Correlation Analysis for Alleviating Effects of Inserted Data in Wireless Sensor Networks

Sapon Tanachaiwiwat and Ahmed Helmy
Department of Electrical Engineering
University of Southern California, Los Angeles, CA, 90089
{tanachai, helmy}@usc.edu

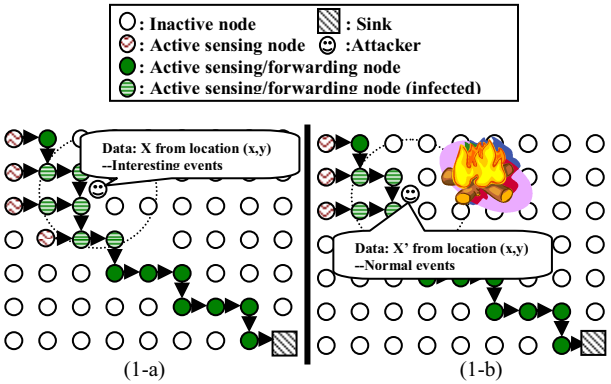
Abstract

This paper introduces a new approach that addresses data contamination problems from attacks in unattended wireless sensor networks. We propose a sliding-window based spatio-temporal correlation analysis called "Abnormal Relationships Test (ART)" to effectively detect, respond and immune to inserted spoofed data from both various-ID impersonators and compromised nodes. Also a systematic approach is given to identify the appropriate sliding window size and correlation coefficient threshold. Our study shows that correlation property of observed phenomenon is not always transitive, different phenomenon from same set of nodes at the same or different period of time can have different correlation coefficients. Our simulation results reveal interesting relationships of outlier percentage and correlation coefficient. With proper parameter setting ART achieves high attack detection rate (90% for correlated attacks and 94% for random attacks even with 100% data insertion).

1. Introduction

Wireless sensor networks (WSN) are designed to collect information in unattended environments; e.g., natural phenomena from remote forests [7]. Several research studies have focused on improving the energy efficiency of routing by using their highly correlated spatio-temporal properties for effective compression [8,14]. However, only a few studies addressed issues of data integrity and accuracy in data gathering in WSN [16]. This paper addresses the data integrity and accuracy problem caused by sybil nodes and compromised nodes. Sybil nodes are forged identities generated by an attacker for malicious purposes, such as creating unfair share of resources leading to denial of service, or forging numerous nodes to have significant biased data reading in WSN. For the latter, inaccurate data readings can cause undesirable effect on critical operations such as remote monitoring of military sites or medical infrastructure. Using median or other robust statistics alone does not avoid large data insertion problem. Hence we propose correlation-based ART to minimize the ill effects caused by unauthorized inserted data.

The effect of data insertion attack is not well studied; some aspects such as data contamination level relating to severity of the attacks is worth deeper analysis. The spoofed data especially from sybil nodes can manipulate or hide useful/interesting data from observers. The attacker can either spoof numerous random or correlated data. Hence it can drastically change the aggregate data such as average, standard deviation; it can also hide the extreme readings when genuine outlier (an extreme deviation from means) occurs. The example scenarios are shown in fig.1.



(a) With normal scenario, attacker inserts abnormal data to cause false alarm.
(b) In real unusual scenario, attacker attempts to insert looked-normal data instead to hide interesting event.

Figure 1 shows example scenarios of one attacker trying to insert data to active sensing/forwarding nodes while pretending to be a set of other valid nodes (especially inactive/destroyed nodes)

Outliers are normally discarded from majority of data. However, due to the unpredictable nature of observed phenomena, the detected outliers should not be just rejected without proper justification. The outliers can lead to important observations. Furthermore, outliers are not always malicious and vice versa. In the case of trace covering, malicious

Proceedings of the Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous'05)
0-7695-2375-7/05 $20.00 © 2005 IEEE

data can actually look normal (static threshold detection technique will fail here).

The challenge is to distinguish between inserted data, failed, un-calibrated and real data with appropriate investigation and effectively react to different scenarios. Furthermore, what makes the problem even more challenging is that the observant sensor nodes have limited access to non-neighbors' data and limited capability in processing and storing a large data set. These problems are extensively discussed in this work.

The rest of paper is organized as follows: in Section 2 we describe some of related work in the literatures. We provide attack model and basic assumption in Section 3. In Section 4, we provide the basic structure and deployed architecture of ART. We also provide description of the ART statistical analysis and authentication mechanisms in Section 5. In Section 6, we explain how sink can statistically analyze anomaly reports. We provide the interpretation of set of ART results in Section 7. In Section 8, we address some limitation issues of the ART mechanisms. We show important observations from "Great Duck Island" data characteristics in Section 9. We simulate ART mechanisms with these data and discuss the results in Section 10. In Section 11, we present our concluding comments and describe the future work.

2. Related Works

Prior work on sybil attacks in sensor networks [11] classifies types and possible effects of such attack. The paper proposes sensor node validation using radio signal and keypool test. It mentions that the misbehavior detection will likely have some false positives caused by sybil nodes' spread blames. Our ART mechanism can minimize such ill effect by using correlation analysis and distributed interactive proof.

For internal attackers, compromised nodes, and faulty nodes, cryptography alone [15, 17, 21, 22] is inadequate. An alternate reputation-based approach [6] relies on reports from other nodes can also help in reducing the chance of accepting and forwarding false reports. However, pretend-to-be high rating nodes (sybil node or compromised node) can still silently insert malicious information. By contrast, our work directly tests the content of data. The test results can be interpreted as ratings in reputation approach.

Data contamination can sometimes be identified by other detection techniques. Outliers are generally based on data neighborhood or spatial outliers which are discussed in [1, 9, 13]. Those works determine data neighborhood based on the data values while our work relies on the physical location of the data owner. Some spatial outlier detection mechanisms are sensitive to specific pattern of data [9].

In our approach, outliers are determined by the correlation-based deviation (described in section 5.1.1) as opposed to distribution-based outlier detection approach [12]. In that approach, a sensor node performs expensive tests to determine which model fits the data best. The sliding-window correlation-based mechanism in our ART requires fewer resources to capture abnormal activities. Our approach uses dynamic threshold to detect outliers because of the highly dynamic of the observed phenomenon. To minimize the false positive, an adaptive threshold magnitude verification or T*-test (described in section 5.1.2) is proposed coupling with dynamic sliding-window correlation analysis.

Our work also relates to data deviation detection in WSN [12]. However that paper mainly focuses on non-phenomena model and non-security problem. Under heavy attack, their mechanism will incorrectly model the data.

Work in [10] uses history table to have accurate data prediction for modeling specific natural phenomenon and ultimately to recognize types of errors. Again, once the attack starts, it will be quickly filled up with false data.

Some applications necessarily depend upon correlation among neighbors' reading such as the in-place calibration in WSN [2]. The calibration detects the outliers and will drop the outliers without profiling it. Because the calibration process heavily relies on data redundancy; the numerous inserted false references from sybil attacks can disrupt the entire calibration process.

3. Attack Model and Assumption

Our paper focuses on active attackers in wireless communication environment. The attack in our model is either compromised-node or sybil attacks. A compromised node is a captured and manipulated node under full control of attacker with exposed keys. The other attack type, "multiple identity" or sybil attacks, are described below.

Sybil attacks can be classified as (i) direct or indirect communication (i.e., immediate neighbor within communication range or non-immediate neighbor); (ii) fabricated or stolen IDs (i.e., existing or non-existing IDs); and (iii) simultaneous or non-simultaneous attacks; i.e., single or multiple attackers (or single attacker with multiple communication channels)[11].

We only concentrate on direct communication, stolen IDs and both simultaneous and non-simultaneous sybil attacks. However, our proposed

mechanism is also able to handle the compromised node problem.

Due to observer-neighbor range (an observer is a sensor node monitoring immediate neighbor's data for the abnormal activities), packet interception alone does not produce an effective attack; because the observer could receive and detect multiple copies of same sampled data packet (it could also receive replicas of its own data from other sybil nodes).

We assume that cryptography keys of sensor nodes, aggregators and sink cannot be compromised by attacker's eavesdropping (unless the node is captured) and can be used for interactive proof. However traditional cryptography in distributed interactive proof and other sybil attack prevention methods [11] will not solve the compromised-node problem, which we address in this work.

4. Abnormal Relationships Test (ART)

We propose the Abnormal Relationships Test (ART), misbehavior detection mechanism, to alleviate malicious data insertion problem. The ART distributively analyzes integrity of data set relationships as well as verifies data ownership among neighbors in WSN. It immunes to spread blame from sybil nodes and not suffer from high false positive. Furthermore, it is able to examine small data set with minimal bias.

The ART has 2 main modules

1. Statistical Analysis Module
2. Authentication Module

(as shown in fig.2):

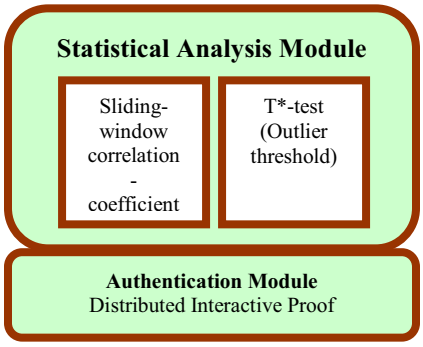


Figure 2 Abnormal Relation Test Modules in each sensor node

Details of each module are examined in section 5. The next section will discuss the pros and cons of 3 types of architectures.

4.1 Architectures for ART deployment

In this section we compare 3 different architectures to obtain the most effective ART deployment: fully distributed architecture, centralized architecture and aggregator/hierarchy architecture.

- **Fully distributed architecture**: If the ART is deployed in every sensor node, each node can be actively involved and quickly response to malicious activity. However, there would be tremendous redundant analysis and possible insufficient information for some nodes. Furthermore, numerous issued reports (unless it is aggregated) from individual nodes to sinks can cause bottleneck and unnecessary energy waste especially in pushing model.
- **Centralized architecture:** This approach requires all data submitted to sink. It does not yield effective response time (if transmission time is much higher than processing time). However sink always have better understanding for overall data characteristic.
- **Aggregate/Hierarchical architecture:** To optimize response time and energy, this architecture is selected for the ART deployment. We assign packet-forwarding nodes in geographic routing to be aggregators to verify the data from neighbors with the ART mechanisms. However if any aggregator has large number of nodes under its supervision, it will have the same problem as a sink in centralized approach does.

5. ART Mechanisms

Let neighbor set $\Xi=\{\eta_i \mid \eta_i$ is the i^{th} neighbor$\}$. By using the ART mechanisms, each neighbor η_i is classified into one of behavior set $\mathbf{B}$ = { normal, compromised, failed, uncalibrated, sybil } or { β_j }. Let define $\mathbf{A}(T_i) = \beta_j$ where T_i is a tuple { q_{sw}, q_{t^*}, q_{dip} } representing node i's test results "q_{sw}"and "q_{t^*}" from the statistical analysis and "q_{dip}" from the distributed interactive proof. In Section 5.1, we show description of the ART protocol. Details of each mechanism are explained in Section 5.2 and 5.3.

5. 1 ART protocol

The ART protocol is shown step by step in table 1. Before starting the ART protocol, sensor nodes need to collaboratively determine their appropriate roles i.e. aggregator, sensing nodes, packet-forwarding nodes or combination of those based on their locations. In step 2, aggregator can verify the data authenticity simply checking the data timestamp against known schedule. If timing is right, we require content verification using ART statistical analysis in step 3. The statistical analysis mechanisms, the

Sliding window correlation coefficient" and "T*-test", are thoroughly discussed in Section 5.2.

In step 5, we minimize authentication overhead by only using authentication module when the violations in step 3 are found. ART will report to sink if abnormality counter exceeds thresholds in step 6 or/and identity verification is failed in step 7. Then, in step 8, ART sliding window size is dynamically adjusted according to number of unproven abnormalities. Random authentication is required for minimizing impact of inactive malicious nodes. We describe the details of authentication mechanism, distributed interactive proof, in Section 5.3.

The core of ART protocol has been tested in our simulation and showing its performance in section 9.

Table 1 ART operations

ART protocol:

1. Obtain the sensor readings of all neighbors of node i
2. Verify data timestamp if schedule is known
3. Calculate the correlation coefficient and t*-value for each neighbor.
4. Request additional data if number of data is insufficient.
5. If either test result drop below predefined threshold, authenticate the suspected node(s). Otherwise, forward data to next hop.
6. If suspected node(s) can verify their identity, increment the associated counter by one and forward the packet to next hop. If counter exceeds threshold, drop packet and report to sink.
7. If suspected node(s) fails to verify its identity, drop packets and report to sink.
8. Adjusting window size based on observed outlier percentage and correlation level until reaching min/max window size
9. Randomly authenticate good neighbors with probability p.

5.2 ART Statistical analysis

Authentic data should spatially correlate to observer's data and have approximately the same magnitude as observer's data. This is tested using the following mechanisms.

5.2.1 Sliding-window-based correlation analysis

The correlation analysis is generally used to define the linear relationships of data sets [19]. Because of its high sensitivity to abnormal readings, it is very attractive to use this approach to capture the inserted outlier. This approach is inspired from [2].
An equation for the correlation coefficient is given by:

$$r_{X,Y} = \frac{Cov(X,Y)}{S_X * S_Y}$$

where:

- $r_{X,Y}$ is a correlation coefficient between X and Y data sets.
- **Cov(X,Y)** is a covariance between X and Y data sets.
- S_X and S_Y are sample variances of X and Y data sets.

Data set cannot have zero standard deviation. The range of $r_{X,Y}$ is from -1.0 to 1.0. The higher $|r_{X,Y}|$ is, the closer relationship between X and Y is.

Let **R** be the observer's data set and let **k** be the neighbor k's data set. The number of $r_{R,k}$ values at any given window period of one neighborhood is $\left\lceil \frac{j}{2} \right\rceil$ where **j** = number of nodes in the neighborhood.

For window size **W**, total data **N**, and **j** neighbors, the storage complexity is **O(Wj)** and computational complexity is **O(WNj)** for each aggregator. The effect of different **W** is examined in the next section

❖ **Default sliding window size**

Default sliding window size (W_d) can be chosen based on (1) physical limitation of sensor nodes such as memory size and processing capability or (2) application's interest. Both choices involve sensor nodes' data sampling rate and observation period which are assumed to be static. The application's interest is selected for this purpose.

To assign W_d and sliding distance **D**, we use following parameters:

1. Sampling rate (**F**) or sample per second (1/**F** = data granularity)
2. Precision granularity (**P**) or application's interest granularity.
3. Spanning period (**S**) or total data collection time.
4. Overlapping factor (**O**) or history weight i.e. how history is weighed in the window.

W_d is simply **FP** which is the number of readings in one window, and total number of sliding windows T_w for entire observation period is calculated from $\left\lceil \frac{S}{P} \right\rceil / (1-O)$; hence **FP(1-O)** is the **D** for one window.

If **F** = 0.2 sample per minutes or **1/F** = 5 minutes per sample; **S** = 120 minutes period; **P** = 30 minutes granularity and **O** = 0.75. Then W_d is 6, T_w is 16 (or 0.13 window per minute) and **D** is 1 (1.5 round down) and real **O** is 5/6 = 83.3%.

W_d should be small enough based on earlier explanation. However it should have enough data to maintain certain level of accuracy. Only the

appropriate window size yields minimum false negatives and/or false positives (as shown in section 9).

❖ **Dynamic sliding window**

The ART can dynamically adjust **W** based on the outlier percentage in initial $\mathbf{W_d}$. Outlier percentage is the abnormal data percentage in one window.

The authenticity of outlier values are proven in the distributed interactive proof mechanism. If number of unproven outlier rises, then the ART increases the window size by one until reaching its maximum defined window size $\mathbf{W_U}$ (and vice versa for minimum defined $\mathbf{W_L}$). Because we need more data to either support or reject the correctness of outlier. The appropriate **W** is expected to yield highest accuracy. The paper compares efficiency of both dynamic-window-size and static-window-size approach in section 9.

As mention earlier, all the values are collected over time and used to find the correlation coefficient pair by pair. Let define time-series data for observing sensor node as $\mathbf{T_R}$ and time-series data or data set for neighbor node 0 to **j** are denoted as $\mathbf{T_k}$ where **k** =0 to **j**. The correlation coefficient of time-series data pair between observing sensor node and node **k** is represented by $\mathbf{r_{R,k}}$.

We use relationships of $\mathbf{r_{R,k}}$ and outlier percentage for analyzing the accuracy of sliding window size. Dropping in outlier percentage should increase the level of correlation coefficient and vice versa. If this trend does not hold, the window size is inappropriately set up (or having low accuracy).

❖ **Notations for dynamic window adjustment mechanism**.

- $\Delta_{\mathbf{rR,k}}$: Deviating correlation coefficient of $\mathbf{r_{R,k}}$ **[t]** and $\mathbf{r_{R,k}}$**[t+1]** where **t** defines time.
- $\Delta_{\mathbf{o}}$: Deviating outlier percentage of $\mathbf{O_{R,k}}$**[t]** and $\mathbf{O_{R,k}}$**[t+1]**.
- $\Delta\mathbf{w}$: Deviating sliding window size of $\mathbf{W_t}$ and $\mathbf{W_{t+1}}$

The following assumptions are used as our guidelines to dynamically adjust window size:

1. The $\Delta_{\mathbf{o}}$ and $\Delta_{\mathbf{rR,k}}$ are not supposed to point to the same direction.

2. The larger **W** will reduce the impact of improper small sample size ($\mathbf{W_U}$ and $\mathbf{W_L}$ must be defined).

Hence, for the appropriate **W**, we define the following function:

$$\textbf{Deviating window ratio } (\gamma) = \Delta w/W_t = \frac{\Delta_{r_{RK}} * \Delta_o}{2}$$

where $\{-1.0 \le \gamma \le 1.0$ **and** $W_U \le w \le W_L\}$ and

$$W_{t+1} = (1+\gamma)\, W_t$$

We can observe from the equation that **W** will increase only if both $\Delta_{\mathbf{O}}$ and $\Delta_{\mathbf{r\ R,k}}$ are positive or negative and only decrease when their signs differ. The relationship of this equation is shown in fig. 3.

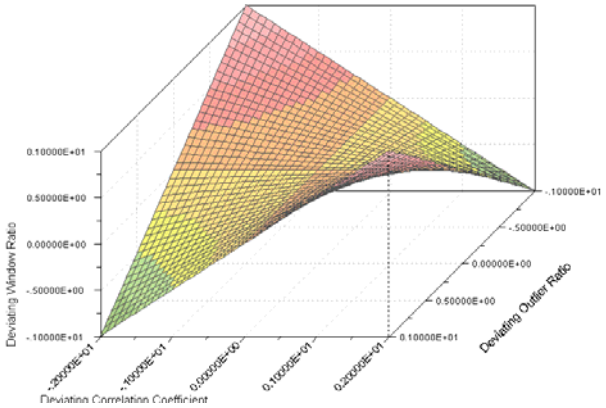


Figure 3 Relationships of deviating outlier percentage (ratio), deviating correlation coefficient and deviating window size ratio.

The threshold for deviating correlation coefficient $\boldsymbol{\pi}_{\Delta}$ is the other important factor to optimize the accuracy of ART algorithm. For simplicity purpose, $\boldsymbol{\pi}_{\Delta}$ is kept constant in the algorithm. $\boldsymbol{\pi}_{\Delta}$ can range from -2.0 to 2.0, the suggested $\boldsymbol{\pi}_{\Delta}$ is between -0.2 and -0.1. The lower $|\boldsymbol{\pi}_{\Delta}|$ yields lower false negative but may increases false positive and unnecessary overheads (more communications and authentications).

Alternatively we can set the threshold based on confidence interval of $\mathbf{r_{R,k}}$. We need to use Fisher z-transform to normalize $\mathbf{r_{R,k}}$.

$$\mathbf{Z_{R,k}} = 0.5 \log \frac{1+\mathbf{r_{R,k}}}{1-\mathbf{r_{R,k}}}$$

This transformed correlation $\mathbf{Z_{R,k}}$ is normally distributed with variance 1/(**W**-3). Hence the 95% confidence interval are given by

$$\mathbf{Z_{R,K}} \pm \frac{\mathbf{Z_{.025}}}{\sqrt{\mathbf{W}-3}} = \mathbf{Z_{R,K}} \pm \frac{1.96}{\sqrt{\mathbf{W}-3}} = (\mathbf{Z_L}, \mathbf{Z_U})$$

To retrieve the format of r, we need to back-transform this limit to r using following equations.

$$\mathbf{r_L} = \frac{\mathbf{e}^{2\mathbf{Z_L}}-1}{\mathbf{e}^{2\mathbf{Z_L}}+1} \textbf{ and } \quad \mathbf{r_U} = \frac{\mathbf{e}^{2\mathbf{Z_U}}-1}{\mathbf{e}^{2\mathbf{Z_U}}+1}$$

With this approach, the threshold is the value of newly calculated $\mathbf{r_{R,k}}$ and not the deviation of $\mathbf{r_{R,k}}$. It is expected to give low false positive and high false negative due to its broad range.

❖ **Predicted performance for ART window size**

We want to use discrete probability analysis to theoretically predict the performance of different

window size. We expect the ART to detect the malicious activity as only small fractions of outliers are inserted (e.g. 10%). We model this scenario as followings: there are 2 types of data: normal data and inserted data. The inserted data are randomly replaced the normal data. Hence it can be approximated from the p.m.f. of hypergeometric distribution (Total data **N** is small and sampled data **W** is large).

Our mechanism uses window size **W** to detect malicious behavior that has at least **x** inserted data in that window.

Let define $\mathbf{D_r}$ = Detection rate or detection probability which is the probability that at least 0.1**W** inserted data is found in one window.

$$\mathbf{D_r} = \sum_{x_i} \frac{\binom{N_1}{x_i}\binom{N_2}{W - x_i}}{\binom{N}{W}}$$

Where

N = total sampled data

$\mathbf{N_1}$= total inserted data

$\mathbf{N_2}$= total good data (N_2=N-N_1)

($\mathbf{x_i \le n}$, $\mathbf{x \le N_1}$ and $\mathbf{n-x \le N_2}$)

Fig.4 shows the $\mathbf{D_r}$ based on following parameters:

N =1000, $\mathbf{N_1}$=100 to 900

$\mathbf{x_i}$ = 1, 2, …,**W** where $\mathbf{W \ge x_i \ge 0.1W}$

W = 3, 6, 9, 12

The highest $\mathbf{D_r}$ can be achieved with **W** = 9 and the worst detection rate is with **W** = 3. While **W** = 12 is slightly better than window size =6 when outlier percentage is higher than 30%. In section 9, we compare this expected theoretical result with simulation results. Please note that this rough estimation assumes full reachability and perfect correlation coefficients among nodes.

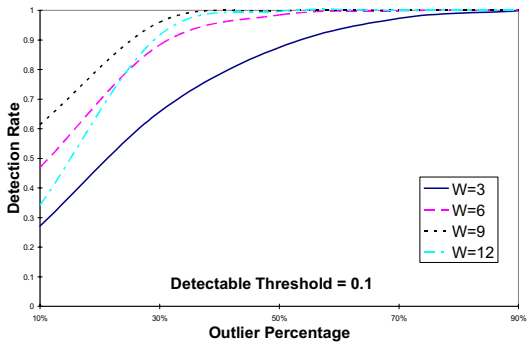


Figure 4 Probability plotting for 4 different sizes of window: (3, 6, 9 and 12) basing on hypergeometric distribution model. The plot is based on 1000 readings of one node with outlier range from 10% to 90%.

❖ **Correlation inference**

Let $\mathbf{r_{k,l}}$ be a correlation coefficient between neighbor **k** and **l** where $\mathbf{k \neq l}$. Hence $\mathbf{r_{R,k}}$ and $\mathbf{r_{R,l}}$ are correlation coefficients between observer and neighbor **k** and **l**.

Random attack from sybil node **k** and **l** generates low $\mathbf{r_{R,k}}$, $\mathbf{r_{R,l}}$, and $\mathbf{r_{k,l}}$. While coordinated attack from sybil node **k** and **l** produces high $\mathbf{r_{k,l}}$ but low $\mathbf{r_{R,k}}$, $\mathbf{r_{R,l}}$. A special case attack that have high $\mathbf{r_{k,l}}$, $\mathbf{r_{R,k}}$, and $\mathbf{r_{R,l}}$ when the inserted data are derived from authentic neighbors' readings. Hence using correlation analysis solely can be insufficient. Observer requires additional data magnitude test. Section 5.1.2 describes the indispensable t*-test which the ART uses with correlation analysis.

5.2.1 Modified t-test (Outlier threshold)

Student's t-test is a statistical method designed for comparing means of 2 populations with small sample set to compensate the possible deviation from real population mean (**μ**) and population standard deviation (**σ**). The test is used to reject null hypothesis if the calculated t is

- higher than $\mathbf{t_\alpha(\upsilon)}$ for upper-tailed hypothesis
- lower than $\mathbf{-t_\alpha(\upsilon)}$ for lower-tailed hypothesis
- $|\mathbf{t}|$ is higher than $\mathbf{t_{\alpha/2}(\upsilon)}$ for two-tailed hypothesis

where **α** is significance level and **υ** is degree of freedom (= n-1).

T-test has an assumption that the data are randomly sampled from approximated normal distribution. The t-test equation is shown below:

$$t = \frac{\overline{\mathbf{X}}_\mathbf{i} - \mu}{\mathbf{S}/\sqrt{\mathbf{n}}}$$

where $\mathbf{X_i}$ is the individual reading, μ is sample mean, **S** is sample standard deviation, and **n** is sample size.

Our mechanism, however, determines outlier by modified "t-test" mechanism or t*-test. Instead of using neighbors' mean, t*-test compares individual neighbor's reading with observer mean (similar to quality control approach). T*-test yields the "t*-score" indicating the level of outlier. The (t*-score) tests are defined as:

$$\text{t*-score} = \frac{\mathbf{X_i} - \mathbf{E(X_R} \pm \delta)}{\sqrt{\mathbf{S_R^2}(\mathbf{X_R} + \delta)/\mathbf{W}}}$$

Where $\mathbf{X_i}$ is neighbor's readings, $\mathbf{X_R}$ is observer's readings, δ is maximum gradient difference (based on long-term observation), $\mathbf{S_R}$ is the observer's sample standard deviation (or standard error).

Based on mean and variance properties:

$$E[X+c] = E[X]+c$$

and

$$\sigma^2(X+c) = \sigma^2(X)$$

where **X** is the random variable and **c** = constant which referring to **δ** in this case.

$$\text{t*-score} = \frac{X_i - (E(X_R) \pm \delta)}{\sqrt{S_R^2(X_R)/W}} = \frac{X_i - (\mu_R \pm \delta)}{S_R/\sqrt{W}}$$

where μ_R is the observer's sample mean.

Using μ_R can avoid inserted-data bias while **δ** accounts for different readings based on diffusion law i.e. $\frac{1}{d^{\phi}}$ of observer's data where **d** is the distance between nodes and **φ** is the specific phenomenon factor e.g. for temperature **φ** =1 and light intensity **φ** =2.

We select t*-score ≥ 3 for outliers identification (outside confidence interval). In normal distribution, ±3σ would account for approximately 99% of population. However, for t-distribution with **W** =6, 9, 12, this would contain approximately 90%, 95% and 95% of population accordingly.

5.3 ART Authentication: Distributed Interactive Proof

The observing sensor nodes are the forwarding nodes in geographic routing and should be able to distinguish between "phenomenon" and "malicious inserted data". By pair-wise authenticating its suspicious neighbors (either symmetric key sets or asymmetric keys [15, 17, 21, 22]), the real neighbors should be able to prove themselves with the correct readings and valid credentials. Ultimately, the authentication process should be accurate and the overhead kept to a minimum. Please note that, for the compromised node attack, it is not possible to directly verify the data via authentication. However, the number of spoofed packets is now limited to the number of compromised nodes (as well as failed nodes). Thus the ART should be able to raise a flag or simply ignore those data (if outlier persists) in future data processing.

Moreover if neighbors' sampling rate and time are known to every sensor node, wrong timing and sampling rate of data reporting indicates the possible sybil attack.

If at least one of neighbors fails to authenticate itself (themselves) or able to authenticate but with different data readings (from the same time stamp), then sensor node can conclude that there is a possible sybil attack in the network. Sink can aggregate and correlate the reported information from several areas and track possible compromised areas as well as the number of attackers.

However it is possible that the report can be intercepted from the attackers, the sensor node can perform robust routing [18] or multi-path routing to prevent the lost of reports.

The other types of data acquiring process, the sensor node reports readings after receiving request only e.g. diffusion routing [7], in that case there will be insufficient data to be learned from neighbor nodes and might not be enough to calculate the spatial/temporal correlation. Alternatively, aggregator can explicitly request data from neighbors. However, sink, by nature, will have more information in hand than other aggregators and can calculate the correlation based on its requested data.

6. Global (sink) view: Spatial and Temporal Characteristics

After sink retrieves number of reports from aggregators, sink can now analyze the global behavior based on spatial and temporal characteristic and determine whether there is serious compromised security in the field.

6.1 Spatial distribution

We assume that sink knows faulty percentage p_0 of observed area. If proportion of anomaly nodes **p** is significant larger than p_0, we can conclude with confidence level (1-**α**) that there is an irregular pattern of abnormality (dense faulty) in the observed area. Again we can define this as binomial distribution which

$$f(x) = \binom{N}{y} p_0^{\,x}(1-p_0)^{N-y}$$

and $H_0 : p = p_0$ vs $H_a : p > p_0$

Where **N** is the number of nodes in specified area and **y** is number of abnormal nodes. We only reject the null hypothesis if $Z = \frac{y/N - p_0}{\sqrt{p_0(1-p_0)/N}} > Z_{\alpha}$

From this analysis, sink can decide not to collect data from any specific area that has dense faulty.

6.2 Temporal Distribution

If a number of total abnormalities (ϑ) for a particular node in a specific time period violates the known or expected failure rate, there is a possibility of high abnormal rate (ϖ) for such node.

$$\varpi = \frac{\vartheta}{\Delta t} = \frac{\vartheta}{\Omega W \Delta s}$$

where Δs is the sampling interval and Ω is the number of windows . If we assume that $\varpi \sim N(\varpi_0, \sigma^2)$, we can again test our hypothesis with following statistic:

$$Z_\upsilon = \frac{\varpi - \varpi_0}{\sigma / \sqrt{\Omega}}$$

where $H_0 : \varpi = \varpi_0$ **vs** $H_a : \varpi > \varpi_0$ which we will reject null hypothesis only when $Z_\upsilon > Z_\alpha$. We can conclude that

7. Behavior Identification

At beginning of section 5, we use $\mathbf{A(T_i)}$ to identify behavior type **B**. The $\mathbf{q_{sw}}$, $\mathbf{q_{t^*}}$ and $\mathbf{q_{dip}}$ can be either "passed" or "failed". For example, node i can have an associated test result $\mathbf{T_i}$ = {Passed, Passed, Failed}. To better classify node, we can also use quantitative value of correlation coefficient and t*-score.

Hence based on $\mathbf{T_i}$, the ART can effectively identify different scenarios including normal event, interesting event, compromised/faulty node, and sybil attack.

Please note that faulty node is a node having problem with a sensing unit or a data processing unit, not a cryptographic module. If the cryptographic module fails, we will treat it as a non-existing node or a sybil node. We assume that captured /compromised nodes' keys will not expose other nodes' keys secrecy otherwise the attackers can assume most of identities for the entire network.

Active compromised or sybil nodes cannot pass statistical tests without being noticed. The main difference between compromised node and sybil node is the ability to authenticate itself with proper credential. Please note that there is possibility that compromised node can act as the failed or un-calibrated node. To distinguish them would require long-term observation of correlation and other necessary parameters.

For abnormal or interesting event case, phenomenon will usually follow diffusion law and have spatial correlated readings with observer. Hence our mechanism will correctly distinguish between unusual phenomenon and attacks. However the closer the distance from observer and smaller number of phenomenal sources is the better accuracy the tests will be.

Table 2 Summary of scenarios from results of our three tests

Behavior Type (B)	Passing Correlation test (Trend)	Passing T*-test (Amplitude)	Passing Distributed Proof (Cryptography)
Normal /Stealth Compromised Node	Y	Y	Y
Abnormal/ Interesting events ≠	Y	Y	Y
Un-calibrated (micro errors)	Y	N (small t*-score)	Y
Compromised / Faulty w/ sensing parts (large errors)	N	Y	Y
	Y	N (large t*-score)	Y
	N	N (large t*-score)	Y
Stealthy Sybil	Y	Y	N
Active Sybil**	N	Y	N
	Y	N	N
	N	N	N

8. Limitation Issues in ART

Even ART can generally handle specified attack model. There are some known limitations of the approach.

8.1 Insufficient Data

Since ART relies on data collected from neighbors, there are 3 possible causes for the insufficient data:

- **Data is aggregated**

With aggregated data e.g. average of temperature over last 30 minutes, ART has 2 choices to operate either use only t*-test for entire window or concatenate multiple windows to form a large scale window e.g. instead of 5-minute interval then we use 30 or 60 minute interval.

- **Neighbor only forwards data and not produces any data**

In this case, analyzing such neighbor is unnecessary. Moreover the node should not be an aggregator if all neighbors only forward the data.

**For sybil attack case, the observer nodes might be able to detect by receiving unusual duplicate or non-neighbor report ≠ test accuracy depends on distance and number of sources.

However for sybil attack which attacker can randomly generate data without priori knowledge of node's role, those data will simply be dropped by aggregator and the attempt will be reported to sink.

- **Limited data generated based on issued query**

Data can be continuously generated or generated per request. In the latter, aggregator can explicitly request extra responses from suspicious neighbor.

8.2 Non-normal or unknown distribution

In some cases, observed data distributions are not Normal such as Poisson, Binomial or Multiplicative. Before performing correlation analysis, we have to transform the observed data (**y**) to following forms:

- Poisson : $\mathbf{y'} = \sqrt{\mathbf{y}}$
- Binomial : $\mathbf{y'} = \mathbf{sin}^{-1}\sqrt{\mathbf{y}}$
- Multiplicative: $\mathbf{y'} = \mathbf{log\ y}$

For unknown distribution, we can always use non-parametric (model-less) statistic such as Spearman rank correlation coefficient. Even though, it is easier to compute the correlation coefficient but information of data is lost in this process.

Otherwise, we can use t*-test to verify the average of readings instead (with **W** > 30), since Central Limit Theorem (CLT) is applicable to any type of distribution.

8.3 Spatial co-operate attackers

If multiple sensor nodes in the same area are compromised, they can co-operate and can form the credible spatial correlated data. Based on ART alone, sink cannot distinguish the good or bad data especially when the compromised area are significantly large e.g. more than half of the total coverage. However if sink is close enough to compromised area (< d unit of distance), it can verify those data against gradient property. The other proposed solution is to have multiple reliable tamper-resistant sensor nodes (TRSN) uniformly distributed in the network and only nearest TRSN probes suspicious nodes upon sink's request.

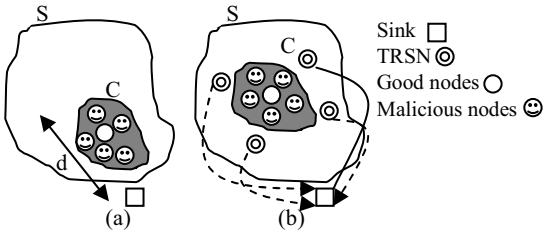
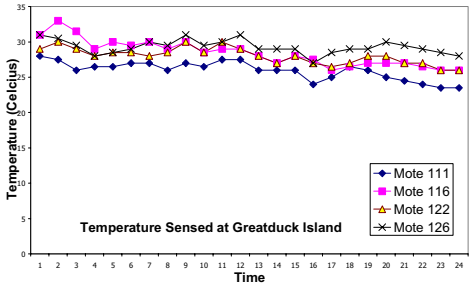


Figure 5 Strategy to alleviate coordinated compromised node attacks (a) using sink's sensor (within distance d) (b) using k-tamper resistant nodes (assume area compromised area C << total area S)

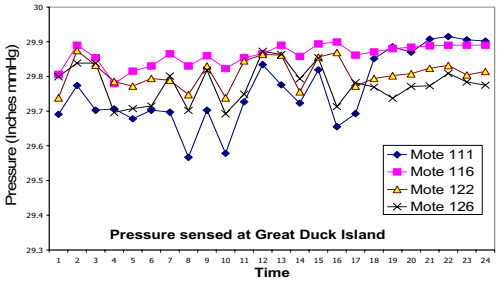
9. Example Scenario: Great Duck Island Project

The Great Duck island project [3] was developed for evaluating the sensor networks' ability to monitor the microclimate on the Great Duck Island, Maine. The sensors periodically report and relay temperature (Celsius), light (Lux), humidity (%), barometric pressure (Hg), video and voice to the base station on the island. This data model is fit to our model perfectly.

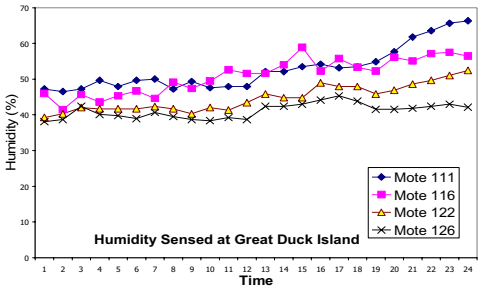
Fig 6 shows the example of data including temperature, pressure and humidity collected from Great Duck Island project.



(6-a)



(6-b)



(6-c)

Figure 6 The data series from nodes 111, 116, 122 and 126. The data were collected from 1:50 PM to 3:50 PM during August 2003. The sampling rate is 0.2 sample/min (24 5-minute interval).

Proceedings of the Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous'05)
0-7695-2375-7/05 $20.00 © 2005 IEEE

9.1 Normality analysis for Great duck island data set

Since our t*-test relies on approximated normal distribution assumption. We use 2 well-known normality tests: P-P Plot and description statistic: Skewness and Kurtosis. Fig. 7 and table 3 show that example mote's temperature reading has slightly deviation from normal distribution. The ideal normal distribution would show straight diagonal line in P-P plot and our distribution is slightly flat, right skewed curve based on calculated kurtosis and skewness.

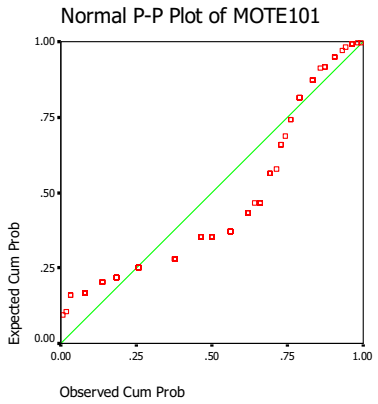


Figure 7 P-P Plot to test normality of mote 101

Table 3 Normality indication of mote 101

Skewness		Kurtosis	
Statistic	Std.Error	Statistic	Std.Error
1.188	.246	.436	.488

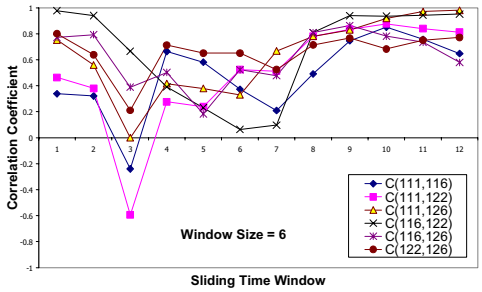
9.2 Correlation coefficients

The correlation coefficient in 2-hour window in Table 4 shows the relationships of data for all nodes.

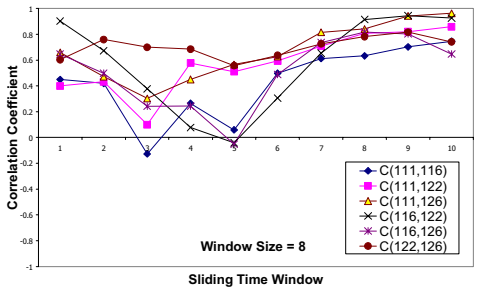
Table 4 Correlation coefficient for 2 hours at Great Duck Island (Temperature/Barometric Pressure/Humidity)

111	1/1/1			
116	0.74/0.64/0.74	1/1/1		
122	0.83/0.42/0.91	0.84/0.67/0.80	1/1/1	
126	0.67/0.41/0.56	0.55/0.50/0.64	0.70/0.55/0.77	1/1/1
Node id	111	116	122	126

As shown in fig.6, correlation coefficients in different window size 6 and 8 show different level of fluctuations. With smaller **W**, the correlation coefficient reveals higher fluctuation.



(8-a)



(8-b)

Figure 8 The correlation coefficient for (a) 6 and (b) 8-sized window (5-minute sampling unit)

From table 4 and fig. 8, followings are novel findings from Great Duck Island's correlation analysis:

- $\mathbf{r_{AB}}$ and $\mathbf{r_{AC}}$ cannot be used to estimate $\mathbf{r_{BC}}$ (non-transitive property) e.g. $\mathbf{r_{111,116}} = 0.74$, $\mathbf{r_{111,126}} = 0.67$ and $\mathbf{r_{116,126}} = 0.55$ (for temperature readings).
- Orders of $\mathbf{r_{AB}}$ for different reading types (temperature, barometric pressure and humidity) are preserved most of the time (at 75% in this observation) e.g. from temperature readings for node 116: $\mathbf{r_{116,122}} > \mathbf{r_{116,111}} > \mathbf{r_{116,126}}$ which is also true for barometric pressure and humidity readings.
- Different time scale and number of collected data will result in different correlation coefficients. We can see this from the different fluctuation of different **W**.

10. Simulations Using Attack Patterns

To evaluate our algorithms, we have designed 2 different types of attacks (both sybil attacks and compromised attacks) as follows:

- Correlated Outlier Attacks (correlated to the set of data)
- Uncorrelated Outlier Attacks (randomly generated within a certain range)

We used 100 sensor nodes in random topology and single sink. Each node samples its own data every 5 minutes. Geographic routing is used to forward the packets. The node forwards data to its selected node.

Our metrics are false positive (incorrectly commit good data) and false negative (incorrectly acquit false data). False positive is Type I error in hypothesis test or the test rejects true null hypothesis (H_0: readings are normal). On the other hand, false negative is Type II error which it accepts false null hypothesis. Detection rate = 1 – false negative.

10.1 Great duck island data set with correlated attacks

We use 8-hour real temperature data set from great duck island to test our **ART**. Fig 9 shows that **W** = 12 has lowest false positives but all **W** have approximately the same false negatives (10%) for any outlier level. When there is no outliers, **W**=12 yields false positive slightly above 10%. This window size has highest false positive around 20% at 30% outlier percentage.

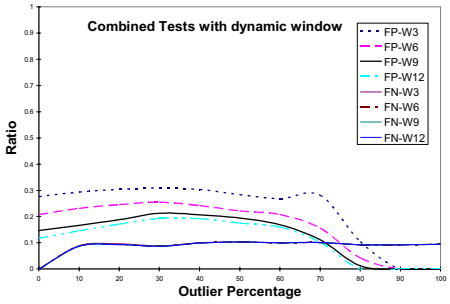


Figure 9 False positive and false negative of four different window size: 3, 6,9, and 12 (threshold = -0.2)

The effect of threshold setting π_Δ is shown in fig. 10. As we expected, the smaller threshold of correlation coefficient causes more false positive. However, they surprisingly have approximately the same false negative. Moreover larger **W** i.e. **W**= 30 cannot further reduce false positive.

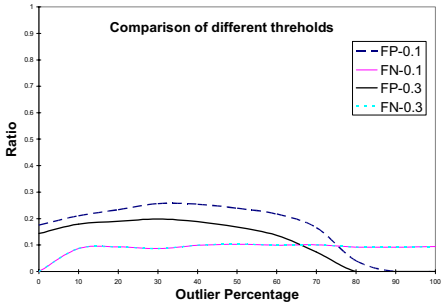


Figure 10 False positive and false negative of three different thresholds: -0.1,-0.2,-0.3 (window size= 30)

As we can see in fig. 11, static-sliding-window-only (SW) approach has very low false positive but incurs very high false negative especially when the outlier percentage is high. T*-test-only approach, however, outperforms every other combination of tests in term of false positive and has equal false negative to the combined approach. Please note that T*-test-only might fail if real outliers from abnormal events are hidden as shown in fig.1 (b).

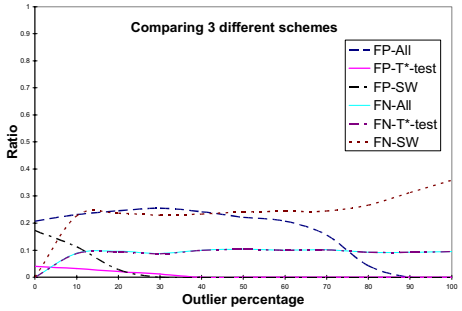


Figure 11 False positive and false negative of 3 different schemes: sliding window (SW) only, t*-test only and all tests (combined) (window size = 6, threshold = -0.2)

10.2 Great duck island data set with random attacks

Fig 12 suggests that random attacks are more detectable than correlated attacks. The false negatives in random attacks are approximately only 6% but the false positives remain the same. Hence false positive is the characteristic of imperfect data correlation and inaccuracy of diffusion property estimation.

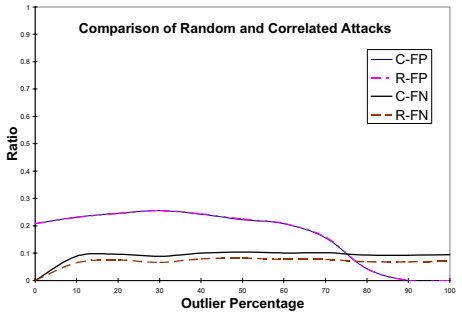


Figure 12 False positive and false negative of 2 types of attacks: Correlated attacks (C) and Random attack (R) (window size = 6, threshold = -0.2)

11. Summary and Future Work

This paper addresses an important security problem in data acquisition in WSN. Sybil attacks and compromised nodes can disrupt the entire data

Proceedings of the Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous'05)
0-7695-2375-7/05 $20.00 © 2005 IEEE

processing process; especially mechanisms that rely on accuracy of data; e.g., in-situ calibration or other critical missions. We introduced the ART mechanism that uses a sliding-window-correlation-coefficient analysis coupled with T*-test to produce low false negatives. Window size has to be chosen carefully for high accuracy and low overhead. Lower correlation coefficient threshold always produces less false negatives but also has higher false positives, which results in higher communication and authentication overhead. Dynamic sliding-window sizing based on outlier existence in a window outperform the static approach.

Our future work will focus on coordinated effect of multiple attackers and attacker mobility. Moreover, this work may directly apply to increase overall quality of service in data acquisition based on the outlier and correlation levels.

Acknowledgement

The authors wish to thank the anonymous referees for their insightful comments. Sapon wishes to thank Wei-Jen Hsu for his helpful feedback.

12. References

[1] V. Barnett and T. Lewis, "Outliers in statistical data," 3rd Edition John Wiley & Sons Inc, 1994.
[2] V. Bychkovskiy, S. Megerian, D. Estrin, M. Potkonjak, "A Collaborative Approach to In-Place Sensor Calibration", IEEE/ACM IPSN, 2003.
[3] A. Cerpa, J. Elson, D. Estrin, L. Girod, M. Hamilton, J. Zhao, "Habitat monitoring: Application driver for wireless communications technology", ACM SIGCOMM Workshop on Data Communications in Latin America and the Caribbean, 2001.
[4] J. Chou, et al, "Tracking and Exploiting Correlations in Dense Sensor Networks," in Proceedings of IEEE Asilomar conference on Signals, Systems and Computers 2002
[5] J. Faruque and A. Helmy, "RUGGED: RoUting on fingerprint Gradients in sEnsor Networks,"
[6] S. Ganeriwal, M. B. Srivastava, "Reputation-based framework for high integrity sensor networks," ACM Security for Ad-hoc and Sensor Networks (*SASN 2004*).
[7] C. Intanagonwiwat, R. Govindan, D. Estrin, "Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks," in Proceedings of MobiCom 2000
[8] C. Intanagonwiwat, et al, "Impact on Network Density on Data Aggregation in Wireless Sensor Networks," in Proceedings of ICDS 2002
[9] C. Lu, et al, "Algorithms for Spatial Outlier Detection," in Proceedings of the 3rd IEEE Intl. conference on Data Mining (ICDM) 2003.
[10] S. Mukhopadhyay, et al, "Data aware, Low cost Error correction for Wireless Sensor Networks," in Proceedings of IEEE Intl. conference on Wireless Communications and Networking Conference (WCNC), 2004.
[11] J. Newsome, E. Shi, D. Song, A. Perrig, "The Sybil Attack in Sensor Networks: Analysis & Defenses," in Proceedings of Intl. conference (IPSN), 2004.
[12] T. Palpamas, et al, "Distributed Deviation Detection in Sensor Networks," in Proceedings of ACM SIGMOD Vol.32 Issue 4 Dec. 2003.
[13] S. Papadimitriou, et al, "LOCI: Fast Outlier Detection Using the Local Correlation Integral," in Proceedings of the 19th IEEE Intl. conference on Data Engineering (ICDE) 2003.
[14] S. Pattem, B. Krishnamachari, R. Govindan, "The Impact of Spatial Correlation on Routing with Compression in Wireless Sensor Networks," in Proceedings of IPSN 2004.
[15] A. Perrig, R. Szewczyk, V. Wen, D. Culler, D. Tygar. SPINS: Security Protocols for Sensor Networks. Wireless Networks Journal, September 2002.
[16]B. Przydatek, D. Song, A. Perrig, "SIA: Secure Information Aggregation in Sensor Networks," in Proceedings of the ACM SenSys 2003.
[17]C. Karlof, N. Sastry, D. Wagner. TinySec: Link Layer Encryption for Tiny Devices. ACM SenSys,2004
[18] S. Tanachaiwiwat, P. Dave, R. Bhindwale, A. Helmy, "Location-centric Isolation of Misbehavior and Trust Routing in Energy-constrained Sensor Networks", IEEE Workshop on Energy-Efficient Wireless Communications and Networks (EWCN), in conjunction with IEEE IPCCC, April 2004
[19] M. C. Vuran, et al, "Spatio-temporal correlation: theory and applications for wireless sensor networks" in Computer Networks 45 (2004) 245-259.
[20] F. Ye, H. Luo, S. Lu, L. Zhang. Statistical En-route Detection and Filtering of Injected False Data in Sensor Networks. In Proceedings of IEEE Infocom, 2004.
[21] W. Zhang and G. Cao, "Group Rekeying for Filtering False Data in Sensor Networks: A Predistribution and Local Collaboration-Based Approach," IEEE INFOCOM, March 2005
[22] S. Zhu, S. Setia, S. Jojodia, P. Ning, "An Interleaved Hop-by-Hop Authentication Scheme for Filtering of Injected False Data in Sensor Networks," Proceedings of the 2004 IEEE Symposium on Security and Privacy (S&P'04)