IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 6, 2019, accepted January 19, 2019, date of publication January 24, 2019, date of current version February 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2894956

Data-Driven Dynamic Active Node Selection for Event Localization in IoT Applications - A Case Study of Radiation Localization

AHMED ALAGHA[1], (Member, IEEE), SHAKTI SINGH[1,2], (Member, IEEE), RABEB MIZOUNI[1,2], ANIS OUALI[3], AND HADI OTROK[1,2], (Senior Member, IEEE)

[1]Electrical and Computer Engineering Department, Khalifa University of Science and Technology, Abu Dhabi 127788, United Arab Emirates
[2]Center on Cyber-Physical Systems, Khalifa University of Science and Technology, Abu Dhabi 127788, United Arab Emirates
[3]Emirates ICT Innovation Center, Abu Dhabi 127788, United Arab Emirates

Corresponding author: Ahmed Alagha (ahmed.alagha@ku.ac.ae)

This work was supported by the Khalifa University Internal Research Fund (KUIRF level 2) under Project 8474000012.

ABSTRACT In this paper, the problem of active node selection for localization tasks, on the Internet of Things (IoT) sensing applications, is addressed. IoT plays a significant role in realizing the concept of smart environments, such as in environmental, infrastructural, industrial, disaster, or threat monitoring. Several IoT sensing nodes can be deployed within an area to collect regional information for the purpose of achieving a common contextual goal. Active node selection proves useful in mitigating common IoT-related issues like resource allocation, network lifetime, and the confidence in the collected data, by having the right sensors active at a given time. Current active node selection schemes prove inefficient when adapted to localization tasks, as they- 1) are usually designed for general monitoring, not localization, 2) do not dynamically exploit data readings in the selection process, and 3) are mostly designed for systems with nodes having sensing ranges. To address these challenges, we propose a novel Data-driven active node selection approach that- 1) dynamically uses data readings from current active nodes to select future ones, 2) assesses the area coverage achieved by a group of nodes while considering range-free sensors, 3) considers parameters like residual energy, power cost, and data confidence levels in the selection process, and 4) combines group-based and individual-based selection mechanisms to enhance the localization process in terms of time and power consumption. These considerations are integrated into a two-phase active node selection mechanism that uses genetic and greedy algorithms to select optimum groups for localization tasks. The efficacy of the proposed approach is validated through an example of radioactive source localization by using real-life and synthetic datasets, and by comparing the proposed approach to existing benchmarks. The results demonstrate the ability of the proposed approach to performing faster localization at low energy cost, even with a smaller number of active nodes.

INDEX TERMS IoT, localization, active node selection, data-driven, radiation detection.

I. INTRODUCTION

With the rapid development of information technologies, the Internet of Things (IoT) paradigm has become one of the main attractions for researchers and businesses. It is considered by many to be the next big technological revolution [1]. IoT is based on the ability of different objects to interact and cooperate towards achieving common goals, with minimal human interaction. Such objects include sensors, RFIDs, actuators, and mobile phones [2]. In sensing applications, a vital enabler for the IoT paradigm is Wireless Sensor Networks (WSNs) [3]. WSNs are networks consisting of a large number of small, inexpensive, and battery powered devices, i.e. nodes, that are equipped with sensors. Typically, these nodes can process, store, and communicate data among each other or to a sink node [4], [5]. IoT plays a vital role in applications related to environmental, infrastructural, industrial, disaster, or threat monitoring [5]–[8]. For instance, nuclear mill tailings are constantly monitored for radiological pollution through random deployment of multiple small IoT sensors in the Area of Interest (AoI) [9].

"Target localization" is an important aspect of environmental monitoring. It relies on deploying sensing nodes to

work towards localizing a specific target in a certain AoI. Although few big powerful sensors could be satisfactory for the purpose of detecting the target, that is not always the case for localizing it. It has been shown in the literature that the performance in localization tasks is improved when many small sensors are employed [10], which demonstrates the potential effectiveness of using IoT nodes in such applications. However, using IoT sensors in such tasks is challenging mainly due to the nature of the used sensors, which have limited resources like energy.

A deployment solution that requires all nodes to be active, when a subset of nodes is enough, compromises the network's lifetime and makes the system inefficient. In addition, the correctness of readings provided by these nodes can be affected by several factors like nodes' faultiness and efficiency [11], [12]. This affects the reliability of such systems, and may lead to incorrect results or fake alarms, which decreases the confidence level in the received data and can be fatal for critical monitoring and localization tasks.

A. PROBLEM STATEMENT AND MOTIVATION

The challenge of energy limitations in localization tasks is primarily addressed in literature by optimizing the number of deployed nodes and their placement, while achieving good AoI coverage [13]–[15]. However, such optimization proves difficult since the prior information about the source location is not known. This complicates the placement schemes and makes the system less adaptive and inefficient. Additionally, many of the current schemes rely on taking many readings from multiple nodes for a localization task. These solutions increase the system overhead, generate redundant data, and often lead to wastage of energy and resources.

Instead of optimizing the number of nodes and their placement, the problem of resource and energy limitations could be addressed through active node selection. In this case, the nodes may be arbitrarily deployed and, for a given task, only a subset of nodes are activated at a given time. Active node selection helps in saving energy by choosing the right subset of nodes to be active depending on the task requirements. Such selection could be group-based, where potential groups are assessed as a whole and a best group is chosen, or individual-based, where each node is assessed individually and the best nodes are chosen. This assessment is based on the specified task requirements and constraints.

The current active node selection frameworks are designed for general sensing tasks, where the goal is to collect regional sensing reports [12], [16]–[18]. While efficient for general sensing or monitoring tasks, these frameworks prove inefficient when adapted to localization tasks. Localization tasks are significantly affected by the proximity of active nodes to the source to be localized [10]. Since the source location is unknown, the desired goal of a selection scheme should be to select nodes near potential source locations. The selection parameters in current works are not enough to solve the problem of target localization. Precisely, the current selection schemes cannot be well-adapted to localization tasks due to the following shortcomings:

- The current selection approaches focus on general environmental monitoring, making them less application-specific.
- The collected data is not used in a feedback mechanism to update the group of active nodes, which could be crucial to speed up the localization process.
- Most works in the literature which consider AoI coverage are designed for nodes that have sensing ranges. This may not be true for all sensors, such as for radiation sensors where no such range exists.

These shortcomings call for a dynamic selection scheme that adapts to the localization task and exploits the collected readings in updating the current set of active nodes.

B. CONTRIBUTION

This work proposes a novel dynamic Data-driven Active Node Selection (DANS) mechanism, which is targeted for monitoring and localization applications. In comparison with the current general selection approaches, the main contributions of the proposed work are-

- Introducing a data-based parameter in the assessment of individual nodes;
- Assessing the utilization of a group, leading to better AoI coverage, while considering range-free sensors;
- Integrating group-based and individual-based selection mechanisms in a two-phase approach, using genetic and greedy methods;
- Enhancing the source localization process with fewer number of active nodes through selection optimization.

The testing and evaluation of the proposed approach is done using two datasets; a synthetic dataset and a real-life dataset obtained from [19]. The applicability of DANS is compared to two data-independent selection mechanisms; a group-based one from [16], and an individual-based one from [12] and [17]. A use case scenario of a radiation environment, where a radiation source is to be localized, is used to prove the efficacy of the proposed approach.

C. PAPER OUTLINE

The rest of the paper is organized as follows- Section II reviews published works which are related to similar topics. Section III describes the generic source localization algorithm and its adaptation to the radiation example. Section IV presents the proposed selection model with all the considered parameters and constraints. Section V discusses the proposed selection approach. Section VI presents and discusses simulation results and evaluation metrics comparing the proposed approach with other works from the literature, and finally Section VII draws conclusions from the presented work. For the sake of clarity, the list of the abbreviations used in this paper is provided in Table 1.

TABLE 1. List of abbreviations.

abbreviation	Term
AoI	Area of Interest
$DANS$	Data-driven Active Node Selection
$DIRS$	Data-driven Individual-based Recruitment System
GRS	Group-based Recruitment System
IoT	Internet of Things
IRS	Individual-based Recruitment System
QoL	Quality of Localization
QoS	Quality of Service
WSN	Wireless Sensor Network

II. RELATED WORK

Various models and several solutions have been proposed in the literature targeting source localization. The main approach to this issue targets node placement, which is a critical problem that has been tackled by many researchers aiming to find the best locations to deploy nodes. The problem has been defined as an NP-Hard problem, where most works approach it through heuristics. The work in [13] considers the placement of 9 nodes using greedy and genetic algorithms, with the aim of minimizing the localization time and maximizing the detection accuracy. Another work adds more complexity to the problem by considering mobile nodes in the deployment strategy, which is also solved by a genetic algorithm based approach [14]. In [15], a mathematical formulation of the node placement problem is presented with considerations to the uncertainty of the placed nodes. The work considers an existing localization network and aims to minimize the number of assisting nodes, i.e. nodes deployed to improve the network's localization accuracy.

As mentioned earlier, existing works that target active node selection focus on general sensing tasks, and none considers localization tasks [12], [16]–[18]. Each of these works use specific parameters during the selection process that can be either device-, AoI-, or user- related. In addition, the selection schemes are either one-time selection, where a single group of nodes is selected to perform the task entirely, or dynamic selection, where groups of nodes alternate during the execution of the task.

In IoT sensing applications, the availability of nodes within or around the AoI forms as the main criterion in the selection process. For example, an individual-based greedy selection scheme is proposed in [12] aiming to maximize the application relevance, which depends on parameters like the proximity to the AoI, with constraints on energy consumption. The problem is formulated as a knapsack problem where nodes are greedily put in the knapsack until it is full, i.e. the constraints are met. Other works constrain the nodes to be within the AoI to be eligible for selection, with the aim of maximizing AoI coverage. A group-based coverage assessment of the AoI in [16] is done by first dividing the AoI into sub-regions of equal dimensions, and then labeling sub-regions that contain at least one node as covered. The same work also assesses the uniformity of the group members' distribution within the AoI using the Chi-square test. The works aims to maximize the group's Quality of Service (QoS) based on area coverage, distribution of nodes, residual energy, and sampling frequency with constraints on the location, budget, and reputation. In [18], a distributed game theoretic approach is proposed for the selection of active nodes with the aim of maximizing AoI coverage. The coverage problem is formulated as a non-cooperative game in which nodes, with certain sensing and communication ranges, compete in each round to be active. This competition is assessed based on coverage redundancy, activation cost, the number of active neighbors, and uncovered region.

The active nodes selected to perform a certain task sacrifice their energy resources. In terms of energy, two main parameters are considered in the literature during the selection process. The first parameter is Residual Energy (RE), which is the amount of energy available in the node's battery, and is a measure of the readiness of a node to perform the task. The other parameter is the power consumption. Nodes vary in terms of power consumption depending on characteristics related to sensing, processing, and communication [20]. It is significant to consider these two parameters during the selection of active nodes due to their effect on the network's lifetime. The selection scheme in [12] introduces RE in the utility function, to be maximized using greedy algorithm. The same work considers power cost as a constraint limited by the energy budget as specified by the task requester. Similarly, the group-based selection in [16] considers the group RE as part of the QoS to be maximized.

The reliability of any IoT-based sensing system heavily relies on the data provided by its nodes, hence it is essential to question the correctness of their readings. This can be referred to as the 'trustworthiness' or the 'reputation' of these nodes. The reputation of a node can be affected by several factors, mainly by ones related to the history of the tasks performed by the node (or its holder in case of Mobile Crowd Sensing). According to [21], data trustworthiness is a function of hard and soft reputation. Hard reputation is quantified as the accuracy of the sensor readings, while soft reputation is quantified as the malicious behavior of the participants, i.e. the nodes' holders. In [16] the reputation of participants is calculated based on their historical commitment and their successful completion of tasks out of all the assigned tasks.

Most of the systems proposed in the literature use a subset of the above discussed parameters in the selection process. However, none of these parameters consider the data obtained by nodes. It is important to consider all such parameters, especially data, due to their effect on the localization task. Moreover, works concerned with AoI coverage are designed for systems with nodes that have sensing ranges. In tasks related to applications like radiation monitoring, sensors do not have a fixed sensing range as this range is source dependent, i.e. the stronger the source, the further the distance it can be detected from. Hence, such models cannot be adapted for such applications.

Additionally, most works in the literature assess the expected QoS before the execution of the actual task.

While this is acceptable as an estimation, it does not fully reflect the actual results obtained during the task. The proposed work selects active nodes dynamically while performing the localization task and assesses the selection scheme based on the actual results of the task execution.

III. LOCALIZATION ALGORITHM

Source localization is a typical problem associated with environmental monitoring. The data reports provided by the nodes, which are deployed within the AoI, are usually aggregated at a fusion center or a platform that processes them with an aim to estimate the location of the source. In this work, radiation localization is used as a running example, where the aim is to locate a radiation source. The following sections illustrate the localization algorithm along with the models employed to simulate radiation. Section III-A discusses the generic Bayesian localization algorithm, while Sections III-B and III-C present the radiation model and its integration into the localization algorithm.

A. PROBABILISTIC LOCALIZATION ALGORITHM

The localization algorithm used in this work is a probabilistic one, based on Bayes' Theorem [22]. The theorem describes the probability of an event based on prior knowledge, combined with a likelihood function that is derived from the current observations.

The basic idea behind the Bayesian localization process, obtained from [10], is to assume a source location and assess the probability of this location being true. This assessment is done using the current observations and the prior belief. The process of localization starts with dividing the AoI into M small equally sized grid elements labeled $k \in \{1, 2, 3, \ldots, M\}$. Each k holds a certain probability related to the belief that the source is within that grid element. Initially, all grid elements have equal probabilities. The localization process is iterative, where in each iteration, the new readings provided by the active nodes are used to update the probability in each grid element. To update the probability in k, the source is assumed to be in that grid element, and the expected readings from each of the active nodes due to this source location are generated and compared with the actual readings. Let $E_k = [e_1, e_2, \ldots, e_n]$ be the expected readings from the n active nodes due to the assumed source location in k, and $R = [r_1, r_2, \ldots, r_n]$ be the actual readings from these nodes. To update the probability in k, the degree of similarity, S_k, between E_k and R is used. S_k represents a score between 0 and 1 that indicates how similar R and E_k are, where 1 indicates that they are exactly the same. The mathematical formulation of S_k varies depending on the application. One approach is to represent S_k as a function of the Euclidean distance between E_k and R. Another approach is based on the probabilistic behavior of some applications, such as radiation, in calculating S_k. This is explained in details in Section III-C. The *a posteriori probability* of the source being in grid element k is then given from Bayes Law as [10]:

$$P_k = Pr_k \times S_k \tag{1}$$

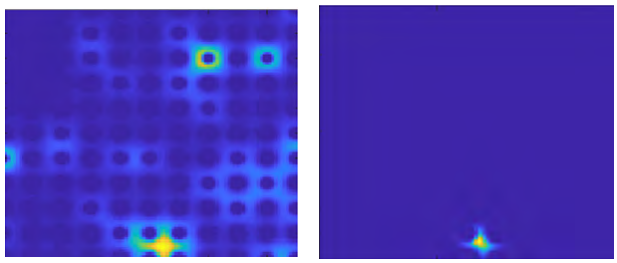

FIGURE 1. Probability distribution (represented via heatmap) of source location after: a. 40 iterations (left), and b. 400 iterations (right).

where P_k is the *a posteriori probability* at k, Pr_k is the *a priori probability* at k, and S_k is the the degree of similarity. This process is repeated for all grid elements with the final result being a probability distribution throughout the AoI. The probability in each grid is normalized to obtain a probability distribution that adds up to 1, as follows:

$$P_{k-normal} = \frac{P_k}{N_s}, \quad N_s = \sum_{k=1}^{M} P_k \tag{2}$$

As the process continues, i.e. more localization iterations have been executed, the probabilities start converging towards a specific area that indicates the source location. The localization process is best illustrated using a heatmap, such as the one shown in Fig. 1. This heatmap represents the different distributions of probabilities after a specific number of localization iterations. As the process progresses with more iterations, the certainty about the source location increases. It can be seen in 1b that after 400 iterations, the hot spots have higher probabilities than the ones after 40 iterations. During the localization process, multiple false hot spots could result due to the probabilistic behavior of the process. For example, Fig. 1a shows multiple false hot spots, which are gradually replaced by a single hot spot, as shown in Fig. 1b. This happens after a certain number of updates, as the process converges towards the source location.

The localization process is terminated once a certain probability is concentrated within a certain area. In this work, a source is declared to be localized when 95% of the probability is contained within 1% of the area. The process of generating the expected readings depends on the phenomenon to be monitored and localized. Section III-B describes the radiation model that is used as a running example in this paper.

B. RADIATION MODEL

This section presents the radiation model used in the localization process. The photons from a radiation source are emitted following a Poisson distribution and the radiation detectors record readings in the form of discrete counts of photons. The expected rate of photon counts per minute (CPM) at

node i due to source S is given as in [13] and [23]:

$$CPM_i = \frac{I_s \times A_i \times \eta_i}{(d_i^s)^2} \tag{3}$$

where I_s is the source strength in photons per minute, A_i is the detector's surface area, d_i^s is the distance between the node and the source, and η_i is the detector's efficiency which is given as [24]:

$$\eta_i = \frac{\#\ of\ photon\ counts\ recorded}{\#\ of\ incident\ photons\ on\ the\ detector} \tag{4}$$

It is assumed that the background radiation, which is same throughout the AoI, is negligible when compared to the source of radiation. Following the Poisson distribution, and given the CPM_i, the probability of node i recording h_i hits from the photons, within time period Δt is given as [10]:

$$p_i^{\lambda_i}(h_i) = \frac{(\lambda_i)^{h_i} \times e^{-\lambda_i}}{h_i!} \tag{5}$$

where $\lambda_i = \Delta t \times CPM_i$ denotes the photons hit rate per Δt, which represents the duration of each localization iteration. In this model, CPM_i is time varying and hence the Poisson process is inhomogeneous. The time varying nature of the process is modeled, as presented in [10], by carrying out calculations in time increments, with the assumption that CPM_i does not change within an increment. This implies that with a proper choice of Δt, h_i can only have a value of 0 or 1. In other words, Δt should be small enough that only 1 photon at the maximum is recorded. Following this consideration, (5) can be simplified to the following equation [10]:

$$p_i^{\lambda_i}(h_i) = \begin{cases} e^{-\lambda_i}, & h_i = 0 \\ 1 - e^{-\lambda_i}, & h_i > 0 \end{cases} \tag{6}$$

This assumption proves accurate provided that the radiation source is stationary and Δt is small enough, but not smaller than the dead time of detector i. The dead time represents the minimum time that the detector needs between two consecutive photon hits to be able to record both. In this work, Δt is set to 20 ms, which is consistent with prior works [10]. To generate readings in simulations, a random probability p_{temp} is generated at node i and is compared with $p_i^{\lambda_i}(0)$, which is the probability that node i gets no hits. If $p_{temp} \leq p_i^{\lambda_i}(0)$ then node i has got 0 hits in this iteration, otherwise it is considered to have had 1 hit.

C. RADIATION LOCALIZATION

The radiation model is incorporated in the localization algorithm described in Section III-A as follows: Given the readings obtained from the active nodes within an iteration, i.e. within a Δt, the probability in grid element k is updated by assuming that the source is in k and then calculating the expected $p_i^{\lambda_i}(0)$ and $p_i^{\lambda_i}(1)$ for each active node, using (6). Hence, the degree of similarity at an active node i given an assumed source location at k is given as [10]:

$$S_k(i) = \begin{cases} p_i^{\lambda_i}(0), & h_i = 0 \\ p_i^{\lambda_i}(1), & h_i > 0 \end{cases} \tag{7}$$

and S_k is then taken as:

$$S_k = \prod_{i=1}^{n} S_k(i) \tag{8}$$

where n is the total number of active nodes. It is assumed that an estimation of the source intensity is already available, which makes the problem a localization one only.

Algorithm 1 summarizes the Bayesian-based localization algorithm with the radiation model.

Algorithm 1 Bayesian-Based Localization Algorithm

Input: priori probabilities (Pr), set of active nodes, Number of active nodes (n)
Output: posterior probability distribution as a heatmap (P), termination flag (SourceLocationFound).

```
for iteration = 1 to NumOfIterations do
    Collect readings (R) from active nodes
    for k = 1 to NumOfGrids do
      for i = 1 to n do
        Calculate d_i^s
        Calculate CPM_i
        Calculate λ_i
        Calculate p_i^{λ_i}(0) and p_i^{λ_i}(1)
        if R(i) == 1 then S_k(i)=p_i^{λ_i}(1);
        if R(i) == 0 then S_k(i)=p_i^{λ_i}(0);
      end for
      P_k = Pr_k × ∏_{i=1}^{n} S_k(i);
    end for
    N_s = sum(P);
    P_norm = P/N_s;
    Pr = P_norm;
    while Sliding Summing Window over P_norm do
      if Sum>0.95 then SourceLocationFound = 1;
      if SourceLocationFound then break;
    end for
    if SourceLocationFound then break;
end for
```

IV. SELECTION ALGORITHM

The active node selection algorithm proposed in this work exploits the probability distributions generated during the localization process. Active nodes are selected based on the probability distribution throughout the AoI, along with other important parameters that are discussed in the following subsections. Performing selection before each localization iteration is time consuming, hence, the iterations of the localization task are grouped into rounds, where each round represents a fixed number of localization iterations that are executed by the same group of active nodes. Fig. 2 describes the full process of selection, data collection, and localization.

A. SELECTION MODEL DESCRIPTION

Given a set of nodes, N, for a task of source localization, the goal is to select the best starting group of active nodes,

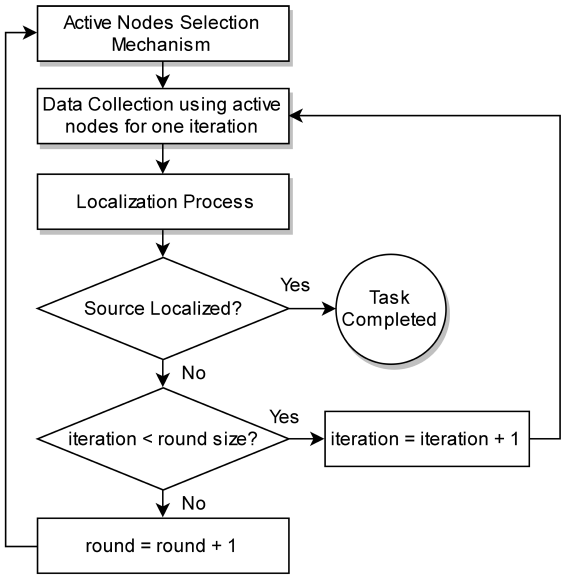

FIGURE 2. High level diagram of the proposed process starting from the active node selection until the completion of the localization task.

TABLE 2. List of used symbols.

Symbols	Definition
ID_i	ID number of node i
l_i	Location of node i in x-y coordinates
RE_i	Residual energy in node i
SA_i	Surface area of the detector at node i
C_i	Power cost of node i
PDR_i	Potential data relevance of node i
η_i	Sensing efficiency of node i
T_i	Trustworthiness of node i
FA_i	Faultiness level at node i
Con_i	Confidence level of node i
P_i	Probability that the source is near node i
d_i^{sp}	Distance between node i and hottest spot
$RE(g)$	Group residual energy
$C(g)$	Group cost
$Con(g)$	Group confidence
$U(g)$	Group utilization
F_i	Individual fitness score
$F(g)$	Group fitness score

$g_1 \in N$, and then dynamically alter the group during the execution of the task, to achieve faster localization with high data confidence and low power cost.

The selection algorithm should comply with all constraints specified by the task requester, such as location or energy constraints. In the proposed approach, each node, i, is characterized with $< ID_i, l_i, RE_i, SA_i, C_i, PDR_i, \eta_i, T_i >$, as defined in Table 2. Using these parameters, a two-phase selection mechanism is proposed, where the first phase employs a group-based selection using genetic algorithm, while the second phase employs an individual-based selection using greedy methods. A group-based selection assesses each potential group as a whole, gives it a 'group fitness' score, then selects the best group with the highest score. On the other hand, an individual-based selection assesses each node individually using a 'fitness' score, ranks nodes accordingly, and finally selects the set of nodes with the highest scores. Group-based selection has an advantage when assessment includes parameters that cannot be assessed for an individual node, such as AoI coverage [16]. Both individual and group parameters that are used for assessment, are further in the following subsections.

B. INDIVIDUAL PARAMETERS

This section describes the different individual parameters used in the selection model. Individual parameters are ones that are assigned to each, independently from other nodes, based on attributes and traits that the node has. These individual parameters are described as:

1) NODE LOCATION (l_i)

The location coordinates of node i.

2) COST (C_i)

The cost of selecting a node is defined as the power consumption in the active mode per second. This reflects the power spent on sensing, processing, and transceiving. On the other hand, if a node is in sleep mode, i.e. not selected to be active, its power cost is negligible since it is much smaller compared to the active mode [25].

3) RESIDUAL ENERGY (RE_i)

Residual energy is defined as the remaining energy in the node's battery. It is a measure of the readiness of a node to participate in the localization task. After each round r, RE_i is updated according to the following expression:

$$RE_i^{r+1} = RE_i^r - (tr \times C_i) \tag{9}$$

where tr is the round duration. The process of either recharging the node's battery or its replacement is not considered in this work.

4) CONFIDENCE (Con_i)

Confidence is a measure of the expected correctness of the data provided by node i. It is evaluated based on two attributes: node's sensor efficiency (η_i) and node trustworthiness (T_i). The efficiency η_i depends on inherent properties of the sensor, and its definition could vary from one application to another. For photon-counting sensors, like radiation detectors, η_i is given as presented in (4). T_i is related to the expected faultiness (FA_i) of node i, which is based on the historical performance of the node. Faultiness can be defined as:

$$FA_i = \frac{set\ of\ tasks\ where\ i\ showed\ faultiness}{set\ of\ tasks\ assigned\ to\ i} \tag{10}$$

Hence, T_i is given as:

$$T_i = 1 - FA_i \tag{11}$$

Con_i is then given as the weighted geometric mean of E_i and T_i:

$$Con_i = \sqrt[3]{\eta_i \times T_i^2} \tag{12}$$

T_i is given more weight since a faulty node, not only fails to deliver the expected readings, but also delivers misleading readings that could significantly affect the localization process. Several methods for faultiness detection have been proposed in the literature, which can be used to determine FA_i. Most of these works are statistical and mathematical methods for anomaly detection [26]–[29].

5) POTENTIAL DATA RELEVANCE (PDR_i)

Potential data relevance is the expected usefulness of the readings provided by node i to the localization process. It is an estimation that is based on the readings obtained during previous rounds of the same localization task and on some node parameters like the detector's surface area. The nodes that are around hot spots are expected to report more useful readings than those which are far. Additionally, sensors with more surface area tend to contribute more to the localization process since they have higher probability of being hit by photons.

PDR_i depends on three attributes: (i) the probability of the source being around the node (P_i), (ii) the distance between the node and the hottest spot (d_i^{sp}), and (iii) the node's sensor surface area (A_i). As discussed in Section III, several hot spots could result during the localization process, which later converge to one spot indicating the source location. To fasten up the process, it is important to select nodes which are around these spots to either verify or deny the existence of a source. This is realized using the parameter P_i, which is given as the sum of all probabilities in a square window of width w_i centered at the node's location. The width (w_i) is chosen based on the surface area (A_i) of the node's detector, where larger w_i is chosen for nodes with higher A_i. This is because larger A_i has more effect on the Bayesian probability distribution, as shown in (3) and (6).

In many cases, a node might not be around any of the hot spots. To differentiate between nodes in such a scenario, those that are closer to the single hottest spot should be given higher importance than those which are far, since they contribute more to the localization process based on (3). This is achieved through the parameter d_i^{sp}, which is the Euclidean distance between node i and the hottest spot that represents a potential source.

Following the localization process executed in a round, a probability distribution is obtained as explained in Section III, which is used to obtain P_i and d_i^{sp} for each node as illustrated in Fig. 3.

PDR_i is then given as:

$$PDR_i = \sqrt[3]{(P_i)^2 \times \delta_i^{sp}} \tag{13}$$

where δ_i^{sp} is a function of the distance d_i^{sp} that calculates the discount to the node's depending on its proximity to the hottest spot; the smaller the distance the higher the score. δ_i^{sp} is given as [30] and [31]:

$$\delta_i^{sp} = 1 - max(0, min[log_D(d_i^{sp}), 1]) \tag{14}$$

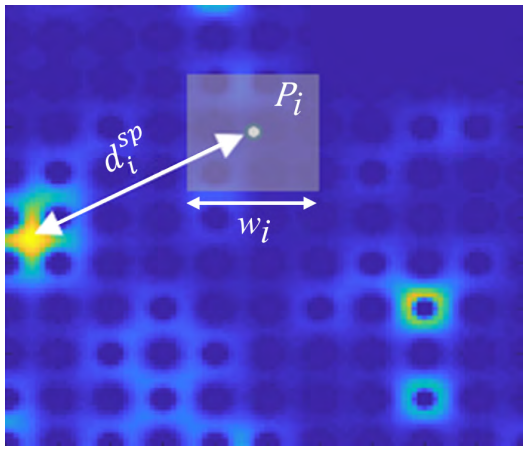

FIGURE 3. The PDR_i of node i depends on its proximity to the hottest spot, d_i^{sp}, and the probability of it being around the source, P_i. P_i is taken as the summation of all probabilities within a square window of width w_i that is determined based on A_i.

where D is the maximum possible distance between a node and the hottest spot, taken as the diagonal of the AoI. PDR_i is then normalized as shown in (15), so that the total potential relevance of all nodes adds up to 1.

$$PDR_{i-norm} = \frac{PDR_i}{\sum_{j=1}^{n} PDR_j} \tag{15}$$

To illustrate the effectiveness of PDR_i and its sub-parameters, Fig. 4 considers an example of 5 nodes and a probability distribution generated during the localization process. The heatmap shows two hot spots, with the one in the bottom left corner having higher probabilities. Table 3 presents the different nodes' characteristics along with their PDR scores. The significance of P_i, D_i, and A_i is discussed below:

- P_i: It can be seen that node 1 has the highest P_i score since it is at the hottest spot, which leads to it having the highest PDR. Additionally, when comparing node 4 to node 5, it can be seen that although node 5 is far from the hottest spot, it has a higher P_i due to the fact that it is around another hot spot. This leads to node 5 having higher PDR than node 4, given that both have same A_i
- D_i: when comparing node 6 to node 2, though both have the same A_i and both are not around any hot spot, node 2

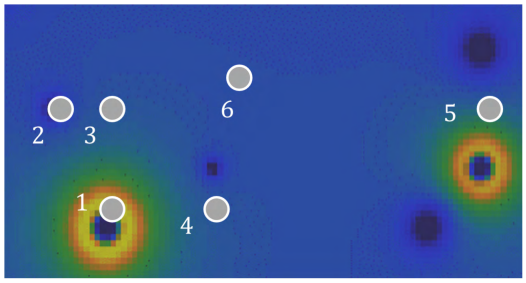

FIGURE 4. Potential data relevance (PDR) example.

TABLE 3. PDR_i scores for 5 different nodes.

Node #	A_i	P_i	δ_i^{sp}	PDR_i
1	93	0.0427	0.734	0.110
2	89	0.0184	0.276	0.0454
3	95	0.0241	0.286	0.0550
4	81	0.0161	0.286	0.0420
5	81	0.0306	0.101	0.0316
6	89	0.0181	0.195	0.0400

has a higher *PDR* score since it closer to the hottest spot; i.e. its D_i score is higher.

- A_i: It can be seen that, although nodes 2 and 3 are at nearly the same distance from the hottest spot, node 3 has a higher P_i score due to its bigger surface area, which leads to a higher *PDR* score.

C. GROUP PARAMETERS

The section describes the different group parameters considered in the selection model. The group parameters are used to collectively assess a group of nodes by combining their individual features into a group score. These parameters are described as:

1) UTILIZATION $U(g)$

This metric evaluates how the members of a group are utilized based on their locations within the AoI. It is a measure of how significant a node is within its group. Bad utilization could be a result of either (a) nodes located around the AoI boundaries, hence covering areas outside the AoI while leaving areas within the AoI uncovered, or (b) multiple nodes redundantly covering same areas. These two scenarios are illustrated in Fig. 5. Good utilization, in which AoI coverage is maximized and redundant coverage is minimized, reflects good distribution of nodes within the AoI.

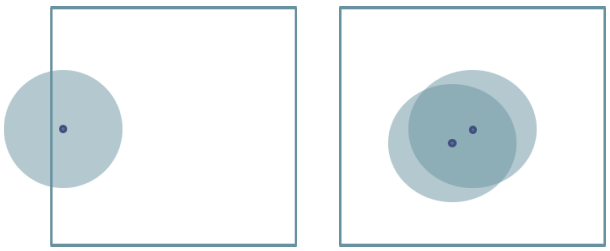

FIGURE 5. Two scenarios showing bad utilization of a node. In (a) the node is covering areas outside the AoI. In (b) the node is redundantly covering areas that are covered by another node.

As discussed earlier, in some applications like radiation, nodes do not have sensing ranges. This introduces difficulties in assessing the utilization, i.e. the sensing coverage, of a node. To circumvent this issue, the solution for a set of problems called Covering Circles, which falls under the Set Cover Problem from combinatorics, is used [32]. Covering Circles is a set of problems which aim to cover certain shapes with either the smallest number of fixed size circles or a fixed number of circles with the smallest size. In [32], the problem of fully covering a square with n identical circles is addressed, where for each n, the aim is to find the smallest radius r_n of the identical circles that will entirely cover the square. This is extrapolated in this work to obtain the radius of circles for each group size n. In other words, for a group of size n, a circle of radius r_n is centered at each of the group members' locations, and $U(g)$ is then calculated as:

$$U(g) = \frac{\#of\ subareas\ covered}{total\ \#\ of\ subareas} \tag{16}$$

A subarea is considered covered if it is within one or more circles. The dependence of r_n on the group size is critical to ensure good distribution of nodes. r_n is smaller for big group sizes, whereas it is bigger for small group sizes, to minimize redundant coverage. This forces the selection scheme to choose groups with well-distributed members. Fig. 6 shows $U(g)$ score for three different five-member groups. It can be seen that a group that is well distributed, as shown in Fig. 6c, has a better $U(g)$ score compared to the other two groups where nodes are either covering areas outside the AoI or redundantly covering areas inside the AoI.

2) GROUP RESIDUAL ENERGY $RE(g)$

To collectively assess $RE(g)$, the arithmetic mean of residual energies alone is not enough as it only reflects the central tendency of the distribution. The standard deviation, along with the arithmetic mean, are used to calculate $RE(g)$ of a group of size n, at round r as:

$$RE^r(g) = \frac{\sum_{i \in g} RE_i^r}{n} \times e^{-\sigma(members' RE^r)} \tag{17}$$

where $\sigma(members' RE^r)$ is the standard deviation of group members' residual energies. This results in uniformity of the selected group in terms of RE. Using (17), $RE(g)$ will be equal to the average only if all group members have the same RE_i. Otherwise, this value will be reduced based on the deviation of members' residual energies from the mean.

3) GROUP CONFIDENCE $Con(g)$

A single member producing false readings could severely drop the value of the group's data and give misleading outcomes. Hence, the group confidence is taken as the lowest confidence value among the group members:

$$Con(g) = min(Con_i) i \in g \tag{18}$$

4) GROUP COST $C(g)$

The cost of a group represents the total power consumption of group members, and is simply computed as:

$$C(g) = \sum_{i \in g} C_i \tag{19}$$

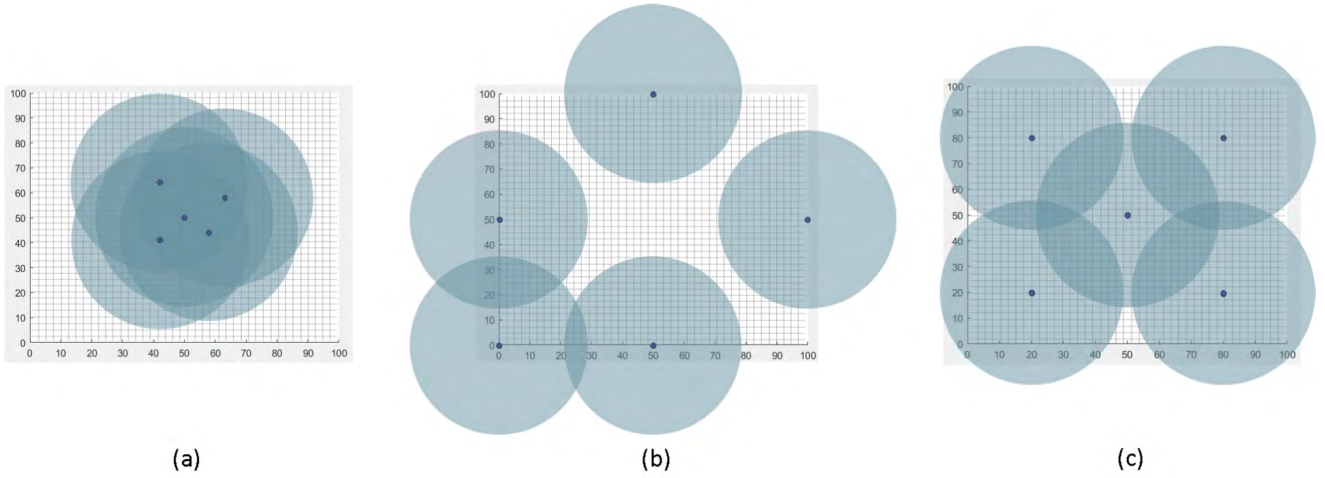

FIGURE 6. Three cases with different $U(g)$ scores: (a) $U(g) = 0.6127$, (b) $U(g) = 0.7065$, (c) $U(g) = 0.9882$.

D. OPTIMIZATION PROBLEM DEFINITION

To assess a group of nodes for selection, the group is given a score based on the following group fitness metric:

$$F(g) = \sqrt[4]{U(g) \times RE(g) \times Con(g) \times C(g)} \tag{20}$$

Similarly, to evaluate each individual node for selection, it is given a score based on the following individual fitness metric:

$$F_i = \sqrt[5]{C_i \times RE_i \times Con_i \times PDR_i^2} \tag{21}$$

Here, PDR_i is given a higher weight to signify the data-driven selection. Finally, in both selection processes, a candidate node must not violate any of the following constraints:

- X and Y coordinates are within the AoI
- $RE_i \geq tr \times C_i$

V. PROPOSED APPROACH

As discussed earlier, the active node selection is done in rounds. Before each round, the selection scheme chooses the best set of active nodes to perform the localization for that round. After the execution of each round, RE_i is updated for nodes that were active, while PDR_i is updated for all nodes.

For the selection of the first group, genetic algorithm is used to select nodes that maximize the group's fitness function ($F(g)$). Since no readings are available prior to the first round, a data-driven selection is not possible. Hence, it is best to have the group distributed throughout the AoI, which is achieved by the $U(g)$ parameter. Therefore, a group-based assessment is used for the first group. The way GA used is similar to that in [16]. Here, GA is used primarily because (i) $U(g)$ cannot be assessed on an individual basis hence algorithms like greedy cannot be used, and (ii) GA is efficient and scalable in searching for the optimal solution in the search space [16]. It should be noted that this only needs to be done once, since PDR will always be available after the first round.

For the subsequent rounds, the selection is an individual-based one, where F_i for each node is computed, and the nodes with the highest scores are selected. Here, since $U(g)$ is not required anymore, greedy algorithm is used because it requires less computation time compared to GA, which is important since the selection is done dynamically during the localization task.

A detailed description of the data-driven active node selection algorithm is given below, and the corresponding flowchart is shown in Fig. 7. The pseudo-code of the full algorithm is detailed in Algorithm 2.

1) Given a dataset of available nodes and their characteristics, nodes' IDs are rearranged randomly, and groups of the specific size are formed. This set of groups is referred to as the initial population.
2) Each group in the population is checked against the constraints, as presented in Section IV-D. The groups that violate these constraints are removed from the selection process.
3) $F(g)$ of each group in the population is evaluated using (20).
4) A roulette wheel is spun to select the parent groups for the next generation based on the fitness scores. The wheel is spun multiple times, where the number of times it is spun is equal to the population size, i.e. the number of groups. For each spin, the groups with higher fitness scores have higher probability of being chosen as parents.
5) Each pair (parents) of the selected groups, chosen in order from step 4, might undergo a one-point crossover process, that is determined by a crossover probability, to generate children. If the crossover probability is 0.6 then there is a 60% chance for a pair to undergo a crossover operation. The point at which the crossover operation for a pair of groups will be done is chosen randomly. The result of this process is a set of new generation of children groups.
6) Some groups selected at random are mutated, where each group member might be replaced with another

TABLE 4. List of parameters used in *QoL*.

Parameter	Explanation
T_{Loc}	number of localization iterations it takes to localize the source
C_{avg}	the average power cost score per round
RE_{avg}	the average residual energy score per round
Con_{avg}	the average confidence score per round
$TotalCost$	The cost of the entire localization process, i.e. the summation of C_{avg} for all rounds

member from the set of nodes, depending on a mutation probability.

7) Steps 2-6 are repeated for the new generated population. The process keeps repeating until i) the maximum number of GA iterations is reached, ii) a $F(g)$ convergence occurs, or iii) a group with the best possible $F(g)$ is found. Convergence occurs when $F(g)$ remains the same for a certain number of GA iterations.
8) The group with the best $F(g)$ executes the first round of the task, i.e. the first set of localization iterations. The result is a probability distribution as discussed in Section III-A, which is used in the selection of the subsequent sets of active nodes. After the end of the first round, RE_i is updated for nodes that were active, while PDR_i is updated for all nodes.
9) F_i for each node is evaluated, and the nodes are sorted in a descending order.
10) The nodes with the highest F_i are selected until the group size (n) is fulfilled.
11) The set of active nodes perform the localization task for the next round and update the probability distribution obtained previously. RE_i is updated for nodes that were active, while PDR_i is updated for all nodes.
12) Steps 9-11 repeat for all subsequent rounds until the stopping criteria of the localization process is met.

VI. SIMULATION AND EVALUATION

A. EVALUATION METRICS & BENCHMARK

To assess the performance of the proposed approach and to compare it with other selection schemes, a Quality of Localization (QoL) evaluation metric is developed that combines several parameters. The QoL is given as:

$$QoL = \sqrt[4]{\frac{1}{T_{Loc}} \times \frac{1}{C_{avg}} \times RE_{avg} \times Con_{avg}} \tag{22}$$

Table 4 explains the various parameters used for QoL.

The proposed approach, henceforth called DANS, is also compared to the following familiar selection schemes for IoT sensing applications:

1) A data-independent, Group-Based Recruitment system (GRS) in [16].
2) A data-independent, Individual-Based Recruitment system (IRS), like the ones used in [12] and [17].

Algorithm 2 Data-Driven Active Node Selection

Input: localization task requirements and constraints, nodes' dataset (N), group size (n).
Output: posterior probability distribution (P), time to localize (T_{Loc}), avg group cost (C_{avg}), avg group confidence (Con_{avg}), avg group residual energy (RE_{avg}).

```
Pr = initialPriori();
for round=1 to MaxNumOfRounds do
  if round == 1 then // GA Process
    seed = []; //holds GA seed
    bestG = []; //holds the best group
    bestFit = 0; //holds the best fitness score
    G = initialPopulation(N,n);
    while GAIteration ≤ MaxGAIteration
        & Fit_Converge ≤ MaxFitConverge do
      for all g ∈ G do
        Calculate F(g)
        if g violates any constraint then delete g
        from G
        if F(g) > bestFit then bestFit=F(g) &
        best G = g;
      end for
      R = rouletteWheel(G);
      C = crossover(R);
      G = mutation(C);
      iteration = iteration + 1;
    end while
  else if round > 1 //Greedy Process
    bestG=[];
    for all i ∈ N
      Calculate F_i;
      Check if node i violates any constraint
      if i violates any constraint then discard i from N
    end for
    SN = Sort(N,F_i,descending); //sorted nodes
    bestG = SN(1:n); // take the top n nodes
  end if
  [P,SourceLocationFound] =
      Localization(Pr,bestG,n);
  Pr = P;
  UpdateRE(N,bestG); //update RE for active nodes
  UpdatePDR(N,bestG,P); //update PDR for all nodes
  if SourceLocationFound then break;
end for
return [T_Loc, Con_avg, C_avg, RE_avg];
```

3) An adapted version of IRS, that is data-driven (DIRS), but without consideration to AoI coverage in the initial round.

Both DANS and DIRS reflect the proposed work. DIRS shows the importance of introducing only the PDR parameter in IRS, while DANS represents the full approach. Table 5 summarizes the methodology behind each of the four models.

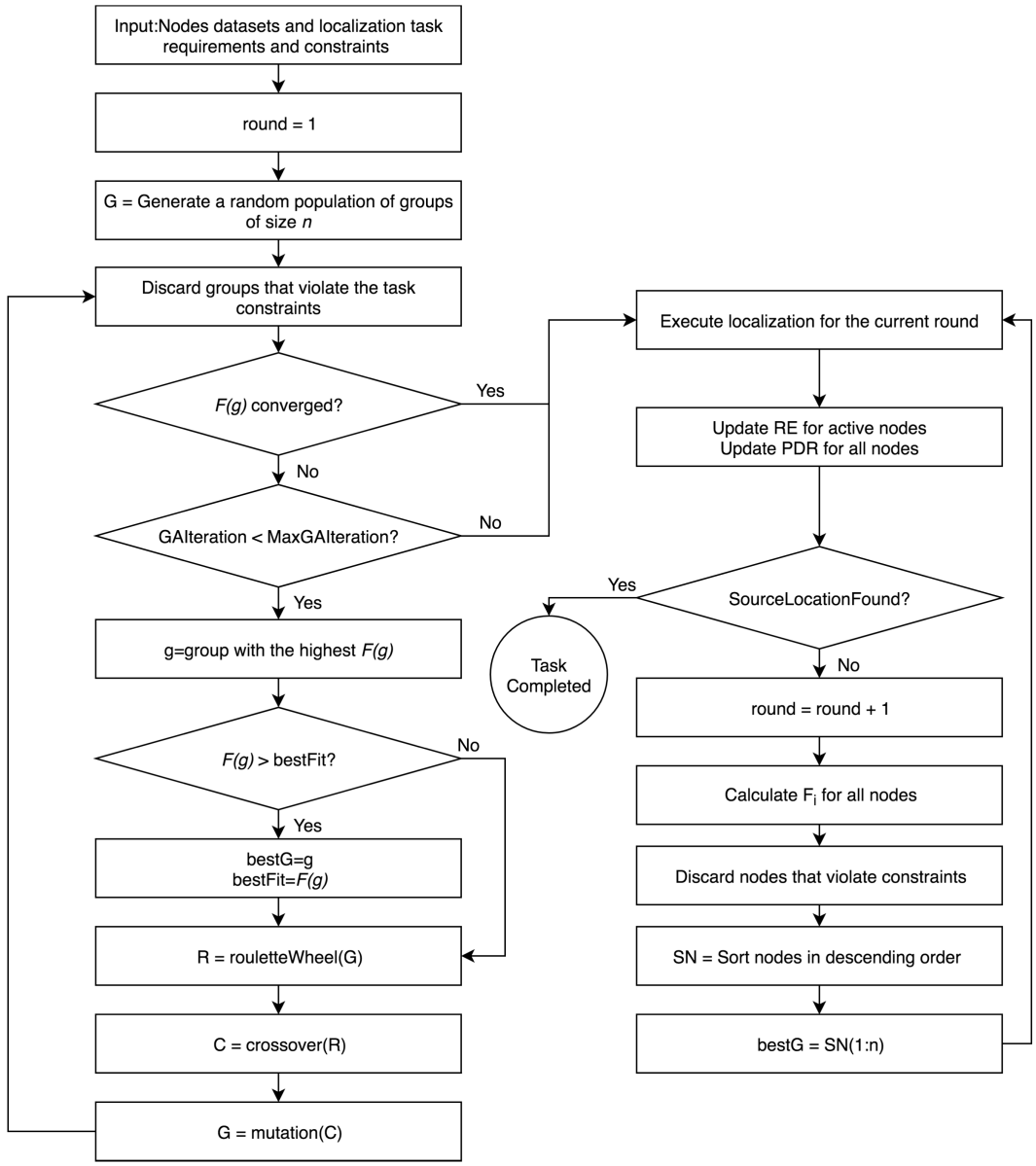

FIGURE 7. Flowchart of the data-driven active node selection scheme.

TABLE 5. Summary of the simulated selection models.

DANS (GA, Greedy, Data-Driv)	**DIRS** (Greedy, Data-Driv)	**IRS** (Greedy)	**GRS** (GA)
1) Form **groups** of the specified group size. 2) Assess $F(g)$ of each **group**. 3) Select the best group according to the first part of Algorithm 2. 4) Perform the first round of the localization task. 5) Update RE and PDR. 6) Assess F_i of each **individual** node. 7) Select best nodes to fill the group according to F_i. 8) Perform the following round of the task. 9) Repeat steps 5-8 until source is localized.	1) Assess F_i of each **individual** node. 2) Select best nodes to fill the group according to F_i. 3) Perform the following round of the task. 4) Update RE and PDR. 5) Repeat steps 1-4 until source is localized.	1) Assess $F_{IRS} = \sqrt{C_i \times RE_i \times Con_i}$ of each **individual** node. 2) Select top nodes to fill the group according to F_i. 3) Perform the following round of the task. 4) Update RE. 5) Repeat steps 1-4 until source is localized.	1) Form groups of candidate nodes. 2) Assess F(g) of each **group**. 3) Select the best group according to the first part of Algorithm 2. 4) Perform the first round of the localization task. 5) Update RE. 6) Repeat steps 1-5 until source is localized.

B. DATASET

The datasets used in simulations consist of 8 parameters, as detailed in Table 2 and as discussed in Section IV-A, that characterize each of the nodes. In this work, two datasets, are used: Dataset 1 - a synthetic one with uniformly distributed nodes throughout the AoI in a grid layout, and Dataset 2 - a non-uniform distribution of nodes which is obtained from the Sarwat Foursquare Dataset in [19]. The Sarwat

Foursquare Dataset includes several parameters, however only nodes' locations are used from this dataset. For both datasets, the residual energy, sensor surface area, trustworthiness, and efficiency are randomly generated and assigned to the nodes following a uniform distribution. The power cost of each node is generated based on the efficiency and sensor surface area, where nodes with higher values for these attributes are expected to consume more power. All these attributes are given a value between 0 and 1.

C. PERFORMANCE EVALUATION AND COMPARISON

Multiple experiments are executed to evaluate and validate the performance of the proposed (DANS) scheme. During these experiments, a Dell Intel Xeon workstation equipped with 256 GB RAM and 300 GB hard disk is used. The simulations are performed for all selection models, to localize a nuclear radiation source in the same environment.

The source strength, population size (number of available nodes), and the size of the AoI are also varied to study their effect on the localization task. The results are presented as *QoL* for different group sizes, where the group size is given as the % of active nodes. For each of the proposed and benchmarked algorithms, the simulations are carried out 5 times in 9 different source locations, with a total of 45 runs. The final result is presented as an average of all the runs.

For the following results and discussion, a uniform dataset, i.e. Dataset 1, is used for Sections VI-C1 and VI-C2, whereas a real-life dataset, referred to as Dataset 2, is used in Section VI-C3.

1) EFFECT OF GROUP SIZE ON TIME, TOTAL COST, AND *QoL*

This experiment shows the effect of group size, i.e. the % of active nodes, on localization time, total cost, and *QoL*, where the group size is varied from 10% to 100% in decade steps. A population size of 121 nodes, an AoI size of $300m \times 300m$, and a source strength of 10^6 photons/minute are used in this experiment. As seen in Fig. 8, the models which include a data-based parameter (*PDR*), namely DANS and DIRS, outperform the data-independent models, namely GRS and IRS, in terms of localization time for different group sizes. The difference is more significant for smaller group sizes. As the group size increases beyond 50-60%, the localization time converges for all models. This is expected as the selection schemes become inconsequential since almost all nodes are becoming active.

As the group size increases, more nodes become active, with expectations that the localization process is to finish faster. However, a question remains: what is the tradeoff between the localization time and the cost of employing more active nodes? Intuitively, employing more active nodes should result in higher average cost (C_{avg}) per round. However, since the nodes remain active for less time, as the localization task finishes faster, the total cost for the entire localization process is reduced. Here, C_{avg} reflects the cost related to the number of active nodes per round, whereas the total cost reflects both C_{avg} and the duration of the localization task. In other words, total cost is given as:

$$TotalCost = C_{avg} \times No.ofLocalizationRounds \quad (23)$$

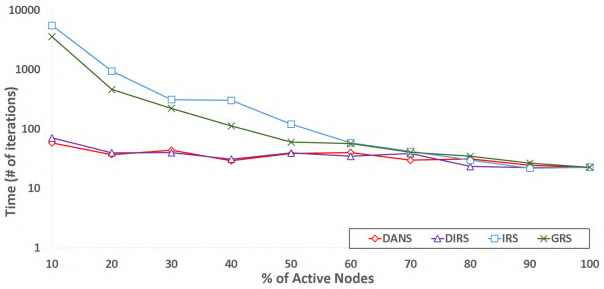

FIGURE 8. The effect of varying the group size on localization time, using dataset 1.

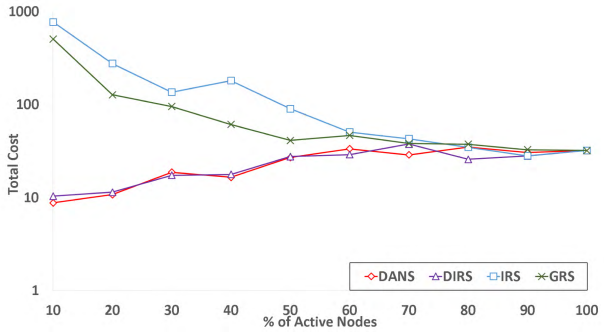

FIGURE 9. The effect of varying the group size on the total power cost, using dataset 1.

Fig. 9 shows that both data-driven approaches, DANS and DIRS, cost less compared to GRS and IRS, especially for low group sizes. Faster localization process by DANS and DIRS, due to the selection of informative nodes, results in significant preservation of energy. It can also be seen that for DANS and DIRS, increasing the % of active nodes just adds more cost without having significant effect on localization time, resulting in a higher total cost. This is because these two models succeed in selecting the nodes that contribute the most to the localization process, which represent a small % of the total number of nodes. When the group size is increased, more nodes with less contribution are added, resulting in an increase in C_{avg} with no significant effect on localization time. This is important as it reflects the ability of both approaches to exploit informative nodes leading to less number of nodes being active.

Fig. 10 shows the *QoL* of the different selection algorithms as a function of group size. DANS and DIRS can be seen to have better *QoL*, until all models converge at high % of active nodes. Specifically, DANS is seen to outperform GRS and IRS by an average of 48.1% and 52.0%, respectively. Additionally, it can be seen that DANS and DIRS have similar performance, although a marked difference between them can be seen at smaller group sizes. Since a selection scheme is more significant when small groups of nodes are selected to perform a given task, thus, Fig. 11 shows the

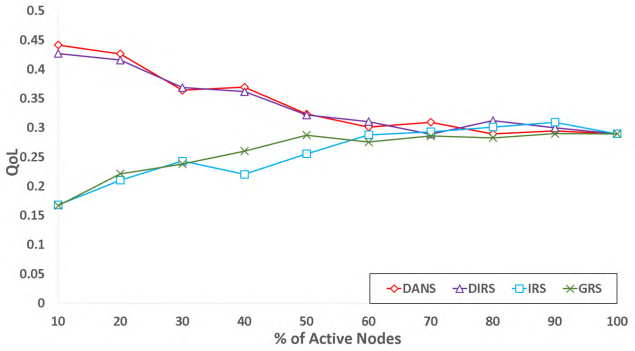

FIGURE 10. The effect of varying the group size on *QoL*, using dataset 1.

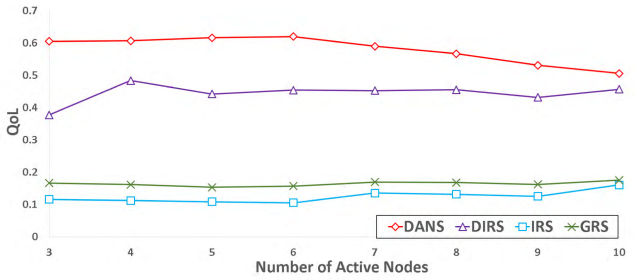

FIGURE 11. The effect of varying the small group sizes on *QoL*, using dataset 1.

comparison of the different selection algorithms for smaller group sizes. The figure shows a noticeable difference between both data-driven approaches, DANS and DIRS, which reflects the significance of the $U(g)$ parameter in the selection of the first group. DANS, which considers this parameter, performs on average 34.2% better than DIRS, which does not include $U(g)$ at all. The selection of the first group is significant, especially for small group size, because it is the group the generates the first probability distribution that is the base *a priori* to the localization process. Since no readings are available prior to the first round, it is shown that considering $U(g)$ in the DANS model, which results in good AoI coverage, improves the *QoL* when compared to DIRS. Additionally, for small group sizes, DANS is found to significantly outperform GRS and IRS by an average of up to 2.7 and 4 times, respectively.

For both DANS and DIRS, the *QoL* initially increases with group size but eventually drops as group size keeps increasing. As the number of active nodes increases, AoI coverage becomes easier to achieve for DIRS even though it does not consider it during the selection process. Hence, the effect of $U(g)$ becomes less significant, leading to DIRS having similar behavior as DANS. This experiment validates the effectiveness of the proposed approach, where the highest *QoL* can be achieved for a small group of size 6, which represents only 5% of the total number of available nodes.

2) SCALABILITY

To prove the scalability of the proposed model, simulations were conducted for different population sizes (varying from

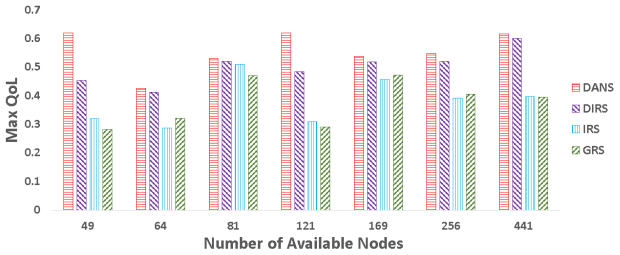

FIGURE 12. The effect of varying the population size on the maximum achieved *QoL*, using dataset 1.

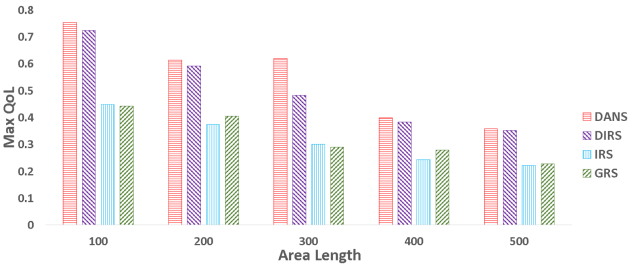

FIGURE 13. The effect of varying the area size on the maximum achieved *QoL*, using dataset 1.

49 to 441), and on different area sizes (varying from $100m \times 100m$ to $500m \times 500m$). Population size was varied for a fixed AoI size of $300m \times 300m$, while AoI size was varied for a fixed population size of 121. A source strength of 10^6 photons/minute was used in all simulation. It should be noted that since the nodes are uniformly distributed throughout a square AoI in a grid layout, the population sizes had to be square numbers for simplicity. For each simulation, for a specific population or area size, the group size is varied and the maximum *QoL*, per population or area size, is recorded. The maximum *QoL* referes to the best *QoL* achieved by a group size, which is varied, in the specific population size. As shown in Fig. 12 and Fig. 13, the proposed DANS still performs better than other selection models. On average, DANS is 54.8% better than GRS, 51.2% better than IRS, and 11.6% better than DIRS, in terms of maximum *QoL*, for varying population sizes. On the other hand, in area scalability, DANS performs on average 67.1% better than GRS, 72.8% better than IRS, and 8.4% better than DIRS.

3) ADAPTABILITY

In this experiment, the performance of the system is studied using Dataset 2 which has real-life nonuniform distribution of nodes. A population size of 200 nodes, an AoI of $300m \times 300m$, and a source strength of 10^6 photons/minute are used. Fig. 14, Fig. 15, and Fig. 16 show a similar behavior to the results obtained in Section VI-C1, where data-driven models perform better than data-independent ones. DANS was found to outperform GRS, IRS, and DIRS by an average of 52.4%, 74.8%, and 3.1%, respectively. As evident, all models converge at a larger group size, around 80%, compared to 50-60% in the uniform distribution simulations, i.e. Dataset 1.

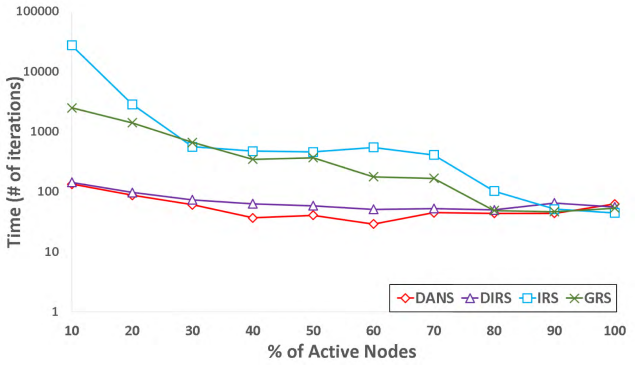

FIGURE 14. The effect of varying the group size on the localization time, using dataset 2.

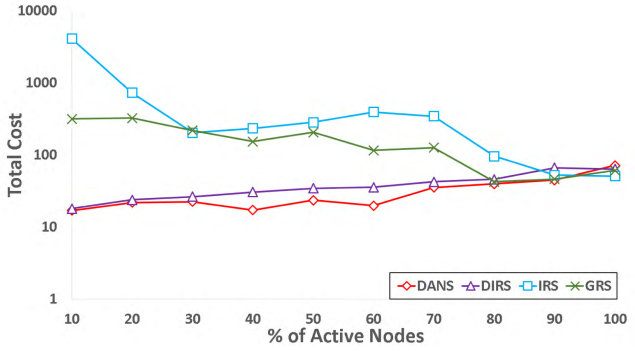

FIGURE 15. The effect of varying the group size on the total cost, using dataset 2.

The sheer non-uniform distribution of the nodes makes a selection process relatively harder when compared to uniformly distributed nodes. This is because the spatial availability of nodes in a uniformly deployed network is better than that in a non-uniformly deployed one, hence increasing the chance of finding informative nodes around the source. This makes the selection scheme, DANS, more significant for non-uniform networks, as other data-independent schemes will struggle more in choosing informative nodes. Additionally, Fig. 17 presents the *QoL* obtained for small group sizes, which shows the significance of DANS over DIRS. DANS is found to perform 7.3% better than DIRS, while performing up to 2.1 and 2.9 times better than GRS and IRS, respectively. Fig. 18 and Fig. 19 show the maximum *QoL* achieved for each model given different population or AoI sizes. It can be seen that DANS adapts to the new dataset and still performs better than all other approaches for different population and AoI sizes.

4) EFFECT OF VARYING SOURCE STRENGTH

This experiment explores the effect of varying the source strength on the localization process in terms of *QoL*. For each selection model, the source strength is varied for different group sizes, and the maximum *QoL* is recorded along with the corresponding group size. As shown in Fig. 20, the proposed approach- DANS, still achieves the highest *QoL* compared to the other models at all source strengths. Additionally, a general trend can be noticed in which the stronger the source

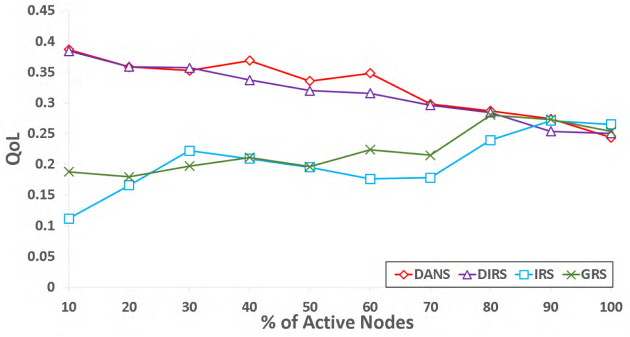

FIGURE 16. The effect of varying the group size on the *QoL*, using dataset 2.

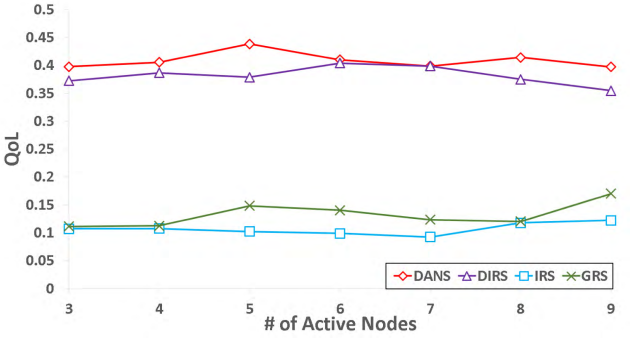

FIGURE 17. The effect of varying small group sizes on the *QoL*, using dataset 2.

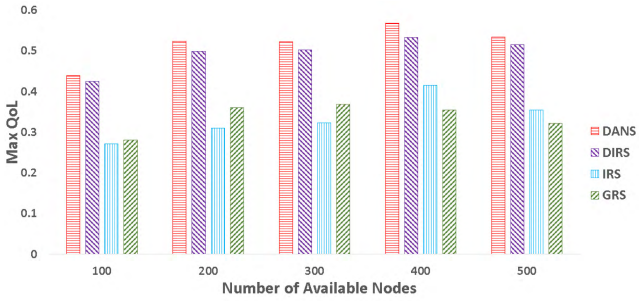

FIGURE 18. The effect of varying the population size on the maximum achieved *QoL*, using dataset 2.

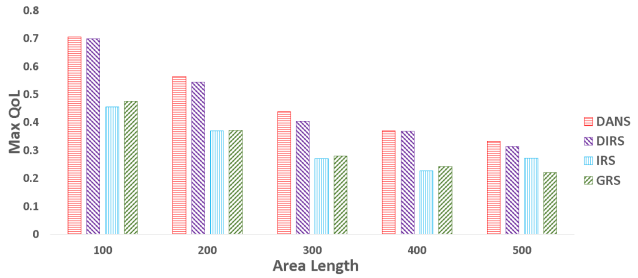

FIGURE 19. The effect of varying the AoI size on the maximum achieved *QoL*, using dataset 2.

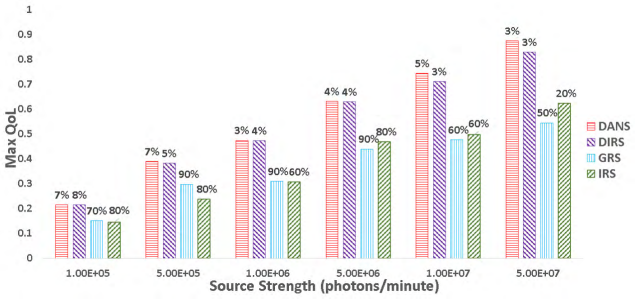

FIGURE 20. The effect of varying the source strength on the maximum achieved *QoL*.

is, the smaller the required group size. For example, at a low source strength of 10^5 photons/minute, DANS achieves the highest *QoL* with only a group size of 7%, compared to 8% for DIRS, 70% for GRS, and 80% for IRS, with the latter two having low maximum *QoL*. This shows the ability of DANS in choosing the informative nodes and performing better than other approaches even with much smaller number of active nodes. For all source strengths, the data driven approaches show high maximum *QoL* with small number of active nodes (3%-8%), whereas other approaches have lower maximum *QoL* even with higher number of active nodes (20% - 90%).

VII. CONCLUSION

In this paper, a two-phase selection mechanism that uses genetic and greedy algorithms has been proposed. The first phase employs a group-based selection using genetic algorithm to collect primitive data about the source. In the second phase, the readings from the active nodes are dynamically used to select the next best set of active nodes through an individual-based greedy selection mechanism. Both phases are integrated in a novel and dynamic Data-driven Active Node Selection framework (called DANS), which tackles localization tasks in IoT sensing applications. Additionally, a coverage assessment method has been developed considering sensors which may not have a sensing range. The effectiveness of the proposed approach is verified by experiments using real-life and synthetic datasets. It is compared to existing data-independent benchmarks, GRS and IRS, in terms of localization time, power cost, and Quality of Localization (*QoL*) metric. The results demonstrated that the proposed approach, DANS, outperforms existing benchmarks in terms of *QoL* by up to 52% using the synthetic dataset, and by up to 75% using the real-life dataset. DANS was especially shown to perform better for small groups of active nodes, where its performance in terms of *QoL* exceeded benchmarks by up to 4 times using a synthetic dataset. It was also found to outperform existing benchmarks by a percentage as high as 74% using a real-life dataset. The same trend is also observed for localization time and power cost, where DANS systematically does exceedingly better than the existing benchmarks. DANS also displayed scalability in terms of population and area sizes and demonstrated superior performance in diverse conditions. The results verify the viability of the presented mechanism and show that by using a novel data-driven approach and the proposed selection algorithm, a faster, more reliable, and lower-cost localization task can be performed even with small number of active nodes, in differing situations and environments.

ACKNOWLEDGEMENT

The authors would like to thank the Research Computing Team, Khalifa University of Science and Technology, for providing the High Performance Computing (HPC) cluster for simulations.

REFERENCES

[1] M. A. Feki, F. Kawsar, M. Boussard, and L. Trappeniers, "The Internet of things: The next technological revolution," *Computer*, vol. 46, no. 2, pp. 24–25, Feb. 2013.
[2] D. Giusto, A. Lera, G. Morabito, and L. Atzori, *The Internet of Things: 20th Tyrrhenian Workshop on Digital Communications*. New York, NY, USA: Springer, 2010.
[3] P. Bellavista, G. Cardone, A. Corradi, and L. Foschini, "Convergence of MANET and WSN in IoT urban scenarios," *IEEE Sensors J.*, vol. 13, no. 10, pp. 3558–3567, Oct. 2013.
[4] V. C. Gungor, B. Lu, and G. P. Hancke, "Opportunities and challenges of wireless sensor networks in smart grid," *IEEE Trans. Ind. Electron.*, vol. 57, no. 10, pp. 3557–3564, Oct. 2010.
[5] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.
[6] I. Lee and K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," *Bus. Horizons*, vol. 58, no. 4, pp. 431–440, 2015.
[7] M. T. Lazarescu, "Design of a WSN platform for long-term environmental monitoring for IoT applications," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 3, no. 1, pp. 45–54, Mar. 2013.
[8] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generat. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, 2013.
[9] L. Yi, X. Deng, M. Wang, D. Ding, and Y. Wang, "Localized confident information coverage hole detection in Internet of things for radioactive pollution monitoring," *IEEE Access*, vol. 5, pp. 18665–18674, 2017.
[10] A. Liu, M. Wu, K. M. Chandy, D. Obenshain, M. Smith, and R. McLean. (2009). *Design Tradeoffs for Radiation Detection Sensor Networks*. [Online]. Available: http://www.cs.caltech.edu/~aliu/documents/IPSN_final.pdf
[11] H. Son, N. Kang, B. Gwak, and D. Lee, "An adaptive IoT trust estimation scheme combining interaction history and stereotypical reputation," in *Proc. 14th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jun. 2017, pp. 349–352.
[12] F. Delicato, F. Protti, L. Pirmez, and J. F. de Rezende, "An efficient heuristic for selecting active nodes in wireless sensor networks," *Comput. Netw.*, vol. 50, no. 18, pp. 3701–3720, Dec. 2006.
[13] A. H. Liu, J. J. Bunn, and K. M. Chandy, "An analysis of data fusion for radiation detection and localization," in *Proc. 13th Int. Conf. Inf. Fusion*, Jul. 2010, pp. 1–8.
[14] A. Mohamed and K. Marzouk, "Optimizing the energy consumption of wireless sensor networks," *Int. J. Appl. Inf. Syst.*, vol. 10, no. 2, pp. 1–5, Dec. 2015.
[15] Z. Liu, W. Dai, and M. Z. Win, "Node placement for localization networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[16] R. Azzam, R. Mizouni, H. Otrok, A. Ouali, and S. Singh, "GRS: A group-based recruitment system for mobile crowd sensing," *J. Netw. Comput. Appl.*, vol. 72, pp. 38–50, Sep. 2016.
[17] H. Xiong, D. Zhang, G. Chen, L. Wang, V. Gauthier, and L. E. Barnes, "iCrowd: Near-optimal task allocation for piggyback crowdsensing," *IEEE Trans. Mobile Comput.*, vol. 15, no. 8, pp. 2010–2022, Aug. 2016.
[18] M. Movassagh and H. S. Aghdasi, "Game theory based node scheduling as a distributed solution for coverage control in wireless sensor networks," *Eng. Appl. Artif. Intell.*, vol. 65, pp. 137–146, Oct. 2017.

[19] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel, "LARS: A location-aware recommender system," in *Proc. IEEE 28th Int. Conf. Data Eng.*, Apr. 2012, pp. 450–461.
[20] B. Martinez, M. Montón, I. Vilajosana, and J. D. Prades, "The power of models: Modeling power consumption for IoT devices," *IEEE Sensors J.*, vol. 15, no. 10, pp. 5777–5789, Oct. 2015.
[21] M. Pouryazdan, B. Kantarci, T. Soyata, L. Foschini, and H. Song, "Quantifying user reputation scores, data trustworthiness, and user incentives in mobile crowd-sensing," *IEEE Access*, vol. 5, pp. 1382–1397, 2017.
[22] M. G. Kendall, *The Advanced Theory of Statistics*. Galveston, TX, USA: Charles Griffin, 1943.
[23] S. Sen *et al.*, "Performance analysis of Wald-statistic based network detection methods for radiation sources," in *Proc. 19th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2016, pp. 820–827.
[24] G. Knoll, *Radiation Detection and Measurement*. Hoboken, NJ, USA: Wiley, 2010.
[25] H. Jayakumar, K. Lee, W. S. Lee, A. Raha, Y. Kim, and V. Raghunathan, "Powering the Internet of things," in *Proc. IEEE/ACM Int. Symp. Power Electron. Design (ISLPED)*, Aug. 2014, pp. 375–380.
[26] J. Choi *et al.*, "Detecting and identifying faulty IoT devices in smart home with context extraction," in *Proc. 48th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, Jun. 2018, pp. 610–621.
[27] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 2, pp. 159–170, 2nd Quart., 2010.
[28] P. Jiang, "A new method for node fault detection in wireless sensor networks," *Sensors*, vol. 9, no. 2, pp. 1282–1294, Feb. 2009.
[29] D. Miljković, "Fault detection methods: A literature survey," in *Proc. 34th Int. Conv. MIPRO*, May 2011, pp. 750–755.
[30] R. Estrada, R. Mizouni, H. Otrok, A. Ouali, and J. Bentahar, "A crowd sensing framework for allocation of time constrained and location-based tasks," *IEEE Trans. Services Computing*, to be published, doi: 10.1109/TSC.2017.2725835.
[31] M. Abououf, S. Singh, H. Otrok, R. Mizouni, and A. Ouali, "Gale-shapley matching game selection—A framework for user satisfaction," *IEEE Access*, vol. 7, pp. 3694–3703, 2018.
[32] K. J. Nurmela and P. R. J. Östergård, "Covering a square with up to 30 equal circles," Dept. Comput. Sci. Eng., Helsinki Univ. Technol., Espoo, Finland, Res. Rep. A62, 2000.

AHMED ALAGHA received the B.Sc. degree in electrical and electronic engineering from the Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, where he is currently pursuing the M.Sc. degree in electrical and computer engineering. His research interests include the IoT, sensing technologies, crowd sensing and sourcing, radiation detection and localization, and semiconductor devices.

SHAKTI SINGH received the B.Sc., M.Sc., and Ph.D. degrees in electrical and computer engineering from Purdue University, West Lafayette, IN, USA. He is currently an Assistant Professor with the Electrical and Computer Engineering Department, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. His research interests include semiconductor devices and integrated circuits, sensors, sensing technologies, crowd sourcing, crowd sensing, and the development of the IoT and wireless sensor networks.

RABEB MIZOUNI received the M.Sc. and Ph.D. degrees in electrical and computer engineering from Concordia University, Montreal, Canada, in 2007 and 2002, respectively. She is currently an Associate Professor in electrical and computer engineering with the Khalifa University of Science and Technology. Her current research interests include the deployment of context aware mobile applications, crowd sensing, software product line, and cloud computing.

ANIS OUALI received the B.Sc. degree in computer engineering from the L'Ecole Nationale des Sciences de l'Informatique, Tunisia, in 2000, the M.Sc. degree in computer science from the Université du Québec à Montréal, Canada, in 2004, and the Ph.D. degree from the Electrical and Computer Engineering Department, Concordia University, Montreal, Canada, in 2011. He joined the Emirates ICT Innovation Center, in 2010, where he is currently working in the network optimization team which focuses on solving network design related problems. His research interests include P2P networks for video streaming, distributed computing, and content adaptation.

HADI OTROK received the Ph.D. degree in ECE from Concordia University. He is currently an Associate Professor with the Department of ECE, Khalifa University of Science and Technology, an affiliate Associate Professor with the Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada, and an affiliate Associate Professor with the Electrical Department, École de Technologie Supérieure, Montreal, Canada. His research interests include the domain of computer and network security, crowd sensing and sourcing, ad hoc networks, and cloud security. He is a Senior Member of IEEE. He co-chaired several committees at various IEEE conferences. He is an Associate Editor of *Ad-hoc Networks* (Elsevier) and the IEEE COMMUNICATIONS LETTERS.