Abstract—To facilitate more efficient control, massive amounts of sensors or measurement devices will be deployed in the Smart Grid. Data collection then becomes non-trivial. In this paper, we study the scenario where a data collector is responsible for collecting data from multiple measurement devices, but only some of them can communicate with the data collector directly. Others have to rely on other devices to relay the data. We first develop a communication protocol so that the data reported by each device is protected against honest-but-curious data collector and devices. To reduce the time to collect data from all devices within a certain security level, we formulate our approach as an integer linear programming problem. As the problem is NP-hard, obtaining the optimal solution in a large network is not very feasible. We thus develop an approximation algorithm to solve the problem. We test the performance of our algorithm using real topologies. The results show that our algorithm successfully identifies good solutions within reasonable amount of time.

I. INTRODUCTION

One of the significant hallmarks of the Smart Grid (SG) initiative is the pervasive data sensing to facilitate a more efficient control. While the existing Supervisory Control And Data Acquisition (SCADA) system already collects data from various sensors, the scale and scope of the data collection in the SG are expected to pose new challenges. Applications of data sensing in the SG include conditional and structural monitoring of Distributed Energy Resources and renewables in the generation; State-of-Charge monitoring; substation, transformer, underground and overhead lines in the transmission and distribution; and collection of information from smart meters in the Advanced Metering Infrastructure (AMI) [1].

In addition to the emerging and new sensing in the SG, data collection from the legacy telemetric devices widely deployed in the field needs to be accommodated as part of the infrastructure, at the very least, in the transitional period. An example of such legacy telemetric collection need is provided in [2]. In this scenario, a mobile data collector (DC) moves along a road, or a certain path, to collect data from measurement devices (MDs). It can only connect directly to the MDs that are within DC’s proximity. We call the MDs that can talk to the DC directly root MDs. Those MDs that are outside the communication range of the DC have to send data to the DC through root MDs in a multihop manner. Figure 1 presents a simple example where only root MDs (MD1, MD2, MD3) can talk to the DC directly. Other remote MDs should send their data to one of the root MDs. Note that the underlying data collection paradigm of the aforementioned case is present in many other emerging sensing and measurement scenarios [1], [3], especially when sequential data collection is needed over a multihop communications topology. To reduce the number of messages needed for the whole data collection, we propose generating several disjoint trees, each rooted at a certain MD, such that the trees reach all MDs. Fig. 2 shows an example, where bold edges show the trees formed from the roots.

In this paper, we are interested in secure, scalable, and optimized data collection in the SG. Our objective is to reduce the total data collection time while ensuring the confidentiality of the data. We encrypt data in a way that while allowing the relaying MDs to verify the message integrity, they cannot read the content under the honest-but-curious model. The data collection time is reduced through identifying data collection trees that have minimum sum of depths. We develop an optimization problem for the secure data collection, which is NP-Hard, and then develop an approximation algorithm.

The rest is organized as follows: Sec. II presents the related work. Tree-based secure data collection is detailed in Sec. III. Sec. IV develops the optimization problem and solution together with an approximation algorithm. Simulation results are given in Sec. V. Sec. VI concludes the paper.

II. RELATED WORK

To the best of our knowledge, confidential data collection with integrity validation by means of forming disjoint trees to minimize the total collection time has not been studied in the literature. We dissect our problem into three constituents in order to present a discussion of the partially related work: (1) Multi-sink data collection in Wireless Sensor Networks (WSNs) and opportunistic networks [4], (2) Tree or cluster formation, and (3) Secure data collection.

There is some similarity in our formulation to data collection via mobile relay nodes [5], except that the objective of our approach is not energy efficiency Data collection by means of mobile sinks [6] is somewhat related in the sense that we are looking to choose the root of trees to relay the data to. The optimization part of our problem is similar to the Base...
Station Problem from [7] where the problem of positioning data collecting nodes in a WSN is studied as a maximum flow problem with the objective of finding the optimal data rate. The NP-completeness proof of the problem is also given in [7]. While retaining the complexity\(^1\), our problem is about the minimization of total time of the data collection and incorporates the security aspect. The tree formation of our approach connotes clustering algorithms, especially in WSNs [9]. However, unlike the goals of WSN clustering on node reachability and network longevity, we focus on security and data collection time minimization.

Finally, as for the security of the data collection task, there are two major approaches: One is to ensure the protection of the data content directly without regard to the data semantics. An approach presented in [10] is based on symmetric data processing (also referred to as cryptography to provide data confidentiality and authenticity. An approach presented in [10] is based on symmetric the data content directly without regard to the data semantics.

The security objective of the data collection is to protect the data collected. First, the PO determines the disjoint trees to be used for data collection. Then, it provides the tree information and the necessary key information to the DC. The DC then talks to each root MD along its path to collect data. Root MD sends the key information along each branch on its tree to collect data. After it receives data from all MDs in its tree, it sends the data to DC.

B. Data Collection on a Branch

We first describe a simple situation that data is collected along a certain branch on the tree. We then describe the data collection of a whole tree. Fig. 3 presents how DC collects data on a tree branch spanning from \( MD_1 \), the root MD, to \( MD_2 \), and then \( MD_3 \). Let the branch be \( MD_1 \to MD_2 \to \ldots MD_i \), where \( MD_1 \) is the root MD. The public DH key of PO is \( g^c \) (mod \( p \) is dropped for brevity), and the group key is \( GK \). The key information created by the PO is denoted by \( C = g^c | GK \), where \( | \) represents the concatenation operation. Each MD on the branch must receive \( C \) to develop a shared key with the PO. The detailed procedure of data collection is as follows:

1) \( DC \to MD_1: DC^+ || SIG(PO, C) || SIG(DC, C) || PKE(MD_1^+, C), 1 \leq k \leq l \)

To protect \( C \) from eavesdroppers, \( C \) is encrypted using the public keys of the MDs. As only the PO knows the public keys of all the MDs, \( PKE(MD_k^+, C) \) for \( 1 \leq k \leq l \) are created by PO and sent to DC, which will be discussed in Section III-C. \( MD_1 \) authenticates \( C \) by verifying the signature \( SIG(PO, C) \). \( MD_1 \) verifies the legitimacy of DC by verifying \( SIG(DC, C) \). \( MD_1 \) then signs \( C \) using its public key and sends the signature together with other information to the next MD along the branch.

2) \( MD_1 \to MD_2: MD_1^+ || SIG(PO, C) || SIG(MD_1, C) || PKE(MD_2^+, C), 2 \leq k \leq l \)

Note that as \( MD_1 \) also sends \( MD_1^+ \) to \( MD_2 \) so that \( MD_2 \) does not need to know any public key of any MD before the communication. When \( MD_2 \) receives the message, it performs similar operations as \( MD_1 \) that it authenticates \( C \) and \( MD_1 \). It

\(^1\)Complexity of the optimization part of our problem may also be obtained from the classical multi-facility location problem [8].

\(^2\)\( p \) is a prime number, and \( g \) is a primitive root mod \( p \). Let A pick a secret 
a and B pick a secret b. A sends \( g^a \) mod \( p \) to B, and B sends \( g^b \) mod \( p \) to A. A can then compute the shared key by \( (g^a \cdot g^b)^p \) mod \( p \). B computes the key by \( (g^b \cdot g^a)^p \) mod \( p \).
then passes the key information to the next MD (more details will be provided in Section III-C).

3) \( MD_i \rightarrow MD_{i+1} : MD_i^+ || \text{SIG}(PO,C) || \text{SIG}(MD_{i+1},C) || [\text{PKE}(MD_j^+,C),i+1 \leq k \leq l] \)

Intermediate MDs on a tree branch keep forwarding the key information to the next MD after retrieving \( \text{GK} \) and \( g^d \) from \( C \).

4) \( MD_{l-1} \rightarrow MD_l : \text{MD}_{l-1}^+ || \text{SIG}(PO,C) || \text{SIG}(MD_{l-1},C) || \text{PKE}(MD_l^+,C) \)

\( MD_l \) is the last MD on the branch. After verifying \( C \), it prepares the reported data. It first generates its own DH half key \( g^d \) and develops the shared key between itself and the PO, which is \( g^{d_l} \). Let \( DATA_l \) be the data encrypted and integrity protected using \( g^{d_l} \). \( MD_l \) then sends \( DATA_l \) to \( MD_{l-1} \) for relaying to DC and finally to PO. To allow \( MD_{l-1} \) to perform authentication check on \( DATA_l \), \( MD_l \) computes the hash of \( DATA_l \) using \( g^{d_l} \). \( MD_l \) should also let \( MD_{l-1} \) and the PO knows what \( g^{d_l} \) is. Let message \( M_k = \text{SKE}(\text{GK},g^{d_l}) || \text{SIG}(MD_k,g^{d_k}) || \text{DATA}_k \) for \( 1 \leq k \leq l \). \( M_l \) contains the DH half key and the encrypted data to be delivered to the PO.

5) \( MD_l \rightarrow MD_{l-1} : MD_l^+ || M_l || \text{HASH}(g^{d_l},DATA_l) \)

\( MD_{l-1} \) first retrieves \( g^{d_l} \) from \( \text{SKE}(\text{GK},g^{d_l}) \) contained in \( M_l \) and verifies \( \text{SIG}(MD_l,g^{d_l}) \). It then can authenticate the message by verifying the hash. Note that although \( MD_{l-1} \) knows \( g^d \) and \( g^{d_l} \), according to the property of DH protocol, \( MD_{l-1} \) still cannot compute \( g^{d_l} \) to decrypt \( DATA_l \). Thus, \( DATA_l \) remains secret to \( MD_k \) for all \( k = 1, \ldots, l-1 \) and the DC. After verifying the message, \( MD_{l-1} \) prepares its own encrypted data \( DATA_{l-1} \) and sends both \( DATA_l \) and \( DATA_{l-1} \) to \( MD_{l-2} \) together with the required key information.

6) \( MD_{l-1} \rightarrow MD_{l-2} : \)

\( MD_{l-1}^+ || M_{l-1} || \text{HASH}(g^{d_{l-1}},DATA_{l-1}) || M_l \)

\( MD_{l-2} \) processes the message in a similar manner as \( MD_{l-1} \) does. Finally, \( MD_1 \) would receive a message from \( MD_2 \) that contains all the data from \( MD_2 \) to \( MD_l \). It can then prepare a message that contains all the data on the branch for the DC.

7) \( MD_1 \rightarrow DC : \)

\( MD_1^+ || M_1 || M_2 || \ldots || M_l || \text{HASH}(g^{d_1},DATA_{l-1} || M_2 || \ldots || M_l) \)

C. Tree Structure Representation

To facilitate a root MD to collect data for its tree, it has to know which neighbors belong to its tree. The root MD also needs to inform its children of their children. In other words, the tree structure has to be embedded in the message from the DC to the root MD, and passed along to the MDs on the tree. \( C = g^t || \text{GK} \) has to be encrypted using the public key of each MD on the tree. We use \( \text{PKE}(MD_j^+,C) \) for all \( MD_j \) on the tree to represent the tree. Let \( E(i) = \text{PKE}(MD_1,C) \), and let \( T(i) \) be the tree representation rooted at \( MD_i \). We further let \( MD_{ch_1}, \ldots, MD_{ch_m} \) be the children of \( MD_i \) if it is not a leaf.

The representation is as follows:

\[
T(i) = \begin{cases} 
E(i), T(ch_1), \ldots, T(ch_m) & \text{if } MD_i \text{ is not a leaf} \\
E(i) & \text{if } MD_i \text{ is a leaf}
\end{cases}
\]

Refer to the tree rooted at \( MD_2 \) in Fig. 2, \( T(2) = [E(2), E(6), [E(8), E(9)]] \).

When \( MD_i \) receives \( T(i) \), it can retrieve \( C \) from \( E(i) \). It can also identify its children to forward key information. Let \( MD_j \) be a child of \( MD_i \), it sends \( MD_j^+ || \text{SIG}(PO,C) || \text{SIG}(MD_j,C) || T(j) \) to \( MD_j \).

An MD should wait for all its children to report data before sending its data to its parent. Data reporting does not have to follow the tree structure. A parent MD can simply append all \( M_k \) of descendant \( MD_k \) together. For example, \( MD_2 \) in Fig. 2 can send DC \( MD_2^+ || M_2 || M_6 || M_8 || \text{HASH}(g^{d_2},DATA_2) || M_6 || M_8 || M_9 \).

D. Completing the Protocol

We now complete the protocol by describing the communication between the PO and the DC. Fig. 4 illustrates the information exchange. They first use the DH protocol to establish a shared secret \( K \) to secure the subsequent conversation. PO picks the group keys to be used. The tree structure and the key information are then encrypted using \( K \). Suppose there are \( k \) root MDs, \( MD_{r_1}, MD_{r_2}, \ldots, MD_{r_k} \). PO can use different DH half keys and group keys to develop shared keys with the MDs on different trees. Let the DH half key and the group key for the tree rooted at \( MD_{r_i} \) be \( g^{d_i} \) and \( G_{r_i} \), respectively. We further let \( C_i = g^{c_i} || G_{r_i} \). For root MD \( MD_{r_i} \), PO sends \( \text{SKE}(K,C_i || \text{SIG}(PO,C_i)) || T(h_i) \) to DC. DC can then verify the signature for each \( C_i \) and talk to the root MDs to collect data as described earlier. After all data are collected, DC encrypts \( M_f \) for each \( MD_i \) it collects using \( K \). Note that \( M_j = \text{SKE}(G_{r_i},g^{d_j}) || \text{SIG}(MD_j,g^{d_j}) || DATA_j \), which belongs to the tree of \( MD_{r_i} \), contains enough information for the PO to retrieve and verify the data.

E. Security Discussion and Its Complexity

In our protocol, the confidentiality of the data is ensured through the Diffie-Hellman keys. The DH half keys are transmitted under encryption. Eavesdroppers cannot read them. The data remains confidential to the honest-but-curious DC and intermediate MDs as well because, although they can read the public DH keys exchanged between the PO and the MD, they cannot establish the DH key themselves. Key information from the root to the leaves are authenticated through signatures, while data from the leaves to the root are authenticated through the hashes in a hop-by-hop manner. Although a tampering on data can be detected by the PO eventually, by authenticating the data in a hop-by-hop manner, tampering can be identified
early so that network resources would not have been spent on transmitting the tampered data from the leaf to the PO. Our protocol is thus secure from both active and passive attacks when the keys are not leaked. (We consider the risk of leaking the group key in our problem formulation in Sec. IV-A).  

We now study the computational complexity of an MD. We only consider public key and DH operations since symmetric key and hash operations are not expensive. The leaf MD receives one message from its parent and sends one message to its parent in each data collection cycle. It has to decrypt the key information and verify two signatures for the message received. To prepare the reply, it generates a DH half key and signs it. Five operations are needed in total for a leaf node. Similar operations are needed for the messages to and from the parent for each non-leaf MD. To send messages to its children, an MD has to sign the key information. For each reply from a child, it has to verify the signature of the half key provided by the child. Therefore, the total number of operations a non-leaf MD needs to perform is 6 + number of children.

IV. Collection Tree Formation

A. Problem Description

In this section, we describe how PO constructs the disjoint data collection trees on the network connectivity graph G of MDs. Since the PO has complete information of the topology, graph G, a connected vertex set of all MDs, is known a priori. We use set $\mathcal{M} = \{1,2,\cdots\}$ to denote the index set of these MDs. Furthermore, if there is a direct connectivity between two MDs then an undirected edge exists between them in graph G. Furthermore, we define the set of all MDs that can potentially be selected as the root MDs candidate set, and we use $\mathcal{R}$ to denote the set of indices of these candidate root MDs. The largest possible candidate set consists of all the MDs that can be physically within the communication range of the DC when it is traveling along the predefined data collection path. For example, in Fig. 1, $\mathcal{R} = \{1,2,3\}$.  

Our objective is to minimize the sum of the data collection time over all trees. We further assume identical link delays. Then, we can use the depth of a data collection tree to represent the time needed to collect data from this particular tree and the summation of the depths over all constructed trees to represent the overall data collection time.

Apart from the time to collect data, we also consider the security level of the group key. As the MDs may not be in a very secure physical environment, there is a risk of leaking the group key. If the group key is stolen, the attacker can decrypt $SK_e(GK, g^{sk})$ in $M_k$ to obtain $g^{sk}$ and create the correct hash of fake data. Although the PO can finally detect the data were not legitimate, network resource will be wasted in transmitting the message. Assume that every MD $i$ ($i \in \mathcal{M}$) leaks the group key with probability $p_i$. Then the probability that the group key of a tree $T$ is leaked is $P_{\text{leak}}(T) = 1 - \prod_{i \in \mathcal{M}_D \in T} (1 - p_i)$.

To ensure the security level of every constructed tree $T$, we limit $P_{\text{leak}}(T)$ to be no larger than some predefined threshold. We assume identical $p_i$’s. Then, if we limit the group key leaking probability to be no larger than some threshold value $P_{\text{threshold}}$, we have:

$$P_{\text{leak}}(T) = 1 - \prod_{i \in \mathcal{M}_D \in T} (1 - p_i) = 1 - (1 - p)^{|T|} \leq P_{\text{threshold}}$$

In other words, the cardinality of every constructed tree should satisfy

$$|T| \leq \frac{\log(1 - P_{\text{threshold}})}{\log(1 - p)} = N_{\text{threshold}} \quad (1)$$

As detailed in the next subsection, we formulate the optimization problem with a Min $- \sum - \max$ objective to minimize the summation of the depths over all constructed trees with the aforementioned security constraint.

B. Mathematical Formulation

We state the optimal collection tree formation (CTF) problem into a binary integer programming formulation as follows:

$$\min \sum_{j \in \mathcal{R}} \left( \max_{i \in \mathcal{M}} \sum_{k \in \mathcal{K}_{ij}} x_{ij}^k L_{ij} \right) \quad (2)$$

s.t. $\sum_{j \in \mathcal{M}} \sum_{k \in \mathcal{K}_{ij}} x_{ij}^k \leq N_{\text{threshold}}, \forall j \in \mathcal{R} \quad (3)$

$$\sum_{j \in \mathcal{R}} x_{ij}^k = 1, \forall i \in \mathcal{M} \quad (4)$$

$$x_{ij}^k \leq x_{mj}^i, \forall i \in \mathcal{M}, m \in \mathcal{N}_i, \mathcal{P}_{mj}^k \subseteq \mathcal{P}_{ij}^k \quad (5)$$

$$x_{ij}^k \in \{0,1\}, \forall i \in \mathcal{M}, j \in \mathcal{R}, k \in \mathcal{K}_{ij} \quad (6)$$

$\mathcal{N}_i$: the set of neighbors of $MD_i$, $\mathcal{P}_{ij}^k$: the $k_{th}$ shortest path from $MD_i$ to $MD_j$ in graph G. There can be multiple equal length shortest paths between any pair of MDs in graph G. For example, in Fig. 1 two shortest paths, $MD_8 \rightarrow MD_2 \rightarrow MD_1$ and $MD_8 \rightarrow MD_6 \rightarrow MD_1$ exist between $MD_8$ and $MD_1$. Then, we have $\mathcal{P}_{18}^1 = MD_8 \rightarrow MD_2 \rightarrow MD_1$ and $\mathcal{P}_{18}^2 = MD_8 \rightarrow MD_6 \rightarrow MD_1$. $K_{ij}$: set of indices of the shortest paths between $MD_i$ and $MD_j$. For example, in Fig. 1 because two shortest paths exist between $MD_8$ and $MD_1$, we have $K_{81} = \{1,2\}$. $L_{ij}$: number of MDs in the shortest paths between $MD_i$ and $MD_j$. In the example in Fig. 1, we have $L_{18} = 3$. With Eq. 1, $N_{\text{threshold}}$: maximum number of MDs allowed in one tree.

We associate a binary variable $x_{ij}^k$ with every shortest path $\mathcal{P}_{ij}^k$ between $MD_i$ ($\forall i \in \mathcal{M}$) and $MD_j$ ($\forall j \in \mathcal{R}$). For all $MD_i$’s out of the candidate set and $MD_j$’s in the candidate set, that is $\forall i \in \mathcal{M} \setminus \mathcal{R}$ and $\forall j \in \mathcal{R}$, we have

$$x_{ij}^k = \begin{cases} 1 & \text{if } MD_i \text{'s data is collected along } \mathcal{P}_{ij}^k \text{ to } MD_j \\ 0 & \text{if } MD_i \text{'s data is not collected along } \mathcal{P}_{ij}^k \text{ to } MD_j \end{cases}$$

Furthermore, for all $MD_i$ belonging to the candidate set, that is $\forall i \in \mathcal{R}$, we have $x_{ii}^k = 1$ if $MD_i$ is selected as a root; otherwise, $x_{ii}^k = 0$.

In the objective function (2), $\max_{i \in \mathcal{M}} \sum_{k \in \mathcal{K}_{ij}} x_{ij}^k L_{ij}$ for some fixed $j \in \mathcal{R}$ represents the depth of the tree rooted at candidate root $MD_j$. Furthermore, $\max_{i \in \mathcal{M}} \sum_{k \in \mathcal{K}_{ij}} x_{ij}^k L_{ij} = 0$ if $MD_j$ is not chosen as the root of any collection tree. Therefore, by summing over $\forall j \in \mathcal{R}$, we have the objective function (2) representing the total depth over all constructed trees.
The inequality constraints (5) ensure that any MD, can send data to MDj over path \( P_{ij} \) only if one of MD’s neighbors MDm chooses to send data to MDj over \( P_{mij} \), a sub-path of path \( P_{ij} \). Refer to Fig. 1, suppose \( P_{51}^1 = MD_5 \rightarrow MD_4 \rightarrow MD_1 \) and \( P_{41}^1 = MD_4 \rightarrow MD_1 \) with \( P_{51}^1 \subseteq P_{41}^1 \), then we have \( x_{51}^1 \leq x_{41}^1 \), which means MD5 can choose to send data to MD1 over path \( P_{51}^1 \) only if MD4 chooses to send data to MD1 along path \( P_{41}^1 \).

In a word, the single path constraints (4) and the sub-path constraints (5) together ensure that we construct multiple disjoint trees rooted at all or a subset of the MDs in the candidate set. Furthermore, the security constraints (3) further ensure that the number of nodes in every constructed tree is no larger than the threshold.

C. Solution and Analysis

To solve CTF, we firstly transform it into the standard format by rewriting (2) as follows:

\[
\min \sum_{j \in \mathcal{R}} z_j \quad (7)
\]

\[
s.t. \quad z_j \geq \sum_{k \in K_{ij}} x_{ij}^k, \forall i \in \mathcal{M}, j \in \mathcal{R} \quad (8)
\]

(7) and (8), together with constraints (3) to (6), form the modified CTF. The modified-CTF problem is equivalent to the original CTF problem in terms of the optimal objective function value. Nevertheless, the mixed-integer programming problem is an NP-hard problem. Thus, we propose to use an approximation algorithm by means of a linear relaxation based iterative rounding (LR-IR) [21] as shown in Algorithm 1. It is obvious that the worst-case number of iterations of Algorithm 1 is \( O(N) \) with respect to the number of decision variables \( x_{ij}^k \). Also, Algorithm 1 terminates when each \( x_{ij}^k \) equals to 0 or 1.

V. SIMULATION

In our simulation, we use the SG data set for Washington, DC [22], which contains the exact positions of all the utility poles in the city. We extract the positions of 300 utility poles from two portions of the map, as shown in Fig. 5 and Fig. 6.

Specifically, there are 8 stars in both figures representing MDs in the candidate root set. In Fig. 5, we assume that the DC is traveling along the southernmost street. In Fig. 6, the DC is assumed to be traveling along the westernmost street. Then, we pick evenly distributed 8 MDs along the two streets to be the candidate root sets. Furthermore, we assume that MDs communicate with each other wirelessly with an identical communication range of 100m. Also, topology II has a large MD density(826/km²) than than topology I (722/km²). We used GUROBI solver [23] for our simulations.

In our simulation, we construct a scenario by choosing \(|\mathcal{M}|\) MDs nearest to the reference streets. For example, in topology I, 50 MDs that are nearest to the southernmost street in Fig. 5 are selected to construct the scenario.

In Fig. 7, we compare the total depth of the constructed trees in scenarios with different number of MDs generated by three algorithms: The optimal algorithm is the solution from GUROBI or the modified-CTF problem. The approximation algorithm refers to Algorithm 1. In the random tree formation algorithm, every MD randomly selects a neighbor which has been included in a tree to be its parent. Also, in Fig. 7, two cases are considered in both topologies, namely \( N_{\text{threshold}} = 40\% \times |\mathcal{M}| \) and \( N_{\text{threshold}} = 80\% \times |\mathcal{M}| \). For example, if we assume that every MD leaks the group key with probability \( p = 0.01 \) and \(|\mathcal{M}| = 100\), then \( N_{\text{threshold}} = 40\% \times |\mathcal{M}| \) means we set the threshold probability \( P_{\text{threshold}} = 1 - (1 - p)^{40\% \times |\mathcal{M}|} = 1 - (1 - 0.01)^{40\% \times 100} = 0.331 \).

In Fig. 7, our approximation algorithm yields a total tree depth which is much smaller than that of the random tree formation algorithm while staying fairly close to the optimal value. As the maximum tree size \( (N_{\text{threshold}}) \) increases, both the optimal and the approximation algorithm tend to yield decreasing values of the total tree depth. This observation captures the trade-off between efficiency and security of our optimization formulation. We will further show it in Fig. 9.
| $|\mathcal{M}|$ | 25 | 50 | 75 | 100 |
|----------------|-----|-----|-----|-----|
| Top. I Optimal | 1.953 | 73.956 | 1418.134 | 6422.439 |
| Top. I Approx. (40%) | 0.048 | 1.892 | 12.309 | 15.319 |
| Top. II Optimal | 2.332 | 1650.730 | 6503.994 | 36069.112 |
| Top. II Approx. (40%) | 0.089 | 1.807 | 4.235 | 5.254 |

TABLE I

Algorithm Completion Time in Seconds.

Furthermore, from Table I, we observe that when $|\mathcal{M}| = 100$, the optimal algorithm already has to take more than 6000s to complete. This indicates that when solving the modified-CTF problem for a network of more than 100 MDs, it is better to turn to our approximation algorithm. Hence, we carry out simulation in scenarios with up to 300 MDs whereby our approximation algorithm terminates in reasonable time period.

From Fig. 8, we can observe that the approximation algorithm outperforms the baseline random tree formation algorithm in terms of the total tree depth when the number of MDs is larger than 100.

In Fig. 9, we fix the number of MDs to be $|\mathcal{M}| = 50$ and vary the maximum percentage of MDs in one tree, $N_{\text{threshold}}/|\mathcal{M}|$ from 20% to 70%. As this ratio increases, we see an increasing trend of the average constructed tree size and a decreasing trend of the total tree depth. This observation reflects the trade-off between efficiency and security of data collection.

Fig. 8. Comparison of Total Tree Depth (Approximation vs. Random).

VI. CONCLUSION

In this paper, we develop a secure and efficient protocol to collect data from measurement devices (MD) in a multi-hop manner through a mobile data collector (DC). MDs report data via trees rooted at the MDs that have direct communication with the DC. We formulate the secure and optimal tree construction problem as an integer linear programming, and develop an approximation algorithm to compute a solution.

The simulations conducted using real topologies reflect that our algorithm performs well and efficiently.

ACKNOWLEDGMENT

This material is based upon work supported in part by the Department of Energy under Award Number DE-OE0000097 and the Ralph and Catherine Fisher grant. King-Shan Liu’s work was funded in part by the Small Project Funds (Project No: 201109176048) of the University of Hong Kong.

REFERENCES