Hiding Contextual Information in WSNs

Alejandro Proano
Dept. of Electrical and Computer Engineering
University of Arizona
Tucson, AZ, USA
Email: aaproano@email.arizona.edu

Loukas Lazos
Dept. of Electrical and Computer Engineering
University of Arizona
Tucson, AZ, USA
Email: llazos@ece.arizona.edu

***Abstract*—We address the problem of preserving the confidentiality of contextual information in wireless sensor networks (WSNs). Such information includes the time and location of events observed by the WSN, the position of the sink, and possible routes to the sink. Contextual information can be extracted via traffic analysis, even when all traffic is encrypted. We consider a global threat model in which the adversary is assumed to be capable of eavesdropping on all communications. Compared to previous works, our method significantly reduces the communication overhead for hiding contextual information. In our approach, we first reduce the number of bogus traffic sources necessary for hiding traffic patterns by finding minimum connected dominating sets that cover the deployment area. We then randomize the traffic distributions observed by eavesdropping nodes.**

I. Introduction

Wireless sensor networks (WSNs) can effectively monitor our physical word at low cost. When collected information is of sensitive nature, its confidentiality is typically secured via cryptographic methods, such as data encryption. However, encryption alone cannot prevent the leakage of contextual information such as the location and time of occurrence of a sensed event, the sink's position, and the routing paths followed by data packets. Passive eavesdroppers can infer contextual information by correlating low-level packet identifiers and performing traffic analysis, even if the contents of the communications remain secret. [2], [3], [8], [9], [13].

The problem of hiding contextual information in WSNs has been addressed under both local and global adversary models. Under a local adversary model, the eavesdroppers are assumed to have a limited presence within the WSN [3], [6], [12], and therefore can intercept only a fraction of the WSN traffic. Hiding methods include random and directed random walks [2], addition of pseudo-sources and pseudo-destinations, creation of routing loops [6], [12], and flooding [2]. These methods fail to provide protection under a global eavesdropper capable of intercepting all communications within the WSN [3], [13]. This global threat model is a plausible scenario given the relatively low acquisition and deployment costs of WSNs. For example, $10,000$ sensors valued at $\$1$ each, can be obtained with an expense of $\$10,000$.

Current techniques for hiding contextual information in the presence of a global adversary, employ bogus transmissions that normalize traffic patterns independent of the occurrence of real communications [3], [8], [9], [13]. However, these techniques introduce significant communication overhead which is proportional to the size of the WSNs. In this paper, *we address the problem of hiding contextual information under a global adversary, in a resource-efficient manner.*

Our Contributions: We propose a resource-efficient hiding scheme which hides contextual information using fake data sources. Our scheme differs from previous approaches in that it employs only a subset of sensors for transmitting bogus traffic. This set is independent of the sensor deployment density. We map the problem of selecting the fake sources to the problem of finding a minimum connected dominating set (MCDS) that covers the deployment area. The MCDS guarantees that every sensor is at most one hop from the MCDS and that every eavesdropper perceives random traffic patterns. We then regulate the transmissions of real sources such that their impact on the observed traffic rates is statistically insignificant. Our simulations show that our method significantly reduces the communication overhead for hiding contextual information.

The remainder of the paper is organized as follows. In Section II, we present related work. In Section III, we state our model assumptions. Section IV presents our privacy preserving scheme. In Section V, we evaluate the performance of our hiding method, and in Section VI, we conclude.

II. Related Work

Methods for hiding contextual information in WSNs can be classified to those considering a local adversary model and those considering a global adversary model. Due to space limitation we focus on the latter class.

In [8], the authors proposed two traffic normalization methods based on the injection of dummy traffic; periodic collection and source simulation. In periodic collection, each sensor generating bogus packets at a constant rate. To transmit real data, sensors substitute fake packets with real ones and transmit them at the same constant rate. This method prevents the extraction of information from traffic analysis even in the presence of a global eavesdropper, at the expense of significant communication overhead due to the global dummy traffic generation. In the source simulation method, the incurred communication overhead is reduced by selecting only a subset of sensors as sources of dummy traffic. These are chosen in such a way so that they simulate the expected distribution of real events which has to be known beforehand.

In [2], [3], the authors propose a fractal propagation scheme based on bogus traffic. Assuming that the adversary is not

978-1-4673-1239-4/12/$31.00 ©2012 IEEE

capable of distinguish real sources from fake ones, a sensor that observes the transmission of real data in its vicinity, generates a fake packet with some probability p and forwards it to its neighbors. The packet is probabilistically flooded in a radius of K hops from the fake source, at each hop the packet is retransmitted with probability p.

In [9], Shao et. al. propose a method to reduce the tradeoff between the event reporting delay and privacy in WSNs. All sensors transmit bogus traffic based on a predetermined distribution. A sensor that wants to insert a real data packet among fake ones, reduces the inter-message delay time by transmitting earlier than the time dictated by its fake packet distribution. It then compensates by delaying the transmission of the next fake packet. In the proposed scheme all sensor nodes are sources of bogus traffic.

In [13], the authors propose methods that reduce the propagation of dummy messages through the network. In the *Proxy-based Filtering Scheme (PFS)*, a subset of sensors are designated as proxies in different parts of the deployment area. Sensors transmit packets (real or bogus) to the closest proxy who filters dummy traffic and forwards real messages towards the base station. At the proxies, real traffic is transmitted along with bogus traffic in order to maintain the uniformity of the traffic patterns in the network. In the *Tree-based Filtering Scheme (TFS)*, the proxies are organized as a tree structure rooted at the base station, which prevents messages from traversing through multiple proxies.

III. Network and Adversarial Models

Network Model: The network consists of a set of $\mathcal{V}$ sensors randomly deployed within an area of interest. The communications among sensors follow the unit disc graph model. Every sensor $v \in \mathcal{V}$ has a fixed communication range γ and a known location ℓ_v. Relevant packet identifiers such as source and destination addresses are assumed to be hidden from eavesdroppers via the application of link level re-encryption. This prevents the correlation of incoming to outgoing traffic on any given sensor based on packet contents. Contention management protocols are assumed to conform to the packet transmission rates imposed by our scheme.

Adversary Model: We assume that the adversary randomly deploys a set of sensors $\mathcal{A}$ with minimum density for 1-covering the sensor field with a desired probability p_c. This density can be calculated using well known analytic formulas [7]. The adversary network collectively eavesdrops on all traffic of the WSN, in order to identify the location and time of an event, or the location of the sink. The adversarial sensors are assumed to have the same device characteristics as the legitimate sensors due to cost limitations. For time sensitive information such as the time and location of an event, adversarial sensors are assumed to individually analyze the intercepted traffic. This is preferred in order to reduce the delay associated with centralized detection schemes. For information where detection delay is not critical such as the sink's location, centralized analysis of all traffic is possible.

An adversarial sensor a intercepting a packet, can pinpoint the location of the transmission originator at a granularity equal to the communication area $C_a(\ell_a, \gamma)$ of sensor a. As a result, when static packet identifiers are hidden and randomized due to the application of encryption, eavesdropping sensors cannot differentiate between packets originating from distinct sources within their communication area. Finally, the adversary is assumed to be a passive external observer that does not launch active attacks (e.g., jamming, packet modification and injection attacks) against the WSN.

IV. Hiding Contextual Information

In this section, we propose a scheme that prevents the leakage of contextual information due to traffic analysis. Our scheme involves two phases: a bogus traffic source selection phase and a rate assignment phase.

A. Design Motivation

To hide contextual information, we inject bogus traffic from fake sources, similar to most schemes that deal with global adversaries [8], [9]. Most prior designs require that every sensor of the WSN injects bogus traffic with some rate that satisfies desired statistical properties. However, we make the observation that it is not necessary that all sensors are active sources of bogus traffic, in order for the observation set collected by the adversarial WSN to maintain statistical uniformity. When link-level re-encryption and ciphertext randomization is applied, correlating successive packets based on the packet contents is not possible. Therefore, every eavesdropping sensor records the collective rate of all the sensors located within its communication area, without being capable of computing the individual rate of each source. This allows us to reduce the number of bogus traffic sources as long as all adversarial sensors observe some traffic.

An example of our design motivation is shown in Fig. 1, where a WSN of five sensors $v_1 - v_5$ coexist with three eavesdroppers $a_1 - a_3$. Instead of broadcasting bogus traffic from all sensors $v_1 - v_5$, it is sufficient to choose a subset of those sensors that covers the locations of a_1, a_2, and a_3. Candidate subsets are $\{v_1, v_2, v_3\}$, $\{v_2, v_3, v_5\}$, $\{v_2, v_4\}$ and others. Transmission from a sensor that does not belong to the set of fake sources must be regulated so that the traffic pattern recorded by eavesdroppers does not change. Our design, reduces to the problems of: (a) finding the appropriate subset $\mathcal{D} \subseteq \mathcal{V}$ of sensors that generate bogus traffic, and (b) assigning transmission rates to sensors. We address both problems in two separate phases.

B. Selection of Bogus Traffic Sources

In this phase, we select the set $\mathcal{D} \subseteq \mathcal{V}$ of bogus traffic sources. First, we set the following four selection principles.

(a) Every sensor $a_i \in \mathcal{A}$ must overhear bogus traffic.
(b) The set $\mathcal{D}$ must be of minimum size.
(c) The transmissions of sensors in $\mathcal{V} \backslash \mathcal{D}$ must be minimized.
(d) Sensors in $\mathcal{D}$ must form a connected network.

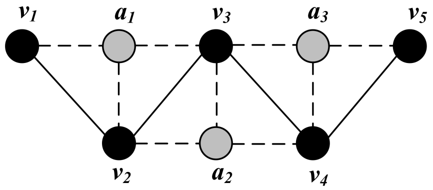

Fig. 1. Small sensor network deployed under the presence of a global adversary.

Principle (a) is a necessary condition in order to guarantee contextual information privacy. If a sensor $a_i \in \mathcal{A}$ is not covered by any sensor in $\mathcal{D}$, it will observe a zero traffic rate in the absence of any real traffic. Hence, if a sensor $v_j \in \mathcal{V} \backslash \mathcal{D}$ within the communication range of a_i broadcasts real traffic, it will always be detected as a real information source. Given that the positions of sensors in $\mathcal{A}$ are unknown (the adversary is passive), to satisfy principle (a), sensors in $\mathcal{D}$ must cover the sensor field.

The second principle minimizes the number of bogus traffic sources, in order to reduce the communication overhead. The third principle minimizes the number of sensors in $\mathcal{V} \backslash \mathcal{D}$ that relay real traffic. This principle minimizes the exposure of real traffic sources to detection, while improving communication efficiency by providing the opportunity to substitute bogus traffic with real traffic (real traffic is relayed only by sensors in $\mathcal{D}$). Finally, principle (d) is a direct consequence of principle (c). To minimize the number of real traffic relays that do not belong to $\mathcal{D}$, sensors in $\mathcal{D}$ must form a connected network. This connected network is responsible for routing real traffic from any sensor to the sink or any other sensor. Based on principles (a)-(d), we reduce the problem of selecting set $\mathcal{D}$ to the problem of *finding a minimum connected dominating set that covers the sensor deployment area*. To show this mapping, we first provide a relevant definition.

Definition 1: Given a graph $G = (\mathcal{V}, \mathcal{E})$ with set of vertices $\mathcal{V}$, and set of edges $\mathcal{E}$, a subset $\mathcal{D} \subseteq \mathcal{V}$ is a *dominating set (DS)* if a vertex $u \in \mathcal{V}$ is in $\mathcal{D}$, or adjacent (within one hop) to some vertex in $\mathcal{D}$. If the set $\mathcal{D}$ induces a connected subgraph, it is a *connected dominating set (CDS)* [5].

In our context, the set of sensors $\mathcal{D}$ that generate bogus traffic must form a minimum size CDS (MCDS) that covers the deployment area. Based on the DS property, every candidate source of real traffic (i.e., sensors in $\mathcal{V}$) is either part of the CDS or within one hop from a sensor that belongs to the CDS. Hence, real traffic requires at most one hop until it is received by a sensor of the CDS. Once in the CDS, traffic is routed to the destination (e.g., sink) using multi-hop routes consisting of sensors that belong to the CDS, and by substituting bogus traffic with the real data. An example of our system is shown in Fig. 2. The sensors in $\mathcal{D}$ form a CDS that covers the deployment area. Sensor s sends data to the BS via CDS sensors v_5, v_4, v_2 and v_1.

Finding an MCDS in random network topologies is known to be an NP-complete problem [4]. In the absence of a polynomial-time algorithm for creating an MCDS, we employ the two-phase heuristic algorithm presented in [1]. This algorithm is distributed and provides a constant approximation of the MCDS with an approximation factor equal to eight, time complexity $\mathcal{O}(|\mathcal{V}|)$ and message complexity $O(|\mathcal{V}| \cdot \Delta)$ where Δ is the maximum degree of any node in the network. Once the MCDS is created, we execute a test phase that verifies area coverage. The steps for the computation of $\mathcal{D}$ are as follows.

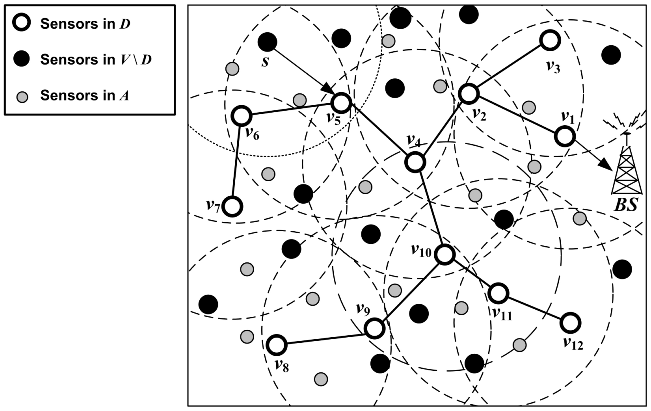

Fig. 2. Sensor S routes information to the base station (BS) via the CDS, by following path $s - v_5 - v_4 - v_2 - v_1 - BS$.

Step 1: DS generation– Given a connected graph $G = (\mathcal{V}, \mathcal{E})$, let $m(v)$ be a marker for $v \in \mathcal{V}$, which can take the values WHITE, BLACK or GRAY. Let also $\mathcal{N}_v$ be the set of neighbors of v, $\delta(v) = |\mathcal{N}_v|$ be the degree of v and $\delta^*(v)$ be the effective degree of v. The latter is defined as the number of neighbors of v for which $m(v) =$ WHITE. Finally, let $b(v)$ denote the number of neighbors of v for which $m(v) =$ BLACK. Initially, $m(v) =$ WHITE, $\delta^*(v) = \delta(v)$ and $b(v) = 0$ for all $v \in \mathcal{V}$. The marking process that outputs a DS is as follows.

- Every node broadcasts its effective degree $\delta^*(v)$.
- A node v changes his marker $m(v)$ to BLACK, if $v = \arg\max_{u \in \mathcal{N}_v \cup \{v\}} \delta^*(u)$. Node v becomes a "dominator" and broadcasts its new marker value.
- A node u with $m(u) =$ WHITE is dominated by a node $v \in \mathcal{N}_u$ if $m(v) =$ BLACK. Node u changes its marker to $m(u) =$ GRAY and broadcasts its new marker value.
- The value $\delta^*(v)$ of a node v decreases by one every time a node $u \in \mathcal{N}_v$ changes its marker to GRAY. Node v broadcasts its new effective degree $\delta^*(v)$.
- The marking process is repeated until no nodes are marked as WHITE (i.e., $\delta^*(v) = 0, \forall v \in \mathcal{V}$).

With the termination of the marking process, the set $\mathcal{D} = \{v : m(v) = \text{BLACK}, v \in \mathcal{V}\}$ forms a DS.

Step 2: Approximation of the MCDS–Given the DS $\mathcal{D}$, the MCDS is approximated by expanding $\mathcal{D}$ such that all nodes in $\mathcal{D}$ are connected. The expansion process is essentially equivalent to generating a Steiner tree for the nodes in the DS, by changing the marker values of GRAY nodes to BLACK. In this phase, BLACK nodes become dominated by GRAY nodes. This can be achieved as follows.

- A node v with $m(v) =$ BLACK elected as the leader (random starting BLACK node) finds node $u \in \mathcal{N}_v$ such that $u = \arg\max_{\mathcal{N}_v} b(u)$. Node u changes its marker value from GRAY to BLACK.
- Every node $w \in \mathcal{N}_u$ with $m(w) =$ BLACK, broadcasts that it is dominated by u.
- Node u broadcasts its new marker value. It also sets $b(v) = 0$ (all BLACK neighbors are dominated).
- Every node w marked GRAY that overhears the message of a dominated node, reduces $b(w)$ by one, and broadcasts its new $b(w)$.
- The process is iteratively repeated until $b(v) = 0, \forall v \in \mathcal{V}$.

Step 3: Verifying coverage–A node $m(v)$ marked GRAY changes his marker to BLACK if its communication area $C_v(\ell_v, \gamma)$ is not covered by $u \in \mathcal{D}$. Set $\mathcal{D}$ is expanded by adding node v. With the termination of this step, the sensor field is covered by at-least one node marked with BLACK.

With the termination of the marking process, the set $\mathcal{D} = \{v : m(v) = \text{BLACK}, v \in \mathcal{V}\}$ form a CDS that 1-covers the sensor field.

C. Assigning Transmission Rates

After the CDS is constructed, we assign transmission rates of bogus traffic to sensors in $\mathcal{D}$. Our solution is based on a random rate assignment that satisfies $(\alpha, \epsilon)-$*unobservability* which is defined as follows [9].

Definition 2: (α, ϵ)*-unobservability*–Given a candidate set of events $\mathcal{E}$, an observation set $\mathcal{O}$ ($\mathcal{E} \subseteq \mathcal{O}$), and probability distributions Z and Z' with parameters (moments) $\theta_1, \ldots, \theta_k$, distributions Z and Z' are indistinguishable under the conditions: (a) $f(Z, Z') \leq g(\alpha)$, and (b) $(1-\epsilon)\theta_i \leq \hat{\theta}_i \leq (1+\epsilon)\theta_i$, for $i = 1, \ldots, k$.

Here, Z is the distribution of $\mathcal{O}$ when $\mathcal{E} = \emptyset$ and Z' is the distribution of $\mathcal{O}$ when $\mathcal{E} \neq \emptyset$. Function $f(Z, Z')$ is the distance between the Z and Z' and α is the significance level. Function g yields the tolerance in deviation between Z and Z' and is a function of the significance level α, and ϵ is the allowed deviation for the parameters of the distribution.

Functions f and g depend on the statistical test employed for testing the similarity between two distribution. This could be a χ^2 test or an Anderson-Darling test [10], for example. The selected value of ϵ depends on the desired false alarm rate on behalf of the adversary. Moreover, the adversary selects the desirable sample size n. A larger n lowers the false alarm rate and reduces the deviation tolerance between Z and Z'. However, it increases the time uncertainty with respect to the occurrence of an event, since any of the set of the n samples of Z' could be responsible for the deviation from Z. Moreover, larger values of n lead to lower sensitivity to changes of individual sample values. For given values α, ϵ and n, the rate assignment is as follows.

Rate assignment in $\mathcal{D}$: We divide time into intervals $I_1, I_2, \ldots$ of length T units. At every interval I_i, a sensor $u \in \mathcal{D}$ transmits bogus traffic at a constant rate r_u^i, which is selected from a probability distribution $\mathcal{Y}(\theta_{1\mathcal{Y}}, \ldots, \theta_{k\mathcal{Y}})$, where $(\theta_{1\mathcal{Y}}, \ldots, \theta_{k\mathcal{Y}})$ are the parameters of the distribution (e.g., $\theta_{1\mathcal{Y}} = \mu_{\mathcal{Y}}$ (mean) and $\theta_{2\mathcal{Y}} = \sigma_{\mathcal{Y}}$ (standard deviation)). The sample space of $\mathcal{Y}$ is defined as $\mathcal{S}_{\mathcal{Y}} = [R_{min}, R_{max}]$.

At a given interval I_i and in the absence of real traffic, an adversary sensor a observes a rate r_a^i which is the sum of the rates of all the neighbors of a that belong to the CDS. Let the set of neighbors of a be denoted by $\mathcal{N}_a$. Then, the random variable representing the observed rate at a is $R_a = \sum_{u \in \mathcal{N}_a} R_u$, distributed according to Z_a, which is a sum of independent and identically distributed (IID) random variables. Therefore, the parameters of Z_a take the values of $\mu_{Z_a} = |\mathcal{N}_a|\mu_{\mathcal{Y}}$ and $\sigma_{Z_a} = |\mathcal{N}_a|\sigma_{\mathcal{Y}}$.

If a sensor that belongs to the CDS wants to transmit real traffic, it simply substitutes bogus traffic packets with real ones. This substitution will not affect the statistical properties of the rates observed by the adversarial sensors.

Rate assignment in $\mathcal{V}\backslash\mathcal{D}$: Assume now that a sensor $s \in \mathcal{V}\backslash\mathcal{D}$ wants to transmit real traffic. This sensor must relay its traffic to the CDS via a one-hop transmission. To do so, sensor v initiates its transmission at the beginning of a time interval I_i at a rate r_s^i. The new rate observed by a sensor $a \in \mathcal{A}$ that overhears the transmission of s is $r_a^i = r_s^i + \sum_{u \in \mathcal{N}_a} r_u^i$, which follows distribution Z_a'.

To detect the difference between Z_a and Z_a', sensor a must use rate samples $(r_a^j, r_a^{j+1}, \ldots, r_a^{j+n}), j \leq i \leq j+n$, observed over n intervals. With every n collected samples, the adversary runs the goodness of fit test to determine if there is statistical evidence to conclude that $(r_a^j, r_a^{j+1}, \ldots, r_a^{j+n})$ is not distributed according Z_a. Here, Z_a and its parameters are assumed to be accurately estimated by a, based on long-term traffic observations. In order to preserve $(\alpha, \epsilon)-$unobservability, the transmission rate r_s^i of sensor s must be chosen in such a way that the sample r_a^i collected during interval I_i by any $a \in \mathcal{A}$ located within the communicating disk $C_s(\ell_s, \gamma)$ of s, is not statistically significant to distinguish Z_a' from Z_a. Note that sample r_a^i that includes r_s^i can be part of the estimation process for a window of n intervals, following interval I_i (the sensor s does not know which n samples will be used by a for performing the test). Hence, sensor v must select r_s^i to satisfy the goodness of fit test for Z_a', estimated with sets of samples $\{(r_a^{i-n}, \ldots, r_a^i), (r_a^{i-n+1}, \ldots, r_a^{i+1}), \ldots, (r_a^i, \ldots, r_a^{i+n})\}$.

To compute the appropriate rate r_s^i, sensor s must be aware of all previous and future samples from I_{i-n+1} to I_{i+n}. To do so, all candidate positions of an adversary sensor a must be considered (the positions of sensors in $\mathcal{A}$ are not known to s). For this purpose, sensor s considers all sensors $v_i \in \mathcal{V}$ whose transmissions can be overheard at any part of its communication disk $C_s(\ell_s, \gamma)$. Using a unit disk model and the known sensor positions, s partitions $C_s(\ell_s, \gamma)$ into areas $U_1, U_2, \ldots U_m$, with area U_i denoting the intersection of $C_s(\ell_s, \gamma)$ with the communication areas of a *unique* subset of sensors of the CDS, heard at U_i. These areas indicate the set of possible distinct rates overheard by an adversarial sensor a, located within $C_s(\ell_s, \gamma)$. An example of the partitioning of $C_s(\ell_s, \gamma)$ into distinct areas U_i is shown in Fig. 3. Sensor

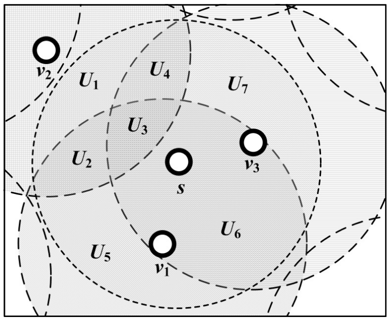

Fig. 3. Sensor $s \in \mathcal{V}\backslash\mathcal{D}$ computes its transmission rate based on the observations made in its communication disk $C_s(\ell_s, \gamma)$. Sensors v_1, v_2, and v_3 belong to $\mathcal{D}$.

$s \notin \mathcal{D}$ is surrounded by sensors $v_i \in \mathcal{D}$. Based on the known coordinates of v_i, sensor s determines U_i.

To compute previous and future rate samples for each U_i, sensor s must be aware of the rates individually selected by each sensor of the CDS, located within its two-hop neighborhood (it can be easily shown via a geometric argument that the communication areas of sensors at most two-hops away from s intersect with $C_s(\ell_s, \gamma)$). This information can be made available during the CDS setup phase. For instance, if the rates are selected from a uniform distribution, each sensor of the CDS can provide the random seed value used for generating random rates to its two-hop neighbors. Moreover, the values of the samples used in the goodness of fit test must be adjusted to account for traffic originating from other co-located real data sources. To allow for accurate computation of the sample rates at all areas U_j, when a sensor chooses its rate r_s^i, it announces this information to its two-hop neighborhood (using the selected rate). Hence, other sensors that are affected by this real traffic transmission, obtain the correct sample values.

Sensor s runs the goodness of fit test for all the sets of samples in the set $\{(r_a^{i-n}, \ldots, r_a^i), (r_a^{i-n+1}, \ldots, r_a^{i+1}), \ldots, (r_a^i, \ldots, r_a^{i+n})\}$ for each of the areas $U_1, U_2, \ldots U_m$. The maximum rate r_v^i that satisfies all the goodness of fit tests is selected.

For example, consider the network presented in Fig. 3. In order to transmit at interval I_i, sensor s (not in the CDS) divides its communication disk $C_s(\ell_s, \gamma)$ in seven non-overlapping areas $U_1, \ldots, U_7$ ($U_1 = C_s(\ell_s, \gamma) \cap C_{v_2}(\ell_{v_2}, \gamma)$, $U_2 = C_s(\ell_s, \gamma) \cap C_{v_1}(\ell_{v_1}, \gamma) \cap C_{v_2}(\ell_{v_2}, \gamma), \ldots,$ $U_7 = C_s(\ell_s, \gamma) \cap C_{v_3}(\ell_{v_3}, \gamma)$). First, sensor s assumes that $a \in \mathcal{A}$ is located in U_1. Using the random seed provided by v_2, s obtains the set of samples $\{(r_a^{i-n}, \ldots, r_a^i), (r_a^{i-n+1}, \ldots, r_a^{i+1}), \ldots, (r_a^i, \ldots, r_a^{i+n})\}$. In this case, $r_a^j = r_{v_2}^j$ (for $j = i-n, \ldots, i+n$ and $j \neq i$) and $r_a^j = r_{v_2}^j + r_s^j$ (for $j = i$). Using the set of samples, s estimates r_s^i that will satisfy the goodness of fit test. For instance, consider the assignment of rates based on the χ^2 test [11]. To compute the distance between Z_a and Z_a', data is divided into q classes defined by dividing $[R_a^{min}, R_a^{max}]$ into q non-overlapping intervals. Each sample $(r_a^{i-n}, \ldots, r_a^i)$ is accordingly assigned to each class. After the data is classified, the number of samples O_j as well as the expected number of samples Q_j in class j are calculated. The values of O_j depend on Z_a' and the values of Q_j depend on Z_a. The test concludes that Z_a' and Z_a are indistinguishable if,

(a) $\sum_{j=1}^{m} \frac{(O_j - Q_j)^2}{Q_j} \leq \chi^2_{(\alpha, q-c)}$.
(b) $(1-\epsilon)\mu_{Z_a} \leq \mu_{Z_a'} \leq (1+\epsilon)\mu_{Z_a}$.
(c) $(1-\epsilon)\sigma_{Z_a} \leq \sigma_{Z_a'} \leq (1+\epsilon)\sigma_{Z_a}$.

Where $\chi^2_{(\alpha, q-c)}$ is the value of the χ^2 distribution for significance level α and $(q-c)$ degrees of freedom, and $(c-1)$ is number of parameters of distribution Z_a. The test is repeated for all the samples obtained in areas $U_2, \ldots U_7$ and the maximum allowable rate is selected.

Computational Complexity: The computation of the transmission rate of a sensor $s \in \mathcal{V}\backslash\mathcal{D}$ requires that node s runs the statistical test for the set of sample rates $\{(r_a^{i-n}, \ldots, r_a^i), (r_a^{i-n+1}, \ldots, r_a^{i+1}), \ldots, (r_a^i, \ldots, r_a^{i+n})\}$ obtained in each of the areas $U_1, U_2, \ldots U_m$. Hence, the total number of tests that are run is $n \times m$. If we consider the χ^2 test described above, the test is done in two stages. In the first stage the samples are grouped in q non-overlapping classes, which can be performed using a sorting algorithm like quicksort or heapsort of time complexity $O(n \log n)$. In the second stage, the values of O_i and $\frac{(O_i - Q_i)^2}{Q_i}$ are calculated for each class, which requires $O(q)$ operations, so the time complexity of the test is $O(n \log n)$. Since for each area U_i we require to test the set of samples $\{(r_a^{i-n}, \ldots, r_a^i), (r_a^{i-n+1}, \ldots, r_a^{i+1}), \ldots, (r_a^i, \ldots, r_a^{i+n})\}$, the total complexity is $O(mn \log n)$. However, note that consecutive sample sets differ only in the first and last sample. Given the values of O_j for the first set of samples, assume that r_a^{i-n} belongs to class k and r_a^{i+1} to class l. The values of O_j for the second set remain the same in all the classes except for k and l, for which the values are $O_k - 1$ and $O_l + 1$, respectively. Following this process, we can perform the test by avoiding to sort the samples every time, and reducing the total number of operations to perform the n test to $O(n \log n) + 2(n-1) = O(n \log n)$.

V. Performance Evaluation

We uniformly deployed a WSN of density $d_{\mathcal{V}}$ within an area of $1,000 \times 1,000$ meters, and computed the MCDS set $\mathcal{D}$ that covers the sensor field. We further uniformly deployed $5,000$ adversarial sensors. For sensors in $\mathcal{D}$, the bogus transmission rate at each interval was selected from a uniform distribution in $(0, 1]$. The real traffic transmission rate r_s^i at interval I_i for a sensor $s \in \mathcal{V}\backslash\mathcal{D}$ was chosen by applying a χ^2 test with parameters $\alpha = 0.05$, $n = 100$, and the mean and standard deviation tests with parameter $\epsilon = 0.1$. Fig. 4(a) shows the fraction of sensors that belonged to the MCDS set $\mathcal{D}$, as a function of the sensor density $d_{\mathcal{V}}$. This metric indicates the fraction of sensors that are active sources of bogus traffic and is proportional to the amount of energy savings achieved compared to methods where all sensors are fake sources. We observe that our scheme satisfies (α, ϵ)-unobservability while reducing the number of fake sources by more than 90%. This percentage is reduced to as much as 1.7% with the increase

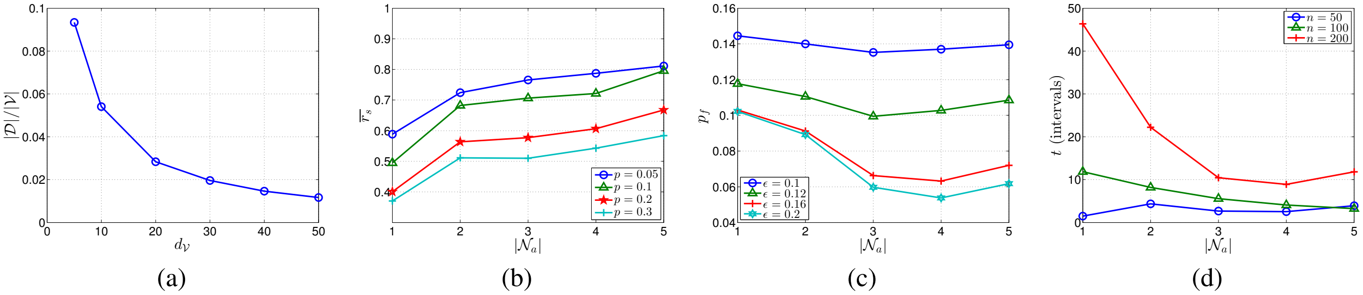

Fig. 4. (a) Fraction of sensors generating bogus traffic as a function of $d_{\mathcal{V}}$, (b) average rate R_v as a function of $|\mathcal{N}_a|$, (c) probability of false alarm p_f in the absence of real traffic (d) average delay introduced by the rate assigment.

of the sensor density. Fig. 4(b) shows the average achievable rate by sources in $\mathcal{V}\backslash\mathcal{D}$ as a function of the number of MCDS neighbors $|\mathcal{N}_a|$ of the adversarial sensors monitoring a real transmission. In this graph, p denotes the probability that a sensor $s \in \mathcal{V}\backslash\mathcal{D}$ transmits real traffic at a given interval. We observe that an increase in $|\mathcal{N}_a|$ allows for higher rates $\overline{r_s}$, since the deviation of Z' from Z is smaller when the bogus rates observed by $a \in \mathcal{A}$ are high compared to $\overline{r_s}$. This allows sensor s to achieve rates as much as 80% of the maximum rate of bogus traffic. Moreover, $\overline{r_s}$ reduces with the increase of p, since more sample values of Z' are part of the same statistical test.

Fig. 4(c) shows the probability of failure of the test in the absence of real traffic, as a function of $|\mathcal{N}_a|$. Note that, this probability considers the probability of false positives of the χ^2 test and the tests on the mean and the standard deviation. We observe that p_f slightly reduces as a function of the $|\mathcal{N}_a|$, which shows a lower rate of failure in areas covered by more MCDS sensors. In this case, we can reduce the number of computations required to obtain the transmission rate, by performing the tests only for the set of samples obtained in the area with the highest p_f. We also observe that p_f decreases with the increase of ϵ. This is expected since increasing ϵ relaxes the tests on the mean and the variance, bringing p_f close to α.

Finally, in Fig. 4(d) we present the average number of intervals a sensor $v \in \mathcal{V}\backslash\mathcal{D}$ must waits until it finds an appropriate interval where all tests are passed for all areas U_j. We observe that the delay reduces with the of $|\mathcal{N}_a|$, which is expected since p_f is higher for small values of $|\mathcal{N}_a|$. On the other hand, the delay increases with the increase of the number of samples n. A larger sample set gives a better estimation of the parameters of the distribution, reducing the probability of false positives and the chances to introduce a new transmission without being detected. However, a larger n increases the adversary's uncertainty with respect to the time of occurence of the observed event.

VI. Conclusions

We addressed the problem of protecting contextual information in WSNs under a global threat model. We proposed a hiding mechanism based on the generation of bogus traffic from a fixed set of fake sources. Our mechanism relies on the computation of a minimum connected dominating set that covers the communication areas of all sensors. We showed that event unobservability can be satisfied by randomizing the transmission rates of the bogus traffic sources and regulating the rates of real ones. Our simulations verified that a significantly smaller number of fake sources is necessary to achieve event unobservability.

Acknowledgments

This research was supported in part by NSF (under grants CNS-0844111, CNS-1016943, and CNS-1145913) Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. Cardei, X. Cheng, X. Cheng, and D. Du. Connected domination in multihop ad hoc wireless networks. In *In Proc. of the Sixth International Conference on Computer Science and Informatics*, 2002.

[2] J. Deng, R. Han, and S. Mishra. Countermeasures against traffic analysis attacks in wireless sensor networks. In *Proc. of the First International Conference on Security and Privacy for Emerging Areas in Communications Networks*, 2005.

[3] J. Deng, R. Han, and S. Mishra. Decorrelating wireless sensor network traffic to inhibit traffic analysis attacks. *Pervasive and Mobile Computing*, 2006.

[4] M. Garey and D. Johnson. *Computers and intractability*. Freeman San Francisco, CA, 1979.

[5] J. Gross and J. Yellen. *Handbook of graph theory*. CRC, 2004.

[6] P. Kamat, Y. Zhang, W. Trappe, and C. Ozturk. Enhancing source-location privacy in sensor network routing. In *In Proc. of the 25th IEEE International Conference on Distributed Computing Systems*, 2005.

[7] L. Lazos and R. Poovendran. Stochastic coverage in heterogeneous sensor networks. *ACM Trans. Sen. Netw.*, 2(3):325–358, 2006.

[8] K. Mehta, D. Liu, and M. Wright. Location privacy in sensor networks against a global eavesdropper. In *In Proc. of the IEEE International Conference on Network Protocols*, 2007.

[9] M. Shao, Y. Yang, S. Zhu, and G. Cao. Towards statistically strong source anonymity for sensor networks. In *In Proc. of the 27th Conference on Computer Communications*, 2008.

[10] M. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 1974.

[11] H. Wadsworth. *Handbook of statistical methods for engineers and scientists*. McGraw-Hill Professional, 1998.

[12] Y. Xi, L. Schwiebert, and W. Shi. Preserving source location privacy in monitoring-based wireless sensor networks. In *In Proc. of the 20th International Parallel and Distributed Processing Symposium*, 2006.

[13] Y. Yang, M. Shao, S. Zhu, B. Urgaonkar, and G. Cao. Towards event source unobservability with minimum network traffic in sensor networks. In *In Proc. of the first ACM conference on Wireless network security*, 2008.