differential unimodula function with respect to \(r\). We need to find \(\partial t_d(K, r, p, q) / \partial r = 0\) with respect to \(r\) so that first consider
\[
\partial \left( c \cdot e^{-(K/r)\log(K/r)} \right) / \partial r = cK \frac{K}{r}\frac{K}{r}(\log(K/r) + 1).
\]
Next, we also obtain that
\[
\partial \left( \frac{K\log(K/r)}{r}\frac{K\log(K/r)}{r}(\log(K/r) - 1) \right) / \partial r = \frac{(K+1)h(p, q) \log(Kr/2) - 2}{K\log(Kr/2) - 1}.
\]
Let the above two equations are equal and after some algebra, we obtain
\[
e^{\left(\frac{2(1-p)(1+q^2)\log K}{e\log(d-1)} \right)} \simeq \log K.
\]
By taking logarithm to both side then
\[
r \simeq 1 + \frac{7dp(3H(p) + 2dH(q))\log\log K}{2(d-1)}.
\]
Therefore, we have \(r^* \simeq 1 + \frac{7dp(3H(p)+2dH(q))\log\log K}{2(d-1)}\) for given \(K\) and \(q\) under \(d\)-regular tree and completes the proof.

**I. Deriving \(r^*\) in (40)**

For given \(K, r\) and \(d\), define
\[
g_d(K, r, p, q) := \\
\exp \left[ - \left( p - \frac{1}{2} \right)^2 \left( \frac{K}{r} \right) \log \left( \frac{K}{r} \right) \right],
\]
where \(J_d(r, q) := 1 - e^{-r(d-1)(q-1/d)^2} \) and \(c = (8d + 1)/d\). Then one can check that this is a differential unimodula function with respect to \(r\). Then, by taking derivative \(g_d(K, r, p, q)\) with respect to \(r\) and by setting \(\partial g_d(K, r, p, q) / \partial r = 0\), we have
\[
\partial g_d(K, r, p, q) / \partial r \\
= \frac{\partial(cJ_d(r, q))}{\partial r} \cdot \exp \left[ - \left( p - \frac{1}{2} \right)^2 \left( \frac{K}{r} \right) \log \left( \frac{K}{r} \right) \right] \\
+ cJ_d(r, q) \frac{\partial}{\partial r} \left( e^{-2(p-1/2)^2(\log(K/r))} \log(\log(K/r)) \right) = 0.
\]
Then, we first see that
\[
\partial(cJ_d(r, q)) / \partial r = \frac{c(d-1)(q-1/d)^2}{d(1-q)} \\
\cdot e^{-r(d-1)(q-1/d)^2} \left( e^{-\frac{r(d-1)(q-1/d)^2}{d(1-q)}} - 1 \right)^2
\]
Next, consider that
\[
\partial \left( e^{-2(p-1/2)^2(\log(K/r))} \log(\log(K/r)) \right) / \partial r \\
= \frac{1}{K(1-2p)^2(K/r)^2} \cdot \frac{K(1-2p)^2(K/r)^2}{4r^2} (\log(K/r)+1).
\]
Using these facts and after some algebra, we obtain
\[
e^{\left(\frac{2(1-p)(1+q^2)\log K}{e\log(d-1)} \right)} \simeq \log K.
\]
By taking logarithm to both side then
\[
r \simeq 1 + \frac{7d^2(2(1-p)^3 + (1-q)^2)}{3(d-1)} \log\log K.
\]
Therefore, we have \(r^* \simeq \Theta \left(\frac{7d^2(2(1-p)^3 + (1-q)^2)}{3(d-1)} \log\log K \right)\) for given \(K\) and \(q\) under \(d\)-regular tree and completes the proof.