Segment-Based Anomaly Detection with Approximated Sample Covariance Matrix in Wireless Sensor Networks

Miao Xie, Jiankun Hu, and Song Guo, *Senior Member, IEEE*

Abstract—In wireless sensor networks (WSNs), it has been observed that most abnormal events persist over a considerable period of time instead of being transient. As existing anomaly detection techniques usually operate in a point-based manner that handles each observation individually, they are unable to reliably and efficiently report such long-term anomalies appeared in an individual sensor node. Therefore, in this paper, we focus on a new technique for handling data in a segment-based manner. Considering a collection of neighbouring data segments as random variables, we determine those behaving abnormally by exploiting their spatial predictabilities and, motivated by spatial analysis, specifically investigate how to implement a prediction variance detector in a WSN. As the communication cost incurred in aggregating a covariance matrix is finally optimised using the Spearman's rank correlation coefficient and differential compression, the proposed scheme is able to efficiently detect a wide range of long-term anomalies. In theory, comparing to the regular centralised approach, it can reduce the communication cost by approximately 80 percent. Moreover, its effectiveness is demonstrated by the numerical experiments, with a real world data set collected by the Intel Berkeley Research Lab (IBRL).

Index Terms—Wireless sensor network, anomaly detection, distributed computing, spatial analysis, Spearman's rank correlation coefficient, differential compression

1 Introduction

Wireless sensor networks (WSNs) have found many critical applications in harsh or even hostile environments, such as forest fire detection and battlefield surveillance [1]. The size and cost constraints imposed on a sensor node result in it having scarce resources, such as a small memory capacity, weak computing power, narrow communications bandwidth and limited energy [2]. Moreover, a network self-organised by sensor nodes is very susceptible to communication failures because of the unreliable communication paradigm [3]. Therefore, WSNs are extremely vulnerable to random faults and cyber attacks, and inevitably subjected to their resultant anomalies. According to the literature [4], it has been generally recognised that anomaly detection is usually an effective means against these anomalies.

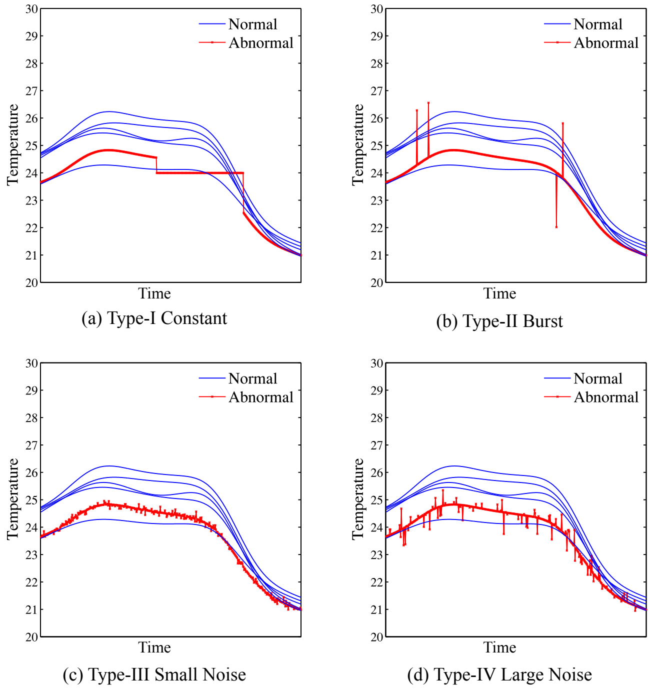
As most abnormal events (either random faults or cyber attacks) tamper with a victim node for a long period, this node is often found to exhibit a long-term abnormal pattern in terms of the sensed measurement or network traffic; for example, an exhausted node will produce measurements with a large variance due to low battery voltage [5] and, consequently, its measurements behave very noisily for a certain period of time. Second, a node that suffers from a calibration error may continuously produce unusually large or small constant measurements. In the physical layer, a deceptive jammer may transmit a random signal or constant stream of bytes into the network [6], [7] in which case a long-term abnormal pattern will occur in the network traffic of every jammed node.

However, existing anomaly detection techniques, as detailed in Section 2, usually operate in a point-based manner that handles each observation individually. For a long-term anomaly, they are not directly applicable as further analysis has to be conducted to make a final decision even if all the observations have been handled separately. What is worse, a long-term anomaly does not necessarily mean that every observation is distinctly abnormal and, sometimes, none of the observations is abnormal by itself but when they occur together as a data segment, it is abnormal and, conceptually, identical to the collective anomaly defined in [8]. Therefore, we have to exploit innovative techniques for specifically addressing long-term anomalies.

In terms of long-term anomalies, we suppose that techniques working in a segment-based manner will outperform conventional point-based techniques. First, making a decision with a data segment that contains multiple observations contaminated by an abnormal event is often easier. Second, it is possible to reduce computational and communication costs by eliminating the information redundancy existing in the data segments. Formally, a collection of continuous-time observations is defined as a data segment whether it is of sensed measurements or network traffic

- *M. Xie and J. Hu are with the School of Engineering and Information Technology, University of New South Wales, Australian Defence Force Academy, Canberra, ACT 2600, Australia. E-mail: clifford1984621@gmail.com, j.hu@adfa.edu.au.*
- *S. Guo is with the Performance Evaluation Lab, School of Computer Science and Engineering, The University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamastu City, Fukushima 965-8580, Japan. E-mail: sguo@u-aizu.ac.jp.*

Manuscript received 8 Aug. 2013; revised 23 Dec. 2013; accepted 30 Dec. 2013. Date of publication 24 Feb. 2014; date of current version 9 Jan. 2015.
Recommended for acceptance by H. Frey.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPDS.2014.2308198



(a) Type-I Constant

(b) Type-II Burst

(c) Type-III Small Noise

(d) Type-IV Large Noise

Fig. 1. Anomaly types.

and, in the rest of this paper, unless stated otherwise, a data segment in which multiple contaminated observations are involved is referred to as an 'anomaly'.

According to the literature, the following are the four most common types of anomaly.

- *Type-I Constant*. Some successive observations in the data segment are constant.
- *Type-II Burst*. A few observations in the data segment are extremely larger or smaller than usual.
- *Type-III Small Noise*. Some observations in the data segment are disturbed by small noise of which the variance may be influenced.
- T*ype-IV Large Noise*. Some observations in the data segment are disturbed by large noise which leads to a significant increase in the variance.

Fig. 1 illustrates these types by using an idealised case in which there is a total of six data segments collected from a cluster of neighbouring nodes, with each represented by a solid line (red- abnormal, blue- normal). Due to the linear relationships existing among spatially proximal sensor nodes [9], [10], the neighbouring measurements should exhibit very similar patterns over the same period of time. Consequently, we are motivated to identify the anomalies by measuring the minimum prediction variance of each data segment with respect to the rest and, if a data segment is distinct in terms of its prediction variance, it is identified as an anomaly.

To exploit the spatial correlations, a prediction variance detector is proposed in this paper. In a cluster, the data segments collected by the member nodes (MNs) during a period of time are considered random variables and, by predicting each variable with the others in turn, the cluster head (CH) can separately obtain a set of prediction variances. Then, an anomaly is detectable through constructing a statistical quantity that follows a chi-squared distribution, where a confidence interval is established as the threshold. Moreover, this statistical quantity is updated in real time to track the dynamics of the measurements. Essentially, the proposed detector depends upon the sample covariance among the variables. If the sample covariance matrix is obtained through collecting all the local data segments centrally, the communication cost will be prohibitively expensive. Instead, each MN is allowed to transmit the compressed difference sequence and sample standard deviation corresponding to its local data segment, and then the sample covariance matrix can be approximately retrieved by taking advantage of the Spearman's rank correlation coefficient. In comparison with conventional raw data segment transmission, retrieving the approximated sample covariance matrix can reduce the communication cost by 80 percent on average. Finally, we evaluate the proposed detector with a wide range of numerical experiments for which the data set is generated by the Intel Berkeley Research Lab (IBRL) [11].

The rest of this paper is organised as follows: Section 2 introduces the related work; derivation of the prediction variance detector is detailed in Section 3; in Section 4, the approach that approximates the sample covariance matrix with the Spearman's rank correlation coefficient is introduced; the numerical experiments and evaluation are discussed in Section 5; and, finally, Section 6 provides a summary of this paper. Moreover, in the supplementary file, Section 1 presents the network model and Section 2 summarises all the algorithms as well as the full scheme.

2 Related Works

In WSNs, the anomaly detection techniques often make decisions by analysing the sensed measurements and/or network-related information which can be roughly classified as statistical [12], [13], [14], [15], support vector machine (SVM) [16] and cluster analysis [17], [18]. Furthermore, a statistical technique can either be parametric or nonparametric, according to whether the underlying distribution is known (assumed) a priori or not respectively. Liu et al. proposed a parametric technique by assuming that, as the measurements collected from the neighbourhood of a monitor node follows a multivariate normal distribution, the Mahalanobis squared distance follows a chi-squared distribution. The monitor node reports a measurement as abnormal if it falls outside the confidence interval constructed with the probability density function (PDF). On the contrary, a nonparametric techniques is able to estimate the PDF without any prior-knowledge, whereby an observation is identified as abnormal if its probability (referring to the estimated PDF) is smaller than a threshold. The two typical categories of the nonparametric techniques are histogram [13], [14] and kernel density estimation (KDE) [15]. The techniques based on SVM and clustering analysis exploit the principle of classification similarly, in which a data set is partitioned into single or multiple maximally dense regions and an observation falling outside them is an anomaly.

In addition, anomaly detection is closely related to fault detection [5], [19], [20], [21]. Sharma et al. [5] divided the fault detection techniques into four categories, i.e., rule-based, estimation, time-series analysis, and learning-based. In the technique proposed by Chen et al. [19], a measurement is considered a fault if it deviates significantly from its previous measurements and more than half of the neighbouring measurement. Ding et al. [20] attempted to detect faults with a lower communication cost in which a measurement is compared only with the median of its neighbouring measurements. Guo et al. [21] implemented fault detection by analysing a type of network-related information, received signal strength (RSS), rather than sensed measurements. It ranks the sensor nodes in terms of the RSS and their respective distances to the event, and a fault is raised when a significant mismatch between the detected sequence and the estimated sequence occurs. Since anomalies and faults can be detected using similar techniques, in the remaining text, a fault is considered as the same as an anomaly.

3 Prediction Variance Detector

In this section, we will present the detailed derivation of a prediction variance detector. First, the concept of prediction variance is detailed, while a simplified form is derived to facilitate its implementation in practice. Second, it shows that the prediction variance can be constructed as a statistical quantity that follows a chi-squared distribution based on Cochran's theorem, which enables to identify an anomaly with the concept of interval estimation and, therefore, leads to the detector.

3.1 Prediction Variance and Its Simplified Form

Although there are no explicit evidences demonstrating that spatial correlations exist in all the types of the data collected from a cluster of spatially proximal sensor nodes, it has been widely recognised that most types of sensed measurements are spatially correlated [9], [10]; for example, the climate data such as temperature, humidity, atmospheric pressure and wind obtained in a close neighbourhood by a cluster of sensor nodes will appear to be very similar. Therefore, in this paper, we focus mainly on those spatially correlated sensed measurements. Table 'Notation list' summarises the notations used in this section, which is enclosed in the supplemental file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2014.2308198.

Given a cluster of sensor nodes, their measurements during a period of time are considered as a set of real-valued random variables $Z = \{X_1, X_2, \ldots, X_m\}$, where the subscripts denote the indices of the sensor nodes and m the total number. Any variable $X \in Z$ can be estimated by a linear combination of the remaining variables $\hat{Z} = Z - \{X\}$, i.e.,

$$X = \overline{X} + \varepsilon = \sum_{i=1, X_i \in \hat{Z}}^{m-1} w_i X_i + \varepsilon, \tag{1}$$

where $\overline{X}$ is the estimator, w the weight and ε the estimation error.

According to the ordinary kriging [22], a constraint committed to this estimator is unbiasedness, i.e.,

$$E(\varepsilon) = 0. \tag{2}$$

Supposing that $EX_1 = EX_2 \cdots = EX_m = \mu_X$ and $\mu_X \neq 0$, this constraint yields that

$$\sum_{i=1}^{m-1} w_i = 1 \tag{3}$$

due to

$$E(\varepsilon) = \left(1 - \sum_{i=1}^{m-1} w_i\right)\mu_X = 0. \tag{4}$$

Then, the weights $W = [w_1 \; w_2 \cdots w_{m-1}]^T$ in the form of a vector can be solved by minimising the variance of the estimation error, i.e.,

$$W = \arg\min_W \; Var(\varepsilon). \tag{5}$$

Suppose that the sample covariance matrix is denoted by

$$Q = \begin{bmatrix} c_{X_1X_1} & \cdots & c_{X_1X_m} \\ \vdots & \ddots & \vdots \\ c_{X_mX_1} & \cdots & c_{X_mX_m} \end{bmatrix},$$

where c stands for the sample covariance between two variables. By removing the row and column from Q in which X participates, we obtain the submatrix A. Furthermore, we obtain a subvector B by eliminating c_{XX} from the column of Q in which X participates and write c_{XX} as C for short. Therefore, we have

$$\begin{aligned} Var(\varepsilon) &= Var\left([X_1 \; X_2 \; \cdots \; X_{m-1} \; X]^T \begin{bmatrix} W \\ -1 \end{bmatrix}\right) \\ &= \begin{bmatrix} W \\ -1 \end{bmatrix}^T Var\left([X_1 \; X_2 \; \cdots \; X_{m-1} \; X]^T\right) \begin{bmatrix} W \\ -1 \end{bmatrix} \\ &= \begin{bmatrix} W \\ -1 \end{bmatrix}^T \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \begin{bmatrix} W \\ -1 \end{bmatrix} \\ &= W^T A W - B^T W - W^T B + C. \end{aligned} \tag{6}$$

Formally, the optimisation problem is expressed by

$$\begin{aligned} &\underset{W}{\text{minimise}} \;\; W^T A W - B^T W - W^T B + C \\ &\text{subject to } \mathbf{1}^T \cdot W = 1, \end{aligned} \tag{7}$$

which can be solved by the method of Lagrange multipliers with the Lagrange function constructed as

$$\Lambda(W, \theta) = W^T A W - B^T W - W^T B + C + 2\theta\left(\mathbf{1}^T \cdot W - 1\right), \tag{8}$$

where θ is a Lagrange multiplier. Next, $\frac{\partial \Lambda}{\partial W} = 0$ produces

$$\begin{aligned} &2W^T A - 2B^T + 2\theta \cdot \mathbf{1}^T = 0 \\ &\rightarrow AW + \mathbf{1} \cdot \theta = B. \end{aligned} \tag{9}$$

Together with the constraint $\mathbf{1}^T \cdot W = 1$, they give

$$\begin{bmatrix} A & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} W \\ \theta \end{bmatrix} = \begin{bmatrix} B \\ 1 \end{bmatrix}. \tag{10}$$

Therefore, the weights can be expressed by

$$\begin{bmatrix} W \\ \theta \end{bmatrix} = \begin{bmatrix} A & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} B \\ 1 \end{bmatrix}. \tag{11}$$

According to equation (6), we have

$$Var(\varepsilon) = W^T(AW - 2B) + C. \tag{12}$$

Substituting equations (3), (9) and (11) into equation (12),

$$\begin{aligned} Var(\varepsilon) &= C - W^T(\mathbf{1} \cdot \theta + B) \\ &= C - \begin{bmatrix} B \\ 1 \end{bmatrix}^T \begin{bmatrix} W \\ \theta \end{bmatrix} \\ &= C - \begin{bmatrix} B \\ 1 \end{bmatrix}^T \begin{bmatrix} A & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} B \\ 1 \end{bmatrix}. \end{aligned} \tag{13}$$

$Var(\varepsilon)$ is exactly the so-called prediction variance. During each period of time, there is a covariance matrix that retains the relationships among the MNs, according to which the prediction variance of a MN actually reflects its minimal degree of deviation with respect to the others of MNs. In addition, the measurements taken from a small neighbourhood are strongly correlated and often expose quite similar patterns. Thus, the prediction variance is a good indicator of the abnormality of a MN and, if a MN's $Var(\varepsilon)$ deviates greatly from the others, its current data segment should be reported as abnormal.

While obtaining $Var(\varepsilon)$ from Equation (13), as a $m \times m$ matrix needs to be inverted, the second term is a little complicated. It should be noted that its special form allows it to be simplified by the blockwise inversion (BI) algorithm [23] which inverts a matrix according to the relationship

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} a^{-1} + a^{-1}b(d - ca^{-1}b)^{-1}ca^{-1} & -a^{-1}b(d - ca^{-1}b)^{-1} \\ -(d - ca^{-1}b)^{-1}ca^{-1} & (d - ca^{-1}b)^{-1} \end{bmatrix},$$

where a, b, c and d are submatrices, a and d must be square, a is nonsingular and the Schur complement $d - ca^{-1}b$ must be nonsingular. It is easy to validate that $\left[\begin{smallmatrix} A & \mathbf{1} \\ \mathbf{1}^T & 0 \end{smallmatrix}\right]$ meets all the above constraints. Inserting $a = A$, $b = \mathbf{1}$, $c = \mathbf{1}^T$ and $d = 0$ into Equation (13), produces

$$Var(\varepsilon) = C - \beta_0 - \beta_1(\beta_2 - 1)^2, \tag{14}$$

where

$$\begin{cases} \beta_0 = B^T A^{-1} B, \\ \beta_1 = -\left(\mathbf{1}^T \cdot A^{-1} \cdot \mathbf{1}\right)^{-1}. \\ \beta_2 = \mathbf{1} \cdot A^{-1} B \end{cases} \tag{15}$$

Equations (14) and (15) provide a simplified form for the prediction variance, and the dimension of the matrix inversion in equation (13) is reduced from m to $m - 1$ in Equation (15).

3.2 Detector

In a cluster, a detector is only located in the CH and, providing the covariance matrix is available, the prediction variance can be employed as the indicator of abnormality because using the covariance matrix, the CH is able to separately obtain the prediction variance of each MN. Subsequently, an automated algorithm is expected to identify anomalies which has a fundamental challenge: how to select an appropriate threshold.

We make an assumption about the estimation error ε that $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ and the observations of ε are independent, where σ_ε^2 is the variance of ε. At this time, according to Cochran's theorem [24], $Var(\varepsilon)$ can be constructed as a statistical quantity that follows a chi-squared distribution. If we consider $Var(\varepsilon)$ as a variable, say $Y = Var(\varepsilon)$, then

$$(m - 1)\frac{Y}{\sigma_\varepsilon^2} \sim \chi_{m-1}^2. \tag{16}$$

It should be noted that $Var(\varepsilon)$ must be the sample variance (unbiased) of ε and, in fact, this constraint is met by using the sample covariance matrix Q which can be easily verified through the simplest $m = 2$ case. As the true variance σ_ε^2 is not known, we treat it alternatively with

$$E(Y) = E\left(\frac{\sigma_\varepsilon^2}{m - 1}\chi_{m-1}^2\right) = \sigma_\varepsilon^2. \tag{17}$$

Substituting equation (17) into equation (16) yields

$$(m - 1)\frac{Y}{E(Y)} \sim \chi_{m-1}^2, \tag{18}$$

where $E(Y)$ will be replaced by its unbiased estimator $\overline{\mu}_Y$ (the mean value of Y) in practical applications. Given a small probability α ($\alpha \in (0, 0.5)$), e.g., 0.005, the lower and upper confidence bounds for a chi-squared distribution can be acquired by $F_{\chi_{m-1}^2}^{-1}(\alpha)$ and $F_{\chi_{m-1}^2}^{-1}(1 - \alpha)$ respectively, where F denotes a cumulative density function (CDF). A confidence interval can be established as the threshold that distinguishes between the normal and abnormal, in the form

$$\left[F_{\chi_{m-1}^2}^{-1}(\alpha), F_{\chi_{m-1}^2}^{-1}(1 - \alpha)\right].$$

A normal observation of Y should reside in this interval as a result of the high probability $(1 - 2\alpha)$.

The above analysis does not yet take into account the period of time. Since the dynamics of sensed measurements change over time, the variance in the estimation error is correspondingly time-varying, as is its mean value $\overline{\mu}_Y$. Consequently, we have to track $\overline{\mu}_Y$ in real time. Supposing that the period of time is denoted by $t, t \in \mathbb{Z}$ and it is assumed that no anomalies exist in the network at the very beginning, the detector can start with an initial $\overline{\mu}_Y^t$ that reflects a normal pattern. Alternatively, we can also specify an initial $\overline{\mu}_Y^t$ according to the past experience (if known). During the $(t + 1)\text{th}$ period of time, correspondingly, the set of the predication variances obtained by the CH is denoted by Y^{t+1} and

$$Y^{t+1} = \{y_1^{t+1}, y_2^{t+1}, \ldots, y_m^{t+1}\},$$

where y^{t+1} stands for a specific observation of Y^{t+1}.

If Y^{t+1} is contaminated by some abnormal observations, the mean value of Y^{t+1} itself is not reliable. Detecting it directly with $\overline{\mu}_Y^{t+1}$ may unexpectedly result in either a high false positive rate (FPR) or low detection accuracy (ACC). However, $\overline{\mu}_Y^{t+1}$ must be relative to $\overline{\mu}_Y^{t}$ to some extent as the dynamics of sensed measurements often change smoothly. Thus, we neutralise $\overline{\mu}_Y^{t+1}$ with $\overline{\mu}_Y^{t}$ by employing

$$\hat{\mu}_Y^{t+1} = \lambda\overline{\mu}_Y^{t} + (1-\lambda)\overline{\mu}_Y^{t+1} \tag{19}$$

as the estimate of mean value, where $\lambda \in (0,1)$ is a neutralising factor. For relatively fast-changing dynamics, λ should be specified as being smaller than 0.5 or, conversely, lager than 0.5. If y_i^{t+1} $(i = 1, 2, \ldots, m)$ satisfies

$$(m-1)\frac{y_i^{t+1}}{\hat{\mu}_Y^{t+1}} \in \left[F^{-1}_{\chi^2_{m-1}}(\alpha), F^{-1}_{\chi^2_{m-1}}(1-\alpha)\right], \tag{20}$$

the data segment obtained by ith MN is regarded as being normal; otherwise abnormal. Finally, $\overline{\mu}_Y^{t+1}$ is reset as the mean value of Y^{t+1} after removing the abnormal observations.

4 Approximating Sample Covariance Matrix

All a prediction variance detector depends on is the covariance matrix which is not known but has to be estimated. The sample covariance matrix is the simplest unbiased estimator and can be easily acquired by a centralised approach. However, it is well known that in a WSN the communication cost is often several orders of magnitude higher than the computational cost [1]. It will be very expensive for the CH to collect all the local data segments from the MNs and obtain the sample covariance matrix centrally. Thus, we need to explore an alternative estimate for the covariance matrix, especially for communication efficiency.

Distributed computing often helps to reduce the communication cost; for example, Wiesel and Hero [25] proposed a distributed covariance matrix estimator based on a very natural intuition: each node estimates its local covariance with respect to its neighbours and the covariance matrix is finally obtained by aggregating these local estimates. Although its communication cost is low, this estimator is essentially derived from a known graphical model, i.e., the conditional independence graph topology must be known a priori. Similar distributed estimators can be found in [26], [27]. However, as its advantage results from the sparsity of the covariance matrix, which is not available everywhere, such a distributed estimator may be applicable to only typical 'large p small n' problems.

Without requiring any prior knowledge, we have to exploit a nonparametric estimator while meeting the constraint in the communication cost. As the proposed detector is located in the application layer, we count the communication cost in terms of the number of transmitted bytes, with 4 bytes usually required to hold a real-valued sensed measurement (i.e., double data type [28]); for example, supposing that each MN collects equal-sized n measurements during a period of time, given $n = 20$, each MN has to spend $20 \times 4 = 80$ bytes on transmitting its local data segment to the CH in order to obtain the sample covariance matrix using a regular centralised approach.

It has been widely reported that spatially proximal sensed measurements are highly correlated [9], [10], [29] due to the dense deployment of sensor nodes that continuously monitor physical phenomena. The near-perfect linear correlation also implies significant monotonicity which motivates us to replace the Pearson's correlation coefficient r with its rank correlation coefficient, particularly the Spearman's correlation coefficient $\overline{r}$. Then, transmitting the rank sequence corresponding to a data segment will achieve an immediate 75 percent reduction in the communication cost, as 1 byte can exactly accommodate a rank (i.e., an unsigned 8-bit integer data type [28]). However, the size of a rank sequence can be further reduced through differential compression and, according to the numerical experiments, an additional approximate 10 percent saving can be achieved, as detailed in the next two sections.

4.1 Retrieving Sample Covariance Using Spearman's Rank Correlation Coefficient

The Pearson's r between two variables X_i and X_j is defined as

$$r_{X_iX_j} = \frac{c_{X_iX_j}}{\overline{\sigma}_{X_i}\overline{\sigma}_{X_j}}, \tag{21}$$

where $c_{X_iX_j}$ is the sample covariance between X_i and X_j, and $\overline{\sigma}_{X_i}$ and $\overline{\sigma}_{X_j}$ denote the sample standard deviations for X_i and X_j respectively. Conversely, we have

$$c_{X_iX_j} = r_{X_iX_j}\overline{\sigma}_{X_i}\overline{\sigma}_{X_j}. \tag{22}$$

If a reliable substitute for r can be obtained at lower cost, it will enable us to efficiently retrieve the sample covariance without degrading performance.

The Spearman's $\overline{r}$ is known as the special case of the Pearson's r between two ranked variables [30]. However, unlike the Pearson's r, it emphasises the monotonic relationship between two variables as, when X_i is linearly correlated with X_j, it is evidently also monotonic with respect to X_j and, at this time, $\overline{r}$ appears very close to r. Since linear relationships are enforced by grouping spatially proximal sensor nodes together, in practice, it is reasonable to replace r with $\overline{r}$.

Let $S = [s_1\ s_2, \ldots, s_n]$ denote the rank sequence, where $1 \leq s_i \leq n$, $i = 1, 2, \ldots, n$, and tied ranks are permitted. If any tied ranks are found in S, it has to identically reassign the mean rank (the average of their positions in ascending order) to them which yields a reassigned rank sequence $\overline{S}$. Then, the Spearman's $\overline{r}$ can be computed from

$$\overline{r}_{X_iX_j} = r_{\overline{S}_i\overline{S}_j} = \frac{c_{\overline{S}_i\overline{S}_j}}{\overline{\sigma}_{\overline{S}_i}\overline{\sigma}_{\overline{S}_j}}, \tag{23}$$

where $\overline{S}_i$ and $\overline{S}_j$ denote the reassigned rank sequences for X_i and X_j respectively, and $\overline{\sigma}_{\overline{S}_i}$ and $\overline{\sigma}_{\overline{S}_j}$ are their sample standard deviations. When there are no tied ranks, it is simple to obtain them through

$$\overline{r}_{X_iX_j} = 1 - \frac{6\left(S_i - S_j\right)\left(S_i - S_j\right)^T}{n(n^2-1)}, \tag{24}$$

where S_i and S_j are the rank sequences of X_i and X_j respectively. The example given in Table 1 explains how to

TABLE 1
Example: Computing Reassigned Rank Sequences

X_i	0.4	1.3	0.3	0.3	0.3	0.3	0.3	0.1
positions	7	8	2	3	4	5	6	1
S_i	7	8	2	2	2	2	2	1
$\bar{S}_i$	7	8	4	4	4	4	4	1
tied ranks: $(2+3+4+5+6)/5=4$								
X_j	0.4	2.2	1.8	1.3	0.9	0.4	2.7	2.7
positions	1	6	5	4	3	2	7	8
S_j	1	6	5	4	3	1	7	7
$\bar{S}_j$	1.5	6	5	4	3	1.5	7.5	7.5
tide ranks: $(1+2)/2=1.5$ $(7+8)/2=7.5$								

calculate the reassigned rank sequences and, by inserting $\overline{S}_i$ and $\overline{S}_j$ into Equation (23), $\overline{r}_{X_iX_j} = -0.333$.

An exceptional case is that the sample covariance between X_i and any X_j will be 0 if the observations of X_i are constant. Although the Pearson's r and Spearman's $\overline{r}$ are both undefined for this case, we can immediately set $\overline{r}_{X_iX_j}$ as 0 instead. Once the rank correlation coefficients and sample standard deviations are ready, each sample covariance can be approximately retrieved by

$$\overline{c}_{X_iX_j} = \overline{r}_{X_iX_j}\overline{\sigma}_{X_i}\overline{\sigma}_{X_j}, \tag{25}$$

which eventually constructs an approximated sample covariance matrix $\overline{Q}$ by traversing $i, j = 1, 2, \dots, m$.

Moreover, a constant variable X_i will lead to the second exceptional case. Let $\hat{Z}$ denote $Z - \{X_i\}$ and, for any $X \in \hat{Z}$, its submatrix A (defined in the last section) is not invertible as A is singular. A badly scaled (almost singular) A may also occur when n is smaller than or comparable to m because of 'large p small n'. While dealing with such a (almost) singular A, the Moore-Penrose pseudo-inverse $A^\dagger$ is substituted for A^{-1} to ensure the numerical stability [31]. Firstly, A is factorised by the singular value decomposition (SVD) so that

$$A = U\sum V^T,$$

where U and V^T are $m \times m$ real unitary matrices and $\sum$ a $m \times m$ diagonal matrix with only nonnegative entries. The Moore-Penrose pseudo-inverse is defined as

$$A^\dagger = V{\sum}^\dagger U^T,$$

where $\sum^\dagger$ is obtained from $\sum$ by substituting the reciprocal for each positive diagonal entry. Then, as it is easy to verify that

$$AA^\dagger = U\sum V^T V{\sum}^\dagger U^T = U\sum{\sum}^\dagger U^T,$$
$$AA^\dagger A = U\sum{\sum}^\dagger U^T U\sum V^T = U\sum V^T = A.$$

Based on the Spearman's rank correlation coefficient, each MN can only transmit its sample standard deviation and rank sequence to the CH which immediately achieves an approximate 75 percent reduction in the communication cost.

4.2 Transmitting Compressed Difference Sequence

By observing the rank sequences that the MNs transmit, we find that information redundancy is widespread as a result of the tied ranks and intrinsic monotonicity of the sensed measurements. Based on the difference sequence, it is able to completely manifest the redundancy, enabling the number of bytes to be further reduced through differential compression [32]. Given a rank sequence S, we define its (first order) difference sequence $\hat{S}$ as

$$\hat{S} = [s_1|s_2 - s_1 \cdots s_n - s_{n-1}] = \left[s_1|\Delta^1 \cdots \Delta^{n-1}\right]. \tag{26}$$

For the same example cited in the last section, we have

$$\begin{cases} \hat{S}_i = [7|1\ -6\ 0\ 0\ 0\ 0\ -1], \\ \hat{S}_j = [1|5\ -1\ -1\ -1\ -2\ 6\ 0], \end{cases}$$

regarding X_i and X_j respectively. Intuitively, the repeated '0' in $\hat{S}_i$ and the repeated '-1' in $\hat{S}_j$ should be compressed into a shorter sub-sequence before being transmitted.

In order to ensure that 1 byte can accommodate any element of a difference sequence $\hat{S}$, we have to handle negative values in a special way. As, normally, 1 byte is only allowed to hold a non-negative integer in the range from 0 to 255 ($2^8 - 1$), we restrict the largest data segment to 127 (i.e., $n \leq 2^7 - 1$) so that half the space can be reserved for settling negative values. At this time, any element of S ranges from 1 to 127 and that of $\hat{S}$ from -126 to 126. Regarding any $\Delta^i \in \hat{S}$, $i = 1, 2, \dots, n-1$ and $\Delta^i \leq 0$, we can force it to be positive through

$$\Delta^i = \left|\Delta^i\right| + 2^7 - 1 = \left|\Delta^i\right| + 127. \tag{27}$$

Thus, the value range becomes

$$\left\{ \underbrace{1, 2, \dots, 126}_{positive\ \ values: 1,2,\dots,126} \middle| \underbrace{127, 128, \dots, 254}_{nonpositive\ \ values: 0,-1,\dots,-126} \right\}$$

and it should be noted that 0 is no longer used. Following the above example, there are

$$\begin{cases} \hat{S}_i = [7|1\ 134\ 127\ 127\ 127\ 127\ 128], \\ \hat{S}_j = [1|\ 5\ 128\ 128\ 128\ 129\ 6\ 127]. \end{cases}$$

If a value is found to be successively repeated more than twice in $\hat{S}$, these identical values will be summarised by a sub-sequence $[\Delta\ 0\ \tau]$, where Δ is its repeated value, 0 an indicator of its repetition and $\tau \geq 3$ the number of times repeated. Accordingly, as $\hat{S}_i$ and $\hat{S}_j$ will be compressed into

$$\begin{cases} \hat{S}_i = [7|\ 1\ 134\ 127\ 0\ 4\ 128], \\ \hat{S}_j = [1|\ 5\ 128\ 0\ 3\ 129\ 6\ 127], \end{cases}$$

1 byte can be saved for transmitting $\hat{S}_i$. Although no instant benefit is brought to $\hat{S}_j$, the compressor is at least lossless and will not result in a higher communication cost, with the compressed difference sequence denoted by $\hat{S}^*$.

When a $\hat{S}^*$ arrives at the CH, the CH checks whether any sub-sequence exists in the form of $[\Delta\ 0\ \tau]$ and, by expanding the Δ value τ times, can obtain the raw difference sequence $\hat{S}$. Secondly, it replaces any $\Delta^i \in \hat{S}$ $(i = 1, 2, \dots, n-1)$ and $\Delta^i \geq 127$ with $\Delta^i = 127 - \Delta^i$, and retrieves the raw rank sequence S by

$$S = \left[s_1|s_2 = s_1 + \Delta^1 \cdots s_l = s_{l-1} + \Delta^{n-1}\right]. \tag{28}$$

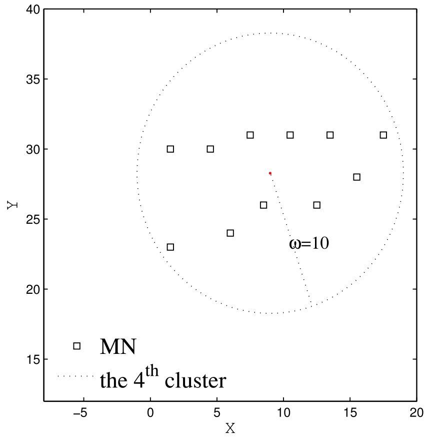


Fig. 2. A cluster in the network.

Each MN reports its $\hat{S}^*$ and local sample standard deviation $\overline{\sigma}$ to the CH which, after working out all the raw rank sequences, acquires $\overline{Q}$ by using either Equation (21) or Equation (22).

5 Numerical Experiments and Evaluation

5.1 Experimental Data Set

The network deployed at the IBRL collected humidity, temperature, light and voltage measurements once every half a minute between 28/02/2004 and 15/04/2004. For the experimental data sets, we select a measurement every 1 minute between 00:00 01/03/2004 and 23:59 02/03/2004, with the very few missing measurements fixed using interpolations, to obtain a total of 2880 data points. Without loss of generality, we use only the temperature and humidity measurements for the following numerical experiments.

Excluding the two damaged nodes, there are 52 active nodes in this network which, as previously mentioned, can be partitioned into seven clusters using fixed-width clustering. All the numerical experiments are conducted using the fourth cluster which covers 11 nodes, as shown in Fig. 2, from which we assume a CH is selected. The two data sets collected from this cluster are denoted by D_1 and D_2 for temperature and humidity respectively.

Supposing that the size of a data segment is given by n, each data set is actually comprised of $\ell \times m$ data segments, where $\ell = \lceil \frac{2880}{n} \rceil$ and $m = 11$ and, during each period of time, there are m data segments corresponding to the MNs. First, we label the anomalies in the raw data set by performing the proposed detector with a manually adjusted λ, with an optimal λ fixed to each data set if we think the resultant labelling matches the 'ground truth' shown by its plot. The labels computed from the optimal λ are kept in a $\ell \times m$ labelling matrix L_0, where '0' and '1' denote normal and abnormal respectively. However, as evaluating the proposed detector with L_0 is meaningless because the labelling is obtained by itself, we evaluate the performance by artificially injecting some anomalies while considering L_0 synthetically. In particular, the anomaly is only randomly injected into a MN during each period of time and the 4 candidate anomaly types are attempted separately. Table 'Anomaly injector' (enclosed in the supplemental file, available online) summarises these deliberately designed injectors by which the low order moments (i.e., the mean and variance) are only slightly influenced; in other words, an injected anomaly is not significantly distinguishable by its mean and variance. Supposing that the labels of the injected anomalies are represented by a $\ell \times m$ matrix M, where '0' and '1' stand for 'as is' and 'injected' respectively (i.e., normal and abnormal). We finally obtain a labelling matrix L_1 by $L_1 = L_0 + M$ which is used for the performance evaluation.

5.2 Experiments and Performance Metrics

We test $n = 20$ and $n = 40$ for D_1 and D_2 respectively and, by setting their optimal λ as 0.1 and 0.2 (as detained in the previous section) respectively, about 5 percent of each data set contains anomalies and five experiments are conducted. Experiments (1)-(4) test the four anomaly types separately to determine how the detector performs for each type of anomaly. Experiment (5) tests a case for which the anomaly type is randomly chosen which may be closer to a real-life setting. The initial $\overline{\mu}_Y^t$ is estimated by averaging the prediction variance obtained from the first period of time after removing the abnormal observations (if any), which are exactly 0.0002 and 0.006 for D_1 and D_2 respectively and, the confidence level α is invariably equal to 0.0001, such that $F^{-1}_{\chi^2(10)}(0.0001) = 0.889$ and $F^{-1}_{\chi^2(10)}(0.9999) = 35$.

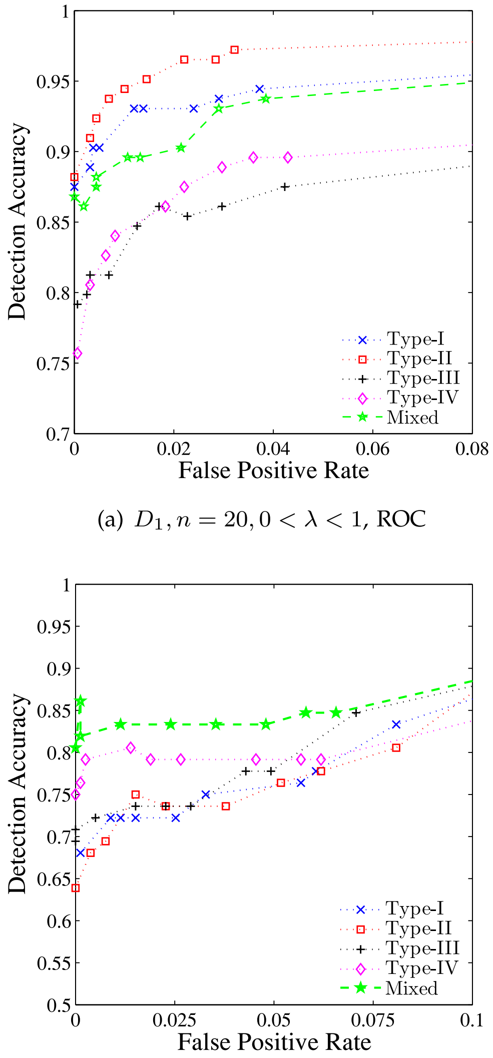
As, if some additional anomalies are injected, the ground truth appearing in the raw data sets will change, L_0 is no longer accurate for labelling and, the increased number of anomalies indicates that a larger λ has to be selected to mitigate the inaccuracy caused by contaminated observations. Therefore, each experiment is repeated with a range of λ (from 0 to 1) and the performance evaluation is concerned mainly with the injected anomalies. The ACC is defined as the rate that the number of injected anomalies can successfully be detected divided by the total number (i.e., ℓ), with a false positive caused if the data segment is neither an injected anomaly nor an anomaly labelled by L_0 but still reported as an anomaly. Given the labelling matrix L produced by the detector, the ACC and FPR can be simply obtained by

$$\begin{cases} ACC = 1 - \frac{number\ of\ 1\ in\ [M-L]}{\ell}, \\ FPR = \frac{number\ of\ 1\ in\ [L-L_1]}{\ell \times m}. \end{cases} \tag{29}$$

Apart from the above two performance metrics, we are very interested in determining by how much the communication cost is reduced through transmitting compressed difference sequences which can be measured using the average saving rate (ASR) acquired by dividing the mean value of the reduced number of transmitted bytes during each period of time by ℓ, i.e.,

$$ASR = 1 - \frac{mean(transmitted\ bytes)}{\ell}. \tag{30}$$

Fig. 3 shows the experimental results for D_1 and D_2 in which the ACCs and FPRs are shown in the form of ROC curves with respect to the values of λ ranging from 0 to 1 and, in particular, the ASRs are presented for every experiment (5) in Fig. 4.



(a) $D_1, n = 20, 0 < \lambda < 1$, ROC

(b) $D_2, n = 40, 0 < \lambda < 1$, ROC

Fig. 3. Experimental results of ROC from D_1 and D_2.

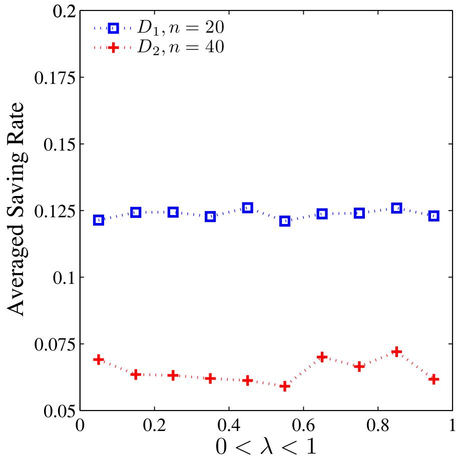


Fig. 4. Experimental results of ASR from D_1 and D_2.

Furthermore, we conduct two experiments for comparison; in the former, the algorithm proposed in [20] is employed as a representative of point-based anomaly detection techniques and, in the latter, it examines the performance degradation caused by approximating a sample covariance matrix. The experimental configurations are same as that of the aforementioned experiment (5), except some parameters specially adjusted. The algorithm is described in Algorithm 6 *point-based median detector* (PMD), where the threshold η ($\eta > 1$) determines the deviation of an observation from the median of its neighbouring observations to be considered as abnormal. After the labels are produced for all the observations using Algorithm 6, they are rearranged to coincide with the data segments and, if more than $0.1 \times n$ abnormal observations are involved in a data segment, it is identified as an anomaly. Based on the metrics given in Equation (29), in the first experiment, η is evaluated from 1 to 2 with a step width of 0.1, which produces two RoC curves, as shown in Fig. 5, for D_1 and D_2 respectively. Then, in the second experiment, the sample covariance matrices are obtained through the regular centralised approach and their performances are compared with those of the proposed distributed approach, with the results shown in Fig. 6.

5.3 Experimental Results and Evaluation

Regarding the **Type-I** anomalies, it can achieve 92 percent ACC with only a 3 percent FPR for D_1 when $\lambda = 0.85$ and, providing that $\lambda \geq 0.5$, the ACC usually stays above 85 percent with a FPR of less than 5 percent. For D_2, the best performance is a 75 percent ACC and 6percent FPR which appears at $\lambda = 0.85$. For the **Type-II** anomalies, the ACC gradually increases from 80 percent to an average of 90 percent for D_1, where the FPR finally reaches 5 percent whereas, for D_2, the ACCs approximately stabilise at 75percent when the FPRs are smaller than 5 percent. The **Type-III** and **Type-IV** anomalies seem much easier to identify whereby, for both D_1 and D_2, the ACCs can exceed 75 percent with a wider range of λ, namely $\lambda \geq 0.6$ and their the highest FPRs are limited to 5 and 8 percent respectively. Fig. 3 a shows that the ACC rises to 75 percent when $\lambda = 0.1$ and then remains above 80 percent when $\lambda \geq 0.5$, where the peak FPR is 5 percent. For D_2, it produces a similar curve but a worse ACC (a stead 75 percent) and higher peak FPR (8 percent). Overall, D_1 outperforms D_2, mainly because a shorter data segment is more sensitive to injected anomalies than difference between

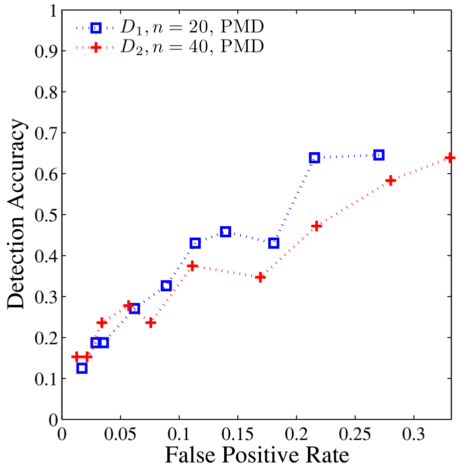
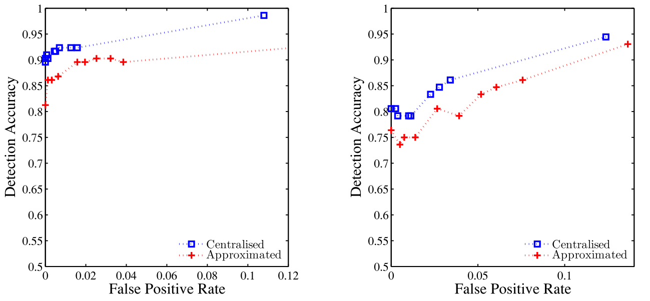


Fig. 5. Experimental results of ROC (PMD) from D_1 and D_2.



(a) $D_1, n = 20$, Mixed, ROC (b) $D_2, n = 40$, Mixed, ROC

Fig. 6. Comparisons between centralised and approximated sample covariance matrices.

temperature and humidity. Finally, Fig. 4 shows that the ASRs are around 12 and 6.5 percent for D_1 and D_2 respectively.

According to Fig. 5, comparing with the 'mixed' shown in Figs. 3 and 4, the proposed scheme outperforms the point-based scheme significantly with 90 and 80 percent ACCs on average for D_1 and D_2 respectively and FPRs lower than 10 percent. However, the best performance of using the point-based scheme is only around 40 percent ACC when the FPRs are smaller than 20 percent. If approximating the sample covariance matrices with the concept of Spearman's rank correlations coefficient, from Fig. 6, it can be observed that there is a slight performance degradation which, in particular, includes a 10 percent reduction in the ACCs on average and an approximate 2 percent increase in the FPRs.

Based on the above observations, several conclusions can be drawn. The proposed scheme is most capable of tackling **Type-III** and **Type-IV** anomalies, but a little weaker in terms of **Type-I** anomalies while **Type-II** anomalies are fairly ambiguous and as sometimes, it is difficult to differentiate them from transient anomalies, we suppose that a slightly lower performance is reasonable when handling **Type-II** anomalies. Nonetheless, the scheme performs satisfactorily with mixed-type anomalies which mostly reflects practical situations. Regarding the selection of λ, a value larger than 0.5 is often preferred when a relatively large number of anomalies are present in a cluster. At the same time, it should be larger for a longer data segment, as confirmed by the optimal values of 0.1 and 0.2 for D_1 and D_2 respectively, as previously mentioned. According to the experimental results, differential compression does not lead to as great an improvement as we expected. We find that repeated elements are widespread in the difference sequences, but only a subsequence covering more than three repeated elements is compressible using the proposed algorithm. Actually, any length of repeated elements can be largely compressed by encoding if it focuses on decreasing the number of transmitted bits. We still assume the repeated elements are summarised by $[\Delta\ 0\ \tau]$, where $\tau \geq 2$. The simplest algorithm is to encode Δ and τ together without counting the indicator symbol '0', which consumes only up to $8 + log_2 n$ bits (definitely less than 2 bytes) and then the reduction in the communication cost is quite considerable. However, this work is outside the scope of this paper as bit-level optimisation is generally a task to be accomplished by the link layer rather than the application layer. The final two experiments demonstrate that, although, the scheme suffers a slight degradation in performance because of approximating the sample covariance matrix, its performance is still within an acceptable level and much better than that of a point-based technique.

6 Conclusions

In this paper, a new segment-based anomaly detection technique for handling long-term anomalies by exploiting the spatial correlation existed among neighbouring sensed measurements, with its detector realised through a trackable parameterised statistical quantity, is proposed. In addition, the sample covariance matrix is approximated according to the concepts of the Spearman's rank correlation coefficient and differential compression such that the computational cost is greatly reduced. The effectiveness and efficiency of this technique is demonstrated using the data set of the IBRL. However, it is highly dependent on the assumption that the data are spatially correlated as this is the only situation in which predictability and information redundancy occur. From such a limitation, several further research question arises: (1) if spatial correlations are not significantly present, it may utilise other metrics to measure the difference between two data segments such as similarity; and, (2) the prediction variance is now derived from a linear estimator but it is entirely possible to realise the same idea with a nonlinear estimator, which may adapt to more generalised cases.

References

[1] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless Sensor Networks: A Survey," *Computer Networks*, vol. 38, no. 4, pp. 393-422, 2002.

[2] D.W. Carman, P.S. Kruus, and B.J. Matt, "Constraints and Approaches for Distributed Sensor Network Security," Technical Report 010, NAI Labs, The Security Research Division Network Assoc., Inc., http://www.cs.umbc.edu/courses/graduate/CMSC691A/Spring04/papers/nailabs_report_00-010_final.pdf, Sep. 2000.

[3] P. Baronti, P. Pillai, V.W.C. Chook, S. Chessa, A. Gotta, and Y. F. Hu, "Wireless Sensor Networks: A Survey on the State of the Art and the 802.15.4 and ZigBee Standards," *Computer Comm.*, vol. 30, no. 7, pp. 1655-1695, http://www.sciencedirect.com/science/article/pii/S0140366406004749, May 2007.

[4] M. Xie, S. Han, B. Tian, and S. Parvin, "Anomaly Detection in Wireless Sensor Networks: A Survey," *J. Network and Computer Applications*, vol. 34, no. 4, pp. 1302-1325, 2011.

[5] A.B. Sharma, L. Golubchik, and R. Govindan, "Sensor Faults: Detection Methods and Prevalence in Real-World Datasets," *ACM Trans. Sensor Networks*, vol. 6, no. 3, pp. 1-39, 2010.

[6] D.R. Raymond and S.F. Midkiff, "Denial-of-Service in Wireless Sensor Networks: Attacks and Defenses," *IEEE Pervasive Computing*, vol. 7, no. 1, pp. 74-81, Jan.-Mar. 2008.

[7] S. Wen, Y. Xiang, and W. Zhou, "A Lightweight Intrusion Alert Fusion System," *Proc. 12th IEEE Int'l Conf. High Performance Computing and Comm. (HPCC)*, pp. 695-700, 2010.

[8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.

[9] M.C. Vuran, Ö.B. Akan, and I.F. Akyildiz, "Spatio-Temporal Correlation: Theory and Applications for Wireless Sensor Networks," *Computer Networks*, vol. 45, no. 3, pp. 245-259, 2004.

[10] A. Jindal and K. Psounis, "Modeling Spatially Correlated Data in Sensor Networks," *ACM Trans. Sensor Networks*, vol. 2, no. 4, pp. 466-499, 2006.

[11] Intel Lab Data, http://db.csail.mit.edu/labdata/labdata.html, 2014.

[12] F. Liu, X. Cheng, and D. Chen, "Insider Attacker Detection in Wireless Sensor Networks," *Proc. INFOCOM '07*, pp. 1937-1945, 2007.

[13] M. Xie, J. Hu, and B. Tian, "Histogram-Based Online Anomaly Detection in Hierarchical Wireless Sensor Networks," *Proc. IEEE 11th Int'l Conf. Trust, Security and Privacy in Computing and Comm. (TrustCom)*, pp. 751-759, 2012.

[14] M. Xie, J. Hu, S. Han, and H.-H. Chen, "Scalable Hypergrid K-NN-Based Online Anomaly Detection in Wireless Sensor Networks," *IEEE Trans. Parallel and Distributed Systems*, vol. 24, no. 8, pp. 1661-1670, Aug. 2013.

[15] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online Outlier Detection in Sensor Data Using Non-Parametric Models," *Proc. 32nd Int'l Conf. Very Large Data Bases*, pp. 187-198, 2006.

[16] S. Rajasegarar, C. Leckie, J.C. Bezdek, and M. Palaniswami, "Centered Hyperspherical and Hyperellipsoidal One-Class Support Vector Machines for Anomaly Detection in Sensor Networks," *IEEE Trans. Information Forensics and Security*, vol. 5, no. 3, pp. 518-533, Sep. 2010.

[17] S. Rajasegarar, C. Leckie, M. Palaniswami, and J.C. Bezdek, "Distributed Anomaly Detection in Wireless Sensor Networks," *Proc. 10th IEEE Singapore Int'l Conf. Comm. Systems (ICCS '06)*, pp. 1-5, 2006.

[18] M. Moshtaghi, C. Leckie, S. Karunasekera, J.C. Bezdek, S. Rajasegarar, and M. Palaniswami, "Incremental Elliptical Boundary Estimation for Anomaly Detection in Wireless Sensor Networks," *Proc. IEEE 11th Int'l Conf. Data Mining (ICDM)*, pp. 467-476, 2011.

[19] J. Chen, S. Kher, and A. Somani, "Distributed Fault Detection of Wireless Sensor Networks," *Proc. Workshop Dependability Issues in Wireless Ad Hoc Networks and Sensor Networks*, pp. 65-72, 2006.

[20] M. Ding, C. Dechang, X. Kai, and C. Xiuzhen, "Localized Fault-Tolerant Event Boundary Detection in Sensor Networks," *Proc. INFOCOM '05*, vol. 2, pp. 902-913, 2005.

[21] S. Guo, Z. Zhong, and T. He, "Find: Faulty Node Detection for Wireless Sensor Networks," *Proc. Seventh ACM Conf. Embedded Networked Sensor Systems*, pp. 253-266, 2009.

[22] R. Webster and M.A. Oliver, *Geostatistics for Environmental Scientists (Statistics in Practice)*. John Wiley & Sons, 2007.

[23] T.H. Cormen, C. Stein, R.L. Rivest, and C.E. Leiserson, *Introduction to Algorithms*. McGraw-Hill Higher Education, 2001.

[24] K. Knight, *Mathematical Statistics*. Chapman and Hall/CRC, 1999.

[25] A. Wiesel and A.O. Hero, "Distributed Covariance Estimation in Gaussian Graphical Models," *IEEE Trans. Signal Processing*, vol. 60, no. 1, pp. 211-220, Jan. 2012.

[26] V. Delouille, R. Neelamani, and R.G. Baraniuk, "Robust Distributed Estimation Using the Embedded Subgraphs Algorithm," *IEEE Trans. Signal Processing*, vol. 54, no. 8, pp. 2998-3010, Aug. 2006.

[27] A. Wiesel and A.O. Hero, "Decomposable Principal Component Analysis," *IEEE Trans. Signal Processing*, vol. 57, no. 11, pp. 4369-4377, Nov. 2009.

[28] D. Gay, P. Levis, D. Culler, and E. Brewer, "NesC 1.3 Language Reference Manual," technical report, Univ of California Berkeley, 2009.

[29] M.C. Vuran and I.F. Akyildiz, "Spatial Correlation-Based Collaborative Medium Access Control in Wireless Sensor Networks," *IEEE/ACM Trans. Networking*, vol. 14, no. 2, pp. 316-329, Apr. 2006.

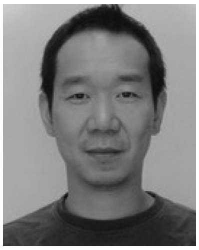
[30] J.L. Myers and A.D. Well, *Research Design and Statistical Analysis*. Lawrence Erlbaum Assoc., 2003.

[31] G. Golub and W. Kahan, "Calculating the Singular Values and Pseudo-Inverse of a Matrix," *J. Soc. for Industrial and Applied Math.: Series B, Numerical Analysis*, vol. 2, no. 2, pp. 205-224, 1965.

[32] Y.W. Nijim, S.D. Stearns, and W.B. Mikhael, "Lossless Compression of Seismic Signals Using Differentiation," *IEEE Trans. Geoscience and Remote Sensing*, vol. 34, no. 1, pp. 52-56, Jan. 1996.

Miao Xie is currently working toward the PhD degree at the School of Engineering and IT, University of New South Wales at the Australian Defence Force Academy (UNSW@ADFA), Canberra, Australia. His research interests include intrusion/anomaly detection in wireless sensor networks, network security, data mining, and forecasting algorithms.

Jiankun Hu received the BE degree from Hunan University, China, and the PhD degree in control engineering from the Harbin Institute of Technology, China, in 1983 and 1993, respectively, and the master's of research in computer science and software engineering from Monash University, Australia, in 2000. He is currently a professor and research director of Cyber Security Lab, School of Engineering and IT, University of New South Wales at the Australian Defence Force Academy (UNSW@ADFA), Canberra, Australia. He has worked in Ruhr University Germany on the prestigious German Alexander von Humboldt Fellowship from 1995 to 1996, research fellow in the Delft University of the Netherlands from 1997 to1998, and research fellow in Melbourne University, Australia, from 1998 to1999. His main research interests include field of cyber security including biometrics security where he has published many papers in high-quality conferences and journals including the *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He has served in the editorial board of up to seven international journals and served as Security Symposium chair of the IEEE flagship conferences of IEEE ICC and IEEE Globecom. He has obtained seven ARC (Australian Research Council) Grants and is now serving at the prestigious Panel of Mathematics, Information and Computing Sciences (MIC), ARC ERA (The Excellence in Research for Australia) Evaluation Committee.



Song Guo (M'02-SM'11) received the PhD degree in computer science from the University of Ottawa, Canada. He is currently a Full Professor at School of Computer Science and Engineering, the University of Aizu, Japan. His research interests are mainly in the areas of protocol design and performance analysis for reliable, energy-efficient, and cost effective communications in wireless networks. He received the Best Paper Awards at ACM Conference on Ubiquitous Information Management and Communication 2014, IEEE Conference on Computational Science and Engineering 2011, and IEEE Conference on High Performance Computing and Communications 2008. He currently serves as Associate Editor of the *IEEE Transactions on Parallel and Distributed Systems*. He is in the editorial boards of *ACM/Springer Wireless Networks*, *Wireless Communications and Mobile Computing*, and others. He has also been in organizing committee of many international conferences, including serving as a General Chair of MobiQuitous 2013. He is a senior member of the IEEE and the ACM.