Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

An efficient lossy link localization approach for wireless sensor networks[*]

Wen-yan CUI[†1,3], Xiang-ru MENG[1], Bin-feng YANG[1], Huan-huan YANG[1,2], Zhi-yuan ZHAO[1]

(*[1]College of Information and Navigation, Air Force Engineering University, Xi'an 710077, China*)

(*[2]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*)

(*[3]PLA of 94543, Jining 272000, China*)

[†]E-mail: cwy_edu@163.com

Received May 10, 2016; Revision accepted Nov. 10, 2016; Crosschecked Apr. 13, 2017

Abstract: Network fault management is crucial for a wireless sensor network (WSN) to maintain a normal running state because faults (e.g., link failures) often occur. The existing lossy link localization (LLL) approach usually infers the most probable failed link set first, and then gives the fault hypothesis set. However, the inferred failed link set contains many possible failures that do not actually occur. That quantity of redundant information in the inferred set can pose a high computational burden on fault hypothesis inference, and consequently decreases the evaluation accuracy and increases the failure localization time. To address the issue, we propose the conditional information entropy based redundancy elimination (CIERE), a redundant lossy link elimination approach, which can eliminate most redundant information while reserving the important information. Specifically, we develop a probabilistically correlated failure model that can accurately reflect the correlation between link failures and model the nondeterministic fault propagation. Through several rounds of mathematical derivations, the LLL problem is transformed to a set-covering problem. A heuristic algorithm is proposed to deduce the failure hypothesis set. We compare the performance of the proposed approach with those of existing LLL methods in simulation and on a real WSN, and validate the efficiency and effectiveness of the proposed approach.

Key words: Lossy link localization; Redundancy eliminating algorithm; Set-covering; Wireless sensor networks (WSNs); Network diagnosis

http://dx.doi.org/10.1631/FITEE.1601247 **CLC number:** TP393

1 Introduction

Wireless sensor networks (WSNs) have a wide range of applications in many fields such as environmental monitoring (Manolov *et al.*, 2004), health assistance (Harris *et al.*, 2016), inventory management (Chipara *et al.*, 2010), home security (Li, 2007), and battlefield situational awareness (Li *et al.*, 2010). Extensive experience reveals that an outage usually occurs in the sensor network as a result of error-prone nodes and lossy links caused by internal losses or environmental interference (Chipara *et al.*, 2010). Compared to wired networks, WSNs do not have efficient monitoring tools since they have a limited bandwidth and suffer from high packet loss rates. Many potential faults that do not actually occur occupy a fairly large proportion of the inferred fault hypothesis set, which may lead to false judgments. Therefore, there has recently been a surge of interest in the design of low-cost and highly accurate WSN failure monitoring tools that can ensure timely detection and fault locations for administrators.

The alarms acquired by a sensor network monitoring module are considered as the external symptoms, and alarms are indicated by faults. According to the symptom set, we can infer the most probable lossy link set and locate the root fault. Due to the bandwidth and energy constraints, an end-to-end measurement is

[*] Project supported by the National Natural Science Foundation of China (Nos. 61401409 and 51577191)

ORCID: Wen-yan CUI, http://orcid.org/0000-0003-3697-5689

used to infer the lossy links in a sensor network. A lossy link localization (LLL) is formulated as a Bayesian inference problem, and a max-product algorithm was proposed to solve it (Zhao and Cai, 2010). Approaching the problem as a huge communication overhead of the sink-based tools, Liu *et al.* (2014) proposed a self-diagnostic approach, which encourages each single sensor to join the fault decision process. To deal with the challenges of bandwidth shortages and the frequently changing routing topologies for monitoring sensor networks, an end-to-end application traffic was used for inferring the performance of the internal network links (Nguyen and Thiran, 2006). Based on this technique, a lossy link diagnostic approach to infer lossy links was proposed using the existing traffic information from the sensor nodes, which reduced the overhead (Zhang *et al.*, 2014). In addition, to address the issue that information collection is independent of root-cause deduction, Gong *et al.* (2015) proposed a directional diagnosis approach where the acquisition of diagnostic information is guided by the fault inference process. However, the methods reported above share the following drawbacks. They use the gathered diagnostic information to infer the fault hypothesis set and locate the lossy links. However, there is too much redundant information, which increases the fault localization time and reduces the localization accuracy. Moreover, the methods consider the lossy links as independent events and ignore the nondeterministic fault propagation phenomenon. Therefore, designing an efficient and highly accurate LLL method is still challenging and open to new approaches.

This paper provides a lightweight solution to locate the lossy links by eliminating most of the redundant information in the raw fault set. We divide the LLL process into three modules, i.e., a lossy link prediction module, a redundancy elimination module, and an LLL module. We first construct a probabilistic correlation based lossy link model to express the relationship among the failures. Then, we model a probabilistic weighted bipartite graph (PWBG), which can represent the nondeterministic causal relationships between lossy links and symptoms. In the redundancy elimination module, conditional information entropy is adopted to calculate the degree of importance of a raw fault. On this basis, we attempt to remove faults that are unrelated or unimportant, and acquire a possible lossy link set. In the LLL module, the LLL problem is formulated and it can be considered as a set-covering problem. It is solved by means of a heuristic algorithm that can be executed correctly and efficiently.

This paper presents the following major contributions:

1. We observe that one failure may be correlated with another due to the dependence of logic or the function of links. Furthermore, we develop a probabilistically correlated failure model, which gives the quantified impact that one failed link has on another one.

2. We construct a PWBG, which can effectively represent the nondeterministic causal relationships between lossy links and symptoms. With PWBG, we can find the most probable lossy link set that involves all the possible lossy links related to the symptoms.

3. We propose a conditional information entropy based algorithm for redundancy elimination. With it, the lossy links that are less likely to occur are removed and a possible lossy link set that involves fewer redundant faults is obtained to realize lossy link filtering.

4. We formulate the LLL problem, and prove that it can be considered as a set-covering problem. Additionally, we propose a heuristic algorithm to solve it efficiently.

2 Related works

There are three kinds of studies that exist related to our work.

2.1 Failure propagation modeling

Many works address the failure propagation modeling problems to research how the local component triggers cascading failures in complex systems or networks, for instance, a propagation model for bank failures (Dias *et al.*, 2015), a cable routing model constructed in the early system design stage to prevent cable failure propagation events (Bossuyt *et al.*, 2016), a new network measure called ‘epidemic survivability’ to characterize a network under epidemic-like failure propagation scenarios (Manzano *et al.*, 2013), and an approach for modeling and quantifying the survivability of telecommunication network systems under fault propagation (Xie *et al.*, 2013). However, these investigations concentrated on analyzing the failure propagation mechanism and attempted to find some methods to control large-scale cascading failures. For integrated fault diagnosis, some works introduced timed failure propagation graphs (TFPGs), which are causal models that describe system behavior in the presence of faults and how observable symptoms map to hidden faults, e.g., an automatic generation approach of TFPGs for fault diagnosis (Priesterjahn *et al.*, 2013), a distributed diagnosis approach using TFPG models for complex systems (Mahadevan *et al.*, 2010), an approach for diagnosing alarm sequences at the system level using TFPG models (Strasser and Sheppard, 2011), and an extended TFPG for fault detection and resolution (Troiano *et al.*, 2015). However, the TFPG models are used mainly for time constrained fault diagnostics and usually determine the mapping. However, in many cases, the mapping relations between symptoms and faults are nondeterministic. In addition, a variety of fault propagation models are available for mapping faults to symptoms. For example, Petri nets (Benveniste *et al.*, 2003), finite state machines (Ntalampiras, 2014), dependency graphs (Urbanics *et al.*, 2014), causal graphs (Nyberg, 2013), and Bayesian nets (Zhang *et al.*, 2010) can all be used to construct fault propagation models. Among them, because the relationships between faults and symptoms can be represented using a perceivable form, a dependency graph based representation is easier to understand. Causal graph based and Bayesian net based models can present a more precise mapping relationship, as they provide an effective representation of causality, which is a concept at the core of fault propagation. However, they have a high computational complexity. Bipartite graphs (Niu *et al.*, 2009) can be considered as a simplification of causal graphs and Bayesian net based models. However, the deterministic bipartite graph cannot show the nondeterministic fault propagation phenomenon. In contrast to these methods, our proposed failure propagation model is a nondeterministic probabilistic fault-symptom mapping model.

2.2 Redundant failure elimination

Most of previous investigations focused on redundancy elimination for network traffic or data to achieve an improvement in link load and network efficiency. Examples include an algorithm for redundancy elimination in network traffic (Xu *et al.*, 2012), a protocol-independent elimination method for data redundancy (Zhang *et al.*, 2014), network redundancy elimination by dynamic buffer allocation (Yang *et al.*, 2014), a network traffic awareness architecture for universal redundancy elimination (Wu *et al.*, 2011), a peer-to-peer (P2P) packet cache router scheme (Yamamoto and Nakao, 2012), and a redundancy-maximizing identification scheme (Zhang *et al.*, 2016) for network-wide traffic redundancy elimination. Compared with wired networks, wireless networks (e.g., opportunistic communications and WSNs) have lower bandwidth due to hardware limitations. Moreover, to ensure a reliable data transmission, wireless networks may generate a considerable redundant traffic load. Therefore, redundancy elimination models are constructed to achieve bandwidth savings for these networks, e.g., a bloom-filter-aided redundancy elimination model for opportunistic communication (Park *et al.*, 2016), a support vector machine (SVM) based redundancy elimination model for WSNs (Patil and Kulkarni, 2013), a lightweight traffic redundancy elimination model for software-defined wireless mesh networks (Kim *et al.*, 2014), and an inter-application redundancy elimination model with compiler-assisted scheduling for WSNs (Gupta *et al.*, 2012). The data aggregation and a false data elimination can cause a computational overhead, and will further cause the battery to drain quickly. To address these problems and ensure data integrity for WSN, an aggregator node selection method and a false data redundancy elimination model were proposed (Sandhya *et al.*, 2015). However, this approach aims to eliminate redundant data/traffic and multiple duplicates, whereas our objective is to eliminate the hidden redundant lossy links. Moreover, our redundant failure elimination requires mapping the failures to the symptoms.

Few works have addressed the redundant failure elimination problem in fault diagnosis. In most cases, the information collection process is independent of root-cause deduction, which results in redundant information. Redundant faults can pose a high communication burden on WSNs (Gong *et al.*, 2015). However, those investigations do not provide a solution for the redundant faults. In Wang *et al.* (2013), a considerable amount of redundant information was involved in the raw fault set and can be compressed. The authors adopted compressive sensing to eliminate the redundant faults. However, they did not consider the correlation between the faults and the symptoms in the redundancy elimination process.

2.3 Failure localization

Localization is an important issue in the application of WSNs. More accurate localization is needed in these networks. A low-cost localization approach was proposed to address this issue (Assaf *et al.*, 2015). It complies with the heterogeneous nature of WSNs to further improve the localization accuracy without causing any additional cost. WSNs and multiple fuzzy logic controllers were combined to be applied in dynamic traffic light management, which does not require powerful hardware and can be easily implemented (Collotta *et al.*, 2015). Regarding the issue of indoor localization, a time of arrival (ToA) approach was used to determine the localization, using the time delta to estimate the distance between two WSN nodes, and the results showed that the localization accuracy could be effectively improved (Haute *et al.*, 2016). A scheme for congestion avoidance, detection, and alleviation (CADA) in WSNs was proposed to decrease excessive battery consumption caused by the congestion of WSNs (Fang *et al.*, 2010). However, all these methods concentrate on the localization of nodes in WSNs and try to obtain higher localization accuracy. We focus mainly on the LLL problem, and our objective is to find the failed link rather than the specific location of nodes.

The LLL in WSNs is usually achieved by monitoring the traffic load of the link or the topology of the network, and can be categorized into active monitoring tools (Rajasegarar *et al.*, 2008; Tang *et al.*, 2008; He *et al.*, 2011; Miao *et al.*, 2011; Benhamida *et al.*, 2014) and passive monitoring tools (Yang *et al.*, 2010; Wang *et al.*, 2011; Haddad *et al.*, 2013; Ali *et al.*, 2014; Ma and Zhang, 2014). With active monitoring methods, end systems send probing packets to each other to measure the delay, loss rate, and path jitter. A network administrator detects and locates the lossy link according to the measured results of the routing path performance. However, because the sensor networks have low bandwidth characters and lose packets easily, they support much less probing traffic, especially when the link is congested. Compared with active tools, the passive monitoring approaches are free of injecting additional probing traffic into the network. They can reflect network performance for the end user's experience more accurately. An exact reasoning algorithm which always outputs an optimal solution was proposed. It is the optimal algorithm for computing the problem of mapping faults to symptoms. However, the computational complexity of the exact reasoning algorithm is exponential (Shen, 2012). Couillet and Hachem (2011) gave a formal description of local failure localization in WSNs. However, they aimed at the local node failure rather than the link failure in the whole network. An iterative belief-updating algorithm which was an approximate reasoning algorithm was proposed. It can reduce computational complexity and obtain a nearly optimal solution for computing the fault localization problem (Tang *et al.*, 2009). However, all these methods adopted the raw fault set, which involves considerable redundant information to locate the lossy links. Moreover, most existing LLL works have been based on a noisy-OR hypothesis (Zhang *et al.*, 2010), which takes the lossy links as independent events and assumes that multiple lossy links do not usually happen simultaneously. In contrast to these methods, our proposed method focuses on the localization of links and does not make assumptions about the independence of the links. In addition, before LLL, we first eliminate most redundant failures, which effectively decrease the complexity.

3 Preliminaries

In this section, we first build the sensor network model, and then construct an LLL infrastructure and define our problems. Finally, the notations that are frequently used are given.

3.1 Network model

We model the sensor network as static with a single sink, and analyze the sensor network in a time window W of length T. Supposing that the sensor network routes data to the sink with a tree topology (Fig. 1), which can be constructed according to methods similar to that proposed by Woo *et al.* (2003), empirical research shows that the routing tree topology changes frequently. Therefore, we define a series of small time slots with equal length, and assume that the routing tree topology remains unchanged in each time slot.

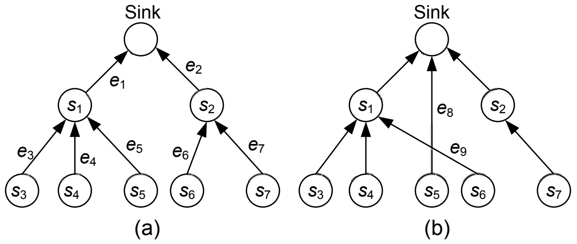
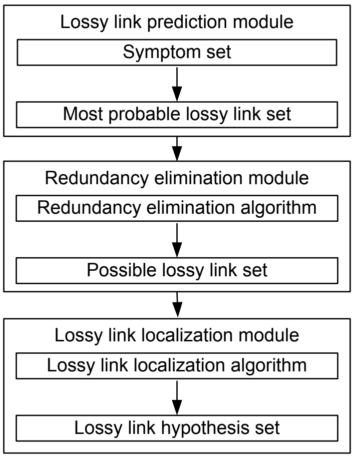


Fig. 1 Routing topologies (a) and (b) for a sensor network in a given period of time

3.2 Lossy link localization infrastructure

Our localization approach can be divided into three successive modules (Fig. 2). We first infer the most probable lossy link set according to the symptom set. Then, we use a redundancy elimination algorithm to filter the most probable lossy link set to acquire the possible lossy link set. Finally, we locate the root lossy links according to the LLL algorithm.


Lossy link prediction module

Symptom set

Most probable lossy link set

Redundancy elimination module

Redundancy elimination algorithm

Possible lossy link set

Lossy link localization module

Lossy link localization algorithm

Lossy link hypothesis set

Fig. 2 Lossy link localization infrastructure

With the lossy link prediction module, a network administrator acquires network alarms and regards them as symptoms which are brought about by the lossy links. In this module, the lossy link propagation phenomenon is modeled. Moreover, the most probable lossy link set is deduced according to the symptoms.

With the redundancy elimination module, most of the lossy links that actually do not occur are eliminated. We call the lossy links that do not actually occur the 'redundant lossy links'. Redundant lossy links increase the computational time and complexity, and reduce the LLL accuracy. In this module, we can acquire the possible lossy link set that involves minor redundant lossy links. Reserving the real lossy links and eliminating the redundant lossy links are the key problems in the redundancy elimination algorithm.

With the LLL module, we can deduce the root lossy link according to the LLL algorithm. In this module, what we should do is to find the lossy link hypothesis set that can best explain the symptoms observed.

Next, we illustrate the main content and define our problems:

Step 1: modeling lossy link propagation. When a sensor network link fails, owing to the dependency relationships of structure and function between the links, this may lead to lossy links in the correlative links. Then, a lossy link propagation phenomenon will occur, and each lossy link may produce many alarms (symptoms). In this work, we adopt a PWBG to model the lossy link propagation phenomenon, which can effectively reduce the computational complexity. With the PWBG model, we can gain the most probable lossy link set.

Problem 1 (Lossy link propagation) Given a WSN G, a set of symptoms S, and a set of lossy links F, the objective is to construct a propagation model, so that all symptoms can be mapped to the most probable lossy link set.

Step 2: eliminating redundant lossy links. The objective of redundancy elimination is to eliminate as many redundant lossy links as possible in the most probable lossy link set. The lossy links in the most probable lossy link set usually include lossy links that do not actually occur, and these links occupy a quite large percentage. With redundancy elimination, we can acquire the possible lossy link set with a low redundancy rate. In this work, we adopt a conditional information entropy based method to eliminate the redundant lossy links.

Problem 2 (Redundancy elimination) Given the most probable lossy link set, the objective is to acquire the possible lossy link set, in which the lossy links that do not occur are eliminated as much as possible, so that LLL can be more exact.

Step 3: Locating the lossy links. Once we have achieved the possible lossy link set, the objective in the next stage is to obtain the lossy link hypothesis set. We can achieve the target according to an LLL algorithm. We formalize the LLL problem. Because the problem is NP-hard (Gong *et al*., 2015), we present a greedy heuristic algorithm to solve it.

Problem 3 (Lossy link localization) Given the possible lossy link set, the objective is to acquire the lossy link hypothesis set, so that we can acquire the root lossy links and locate the lossy links.

Table 1 lists the frequently used notations.

Table 1 Notations

Notation	Definition
F	Set of possible lossy links
P_F	Set of prior probabilities of F
S	Set of possible symptoms
$E(F\times S)$	Set of directed edges where F causes S
$\boldsymbol{P}_{F\times S}$	Correlation matrix
p_{ij}	Probability that f_i causes s_j
r_{PSRLG}	Probabilistically shared risk link group about event r
$p(f_i)$	Failure probability of link l_i
RR	Redundancy rate
$F_{\max}$	Set of most probable lossy links
F_{S}	Set of possible lossy links
KES	Knowledge expression system that involves lossy links and symptoms
U, A	Universe, set of lossy links and symptoms
V, f	Attribute set, information function
$\eta(F_{\mathrm{S}})$	Real lossy link coverage rate
$R(F_{\mathrm{S}})$	Lossy link redundancy rate
S_{O}	Set of observed symptoms
OR	Observability ratio
LR(s)	Loss rate
SSR(s)	False positive rate of symptoms

4 Lossy link model and prediction model

In this section, we first model the lossy link based on probabilistic correlation. Then, we present schemes to solve the lossy link propagation-modeling problem and acquire the most probable lossy link set according to the symptoms.

4.1 Probabilistic correlation-based lossy link model

To the best of our knowledge, most existing LLL works are based on a noisy-OR hypothesis (Zhang *et al*., 2010), taking lossy links as independent events and assuming that multiple lossy links usually do not occur simultaneously. However, because sensor networks change their routing topologies quite frequently, lossy links are not really independent of each other. For example, a sensor network has two routing topologies during a given period of time as shown in Figs. 1a and 1b, respectively. In Fig. 1a, when sensor node s_1 fails, sensor nodes s_3, s_4, and s_5 cannot route data to the sink sensor node, and nodes s_3, s_4, and s_5 are correlated with each other. A shared risk link group (SRLG) can be used to represent the group of links sharing the same risk. When one fails, the other nodes in the same group fail simultaneously. However, as shown in Fig. 1b, when sensor node s_1 fails, nodes s_3, s_4, and s_6 disconnect from the sink sensor node, while node s_5 still works well. Nodes s_3, s_4, and s_5 are not absolutely correlated; i.e., there exists a correlation between lossy links with a probability. In considering this problem, we define a probabilistically shared risk link group (PSRLG) model, which can express relations between lossy links with a probabilistic correlation.

Definition 1 (Probabilistically shared risk link group) Let R be the set of SRLG. When event $r\in R$ occurs, the links that fail with a nonzero probability construct a PSRLG about event R, as

$$r_{\mathrm{PSRLG}}=\{l_i\in L:p_r(l_i)\neq 0\},\tag{1}$$

where L is the set of links, and $p_r(l_i)$ is the probability that l_i breaks down when SRLG event r occurs.

Let p_r be the probability that event r occurs. When there is lossy link correlation between links l_i and l_j, the probabilities that they break down are $p(f_i)$ and $p(f_j)$, respectively:

$$p(f_i)=p_r\cdot p_r(l_i),\tag{2}$$

$$p(f_j)=p_r\cdot p_r(l_j).\tag{3}$$

According to Eqs. (2) and (3), we can derive that only when $p_r\neq 0$, are $p(f_i)$ and $p(f_j)$ nonzero. Namely, the failure probabilities of links l_i and l_j are determined by event r. Next, we will solve the LLL problem with the PSRLG model.

4.2 Lossy link propagation model

In a sensor network, when a failure (such as a lossy link) occurs, some symptoms that reflect the degradation in network performance occur. The symptom is a reflection of inherent failures. However, failures are covered by massive symptoms, which make the failures invisible. Because there exist dependency relations of structure and function between network elements, when one network element fails, it may cause failure of the related elements, which means that a lossy link can spread in the network and a lossy link propagation phenomenon may occur. Therefore, to locate the lossy links accurately and quickly, we must take into account the lossy link propagation phenomenon in the sensor network LLL.

In general, the lossy link propagation model (LLPM) can be realized by modeling the lossy link state of a network element, the relation between symptoms, and the relation between the lossy link and the symptom. Causal graph and Bayesian net based models are traditional LLPMs, which can present an effective representation of causality. However, they have a high computational complexity, which makes them impractical in the LLL problem. A bipartite lossy link propagation model (BLLPM) can be considered as a simplification of the causal graph and Bayesian net based models. Because BLLPM preserves the modeling ability and greatly reduces computational complexity, it is widely used in the LLL problem.

BLLPM can model the causality between the lossy links and the symptoms by adopting a bipartite graph. Lossy link vertexes and symptom vertexes constitute two kinds of disjoint vertex sets of BLLPM. The causality between vertexes is denoted by a directed weighted edge. If the edge weight is 0 or 1, BLLPM is the deterministic model. For a deterministic BLLPM, when one lossy link occurs, all the symptoms related with it occur. However, due to the imprecise threshold and the failed alarm, an event where lossy links cause symptoms is a probability event. Therefore, we should construct a reasonable LLPM to reveal the nondeterministic causal relationships. We adopt a nondeterministic BLLPM, namely, a PWBG, which is a probabilistic LLPM.

A PWBG can be characterized by a five-element vector: **PWBG**=(F, P_F, S, $E(F\times S)$, $\boldsymbol{P}_{F\times S}$). The details are as follows:

$F=\{f_1, f_2, \ldots, f_m\}$ is the set of possible lossy links;

$P_F=\{p(f_1), p(f_2), \ldots, p(f_m)\}$ is the set of the prior probabilities of F;

$S=\{S_1, S_2, \ldots, S_n\}$ is the set of possible symptoms;

$E(F\times S)$ is the set of directed edges where F causes S;

$\boldsymbol{P}_{F\times S}=(p_{ij})_{m\times n}$ is the correlation matrix, where $p_{ij}=p(s_j|f_i)\in[0, 1]$ is the probability that f_i causes s_j.

Fig. 3 shows an example of a PWBG that consists of three lossy links and four symptoms, where $F=\{f_1, f_2, f_3\}$ is the lossy link set, $S=\{s_1, s_2, s_3, s_4\}$ is the symptom set, and $P_F=\{0.004, 0.006, 0.005\}$. We can then acquire the conditional probability matrix

$$\boldsymbol{P}_{F\times S}=\begin{pmatrix} 0.5 & 0.2 & 0.8 & 0.0 \\ 0.0 & 0.7 & 0.5 & 0.7 \\ 0.4 & 0.0 & 0.0 & 0.5 \end{pmatrix}.$$

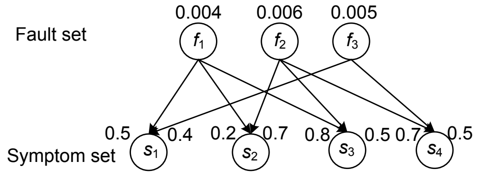


Fig. 3 A probabilistic weighted bipartite graph

With a PWBG, we can find all the possible lossy links related to the symptoms, and acquire the most probable lossy link set. The lossy links in the set are always more than real lossy links. It means that most lossy links in the set do not occur and are acturally redundancies. Therefore, we should screen all the candidate lossy links in the set, and filter the lossy links that are less likely to occur to acquire the possible lossy link set that involves fewer elements. We remove as many redundant lossy links as possible, which can help us realize a more accurate LLL.

We define the redundancy rate (RR) to represent the proportion of redundant lossy links in the most probable lossy link set, formulated as follows:

$$\mathrm{RR}=\frac{|F_{\mathrm{R}}|}{|F_{\max}|}=1-\frac{|F|}{|F_{\max}|}, \tag{4}$$

where $F_{\max}$ is the most probable lossy link set, F_{R} is the set constituted by redundant lossy links, and F is the possible lossy link set.

Redundant lossy links greatly affect the precision of LLL. Next, we will provide an efficient method to eliminate the redundant lossy links.

5 Redundancy elimination

The objective of the redundancy elimination module is to eliminate as many lossy links that do not actually occur as possible, and to acquire the possible lossy link set with a lower redundancy rate. In this section, we propose an approach to achieve this target.

It is known that attributes describing the knowledge are not equally important in the knowledge base, and some of them are redundant. Redundant attributes increase the complexity of decision analysis, and even result in wrong decision making. Knowledge reduction involves removing the uncorrelated or unimportant attributes while maintaining the classification ability. As mentioned above, there are many redundant lossy links in the most probable lossy link set. Redundancy elimination for the set is to remove as many redundant lossy links as possible and reserve the actual lossy links. Therefore, the redundancy elimination problem can be transformed into a knowledge reduction problem. We define a knowledge expression system that involves lossy links and symptoms for redundancy elimination, which is denoted by a four-element vector as follows:

$$\mathbf{KES}=(U, A, V, f), \tag{5}$$

where U is the universe, A is the non-empty finite set, $V = \bigcup_{a\in A} V_a$ is the attribute set, and f is the information function which satisfies f: $U\times A\rightarrow V$, where A is divided into a condition attribute set C and a decision attribute set D, and $A=C\cup D$, C consists of lossy links, and D consists of symptoms.

To better measure the degree of importance that the lossy links have on the symptoms, and filter the lossy links that are less likely to occur, we introduce the information entropy. Next, we calculate the failure probability with the information entropy.

Let P be a subset of C. We denote the binary relation of P by IND(P), which can be formulated as follows:

$$\mathrm{IND}(P)=\left\{(x,y)\middle|\forall a\in P,\ f(x,a)=f(y,a)\right\}. \tag{6}$$

The divisions of universe U by binary relations IND(P) and IND(D) are U|IND(P)={X_1, X_2, …, X_t} and U|IND(D)={Y_1, Y_2, …, Y_s}, respectively. The probability distributions of subsets P and D in U can be formulated as follows:

$$[X:p]=\begin{bmatrix} X_1 & X_2 & \cdots & X_t \\ p(X_1) & p(X_2) & \cdots & p(X_t)\end{bmatrix}, \tag{7}$$

$$[Y:p]=\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_s \\ p(Y_1) & p(Y_2) & \cdots & p(Y_s)\end{bmatrix}. \tag{8}$$

Then, the conditional information entropy that D is relative to P can be formulated as follows:

$$H(D\,|\,P)=-\sum_{i=1}^{t} p(X_i)\sum_{j=1}^{s} p(Y_j\,|\,X_i)\log_2 p(Y_j\,|\,X_i), \tag{9}$$

where $p(Y_j\,|\,X_i)=\left|Y_j\cap X_i\right|/\left|X_i\right|$, i=1, 2, …, t, j=1, 2, …, s.

Therefore, the measure of degree of importance formula for lossy links can be formulated as

$$\begin{gathered}\mathrm{SGF}(a,P,D)=H(D\,|\,P)-H(D\,|\,P\cup\{a\}),\\ \forall a\in(C-P).\end{gathered} \tag{10}$$

According to the degree of importance under conditional information entropy, we iteratively select lossy links that are important to the symptoms. On this basis, we remove as many lossy links that are unrelated or unimportant as possible, and acquire the possible lossy link set that involves fewer redundant lossy links to realize lossy link filtering. The proposed redundancy elimination algorithm is called 'conditional information entropy based redundancy elimination (CIERE)'.

The details of CIERE are shown in Algorithm 1, which works as follows. The divisions of universe U and the conditional information entropy H(D|C) are calculated first. For each lossy link in set C, the core set that involves lossy links contributing to the symptoms is found; consequently, the part of the redundant lossy links is removed initially. Then, an indicator value Temp to estimate set P is defined. Finally, the lossy link that is important to the symptoms is selected iteratively. For each lossy link in set C, it calculates the degree of importance according to Eq. (10) and selects an attribute with the maximum

Algorithm 1 Conditional information entropy based redundancy elimination

```
Input: the most probable lossy link set $C$, symptom set $D$
Output: reduction set $P$
Initialization: Core$\leftarrow\varnothing$, $P\leftarrow\varnothing$
for all $a\in C$ do
    Calculate $U|\text{IND}(C-\{a\})$, $U|\text{IND}((C-\{a\})\cup D)$, and $H(D|C-\{a\})$
    Find all the lossy links that maximize $H(D|C)$ and let them be the set Core
end for
$P\leftarrow$Core
Calculate $U|\text{IND}(P)$, $U|\text{IND}(P\cup D)$, and $H(D|P)$
Temp$\leftarrow H(D|P)$
for each $a\in(C-P)$ do
    Calculate SGF$(a, P, D)$
    Select attribute $a'$ with the maximum SGF$(a, P, D)$ from set $(C-P)$
    Temp$\leftarrow H(D|P\cup\{a'\})$
    $P\leftarrow P\cup\{a'\}$
end for
```

degree of importance to construct set P.

Lemma 1 The conditional information entropy is monotonic.

Proof Suppose $B\subseteq C$ and $a\in C-B$. First, according to the definition of partial order relation $\preceq$, we can derive $U|\text{IND}(B\cup\{a\})\preceq U|\text{IND}(B)$. For $Y\subseteq U$, according to the approximation relation, the B-upper approximation of set Y satisfies $\left|\overline{Y}_{B\cup\{a\}}\right|\le\left|\overline{Y}_B\right|$, and the B-lower approximation of set Y satisfies $\left|\underline{Y}_{B\cup\{\mathrm{a}\}}\right|\ge\left|\underline{Y}_B\right|$. Therefore, the approximation accuracy $\alpha_B(Y)$ of set Y under the binary relation IND(B) satisfies $\alpha_{B\cup\{a\}}(Y)\ge\alpha_B(Y)$.

As defined above, we have $U|\text{IND}(D)=\{Y_1, Y_2, \ldots, Y_s\}$. Then, according to Choi and Park (2014), we can derive the following formula:

$$
\begin{aligned}
&-\sum_{i=1}^{s}\sum_{X\in U|\text{IND}(B\cup\{a\})} p(X)\cdot p(Y_i|X)\cdot\log_2 p(Y_i|X)\\
&\le-\sum_{i=1}^{s}\sum_{X\in U|\text{IND}(B)} p(X)\cdot p(Y_i|X)\cdot\log_2 p(Y_i|X).
\end{aligned}
\tag{11}
$$

According to the definition of conditional information entropy, we can derive that $H(D|B)\ge H(D|B\cup\{a\})$. Thus, the assumption is verified.

Lemma 2 The CIERE algorithm is computationally efficient.

Proof The CIERE algorithm consists mainly of three steps. The calculation of conditional information entropy $H(D|C)$ can be finished first in $O(|U|)$ (Gong *et al.*, 2015), followed by two rounds of iterative search. There are $|C|$ elements in the first loop execution, so the complexity of this step is $O(|C|)$. In the second loop execution, there are $|C–P|$ elements, and this step has a complexity of $O(|C|)$. Therefore, the total computational complexity of CIERE is $O(|C|^2\times|U|)$, which shows that CIERE can be implemented in polynomial time. Thus, the assumption is verified.

The CIERE algorithm has the advantage of low computational complexity. With it, we can eliminate most of the redundant lossy links and acquire the possible lossy link set, which involves fewer false lossy links and little redundant information. Thus, the goal of lossy link selection can be achieved.

Let the selected lossy link set P be F_S. To evaluate the performance of the CIERE algorithm, we define two metrics $\eta(F_\mathrm{S})$ and $R(F_\mathrm{S})$, where $\eta(F_\mathrm{S})$ is the real lossy link coverage rate and $R(F_\mathrm{S})$ is the redundancy rate:

$$\eta(F_\mathrm{S})=\frac{|F_\mathrm{S}|-|F_\mathrm{RS}|}{|F|}=\frac{|F_\mathrm{CS}|}{|F|},\tag{12}$$

$$R(F_\mathrm{S})=\frac{|F_\mathrm{RS}|}{|F_\mathrm{S}|}=1-\frac{|F_\mathrm{CS}|}{|F_\mathrm{S}|},\tag{13}$$

where F_RS is the redundant lossy link set in F_S, and F_CS is the real lossy link set in F_S.

A larger value of $\eta(F_\mathrm{S})$ will make F_S involve more lossy link information. F_S includes all the lossy link information of F_max when $\eta(F_\mathrm{S})=1$. The lower the value of $R(F_\mathrm{S})$, the fewer the redundant lossy links in F_S.

6 Solving the lossy link localization problem

In this section, we first formulate the LLL problem, and then propose a heuristic algorithm to solve it.

6.1 Formulation of the lossy link localization problem

As described above, the LLL problem is about how to choose the most probable candidate lossy link

$H\in F$, which is to find

$$\arg\max_{H\subseteq F} P(H\mid S). \tag{14}$$

From Bayes' rule, we have

$$\arg\max_{H\subseteq F} P(H\mid S)=\arg\max_{H\subseteq F} P(S\mid H)P(H)/P(S). \tag{15}$$

As $P(S)$ is not related with H, the equivalent maximization problem can be formalized as

$$\arg\max_{H\subseteq F} P(H\mid S)=\arg\max_{H\subseteq F} P(S\mid H)P(H). \tag{16}$$

For simplicity, we define an indicator row vector $\boldsymbol{h}$, where h_i=1 if $f_i\in H$, and h_i=0 otherwise. Then we can derive the following formulae:

$$P(H)=\prod_{i=1}^{m} p(f_i)^{h_i}[1-p(f_i)]^{1-h_i}, \tag{17}$$

$$P(S\mid H)=\prod_{s_j\in S} P(s_j\mid H), \tag{18}$$

$$P(s_j\mid H)=1-\prod_{i=1}^{m}(1-p_{ij})^{h_i}. \tag{19}$$

Now, the LLL problem can be formalized as follows:

$$\arg\max_{H\subseteq F} P(S|H)P(H)= \\ \arg\max_{H\subseteq F}\prod_{s_j\in S}\left[1-\prod_{i=1}^{m}(1-p_{ij})^{h_i}\right]\prod_{i=1}^{m}p(f_i)^{h_i}[1-p(f_i)]^{1-h_i}. \tag{20}$$

For simplicity, we take the logarithm of Eq. (20) and acquire:

$$\arg\max_{H\subseteq F}\sum_{s_j\in S}\ln\left[1-\prod_{i=1}^{m}(1-p_{ij})^{h_i}\right] \\ +\sum_{i=1}^{m}\left\{h_i\ln\frac{p(f_i)}{1-p(f_i)}+\ln[1-p(f_i)]\right\}. \tag{21}$$

We eliminate the constant terms and let $c_i=-\ln\dfrac{p(f_i)}{1-p(f_i)}$. Then we can obtain the following problem:

$$\arg\max_{H\subseteq F}\sum_{s_j\in S}\ln\left[1-\prod_{i=1}^{m}(1-p_{ij})^{h_i}\right]-\sum_{i=1}^{m}h_ic_i. \tag{22}$$

Let $y_i=\prod_{i=1}^{m}(1-p_{ij})^{h_i}$. Then, the LLL problem becomes the following minimization problem:

$$\min_{H\subseteq F} Z(h,y)=-\sum_{s_j\in S}\ln(1-y_i)+\sum_{i=1}^{m}h_ic_i \\ \text{s.t. } \ln y_i=\sum_{i=1}^{m}h_i(1-p_{ij}),\ 0\le y_j\le 1, \forall s_j\in S, \text{ and } h_i\in\{0,1\}. \tag{23}$$

From above, we can derive that the LLL problem is actually the 0–1 optimization problem. Since there are multiple variables in Eq. (23), the minimum problem cannot be solved directly and requires some approximations. We adopt the Lagrangian relaxation method to solve it near optimally.

By relaxing the constraints of Eq. (23) and transforming it via Lagrangian relaxation by adopting multiplier $\{\lambda_j\}$, we obtain the Lagrangian function as follows (Shakeri *et al.*, 1996):

$$\min_{H\subseteq F}\Theta(\lambda,h,y)=-\sum_{s_j\in S}[\ln(1-y_j)+\lambda_j\ln y_j] \\ +\sum_{i=1}^{m}h_i\left[c_i+\sum_{s_j\in S}\lambda_j\ln(1-p_{ij})\right]. \tag{24}$$

We denote the first and second expressions in the brackets of Eq. (24) by $f_j(\lambda_j, y_j)$ and $c_i'(\lambda_j)$, respectively. Note that for a fixed λ, the minimization of the above Lagrangian function that is with respect to h and y can be calculated independently. The minimization of $\Theta(\lambda, h, y)$ that is with respect to y is equivalent to the following equation:

$$\min_{0\le y_j\le 1} f_j(\lambda_j,y_j)=\ln(1-y_j)+\lambda_j\ln y_j. \tag{25}$$

We can obtain the extreme value at $y_j^*(\lambda_j)=\dfrac{\lambda_j}{1+\lambda_j}u(\lambda_j)$, where $u(\lambda_j)$ is the unit step function.

The minimization of $\Theta(\lambda, h, y)$ that is with respect to h is equivalent to

$$\min_{0\le y_i\le 1} W(\lambda_j, h_i) = \sum_{i=1}^{m} h_i c'_i(\lambda_j). \tag{26}$$

With the constraints in Eq. (23), the minimization problem in Eq. (26) is a traditional set-covering problem. Next, we propose a heuristic algorithm to solve it.

6.2 Heuristic algorithm design

We call the set-covering problem in Section 6.1 the 'LLL problem'. As the problem is NP-hard (Gong *et al.*, 2015), we propose a heuristic algorithm LLL to solve it.

The LLL algorithm first ranks the lossy links according to their contributions, and then finds the covering set that covers all the symptoms via adding lossy links one by one. The contribution of each lossy link f_i to the occurrence of s_i is defined as $g(f_i, s_i)$, which is the unconditional probability under observation S_O, and can be formulated as follows:

$$g(f_i, s_i) = \sum_{s_i\in S_\mathrm{O}} p(f_i \mid s_i) \Big/ \sum_{s_i\in S_{f_i}} p(f_i \mid s_i), \tag{27}$$

where the numerator represents the sum of the posterior probabilities with the observation S_O, and the denominator is the sum of all the posterior probabilities with the symptoms that result from f_i.

In Eq. (27), we adopt posterior probability $p(f_i|s_i)$ rather than prior probability $p(s_i|f_i)$ to define the contribution function. It is principally because the estimation accuracy of the prior probability is low. This may lead to low discrimination in explaining symptoms for the lossy links, and may result in mistakes of lossy link reasoning. Therefore, we introduce the concept of posterior probability, which is the unconditional probability, and can effectively overcome the shortcomings of prior probability. Posterior probability $p(f_i|s_i)$ can be formulated as follows:

$$p(f_i, s_i) = \frac{p(s_i \mid f_i)p(f_i)}{\sum_{f_i\in F_{s_i}} p(s_i \mid f_i)p(f_i)}. \tag{28}$$

According to the probabilistic correlation based lossy link model, Eq. (28) can be formulated as

$$p(f_i, s_i) = \frac{p(s_i \mid f_i)\cdot p_r \cdot p_r(l_i)}{\sum_{f_i\in F_{s_i}} p(s_i \mid f_i)p_r \cdot p_r(l_i)}, \tag{29}$$

where p_r is the probability that event r occurs, and $p_r(l_i)$ is the probability that l_i fails when an SRLG event r occurs.

With $g(f_i, s_i)$, we can rank all the lossy links in descending order, and take lossy links according to priority to explain the symptoms.

We take the loss rate of a link to determine whether it fails or not. A threshold ϕ_th is defined according to the performance demand of WSNs. If the link loss rate $\phi_{e(i)}\ge\phi_\mathrm{th}$, it is considered as failed, and otherwise as good. The threshold ϕ_th can be set based on the statistical data that can separate good or failed links.

The LLL algorithm details are shown in Algorithm 2, which works as follows. The LLL algorithm first calculates the contribution of each lossy link and ranks the lossy links in descending order (lines 2–6). Then it takes lossy links in F_RS to find the covering set according to the rank to explain the symptoms until all the symptoms are explained or there are no available lossy links for explaining symptoms (lines 7–17). Finally, we can acquire the lossy link hypothesis set Φ. We note that Φ is the minimal subset of F_S that minimizes the number of lossy links and covers all the symptoms, which is a solution of the set-covering problem in Section 6.1 for the LLL problem.

In real sensor networks, there may be no sufficient management codes, which results in missing reports for some symptoms. Therefore, these missing reports cannot be observed. Due to packet loss and application errors, alarms may be lost. Moreover, false alarms may occur when the alarm threshold is mistakenly set or due to network congestion. For example, as shown in Fig. 4, there is a PWBG, in which symptom s_5 gains losses, symptom s_6 is the false positive symptom, and f_4 is the corresponding false positive lossy link. Therefore, to make the network character more realistic and acquire higher LLL accuracy, we define the metrics in Algorithm 2.

Lemma 3 The LLL algorithm can be executed effectively.

Proof First, the selected lossy link set F_S is the reduction set with Algorithm 1, which involves all the

Algorithm 2 Lossy link localization algorithm

```
Input: the selected lossy link set $F_S$, the observed symptom set $F_O$, OR, LR($s$), SSR($s$), and the symptom set including all the explained symptoms $S_{ep}$
Output: lossy link hypothesis set
Initialization: $\Phi \leftarrow \varnothing$
for each $f_i \in F_S$ do
    Calculate $g(f_i, s_i)$
    Rank all the lossy links in $F_S$ in descending order
        according to $g(f_i, s_i)$
    $F_{RS} \leftarrow F_S$
end for
Initialization: $S_{ep} \leftarrow \varnothing$
for each $f_i \in$FRS do
    Take all $f_i \in F_{RS}$ in turn
    while $|S_{ep} \cap S_O|/|S|<1$ or there are available lossy links in
        $F_{RS}$ for explaining symptoms do
      if $S(f_i) \cup S_{ep} - S_{ep} \neq \varnothing$ then
        $\Phi \leftarrow \Phi \cup \{f_i\}$ and $S_{ep} \leftarrow S(f_i) \cup S_{ep}$
      else
        $\Phi \leftarrow \Phi$ and $S_{ep} \leftarrow S_{ep}$
      end if
    end while
end for
```

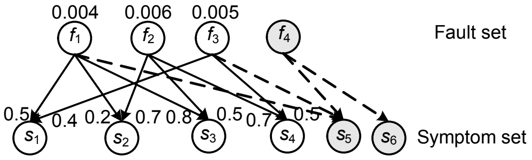


Fig. 4 A probabilistic weighted bipartite graph with OR, LR(s), and SSR(s) where OR=$|S_O|/|S|$, $s \in S_O$

important lossy link information. Then, we select the lossy links that greatly contribute to the creation of symptoms. Moreover, we construct set Φ by choosing the lossy links one by one according to the ranks. The first k lossy links that can explain all the symptoms are chosen as lossy link hypothesis set Φ, which can ensure the minimal Φ and best explain the symptoms.

Lemma 4 The LLL algorithm is computationally efficient.

Proof The LLL algorithm consists of two rounds of loop executions. First, ranking all the lossy links in F_S can be finished in $O(|F_S|\log_2|F_S|)$ based on a quick sort algorithm in the first loop execution. Then, the lossy link hypothesis set can be obtained with a complexity of $O(|F_{RS}|\log_2|F_{RS}|)$ with another for-loop that nests the binary search. Therefore, the complexity of the LLL algorithm is $O(|F_S|\cdot|F_{RS}|\log_2|F_S|\log_2|F_{RS}|)$, which can be implemented in polynomial time. The LLL algorithm is computationally efficient with an acceptable time complexity.

7 Performance evaluation

In this section, we evaluate our proposed method through extensive simulations and experiments. First, we analyze the performance of our redundancy elimination algorithm CIERE. Second, we compare our algorithms with several existing LLL algorithms in WSNs under the same simulation environment. Finally, we test our algorithm and compare it with existing algorithms on real data collected from actual WSNs.

7.1 Simulations

In the simulation, we construct a data-acquisition sensor network with the tree routing topology. The details of the network topology are presented in Section 3.1. We assume that the packet losses for a link are with a Bernoulli process, in which a packet traversing a link is dropped with a probability that is consistent with the loss rate. The transmission rates are distributed with density function $f(\xi)=\lambda\xi^{\lambda-1}$, where $0<\xi\leq 1$ and $\lambda>1$. Similar to Nguyen and Thiran (2006), we set $\lambda=4$, and thus the expected link loss rate is 0.8. The independent failure probabilities are uniformly distributed between 0.001 and 0.01, and can be obtained according to historical statistics over a certain period of time in practice. The conditional probabilities are randomly chosen from range (0, 1). In practice, they may be assigned by expert knowledge (Tang *et al.*, 2008). We repeat each simulation setting 10 times. In each time, 200 packets are sent from each sensor node to the sink. We report the average detection rate (DR) which is the percentage of links that are successfully located, the false positive detection rate (FPR) which is the percentage of links that function properly but are considered as lossy, and the failure localization time which is the time consumed acquiring the lossy link hypothesis set Φ. DR and FPR are defined as follows:

$$\mathrm{DR}=\frac{|F_{CS}\cap\Phi|}{|F_{CS}|}, \tag{30}$$

$$\mathrm{FPR}=\frac{|\Phi \setminus F_{\mathrm{CS}}|}{|\Phi|}, \tag{31}$$

where F_{CS} is the set of actual lossy links.

To evaluate the confidence levels of our simulation results, we also give their confidence intervals in the form of error bars in the comparisons. The detailed computing method is as follows. The population distribution of the overall simulation results is a normal distribution and the population standard deviation is unknown. Additionally, the sample size is less than 30. Therefore, the deviation statistics between results are distributed in a t-distribution, or $X\sim N(\mu, \sigma^2)$, where X is the population sample. For samples X_1, X_2, …, X_n of X, we can obtain $T=\frac{\overline{X}-\mu}{\sqrt{S^2/n}}\sim t(n-1)$, where $\overline{X}$ is the average value of samples X_1, X_2, …, X_n, and S^2 is the sample variance. We can further obtain the confidence interval of μ, which is $\overline{X}\pm\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)$. Based on this, we can calculate the confidence intervals. Note that we let α=0.05. It means that the confidence level of our results is 95%.

7.1.1 Performance analysis of the redundancy elimination algorithm CIERE

In this section, we analyze the redundancy elimination performance of our proposed algorithm CIERE and compare it with the MCSED algorithm in Wang *et al.* (2013) in terms of the failure redundancy rate RR (Eq. (4)) and the real lossy link coverage rate $\eta(F_{\mathrm{S}})$ (Eq. (12)).

Fig. 5 shows the results when the number of nodes varies from 20 to 100. We observe that as the network size increases, the lossy link redundancy rate for raw data remains almost constant, which means that the redundancy rate has a constant proportion in WSNs. The redundancy rate for raw data is approximately 89%. The redundancy rate of the MCSED algorithm is approximately 38%, and our algorithm has a lower redundancy rate of approximately 29%. Therefore, there are many redundant lossy links in set $F_{\max}$. We can effectively eliminate most of the redundant lossy links with our CIERE algorithm.

Table 2 shows the results of the real lossy link coverage rate in the selected set F_{S} with the same network size as in Fig. 5. We observe that $\eta(F_{\mathrm{S}})$ maintains 1 at different network sizes, which means that the selected set with our CIERE algorithm can always reserve real lossy links. The kernel attributes are all reserved and the classification ability is unchanged after redundancy reduction.

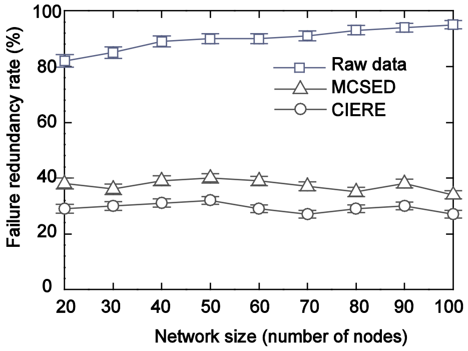


Fig. 5 Comparison of the failure (lossy link) redundancy rates with algorithms MCSED and CIERE

Table 2 Real lossy link coverage rate in set F_{S}

Network size	20	30	40	50	60	70	80	90	100
$\eta(F_{\mathrm{S}})$	1	1	1	1	1	1	1	1	1

7.1.2 Comparison with other methods

In this section, we focus primarily on the comparison of our LLL algorithm with several existing LLL algorithms—DID (Gong *et al.*, 2015), MPA (Zhao and Cai, 2010), TinyD2 (Liu *et al.*, 2014), and MCSED (Wang *et al.*, 2013)—in terms of DR, FPR, and the LLL time. Note that we compare DR and FPR under different network sizes and fractions of lossy links, respectively.

1. LLL scene I (OR=100%, LR=0%, SSR=0%)

Figs. 6a and 6b show the results with different network sizes. Additionally, Table 3 gives the percentage of improvements for LLL compared with the other four algorithms. We observe that as the network size increases, the accuracies of DID, MPA, TinyD2, and MCSED decrease, while our LLL algorithm maintains a steady accuracy. Similar results can be obtained when the number of lossy links increases. We also observe that as the network size increases, the LLL time of the MPA algorithm increases exponentially, and the localization times for LLL, DID, TinyD2, and MCSED slightly increase. Our LLL algorithm has the lowest localization time. We attribute the reason for the above observations to the fact that the raw data (the most probable lossy link set) has

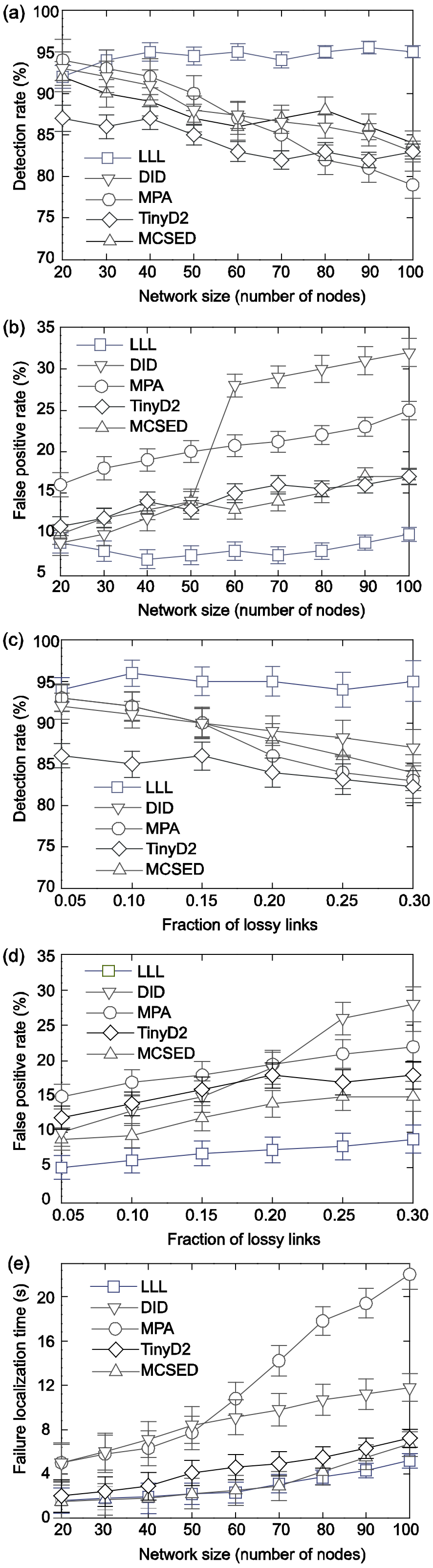


Fig. 6 Comparisons of detection rate, false positive rate, and localization time in scene I

The fraction of lossy links in (a), (b), and (e) is 0.15; the number of nodes in (c) and (d) is 60

Table 3 Average improvements in LLL in scene I

Method	Average improvement				
	DR (%)	FPR (%)	DR (%)	FPR (%)	Failure localization time (s)
DID	5.9	51.8	5.9	59.1	67.9
MPA	7.9	59.4	7.9	62.5	74.2
TinyD2	12.3	41.8	12.3	55.5	33.8
MCSED	6.8	39.5	6.8	42.7	4.3

The fraction of lossy links in Figs. 6a, 6b, and 6e is 0.15; the number of nodes in Figs. 6c and 6d is 60. DR: detection rate; FPR: false positive rate

massive redundancies, which decrease the localization accuracy and increase the computational complexity. However, our redundancy eliminating algorithm can obtain the most probable lossy link set at a low redundancy rate.

2. LLL scene II (OR=50%, LR=0%, SSR=0%)

When OR is 50%, compared with scene I, the accuracies of the five algorithms decrease (Fig. 7). There are two reasons for this. First, as OR decreases, the five algorithms have fewer available symptoms to infer lossy links. The lack of symptoms therefore results in lower accuracy. Second, our LLL algorithm models the lossy link with a probabilistic correlation model, which can reflect more intrinsic correlations between the lossy links. We also observe that as the number of observed symptoms decreases, the LLL time decreases. It is mainly because that there are fewer symptoms in the algorithm execution process. The performance improvements of LLL compared with the other four algorithms are shown in Table 4.

3. LLL scene III (OR=50%, LR=10%, SSR=1%)

Based on scene II, we increase LR to 10% and SSR to 1% in scene III. From Fig. 8, we observe that the accuracies of the five algorithms considerably decrease, because the losses and false positive observations of the symptoms increase the estimation

Table 4 Average improvements in LLL in scene II

Method	Average improvement				
	DR (%)	FPR (%)	DR (%)	FPR (%)	Failure localization time (s)
DID	0.4	30.5	6.8	39.3	55.5
MPA	0.4	34.0	5.3	45.7	67.3
TinyD2	5.3	23.8	14.2	33.4	33.7
MCSED	4.4	13.3	2.9	20.4	9.5

The fraction of lossy links in Figs. 7a, 7b, and 7e is 0.15; the number of nodes in Figs. 7c and 7d is 60. DR: detection rate; FPR: false positive rate

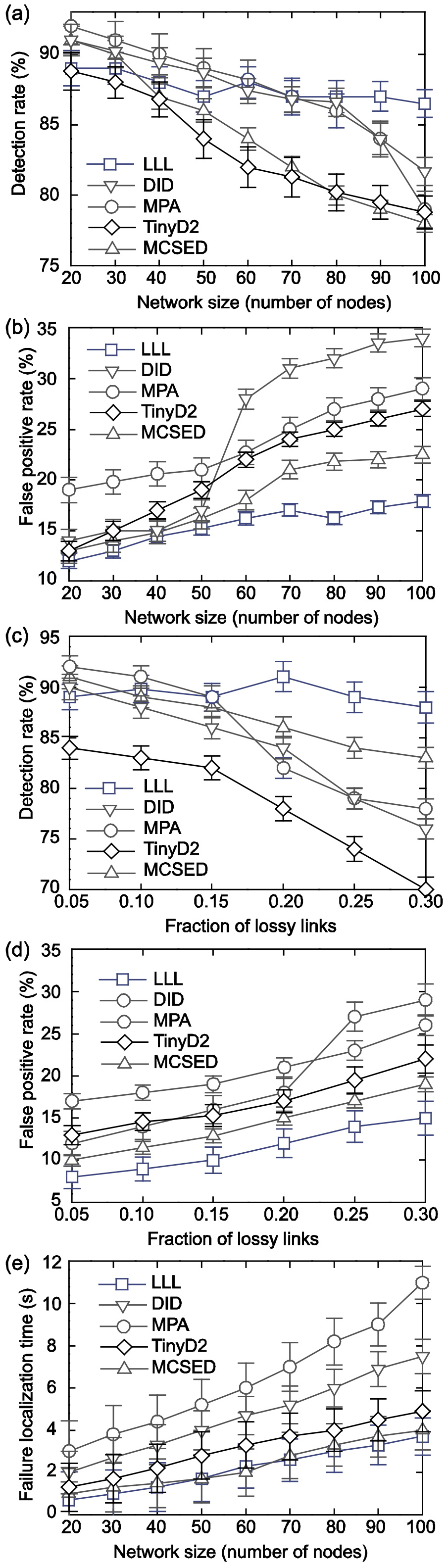


Fig. 7 Comparisons of detection rate, false positive rate, and localization time in scene II

The fraction of lossy links in (a), (b), and (e) is 0.15; the number of nodes in (c) and (d) is 60

errors. The algorithms DID and MPA try to infer exact loss rates, which results in many false inferred packet losses. The TinyD2 algorithm uses multiple nodes to cooperate with each other, which can meet the requirement of smaller loss rates. Our LLL algorithm uses a nondeterministic propagation model and is based on set-covering to locate the lossy links. This can reduce the false inferred faults and avoid missing the real faults; thus, the performance of LLL is better than those of the other algorithms. We also observe that the localization time slightly increases when there are packet losses and false positive symptoms. Explicit performance improvements in LLL are shown in Table 5.

Table 5 Average improvements in lossy link localization in scene III

Method	Average improvement				
	DR (%)	FPR (%)	DR (%)	FPR (%)	Failure localization time (s)
DID	2.2	27.5	2.8	33.7	52.9
MPA	4.5	23.2	3.9	42.9	69.6
TinyD2	4.5	16.0	13.0	29.9	29.6
MCSED	3.6	15.0	1.6	6.4	8.9

The fraction of lossy links in Figs. 8a, 8b, and 8e is 0.15; the number of nodes in Figs. 8c and 8d is 60. DR: detection rate; FPR: false positive rate

The overall detection rate and false positive rate achieved by the proposed LLL method in all three LLL scenes are 89.3% and 12.1%, respectively. Table 6 gives the overall improvements for LLL compared with the DID, MPA, TinyD2, and MCSED methods including the overall accuracy and overall computational complexity reduction.

7.2 Experimental verification

In this section, we compare the LLL, MCSED, MPA, DID, and TinyD2 algorithms using the SensorScope data. Here, we present the evaluation of the data trace collected in SensorScope as in Nguyen and Thiran (2006). We observe the data for 4 h and divide it into 60 slots. We run five algorithms on the data and report the average DR and FPR. For the selection of symptoms, we consider paths that deliver a threshold number of packets t, which is set to 20, 60, and 100. The larger the t, the smaller the errors in the inferences, and the smaller the symptom coverage. Fig. 9 shows the DR and FPR.

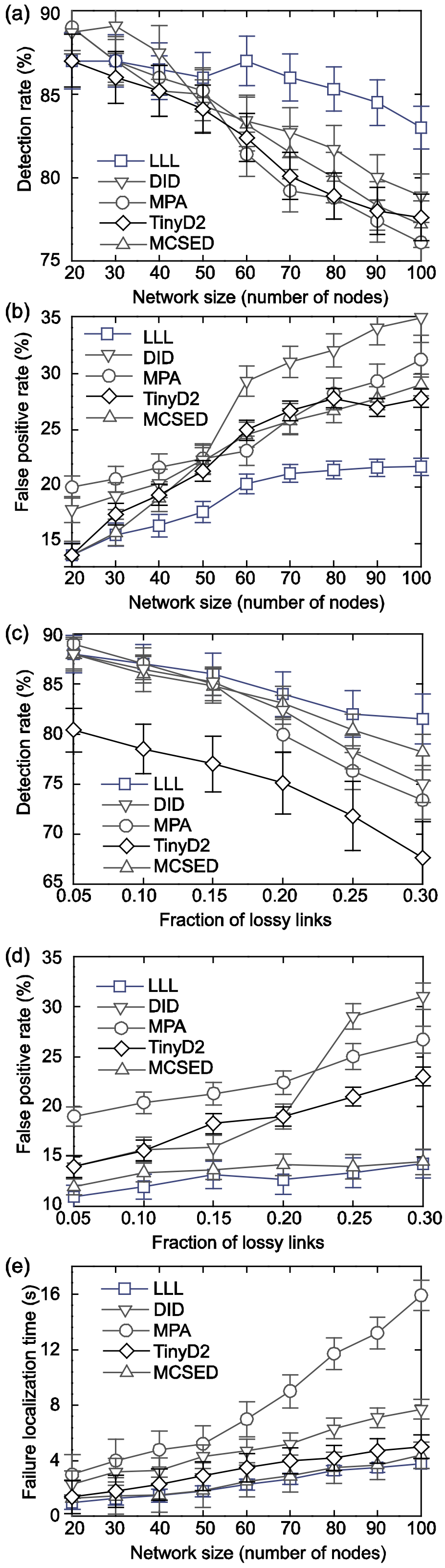


Fig. 8 Comparisons of detection rate, false positive rate, and localization time in scene III

The fraction of lossy links in (a), (b), and (e) is 0.15; the number of nodes in (c) and (d) is 60

Table 6 Overall improvements in lossy link localization

Method	Accuracy (%)	Computational complexity reduction (%)
DID	2.2	4.5
MPA	27.5	23.2
TinyD2	2.8	3.9
MCSED	33.7	42.9

We can observe that our LLL algorithm outperforms the other four algorithms. This proves the validity of LLL. For the paths that deliver the packets with a threshold t=20, we observe a total of 50 links, and 12 out of these 50 links are actually lossy. The LLL and TinyD2 methods correctly identify eight lossy links. However, the MCSED, MPA, and DID methods correctly identify seven lossy links. Moreover, LLL, MCSED, MPA, and DID give two false positives, and TinyD2 gives three false positives. For the paths that deliver packets with a threshold t=100, the LLL algorithm locates 95% of the lossy links and has a negligible FPR (5%). The performance of the LLL algorithm is superior to those of the other ones. The explanation for this is that the increased delivery rate of the paths can decrease the false positive rate of the symptoms, which therefore increases the evaluation accuracy. Note that the seven lossy links identified by MCSED, MPA, and DID are also the seven out of eight lossy links identified by the LLL and TinyD2 algorithms when t=20. It shows that the lossy links identified by the five algorithms are consistent. It is similar to the results in Nguyen and Thiran (2006). Furthermore, we give the percentage of improvement in LLL in Table 7.

8 Conclusions

Existing localization methods assume that multiple lossy links usually do not occur simultaneously. As a result, the failure localization usually has deviations. We propose a probabilistic correlation based lossy link model that can express the association between failures. Our approach effectively represents the nondeterministic causal relationships between lossy links and symptoms due to the constructed PWBG model. CIERE removes the faults that are less likely to occur by adopting conditional information entropy. More importantly, the LLL issue is a set-covering problem after several derivations, and the

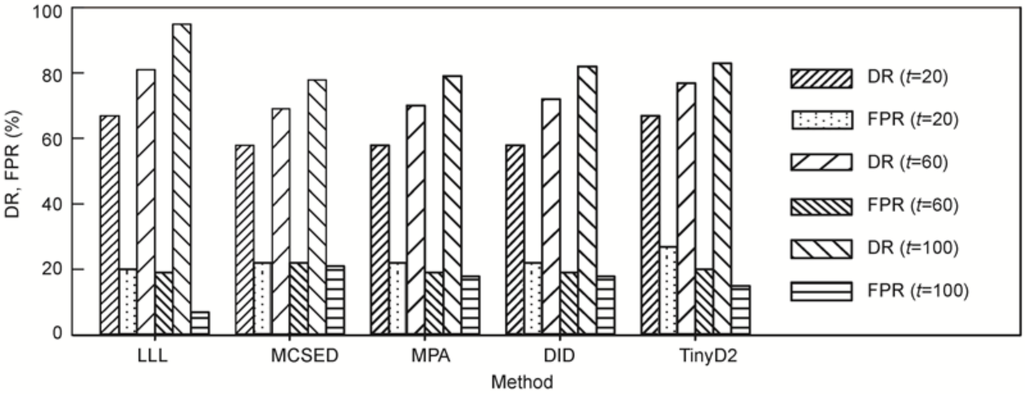


Fig. 9 Comparison of the LLL, MCSED, MPA, DID, and TinyD2 methods in SensorScope
DR: detection rate; FPR: false positive detection rate

Table 7 Average improvements in lossy link localization

Method	Average improvement (%)					
	DR (*t*=20)	FPR (*t*=20)	DR (*t*=60)	FPR (*t*=60)	DR (*t*=100)	FPR (*t*=100)
DID	15.5	9.1	12.5	0	15.8	61.1
MPA	15.5	9.1	15.7	0	20.3	61.1
TinyD2	0	25.9	5.2	5.0	14.5	53.3
MCSED	15.5	9.1	17.4	13.6	21.7	66.6

DR: detection rate; FPR: false positive detection rate

lossy link hypothesis set can be obtained through our LLL algorithm.

The proposed approach ensures that the failure localization is more accurate and efficient, even when multiple links fail simultaneously. Extensive simulations and the experiment show that our proposed method delivers a higher LLL accuracy while significantly reducing computational complexity compared to the existing LLL approaches, i.e., MCSED, DID, MPA, and TinyD2.

We evaluate our approach using only WSNs with tree topologies. For distributed application scenarios in WSNs, the topology may be very complex and may be widely distributed geographically. Therefore, there may be more than one sink, and thus the topology is not a tree structure. Further research should take into account failure localization in WSNs with a non-tree structure topology.

In addition, our work has been implemented in WSNs with the IEEE 802.11 protocol. However, 4G/5G protocols are different from the IEEE 802.11 protocol, and their networking modes are different. Thus, future research into the applicability of our approach should take this into account.

References

Ali, M.L., Ho, P.H., Tapolcai, J., *et al.*, 2014. Multi-link failure localization via monitoring bursts. *J. Optim. Commun. Netw.*, **6**(11):952-964. http://dx.doi.org/10.1364/JOCN.6.000952

Assaf, A.E., Zaidi, S., Affes, S., *et al.*, 2015. Low-cost localization for multihop heterogeneous wireless sensor networks. *IEEE Trans. Wirel. Commun.*, **15**(1):472-484. http://dx.doi.org/10.1109/TWC.2015.2475255

Benhamida, F.Z., Challal, Y., Koudil, M., 2014. Adaptive failure detection in low power lossy wireless sensor networks. *J. Netw. Comput. Appl.*, **45**(4):168-180. http://dx.doi.org/10.1016/j.jnca.2014.07.028

Benveniste, A., Fabre, E., Haar, S., *et al.*, 2003. Diagnosis of asynchronous discrete-event systems: a net unfolding approach. *IEEE Trans. Autom. Contr.*, **48**(9):714-727. http://dx.doi.org/10.1109/TAC.2003.811249

Bossuyt, D.L.V., O'Halloran, B., Papakonstantiou, N., 2016. Cable routing modeling in early system design to prevent cable failure propagation events. IEEE Annual Reliability and Maintainability Symp., p.1-6. http://dx.doi.org/10.1109/RAMS.2016.7448006

Chipara, O., Hackmann, G., Lu, C., 2010a. Practical modeling and prediction of radio coverage of indoor sensor networks. Proc. 9th Int. Conf. on Information Processing in Sensor Networks, p.339-349. http://dx.doi.org/10.1145/1791212.1791252

Chipara, O., Lu, C., Bailey, T.C., 2010b. Reliable clinical monitoring using wireless sensor networks: experiences in a step-down hospital unit. Proc. 8th Int. Conf. on Embedded Networked Sensor Systems, p.155-168. http://dx.doi.org/10.1145/1869983.1869999

Choi, G.S., Park, I.K., 2014. Uncertainty improvement of incomplete decision system using Bayesian conditional information entropy. *J. Inst. Int. Broadc. Commun.*, **14**(6):47-54 (in Korean). http://dx.doi.org/10.7236/JIIBC.2014.14.6.47

Collotta, M., Bello, L.L., Pau, G., 2015. A novel approach for dynamic traffic lights management based on wireless sensor networks and multiple fuzzy logic controllers. *Expert Syst. Appl.*, **42**(13):5403-5415. http://dx.doi.org/10.1016/j.eswa.2015.02.011

Couillet, R., Hachem, W., 2011. Local failure localization in large sensor networks. 45th Asilomar Conf. on Signals, Systems and Computers, p.1970-1974. http://dx.doi.org/10.1109/ACSSC.2011.6190369

Dias, A., Campos, P., Garrido, P., 2015. An agent based propagation model of bank failures. *Lect. Notes Econ. Math. Syst.*, **676**:119-130. http://dx.doi.org/10.1007/978-3-319-09578-3_10

Fang, W.W., Chen, J.M., Shu, L., *et al.*, 2010. Congestion avoidance, detection and alleviation in wireless sensor networks. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **11**(1):63-73. http://dx.doi.org/10.1631/jzus.C0910204

Gong, W., Liu, K., Liu, Y., 2015. Directional diagnosis for wireless sensor networks. *IEEE Trans. Parall. Distr. Syst.*, **26**(5):1290-1300. http://dx.doi.org/10.1109/TPDS.2014.2308173

Gupta, V., Tovar, E., Lakshmanan, K., *et al.*, 2012. Inter-application redundancy elimination in wireless sensor networks with compiler-assisted scheduling. 7th IEEE Int. Symp. on Industrial Embedded Systems, p.112-119. http://dx.doi.org/10.1109/SIES.2012.6356576

Haddad, A., Doumith, E.A., Gagnaire, M., 2013. A fast and accurate meta-heuristic for failure localization based on the monitoring trail concept. *Telecommun. Syst.*, **52**(2): 813-824. http://dx.doi.org/10.1007/s11235-011-9579-0

Harris, P., Philip, R., Robinson, S., *et al.*, 2016. Monitoring anthropogenic ocean sound from shipping using an acoustic sensor network and a compressive sensing approach. *Sensors*, **16**(3):415. http://dx.doi.org/10.3390/s16030415

He, W., Wu, B., Ho, P.H., *et al.*, 2011. Monitoring trail allocation for SRLG failure localization. IEEE Global Telecommunications Conf., p.1-5. http://dx.doi.org/10.1109/GLOCOM.2011.6133707

Kim, W., Park, G., Pack, S., *et al.*, 2014. Lightweight traffic redundancy elimination in software-defined wireless mesh networks. 3rd IEEE Global Conf. on Consumer Electronics, p.723-724. http://dx.doi.org/10.1109/GCCE.2014.7031262

Li, M., 2007. Underground structure monitoring with wireless sensor networks. 6th Int. Symp. on Information Processing in Sensor Networks, p.69-78. http://dx.doi.org/10.1109/IPSN.2007.4379666

Li, Y., Shi, H., Zhang, S., 2010. An optimized scheme for battlefield target tracking in wireless sensor network. 2nd Int. Conf. on Industrial and Information Systems, p.356-359. http://dx.doi.org/10.1109/INDUSIS.2010.5565835

Liu, K., Ma, Q., Gong, W., 2014. Self-diagnosis for detecting system failures in large-scale wireless sensor networks. *IEEE Trans. Wirel. Commun.*, **13**(10):5535-5545. http://dx.doi.org/10.1109/TWC.2014.2336653

Ma, W., Zhang, J., 2014. Algorithm based on heuristic strategy to infer lossy links in wireless sensor networks. *Algorithms*, **7**(3):397-404. http://dx.doi.org/10.3390/a7030397

Mahadevan, N., Abdelwahed, S., Dubey, A., *et al.*, 2010. Distributed diagnosis of complex systems using timed failure propagation graph models. IEEE AUTOTESTCON, p.1-6. http://dx.doi.org/10.1109/AUTEST.2010.5613575

Manolov, R., Guilera, G., Sierra, V., 2004. An analysis of a large scale habitat monitoring application. 2nd Int. Conf. on Embedded Networked Sensor Systems, p.214-226. http://dx.doi.org/10.1145/1031495.1031521

Manzano, M., Calle, E., Ripoll, J., *et al.*, 2013. Epidemic survivability: characterizing networks under epidemic-like failure propagation scenarios. 9th Int. Conf. on the Design of Reliable Communication Networks, p.95-102.

Miao, X., Liu, K., He, Y., 2011. Agnostic diagnosis: discovering silent failures in wireless sensor networks. *IEEE Trans. Wirel. Commun.*, **12**(12):6067-6075. http://dx.doi.org/10.1109/TWC.2013.110813.121812

Nguyen, H.X., Thiran, P., 2006. Using end-to-end data to infer lossy links in sensor networks. 25th IEEE Int. Conf. on Computer Communication, p.1-12. http://dx.doi.org/10.1109/INFOCOM.2006.271

Niu, Q., Xia, S., Tan, G., 2009. A method of fuzzy reasoning based on semantic similarity and bipartite graph matching. IEEE Int. Conf. on Artificial Intelligence and Computational Intelligence, p.141-145. http://dx.doi.org/10.1109/AICI.2009.89

Ntalampiras, S., 2014. Fault identification in distributed sensor networks based on universal probabilistic modeling. *IEEE Trans. Neur. Netw. Learn. Syst.*, **26**(9):1939-1949. http://dx.doi.org/10.1109/TNNLS.2014.2362015

Nyberg, M., 2013. Failure propagation modeling for safety analysis using causal Bayesian networks. IEEE Conf. on Control and Fault-Tolerant Systems, p.91-97. http://dx.doi.org/10.1109/SysTol.2013.6693936

Park, G., Shim, Y., Jang, I., *et al.*, 2016. Bloom-filter-aided redundancy elimination in opportunistic communications. *IEEE Wirel. Commun.*, **23**(1):112-119. http://dx.doi.org/10.1109/MWC.2016.7422413

Patil, P., Kulkarni, U., 2013. SVM based data redundancy elimination for data aggregation in wireless sensor networks. IEEE Int. Conf. on Advances in Computing, Communications and Informatics, p.1309-1316. http://dx.doi.org/10.1109/ICACCI.2013.6637367

Priesterjahn, C., Heinzemann, C., Schafer, W., 2013. From timed automata to timed failure propagation graphs. 16th IEEE Int. Symp. on Object/Component/Service-Oriented Real-Time Distributed Computing, p.1-8. http://dx.doi.org/10.1109/ISORC.2013.6913236

Rajasegarar, S., Leckie, C., Palaniswami, M., 2008. Anomaly detection in wireless sensor networks. *IEEE Wirel. Commun.*, **15**(1):34-40. http://dx.doi.org/10.1109/MWC.2008.4599219

Sandhya, M.K., Murugan, K., Devaraj, P., 2015. Selection of aggregator nodes and elimination of false data in wireless sensor networks. *Wirel. Netw.*, **21**(4):1327-1341. http://dx.doi.org/10.1007/s11276-014-0859-y

Shakeri, M., Pattipati, R., Raghavan, V., 1996. Optimal and near-optimal algorithms for multiple fault diagnosis with unreliable tests. *IEEE Trans Syst. Man Cybern. C*, **28**(3):431-440. http://dx.doi.org/10.1109/5326.704583

Shen, D., 2012. Adaptive fault monitoring in all-optical networks utilizing real-time data traffic. *J. Netw. Syst. Manag.*, **20**(1):76-96. http://dx.doi.org/10.1007/s10922-011-9206-0

Strasser, S., Sheppard, J., 2011. Diagnostic alarm sequence maturation in timed failure propagation graphs. IEEE AUTOTESTCON, p.158-165. http://dx.doi.org/10.1109/AUTEST.2011.6058741

Tang, Y., Al-Shaer, E., Boutaba, R., 2008. Efficient fault diagnosis using incremental alarm correlation and active investigation for Internet and overlay networks. *IEEE Trans. Netw. Serv. Manag.*, **5**(5):36-49. http://dx.doi.org/10.1109/TNSM.2008.080104

Tang, Y., Cheng, G., Xu, Z., 2009. Community-based fault diagnosis using incremental belief revision. IEEE Int. Conf. on Networking, Architecture and Storage, p.121-128. http://dx.doi.org/10.1109/NAS.2009.24

Troiano, L., Cerbo, A.D., Tipaldi, M., *et al.*, 2015. Fault detection and resolution based on extended time failure propagation graphs. IEEE Conf. on Soft Computing and Pattern Recognition, p.337-342. http://dx.doi.org/10.1109/SOCPAR.2013.7054155

Urbanics, G., Gönczy, L., Urbán, B., *et al.*, 2014. Combined error propagation analysis and runtime event detection in process-driven systems. 6th Int. Workshop on Software Engineering for Resilient Systems, p.169-183. http://dx.doi.org/10.1007/978-3-319-12241-0_13

Wang, B., Wei, W., Dinh, H., 2011. Fault localization using passive end-to-end measurements and sequential testing for wireless sensor networks. *IEEE Trans. Mob. Comput.*, **11**(3):439-452. http://dx.doi.org/10.1109/TMC.2011.98

Wang, R., Wu, Q., Xiong, Y., 2013. Multi-parameters link failure localization algorithm based on compressive sensing. *J. Electron. Inform. Technol.*, **35**(11):2596-2601 (in Chinese). http://dx.doi.org/10.3724/SP.J.1146.2013.00265

Woo, A., Tong, T., Culler, D., 2003. Taming the underlying challenges of reliable multi-hop routing in sensor networks. Int. Conf. on Embedded Networked Sensor Systems, p.14-27. http://dx.doi.org/10.1145/958491.958494

Wu, C., Wang, J., Zeng, J., 2011. A network traffic awareness architecture for universal redundancy elimination. Int. Conf. on Electronic and Mechanical Engineering and Information Technology, p.52-55. http://dx.doi.org/10.1109/EMEIT.2011.6022836

Xie, L., Heegaard, P.E., Jiang, Y., 2013. Modeling and quantifying the survivability of telecommunication network systems under fault propagation. International Federation for Information Processing, p.25-36.

Xu, Y., Liu, Y., Liu, Y., 2012. Algorithm for redundancy elimination in network traffic. 2nd IEEE Int. Conf. on Consumer Electronics, Communications and Networks, p.1613-1617. http://dx.doi.org/10.1109/CECNet.2012.6201599

Yamamoto, S., Nakao, A., 2012. P2P packet cache router for network-wide traffic redundancy elimination. IEEE Int. Conf. on Computing, Networking and Communications, p.830-834. http://dx.doi.org/10.1109/ICCNC.2012.6167541

Yang, C., Shi, H., Xue, G., *et al.*, 2014. Network redundancy elimination by dynamic buffer allocation. IEEE 17th Int. Conf. on Computational Science and Engineering, p.1109-1114. http://dx.doi.org/10.1109/CSE.2014.218

Yang, Y., An, Z., Xu, Y., *et al.*, 2010. Passive loss inference in wireless sensor networks using EM algorithm. *Wirel. Sens. Netw.*, **2**(7):512-519.

Zhang, C., Liao, J., Zhu, X., 2010. Heuristic fault localization algorithm based on Bayesian suspected degree. *J. Softw.*, **21**(10):2610-2621 (in Chinese).

Zhang, L., Wang, W., Gao, J., 2014. Lossy links diagnosis for wireless sensor networks by utilizing the existing traffic information. *Int. J. Embed. Syst.*, **6**(2):140-147. http://dx.doi.org/10.1504/IJES.2014.063811

Zhang, N., Yang, X., Zhang, M., *et al.*, 2016. RMI-DRE: a redundancy-maximizing identification scheme for data redundancy elimination. *Sci. China Inform. Sci.*, **59**:089301. http://dx.doi.org/10.1007/s11432-016-5523-y

Zhang, Y., Ansari, N., 2014. On protocol-independent data redundancy elimination. *IEEE Commun. Surv. Tutor.*, **16**(1):455-472. http://dx.doi.org/10.1109/SURV.2013.052213.00186

Zhao, Z., Cai, W., 2010. Passive localizing lossy links in sensor network using max-product algorithm. 3rd IEEE Int. Conf. on Computer Science and Information Technology, p.571-575. http://dx.doi.org/10.1109/ICCSIT.2010.5563652