A Labeled RFS-based Framework for Multiple Integrity Attackers Detection and Identification in Cyber-Physical Systems

Chaoqun Yang, Lei Mo, Xianghui Cao, Heng Zhang, Zhiguo Shi

A Labeled RFS-based Framework for Multiple Integrity Attackers Detection and Identification in Cyber-Physical Systems

Chaoqun Yang, *Member, IEEE,* Lei Mo, *Member, IEEE,* Xianghui Cao, *Senior Member, IEEE,* Heng Zhang, *Member, IEEE,* Zhiguo Shi, *Senior Member, IEEE*

***Abstract*—The problem of multiple integrity attacks (attackers) detection and identification (MIADI) in cyber-physical systems (CPSs) is still a challenging problem to date. The goal of this paper is to develop a knowledge-based method capable of simultaneously detecting and identifying multiple integrity attacks aiming at different sensors in a CPS. In this paper, with the help of labeled random finite set (RFS) theory, a new solution to solve the MIADI problem is proposed. The main contributions of this paper, lie in the the following two aspects, the first is the novel formulation of the MIADI problem, in which labeled RFSs are used to model the behaviors of multiple integrity attackers for the first time, and the second is the proposed labeled RFS-based solution, which provides an elegant framework to cope with the MIADI problem. Numerical experiments are conducted and experimental results demonstrate the effectiveness of the proposed solution. This proposed solution further extends the feasibility of the labeled RFS theory in the context of CPSs cybersecurity.**

***Index Terms*—Cyber-physical systems, cybersecurity, integrity attacks, multiple attackers detection, random finite set.**

I. Introduction

As cyber-physical systems (CPSs) become more and more ubiquitous in vertical industries and critical infrastructures such as the industrial Internet of Things, industrial control systems, and electrical power systems, and their risks of exposure to cyber attacks are dramatically increasing [1–3]. Due to the frequent interaction in cyberspace, and the access to numerous devices with security vulnerabilities, CPSs are vulnerable to malicious attacks [4–6]. These malicious attacks include, but not be limited to, availability, integrity and confidentiality attacks. Among them, integrity attack, which aims at compromising sensor or actuator data by using malicious signals, has recently become a major threat to CPSs [7].

Therefore, as one of the effective strategies to protect CPSs from integrity attacks, integrity attack (attacker) detection is becoming more and more paramount. Fortunately, integrity attack (attacker) detection for the CPSs has attracted great interest in considerable literatures, as demonstrated by references [8–10]. However, most of the existing detection methods are designed to cope with the problem of detecting only one integrity attack (attacker), and the problem of multiple integrity attacks (attackers) detection (MIAD), has been rarely considered by the existing literatures up to now.

To improve the chance of success and the power of integrity attacks, multiple attackers might simultaneously attempt to launch multiple integrity attacks in a cooperative or uncooperative way. According to the recent report published by the Sophos company, the issue of multiple attackers is becoming a clear and present danger [11, 12]. However, in comparison with the problem of detecting only one integrity attack (attacker), the MIAD problem is much more challenging. In this regard, according to the different levels of multiple integrity attacks detection, the following three objectives need to be tackled:

1) The accurate detection of the number of integrity attacks (attackers) is a challenging objective. On the one hand, due to the attack strategy, stealthiness demand, and energy limit of each attacker, the number of integrity attacks (attackers) is usually dynamic versus time. On the other hand, false alarms produced by attack detectors will cause errors in accurately detecting the number of attacks (attackers).
2) The accurate detection of attacked sensors, nodes, channels, or links in a CPS. This vital objective is an essential precondition for isolating the attacked sensors, nodes, channels, or links, further reducing the damage caused by integrity attacks. However, it is not easy because of the following two reasons. First, it is natural that the simultaneous launch of multiple integrity attacks will increase the chance of escaping from detection. Second, it is inevitable that the attack detector equipped by each attacked sensor, node, channel, or link will produce false alarms.
3) The accurate identification of each integrity attacker, i.e., to confirm that each attacked sensor, node, channel, or link is attacked by which integrity attacker. This

This work was supported in part by the Start-up Research Fund of Southeast University under Grant RF1028623002, in part by National Key Research and Development Program of China under Grant 2022YFF0902800, in part by the National Natural Science Foundation of China under Grant 92067111 and 61873106, in part by the "Zhongying" Young Scholars Project, and in part by "Zhishan Scholars" Projects under Grant 2242021R40003.

C. Yang and L. Mo are with the School of Automation, Southeast University, Nanjing 210096, China (e-mail: ycq@seu.edu.cn; lmo@seu.edu.cn). X. Cao is with the School of Automation, Southeast University, Nanjing 210096, China, he is also with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: xhcao@seu.edu.cn). H. Zhang is with the School of Computer Engineering, Jiangsu Ocean University, Lianyungang 222000, China (e-mail: ezhangheng@gmail.com). Z. Shi is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: shizg@zju.edu.cn). (*Corresponding author: Xianghui Cao.*)

objective is helpful in understanding and profiling the attackers' behaviors, which can not only guide in discovering, visualizing, and predicting attacks, but can also further alleviate the damage to CPSs [13].

Reference [14] addressed the MIAD problem, wherein multiple integrity attacks that can be detected by the χ^2 detector were considered, and introduced the random finite set (RFS) theory to cope with the first two kinds of objectives of the MIAD problem. In reference [15], to simultaneously detect the jamming attack and the false data injection attack which is one of the integrity attacks, a resilient attack detection estimator was proposed. However, as far as we know, very few existing literatures have dealt with the MIAD problem involving the third objective. For briefness, in the rest of this article, this kind of the MIAD problem is re-called as the problem of multiple integrity attacks (attackers) detection and identification (MIADI).

Motivated by this, the MIADI problem is studied in this article. Specifically, this article deals with the problem of how to simultaneously detect and identify multiple integrity attacks (attackers) aiming at different sensors in a CPS. We take advantage of the latest advances in the labeled RFS theory and formulate the behaviors of multiple attackers as labeled RFSs. Differing from [14] in which the behaviors were modeled by using common RFSs, the labeled RFS-based problem formulation integrates the unique label information of each attacker, which makes it possible to identify each integrity attacker. Furthermore, the MIADI problem is posed in a Bayesian framework, and a solution based on the δ generalized labeled multi-Bernoulli (δ-GLMB) filter which simultaneously achieves all objectives of the MIAD problem is developed. To our best knowledge, this work is the first work on the MIADI problem and is also the first attempt to exploit the labeled RFS theory in the context of CPSs cybersecurity.

The main contribution of the article is the proposed framework for the MIADI problem, which includes two steps, i.e., the labeled RFS-based problem formulation and the solution based on the δ-GLMB filter. The proposed framework poses the following advantages:

- It provides a systemic labeled RFS-based problem formulation for the MIADI problem, in which both the behaviors and label information of multiple integrity attackers are capsuled, making it feasible to detect and identify multiple integrity attacks (attackers).
- It develops a δ-GLMB filter-based solution to the MIADI problem, which simultaneously achieves the detection of the number of attackers, the detection of each attacked sensor, and the identification of each attacker.
- Since the δ-GLMB filter is an analytic solution to the multi-object Bayesian filter [16], whereas the probability hypothesis density (PHD) filter in [14] is the first moment approximation to the multi-object Bayesian filter, it poses a better detection performance than the existing method in [14], leading to smaller joint detection errors.

The rest of this article is organized as follows: Section II presents an overview of the related work. Section III states the considered problem. Section IV and Section V present the two steps of the proposed framework, i.e., the labeled RFS-based problem formulation and the solution based on the δ-GLMB filter, respectively. Section VI presents numerical experiments. Finally, Section VII concludes this article.

II. Related Work

This section presents a brief review of the recent advances in the issue of attack detection for CPSs. Recently, considerable attention to this issue has been attracted, and various detection methods have been reported [17–21]. Generally speaking, the existing attack detection methods can be classified into two categories, i.e., knowledge-based and data-driven methods [18]. For the knowledge-based methods, the authors in [19] and [20] summarized four main detection methods, i.e., Bayesian detector, WLS-detector, χ^2-detector, and Quasi-FDI detector, and recent advances in these methods were reviewed. In addition, some remarkable advances referring to the knowledge-based methods include a distributed strategy to detect attacks on the communication networks in the large-scale CPSs [22], a real-time resilient CPS framework to detect and isolate the pole-dynamics attack [23], and a control architecture capable of detecting the setpoint attack [24], to name a few.

The data-driven methods fall into two categories, machine/deep learning-based, and graph-based methods. Thanks to the booming of machine/deep learning technologies, it is observed that many machine/deep learning-based methods have been proposed to detect attacks. For instance, in [25], a new machine learning-based method and the review of both semisupervised and supervised machine learning-based methods for attack detection were presented. In [26, 27], dictionary learning, as a powerful machine learning method, was demonstrated to perform well in process monitoring and attack detection. In [17], a deep learning-based method to detect the distributed denial-of-service attack was proposed. In [21], a comprehensive review of deep learning-based attack detection methods in the context of CPSs was presented. Compared with machine/deep learning, a graph neural network (GNN) is better at coping with irregular and imbalanced data [28]. Thus, graph-based attack/intrusion detection methods are also attracting increased attention. For instance, Zhang *et al.* proposed a new GNN-based algorithm that can efficiently exploit the rare and imbalanced training data for intrusion detection with high accuracy [28]. Deng *et al.* proposed a graph deviation network that can detect anomalous events, including attacks and system faults from high-dimensional time series data [29].

As mentioned before, although considerable literatures have concentrated on the issue of attack detection, it is worth noting that rare work has paid attention to the problem of how to detect multiple attacks simultaneously. Thus, in this article, we represent a step forward in the direction of multiple attacks detection. Particularly, the problem of detecting multiple integrity attacks is addressed, and a novel framework that simultaneously achieves the detection of the number of attackers, the detection of each attacked sensor, and the identification of each attacker is proposed.

III. Problem Statement

A. Notations

Throughout this article, the following notations are used. Scalars or random vectors are denoted by lowercase letters (e.g., $x, \boldsymbol{x}$). RFSs are represented by uppercase letters (e.g., X). Especially, labeled RFSs are represented by bold uppercase letters (e.g., $\boldsymbol{X}$). Blackboard bold uppercase letters such as $\mathbb{X}$ stand for spaces. Additionally, the following operators are defined. $\lfloor v \rfloor$ represents the floor function that outputs the greatest integer, which is no more than v. The inner product is defined as $\langle f, g \rangle \triangleq \int f(x)g(x)dx$, and $\delta.(\cdot)$ is the Kronecker delta function. The following indicator function is defined as

$$1_Y(X) \triangleq \begin{cases} 1, & \text{if } X \subseteq Y, \\ 0, & \text{otherwise.} \end{cases}$$

For a real-valued function h and an RFS Y, the notation h^Y is defined by

$$h^Y = \prod_{y \in Y} h(y),$$

with $h^{\emptyset} = 1$.

B. System Model

Consider a CPS that consists of the physical part, the cyber part and the communication networks between the two parts. Thereinto, the cyber part consists of a remote estimator and a remote controller to ensure the operation and security of the CPS, and the physical part usually includes numerous sensors. For illustration purposes, we present a CPS deployed in a smart manufacturing workshop. As shown in Fig. 1, large-scale industrial facilities in the physical part are connected through industrial networks, and scheduled and controlled by the cyber part via the communication networks. Although these industrial facilities are various, including manufacturing facilities, logistics facilities, monitoring facilities, and so on, for the sake of analysis, these facilities are taken as sensors, and each sensor is identified by an ordered index x, where $x \in \mathbb{X}$ and $\mathbb{X}$ is the discrete space of all sensors' indices.

To prevent the CPS from potential integrity attacks, similar to [22, 30, 31], we assume that each sensor is equipped with a χ^2 detector, and the remote estimator will collect detection reports for further analysis. It is worth noting that further analysis is necessary since the detection of multiple integrity attacks is not equivalent to a simple count of the collected detection reports. In fact, due to the inherent characteristics of a χ^2 detectors, both missing reports and false alarms exist in the collected detection reports.

C. Attack Model

To degrade system performance or even cause disastrous consequences [32], assume that multiple attackers are employed by a malicious party to launch integrity attacks on the above CPS, as illustrated in Fig. 1. Although the assumptions of attackers' abilities are various in the existing literatures, ranging from omniscient to ignorant, to better meet the needs of real scenarios, the following reasonable assumptions of the limited abilities are considered.

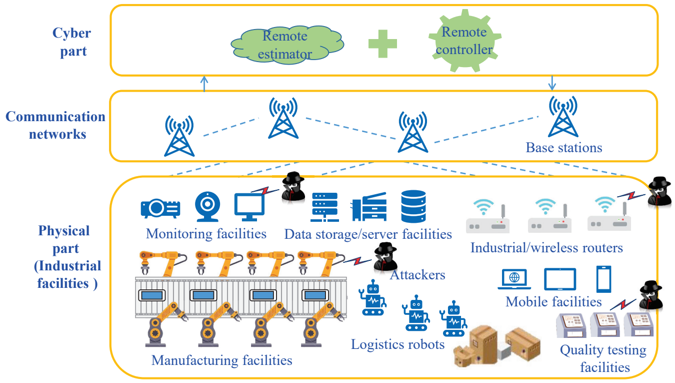

Fig. 1: Illustration of a CPS for a smart manufacturing workshop in the presence of multiple integrity attacks.

- All of the launched integrity attacks aim at the sensors in the physical part, rather than the cyber part or communication networks. The reasons are two-fold. First, the protection of the cyber part or communication networks is usually much more elaborate than that of the sensors. For example, encryption communication is often used in communication networks [33]. Second, the integrity attacks aiming at the sensors are ubiquitous. For instance, many types of integrity attacks such as false data injection attacks and replay attacks can be easily launched on the sensors [32].
- Similar to [34, 35], the energy resources of all of the attackers are limited, which means that they can not launch attacks all the time.
- Due to the limited abilities of the attackers, similar to [14], each attacker can only simultaneously attack no more than one sensor at the same time.

D. Problem Statement

Intuitively, from the view of the CPS, it needs to detect all of the launched multiple attacks at each time. As mentioned before, the detection includes three objectives, while all of them are considered in this article. Specifically:

1) The first is to detect the time-varying number of the active attackers. It is worth noting that the number of active attackers does not simply equal to the number of detection reports, due to missing detection and false alarms.
2) The second is to detect the attacked sensors among all of the detection reports, which are the superposition of real reports and false alarms. To some extent, attack detection has been simply completed by the detector equipped with each sensor. However, it is inaccurate to directly take the collection of reports as the attacked sensors. Further analysis will be carried out in the remote estimator to overcome the negative effects caused by missing detection and false alarms, leading to better detection performance.
3) The third is to identify each active attacker, i.e., to know each attacked sensor is attacked by which attacker. Identifying multiple attackers is helpful in understanding

and profiling the attackers' behaviors, which can not only provide guidance on discovering, visualizing, and predicting attacks but also further alleviate the damage to the CPS.

In summary, the considered problem is: *How to achieve the above three objectives simultaneously only based on the reports from each sensor?*

IV. Labeled RFS-Based Problem Formulation

A. Background on Labeled Random Finite Sets

A set, like $X = \{x_1, x_2, \cdots, x_n\}$, where $x_i \in \mathbb{X}$ and $\mathbb{X}$ represents a state space, is an RFS on $\mathbb{X}$ if its cardinality (the number of elements) $|X| = n$ is random, and x_i is random and unordered [36, 37]. Similar to a random vector, the statistical characteristics of an RFS can also be characterized by its probability density function (PDF) (called multi-object PDF). Subsequently, we introduce the following three kinds of RFSs.

(1) Poisson RFS: A Poisson RFS X on $\mathbb{X}$ satisfies: Its cardinality $|X|$ follows a Poisson distribution with mean $\langle v, 1\rangle$ and its elements are independently and identically distributed subjected to the PDF $v(\cdot)/\langle v, 1\rangle$ [16].

(2) Bernoulli RFS and multi-Bernoulli RFS: A Bernoulli RFS X has an existence probability ε to contain only one element x distributed according to $p(x)$, and has a probability $1-\varepsilon$ to be an empty set [16]. Its multi-object PDF is usually denoted by parameter set $\{\varepsilon, p(x)\}$. A multi-Bernoulli RFS X on $\mathbb{X}$ is a union of a fixed number of independent Bernoulli RFSs (components) $X^{(i)}$ with existence probability $\varepsilon^{(i)}$ and PDF $p^{(i)}$ [16], and its multi-object PDF is usually abbreviated as $\{\varepsilon^{(i)}, p^{(i)}\}_{i=1}^{\lambda}$.

By augmenting each state $x \in \mathbb{X}$ with a unique and distinct label $l \in \mathbb{L}$, where $\mathbb{L}$ denotes a discrete label space, we define a labeled RFS $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$, where $\boldsymbol{x} = (x, l)$. Let $\mathcal{L}$ be the projection from $\mathbb{X} \times \mathbb{L}$ to $\mathbb{L}$, i.e., $\mathcal{L}((x, l)) = l$, we define the distinct label indicator $\Delta(\boldsymbol{X}) = \delta_{|\boldsymbol{X}|}(|\mathcal{L}(\boldsymbol{X})|)$.

Labeled multi-Bernoulli (LMB) RFS: Intuitively speaking, an LMB RFS is a multi-Bernoulli RFS on $\mathbb{X}$ augmented with distinct labels corresponding to the non-empty Bernoulli components [16, 38]. An LMB RFS $\boldsymbol{X}$ defined on $\mathbb{X} \times \mathbb{L}$ is distributed according to the multi-object PDF

$$\pi(\boldsymbol{X}) = \Delta(\boldsymbol{X}) w(\mathcal{L}(\boldsymbol{X})) p^{\boldsymbol{X}}, \tag{1}$$

where

$$\begin{aligned} w(L) &= \prod_{l' \in \mathbb{L}} (1 - \varepsilon^{(l')}) \prod_{l \in L} \frac{1_{\mathbb{L}}(l)\varepsilon^{(l)}}{1 - \varepsilon^{(l)}}, \\ p(x, l) &= p^{(l)}(x), \end{aligned}$$

and $\{\varepsilon^{(l)}, p^{(l)}(x)\}$ is the parameter set of the Bernoulli component with label l. For simplicity, $\pi(\boldsymbol{X})$ is usually abbreviated as $\{\varepsilon^{(l)}, p^{(l)}\}_{l \in \mathbb{L}}$.

B. Labeled RFS-Based Formulation for Multiple Attackers' Behaviors

1) Label Space and Multi-state: Due to the attack strategy and limited energy resources of each attacker, at each time step, all of the attackers fall into two categories: active and inactive attackers. The former is the attackers launching attacks at this time step, while the latter is the attackers who do not launch any attack at this time step. For ease of analysis, similar to [14, 39], suppose that both the active and inactive attackers are mutually independent.

Each active attacker is uniquely identified by a unique label $l = (k, i)$, where k is the time step when the attacker becomes active, and i is a unique index to distinguish the attackers becoming active at the same time step. The label space $\mathbb{L}_k$ for newborn active attackers (i.e., the attackers who become active at time step k) is $\{k\} \times \mathbb{N}$, where $\mathbb{N}$ denotes the integer space, and the label space $\mathbb{L}_{0:k}$ for the active attackers at time step k can be recursively formed by $\mathbb{L}_{0:k} = \mathbb{L}_{0:k-1} \cup \mathbb{L}_k$ [16]. Let $\boldsymbol{x} = (x, l)$ denote the sensor attacked by the active attacker with label l, where $x \in \mathbb{X}$ is the index of this sensor. Then, $\boldsymbol{x}$ can be treated as the state of this active attacker, and the labeled RFS $\boldsymbol{X}_k = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{n(k)}\}$ is regarded as the multi-state of these active attackers at time step k, where $n(k)$ is the number of these active attackers.

Accordingly, at time step k, a newborn active attacker, has a state $\boldsymbol{x} \in \mathbb{X} \times \mathbb{L}_k$, and for these active attackers, a multi-state $\boldsymbol{X}_k$ is a finite subset of $\mathbb{X} \times \mathbb{L}_{0:k}$ [16]. For convenience, detailed references to the current time index "k" are omitted, and those for the next time index "$k+1$" are abbreviated as "$+$" in the following part of this section. As a result, $\mathbb{L}$, $\mathbb{B}$, and $\mathbb{L}_+ = \mathbb{L} \cup \mathbb{B}$ represent the label spaces for active attackers at the current time step, for newborn active attackers at the next time step, and for all active attackers at the next time step, respectively.

2) LMB RFS-based Formulation of Surviving Attackers: Each active attacker at the current time step either continues to be active at the next time step with probability p_s or becomes inactive with probability $q_s = 1 - p_s$ [14]. Namely, given the current multi-state $\boldsymbol{X}_k$, at the next time step, each state $\boldsymbol{x} = (x, l) \in \boldsymbol{X}_k$ either continues to survive with probability $p_s(x, l)$, and transfers to a new state (x_+, l_+) with PDF $f(x_+|x, l)\delta_l(l_+)$, or dies with probability $q_s(x, l) = 1 - p_s(x, l)$ [16], where $f(x_+|x, l)$ denotes the attack strategy of the active attacker with label l, and $\delta_l(l_+)$ means that the label of the active attacker is preserved [16]. Thus, the state of each active attacker turns to be a Bernoulli RFS with parameter set $\{p_s(x, l), f(x_+|x, l)\delta_l(l_+)\}$.

Since all of the active attackers are mutually independent, it follows that the multi-state $\boldsymbol{W}$ of surviving active attackers at the next time step is an LMB RFS with parameter set $\{p_s(x, l), f(x_+|x, l)\delta_l(l_+)\}_{l \in \mathbb{L}}$. According to (1), the distribution of $\boldsymbol{W}$ follows

$$f_s(\boldsymbol{W}|\boldsymbol{X}) = \Delta(\boldsymbol{W})\Delta(\boldsymbol{X}) w(\mathcal{L}(\boldsymbol{W})) p^{\boldsymbol{W}}, \tag{2}$$

where

$$\begin{aligned} w(L) &= 1_{\mathcal{L}(\boldsymbol{X})}(L) \prod_{l' \in \mathcal{L}(\boldsymbol{X})} q_s(x, l') \prod_{l \in L} \frac{1_{\mathcal{L}(\boldsymbol{X})}(l) p_s(x, l)}{q_s(x, l)}, \\ p(x_+, l) &= f(x_+|x, l), \quad (x_+, l) \in \boldsymbol{W}, \end{aligned}$$

$\Delta(\boldsymbol{W})\Delta(\boldsymbol{X})$ indicates that both $\boldsymbol{W}$ and $\boldsymbol{X}$ have distinct labels, and $1_{\mathcal{L}(\boldsymbol{X})}(L)$ implies that the labels in $\boldsymbol{X}$ are preserved in $\boldsymbol{W}$.

3) LMB RFS-based Formulation for Newborn Attackers: For each inactive attacker at the current time step, it either becomes active at the next time step with probability ε_b, and the index of its first attacked sensor follows the PDF p_b or continues to keep inactive with probability $1-\varepsilon_b$ [1]. In other words, the state of a newborn attacker can be modeled by a Bernoulli RFS with parameter set $\{\varepsilon_b, p_b\}$. Since each inactive attacker is independent of each other, the multi-state $\boldsymbol{Y}$ of these newborn attackers at the next time step can be modeled by an LMB RFS with label space $\mathbb{B}$. Hence, it follows that from (1), the distribution of $\boldsymbol{Y}$ is captured by the following multi-object PDF

$$f_b(\boldsymbol{Y}) = \Delta(\boldsymbol{Y}) w_b(\mathcal{L}(\boldsymbol{Y})) p_b^{\boldsymbol{Y}}, \tag{3}$$

where

$$\begin{aligned} w_b(L) &= \prod_{l' \in \mathbb{B}} (1-\varepsilon_b^{(l')}) \prod_{l \in L} \frac{1_{\mathbb{B}}(l)\varepsilon_b^{(l)}}{1-\varepsilon_b^{(l)}}, \\ p_b(x,l) &= p_b^{(l)}(x). \end{aligned}$$

It can be seen that the active attackers at the next time step are made up of the surviving active attackers and newborn active attackers. Therefore, the multi-state of the active attackers at the next time step, $\boldsymbol{X}_+$, is the combination of the multi-state of the surviving active attackers $\boldsymbol{W}$ and that of the newborn active attackers $\boldsymbol{Y}$, i.e., $\boldsymbol{X}_+ = \boldsymbol{W} \cup \boldsymbol{Y}$. According to [16], the multi-object PDF of $\boldsymbol{X}_+$ conditioned on $\boldsymbol{X}$ can be written as

$$f(\boldsymbol{X}_+|\boldsymbol{X}) = f_s(\boldsymbol{X}_+ \cap (\mathbb{X}\times\mathbb{L})) f_b(\boldsymbol{X}_+ - \mathbb{X}\times\mathbb{L}). \tag{4}$$

Remark 1: Differing from the work on modeling the behaviors of multiple attackers by using common RFSs in [14], here the attackers' behaviors are formulated by using labeled RFSs. The significant difference is that the state of an attacker not only contains the index of its attacked sensor but also contains the unique label of this attacker. In contrast, only the former is captured in [14]. Benefited from the introduction to the unique labels of attackers, it becomes possible to identify each active attacker.

C. RFS-based Formulation for Attack Detection Scheme

A χ^2 detector, characterized by its detection probability p_d and false alarm probability p_f, is equipped by each sensor to detect whether the sensor is under integrity attack or not. Intuitively, if detected, it will report the sensor's index $z = \mathcal{D}(\boldsymbol{x} = (x,l)) = x$ to the remote estimator, where $\mathcal{D}(\boldsymbol{x})$ is a projection from $\mathbb{X}\times\mathbb{L}$ to $\mathbb{X}$. If undetected, it will report an empty set to the remote estimator [14]. From the mathematical point of view, the formulation of an attack detection scheme can be described as follows.

Firstly, for a given multi-state $\boldsymbol{X}$ of the current multiple active attackers, each active attacker denoted by $\boldsymbol{x} \in \boldsymbol{X}$ is either detected with detection probability p_d and generates a report z with likelihood $g(z|\boldsymbol{x}) = 1$, or undetected with probability $1-p_d$ and generates an empty report $z = \emptyset$. Therefore, the reports generated by $\boldsymbol{x}$ can be modeled as a Bernoulli RFS with parameter set $\{p_d, g(\cdot|\boldsymbol{x})\}$. Accordingly, suppose that the detection process of each attack detector is independent. Then, the set of these reports generated by $\boldsymbol{X}$ is a multi-Bernoulli RFS denoted as D with the following multi-object PDF

$$\pi_d(D|\boldsymbol{X}) = \{(p_d, g(\cdot|\boldsymbol{x})) : \boldsymbol{x} \in \boldsymbol{X}\}. \tag{5}$$

Secondly, apart from the reports from the detectors equipped by the attacked sensors, the remote estimator possibly receives the false alarm reports from the detectors equipped by the un-attacked sensors due to the non-zero false alarm probabilities of the detectors. On the premise of the independence of each detector, the received false reports can be modeled by a Poisson RFS C with mean $<\kappa, 1>$, which follows the multi-object PDF [16]

$$\pi_c(C) = e^{-<\kappa,1>}\kappa^C. \tag{6}$$

The collected reports at the remote estimator, represented by an RFS Z, are the superposition of all sensors' reports, i.e., $Z = D \cup C$. According to [16], the likelihood of Z can be described as

$$g(Z|\boldsymbol{X}) = \sum_{D \subseteq Z} \pi_d(D|\boldsymbol{X}) \pi_c(Z-D). \tag{7}$$

Remark 2: In this section, the behaviors of multiple active attackers are modeled by the labeled RFS $\boldsymbol{X}_+$ distributed according to (4), which not only includes the indices of the attacked sensors but also contains the labels of the active attackers. The collected reports at the remote estimator are formulated by the RFS Z distributed according to (7). Note that (4) and (7) are the centerpiece of the formulation of the MADI problem via the labeled RFS theory.

Remark 3: Since the information about the attackers is completely captured by the labeled RFS $\boldsymbol{X}$ at each time step, the first three objectives of the CPS can be achieved by iteratively estimating the posterior multi-object PDF of $\boldsymbol{X}_k$ conditioned on the collected reports $Z_{1:k}$, i.e., $\pi_k(\boldsymbol{X}|Z_{1:k})$.

V. A Solution to The MADI Problem Via The δ-GLMB filter

Suppose that the posterior multi-object PDF of $\boldsymbol{X}_{k-1}$ at the last time step $k-1$, $\pi_{k-1}(\boldsymbol{X}|Z_{1:k-1})$, is given. After receiving the collected detection reports Z_k, the posterior conditional multi-object PDF at time step k, $\pi_k(\boldsymbol{X}|Z_{1:k})$, can be calculated by the following iterative multi-object Bayesian filter

$$\pi_{k|k-1}(\boldsymbol{X}|Z_{1:k-1}) = \int f(\boldsymbol{X}_+|\boldsymbol{X}) \pi_{k-1}(\boldsymbol{X}|Z_{1:k-1}) d\boldsymbol{X}, \tag{8}$$

$$\pi_k(\boldsymbol{X}|Z_{1:k}) = \frac{g(Z|\boldsymbol{X})\pi_{k|k-1}(\boldsymbol{X}|Z_{1:k-1})}{\int g(Z|\boldsymbol{X})\pi_{k|k-1}(\boldsymbol{X}|Z_{1:k-1}) d\boldsymbol{X}}, \tag{9}$$

where (8) and (9) are the prediction step and update step, respectively, $f(\boldsymbol{X}_+|\boldsymbol{X})$ and $g(Z|\boldsymbol{X})$ have been derived in (4) and (7). In general, it is difficult to analytically solve the above multi-object Bayesian filter due to the abstract and complicated set integration. Hence a tractable solution is necessary.

[1] Note that the inactive attacker who becomes active at the next time step will be treated as a new active attacker even if it ever was active.

A. The δ-GLMB Filter

Recently, the δ generalized labeled multi-Bernoulli (δ-GLMB) filter has attracted increased attention because it presents an analytic solution to the multi-object Bayesian filter shown in (8)-(9). Moreover, suppose that the state transition function and likelihood function have the forms like (4) and (7), respectively. In that case, the above multi-object Bayesian filter can be solved, and $\pi_k(\boldsymbol{X}|Z_{1:k})$ can be iteratively estimated via the δ-GLMB filter. Consequently, multiple attackers can be detected and identified by further extracting the information from $\pi_k(\boldsymbol{X}|Z_{1:k})$.

Before presenting the δ-GLMB filter, we first introduce the δ-GLMB RFS. A δ-GLMB RFS $\boldsymbol{X}$ defined on $\mathbb{X} \times \mathbb{L}$ is distributed according to

$$\pi(\boldsymbol{X}) = \Delta(\boldsymbol{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} w^{(I,\xi)}\delta_I(\mathcal{L}(\boldsymbol{X}))[p^{(\xi)}]^{\boldsymbol{X}}, \quad (10)$$

where Ξ is a discrete space, and $\mathcal{F}(\mathbb{L})$ denotes the class of finite subsets of $\mathbb{L}$ [16, 40], and $\int p^{(\xi)}(x,l)dx = 1$,

$$\sum_{L\in\mathbb{L}} \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} w^{(I,\xi)}\delta_I(L) = 1.$$

The δ-GLMB filter starts with a δ-GLMB initial prior and includes the prediction and update steps, which are as follows. For convenience, in what follows, we omit the subscript of time indices of posterior quantities such as "k" and "$k-1$" and abbreviate time indices "$k+1|k$" as "+".

Prediction: Suppose that the prior multi-object PDF $\pi(\boldsymbol{X}|Z)$ at time step $k-1$ is a δ-GLMB of the form (10), and the state transition function has the form (4). Then, the predicted multi-object PDF is also a δ-GLMB given by [16]

$$\pi(\boldsymbol{X}_+)=\Delta(\boldsymbol{X}_+) \sum_{(I_+,\xi)\in\mathcal{F}(\mathbb{L}_+)\times\Xi} w_+^{(I_+,\xi)}\delta_{I_+}(\mathcal{L}(\boldsymbol{X}_+)) \left[p_+^{(\xi)}\right]^{\boldsymbol{X}_+}, \quad (11)$$

where

$$\begin{aligned}
w_+^{((I_+,\xi)} &= w_b(I_+ \cap \mathbb{B})w_s^{(\xi)}(I_+ \cap \mathbb{L}),\\
p_+^{(\xi)}(x,l) &= 1_{\mathbb{L}}(l)p_s^{(\xi)}(x,l) + (1 - 1_{\mathbb{L}}(l))p_b(x,l),\\
p_s^{(\xi)}(x,l) &= \frac{< p_s(\cdot,l)f(x|\cdot,l), p^{(\xi)}(\cdot,l) >}{\eta_s^{(\xi)}(l)},\\
\eta_s^{(\xi)}(l) &= \int < p_s(\cdot,l)f(x|\cdot,l), p^{(\xi)}(\cdot,l) > dx,\\
w_s^{(\xi)}(L) &= [\eta^{(\xi)}]^L \sum_{I\subseteq\mathbb{L}} 1_I(L)[q_s^{(\xi)}]^{I-L}w^{(I,\xi)},\\
q_s^{(\xi)}(l) &=< q_s(\cdot,l), p^{(\xi)}(\cdot,l) >,
\end{aligned}$$

and $w_b(L)$, $p_b(x,l)$ are given in (3).

Update: Suppose that the predicted multi-object PDF is a δ-GLMB of the form (11). After receiving the collected detection reports Z at time step k, under the likelihood function (7), the posterior multi-object PDF is also a δ-GLMB given by

$$\pi(\boldsymbol{X}|Z)=\Delta(\boldsymbol{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} \sum_{\theta\in\Theta} w^{(I,\xi,\theta)}\delta_I(\mathcal{L}(\boldsymbol{X})) \left[p^{(\xi,\theta)}\right]^{\boldsymbol{X}}, \quad (12)$$

where Θ is the space of mappings $\theta : I_+ \rightarrow \{0, 1, \cdots, |Z|\}$, so that $\theta(i) = \theta(i^{'}) > 0$ implies $i = i^{'}$ [16, 40], and

$$\begin{aligned}
w^{(I,\xi,\theta)} &= \frac{\delta_{\theta^{-1}(\{0:|Z|\})}(I)w^{I,\xi}[\eta_Z^{(\xi,\theta)}]^I}{\sum\limits_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi}\sum\limits_{\theta\in\Theta}\delta_{\theta^{-1}(\{0:|Z|\})}(I)w^{(I,\xi)}[\eta_Z^{(\xi,\theta)}]^I},\\
p^{(\xi,\theta)}(x,l|Z) &= \frac{p^{(\xi)}(x,l)\psi_Z(x,l;\theta)}{\eta_Z^{(\xi,\theta)}(l)},\\
\eta_Z^{(\xi,\theta)}(l) &=< p^{(\xi)}(\cdot,l), \psi_Z(\cdot,l;\theta) >,\\
\psi_Z(x,l;\theta) &= \delta_0(\theta(l))(1-p_d) + (1-\delta_0(\theta(l)))\lambda_Z(\cdot,l;\theta),\\
\lambda_Z(x,l;\theta) &= \frac{p_d g(z_{\theta(l)}|x,l)}{\kappa(z_{\theta(l)})}.
\end{aligned}$$

Since both the predicted and posterior multi-object PDFs are a combination of LMB components, it is unsurprising that the number of components will grow exponentially. Hence, it is necessary to truncate the multi-object PDFs by picking up the components with the most significant weights. For the detailed truncation process, please refer to [16, 40].

B. Attackers Detection and Identification

After estimating the posterior multi-object PDF $\pi(\boldsymbol{X}|Z)$ via the above δ-GLMB filter, the information about the attackers can be extracted as follows:

1) Detecting the number of integrity attacks (attackers). The number of active attackers can be detected by calculating the maximum a posteriori (MAP) estimation of cardinality. According to [16, 40], the estimated cardinality of the δ-GLMB RFS $\boldsymbol{X}$ is

$$\rho(n) = \sum_{(I,\xi)\in\mathcal{F}_n(\mathbb{L})\times\Xi} w^{(I,\xi)}, \quad (13)$$

where $\mathcal{F}_n(\mathbb{L})$ represents the collection of finite subsets of $\mathbb{L}$ with exactly n elements. Then, the number of the active attackers can be detected by

$$\hat{n} = \arg \quad \max \rho(n). \quad (14)$$

2) Among these LMB components whose cardinality equals to $\hat{n}$, picking up the LMB component with the highest weight, i.e.,

$$\hat{h} = \arg \quad \max w^{(h)}\delta_{\hat{n}}(|I^{(h)}|), \quad (15)$$

where $I^{(h)}$ represents the h-th LMB component in the δ-GLMB distributed according to $\pi(\boldsymbol{X}|Z)$.

3) **Identifying each attacker and detecting its attacked sensor**. For each component $\{\varepsilon^{(l)}, p^{(l)}(x)\}$ in the LMB component $I^{(\hat{h})}$, l is the identified label of the active attacker. Accordingly, the nearest integer to

$$\hat{x} = \int xp^{(l)}(x)dx, \quad (16)$$

is taken as the index of the detected sensor attacked by the active attacker with label l.

Remark 4: A natural concern for the proposed framework is the kinds of integrity attacks that can be detected. According to the detection scheme shown in (5)-(7), the integrity attacks that can be detected need to simultaneously subject to the following two conditions: (1) the integrity attacks are aiming at the sensors, rather than the cyber part and communication networks in the CPS, and (2), without consideration of missing detection, the attacks can not escape from the χ^2 detector equipped by each sensor.

Remark 5: In comparison with the existing method in [14], the proposed framework poses the following two differences. First, it uses labeled RFSs to formulate the attackers' behaviors, which jointly models the indices of attacked sensors and the labels of attackers, while the existing method in [14] only considered the formulation of the indices of attacked sensors. Second, it utilizes the δ-GLMB filter, which is an analytic solution to the multi-object Bayesian filter [16], to achieve the multiple detection of attackers, while the PHD filter that just is the first moment approximation to the multi-object Bayesian filter was used in [14]. As a consequence, the proposed framework poses the following two advantages. First, it can simultaneously achieve the detection of the number of attackers, the detection of each attacked sensor, and the identification of each attacker, while the existing method in [14] can only detect the the number of attackers and each attacked sensor. Second, it shows a better detection performance than the existing method in [14], leading to smaller joint detection errors.

VI. Numerical Experiments

In this section, we present the performance of the proposed labeled RFS-based framework in the presence of multiple integrity attackers aiming at different sensors.

A. Simulation Settings

A CPS deployed for surveillance task is considered, whose physical part includes 0.00,0.07,1.00$\rho = 400$ sensors. All of the sensors are equipped with the χ^2 detector to detect potential integrity attacks, and the detection probability and the false alarm probability of each attack detector are $p_d = 0.97$ and $p_f = 0.02$, respectively. The Poisson RFS with $\kappa = \rho p_f \mathcal{U}(1, \rho)$ is used to modeled the received false alarm reports in (6), where $\mathcal{U}(1, \rho)$ is the discrete and integral uniform distribution between 1 and ρ.

Consider that twenty-five attackers are employed to disrupt the above CPS. To achieve this goal, they will launch multiple integrity attacks aiming at the sensors according to their attack strategies. Due to the limited energy resources of these attackers, the number of active attackers at each time step is time-varying. The surviving probability of each active attacker is $p_s = 0.95$. The newborn probability of each inactive attacker is $\varepsilon_b = 0.02$, and the initial distribution of the indices of the attacked sensors, p_b, follows $\mathcal{U}(1, \rho)$.

To demonstrate the advantages of the proposed framework, the proposed framework is compared with the existing method called the probability hypothesis density (PHD) filter presented in [14]. Considering the feasibility under nonlinear cases, both them are implemented with particles. The detailed particle implementation of the δ-GLMB filter is omitted here, and the reader is directed to [16, 40] for detail. The number of particles assigned to each attacker (including each newborn attacker) is 500.

B. Evaluation Metric

To evaluate the joint detection errors of the detection of attacked sensors and the number of attackers, optimal subpattern assignment (OSPA) distance [41] which is widely used to evaluate the difference between two RFSs is adopted. The OSPA distance is defined as:

Let Π_k denote the set of permutations on $\{1, \cdots, k\}$ for any positive integer k. For x, y, let $d_p{}^{(c)}(x, y) = \min(c, ||x - y||)$. Then, for $p \geq 1, c > 0$, $X = \{x_1, \cdots, x_m\}$ and $Y = \{y_1, \cdots, y_n\}$, if $m \leq n$, the OSPA distance between X and Y, $\bar{d_p}^{(c)}(X, Y)$, is

$$\bar{d_p}^{(c)} = \left(\frac{1}{n} \left(\min_{\pi \in \Pi_n} \sum_{i=1}^{m} d^{(c)}(x_i, y_{\pi(i)})^p + c^p (n - m) \right) \right)^{\frac{1}{p}},$$

and if $m \geq n$, $\bar{d_p}^{(c)}(X, Y) = \bar{d_p}^{(c)}(Y, X)$ [41], where cut-off c and order p are usually selected in advance. The selection of cut-off c and order p depends on the relative importance between the detected accuracy of the attackers' number and that of the attacked sensors' indices. The larger the value of p, the more "punishment" is imposed due to the wrong detection of the attacked sensors' indices. In comparison, the larger the value of c, the more "punishment" is imposed due to the wrong detection of the attackers' number. Similar to [14, 42], the OSPA distance with $p = 1$ and $c = 100$ is adopted.

C. Case 1: Linear Case

Suppose that the attack strategy of each active attacker follows[2]

$$x_{i,k} = \lfloor x_{i,k-1} + v_k \rfloor, \tag{17}$$

where $x_{i,k-1} \in \mathbb{X}, x_{i,k} \in \mathbb{X}, i \in [1, 25] \cap \mathbb{Z}, v_k \sim \mathcal{N}(1, 0.1^2)$. The total number of attackers is 25, and the specific details of integrity attacks are listed in Appendix, in which "FDI", "BI" and "RE" represent false data injection attack, bias injection attack and replay attack, respectively.

The detection and identification results over one trail are presented in Fig. 2-Fig. 3, and the detection results over 100 Monte Carlo (MC) trails are presented in Fig. 4-Fig. 5. From Fig. 2, it can be seen that numerous false alarms exist in the detection reports, making it necessary to filter them with the Bayesian framework. The results shown in Fig. 4-Fig. 5 demonstrate that both the proposed framework and the PHD filter can accurately detect the number of attackers and the attacked sensors in most time. However, it is worth noting that the proposed framework performs better since it achieves smaller joint detection errors, as shown in Fig. 5.

Although the PHD filter can also simultaneously detect the number of attackers and the attacked sensors, it does not provide any identity information of each attacker. Taking the local detection results during time steps 38-39, for example, as Fig. 6 shows, since there is no identity information of each attacker, it is difficult for the detected attacked sensor A to distinguish that it was attacked by which attacker among the attackers who attacked D, E, F at the last time step. However,

[2] $f(x_+|x, l)$ can be derived from (17). For simplicity, here we only present (17).

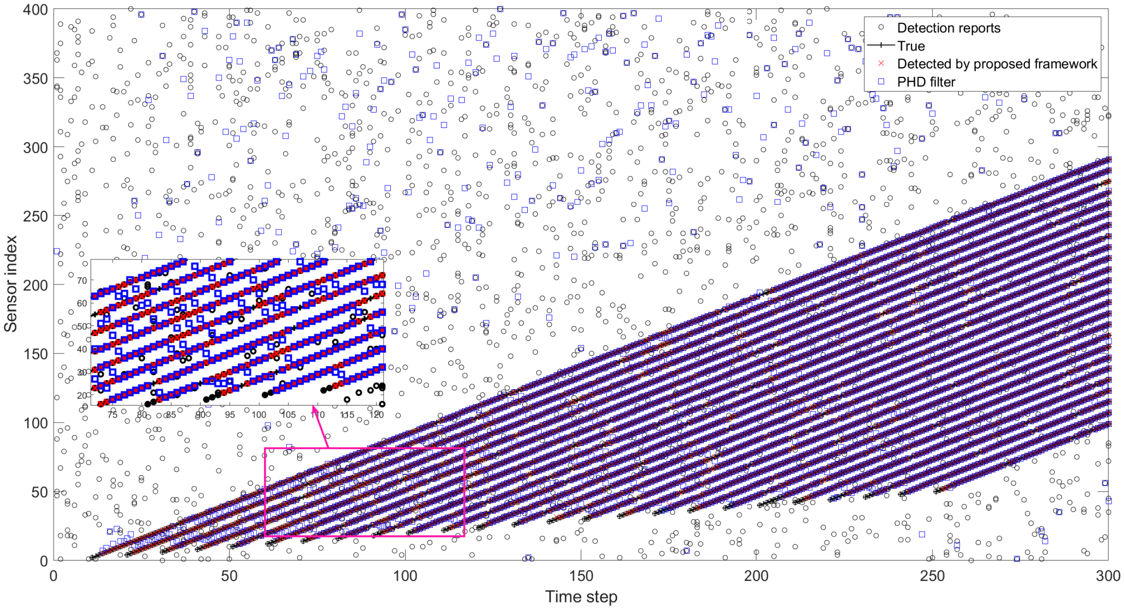

Fig. 2: Sensor reports and the detection results of the attacked sensors in case 1.

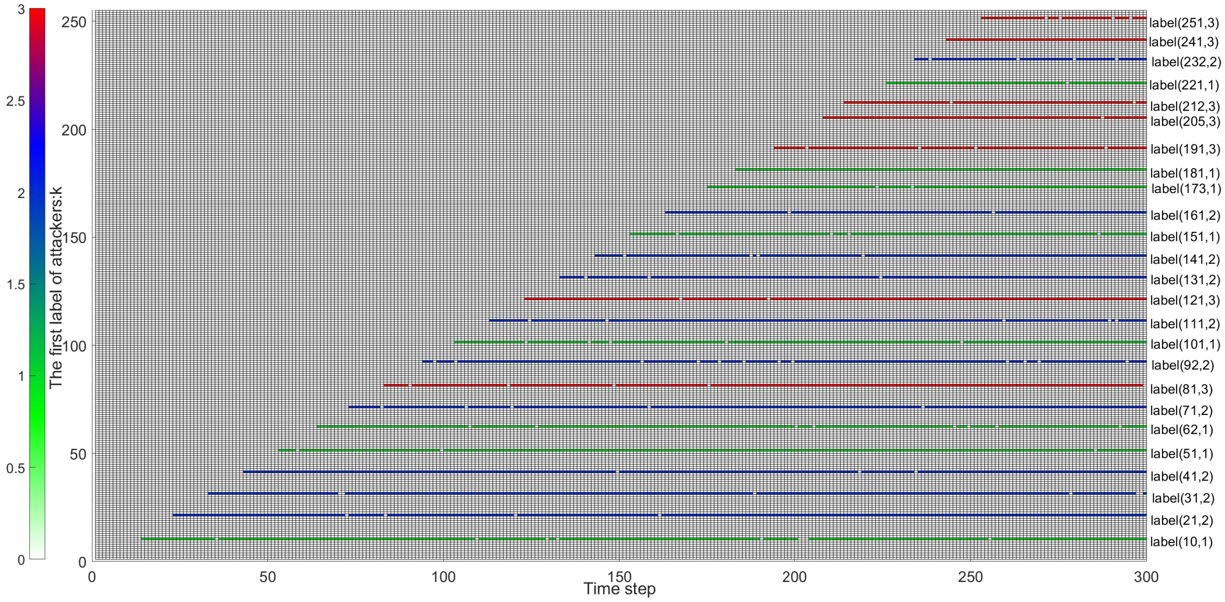

Fig. 3: The identification results of the detected attackers in case 1.

due to the integration of the labels of attackers, the proposed framework can identify each attacker. Fig. 3 presents the labels (k, l) of all of the attackers versus time. The proposed framework can clearly identify that the attackers attacked A and D are the same since both of them share the same label $(31, 2)$.

Remark 6: As shown in Fig. 2, there exist numerous false alarms in the original collected detection reports, while most of them are eliminated after running the proposed framework. The elimination of false alarms can be attributed to the following two procedures. The first is the accurate formulation of the statistical characteristics of both the attackers' behaviors and false alarms. The proposed framework captures the characteristics of the attackers' behaviors via the multi-object PDF $f(\boldsymbol{X}_+|\boldsymbol{X})$, and formulates false alarms via a Poisson RFS. The second is the effective use of these statistical characteristics in the δ-GLMB filter. As one of the knowledge-based methods, the proposed framework utilizes the statistical

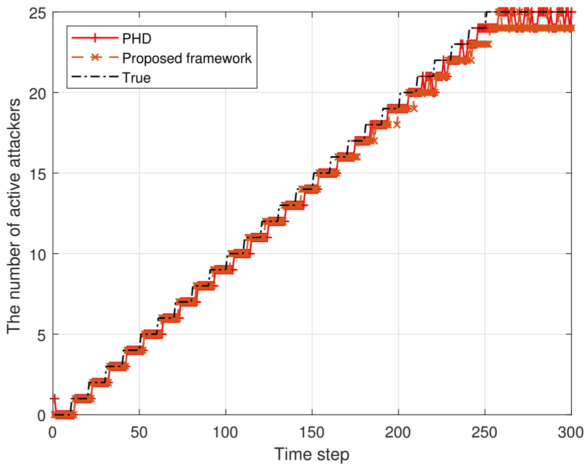

Fig. 4: The detection results of the number of attackers over 100 MC trials in case 1.

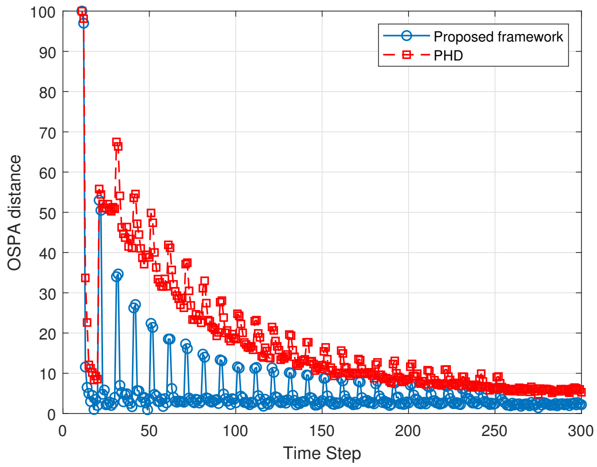

Fig. 5: Joint detection errors over 100 MC trials in case 1.

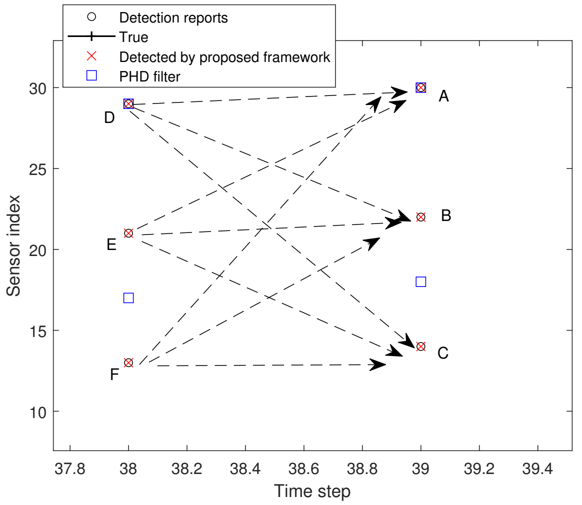

Fig. 6: Local detection results in Fig. 2 during time steps 38-39.

characteristics of both the attackers' behaviors and false alarms to refine the detection reports via the δ-GLMB filter. Since only the behaviors of attackers, rather than false alarms, are subjected to $f(\boldsymbol{X}_+|\boldsymbol{X})$, it is natural that the detection reports caused by the attackers are retained while false alarms are eliminated. If we treat the δ-GLMB filter as the extension of the Kalman filter, it is not surprising that the proposed framework is capable of eliminating false alarms, since the Kalman filter is capable of estimating dynamic target state polluted by noise.

D. *Case 2: Nonlinear Case*

Suppose that the attack strategy of each active attacker follows

$$x_{i,k} = \lfloor 400/x_{i,k-1} + v_k \rfloor, \tag{18}$$

where $x_{i,k-1} \in \mathbb{X}, x_{i,k} \in \mathbb{X}, i \in [1, 25] \cap \mathbb{Z}, v_k \sim \mathcal{N}(1, 0.1^2)$. The total number of attackers is 25, the specific details of attacks are also listed in Appendix, and other settings are the same as case 1.

The detection results over one trail are presented in Fig. 7, and the detection results over 100 MC trails are displayed in Fig. 8-Fig. 9. It can be seen that both the number of the attackers and the attacked sensors are accurately detected even in the case where the attackers pose nonlinear attack strategies. In summary, these results further demonstrate the effectiveness of the proposed framework for the MADI problem with the nonlinear case and confirm that the proposed framework generally outperforms the PHD filter.

E. *Comparison With Data-driven Methods*

From a data-driven perspective, the MAD problem can be treated as a multiple classification problem. Thus, machine learning (ML) classifiers such as support vector machine (SVM) and multilayer perceptron (MLP) can be also used to detect attacks. In this subsection, the detection performance of both SVM and MLP classifiers is investigated. We assume that the total number of attackers is 25, which is known for both SVM and MLP classifiers. Namely, the number of classes is given as 26. The first 25 classes denote different attackers, and the last class denotes false alarms. Then, the objective of the two classifiers is to accurately classify the collected detection reports into 26 classes.

1) Data Construction: We generated the collected detection reports over eight trails, obtaining 32496 samples. There are three features of these samples, i.e., the reported attacked time, the attacked sensor's index, and the kinds of the attack. The labels ranging from 1 to 26 are automatically tagged to these samples. Then, we randomly selected 1000 samples from each class, i.e., total 26000 samples, for training both the SVM and MLP models, and the remaining samples are used for testing. In addition, after arranging the remaining samples in order of time from smallest to largest, we take them as the measurement Z, which is taken as the input of the proposed framework.

2) MLP Network: As shown in Fig. 10, the adopted multilayer perceptron network is a network with 5 hidden layers, and the number of hidden neurons is set to 20 for each hidden layer.

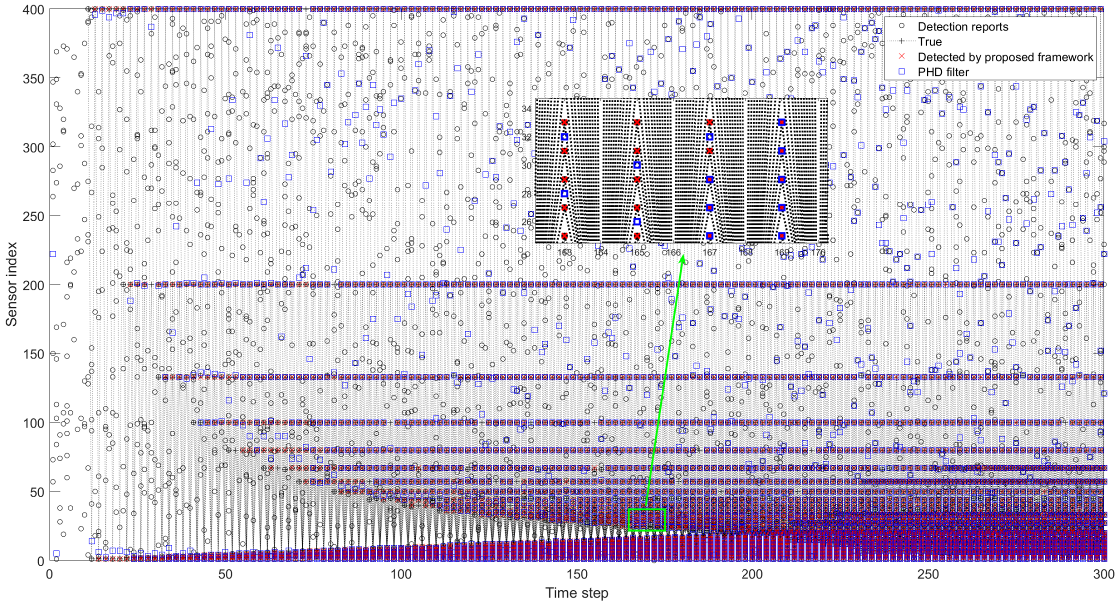

Fig. 7: Sensor reports and the detection results of the attacked sensors in case 2.

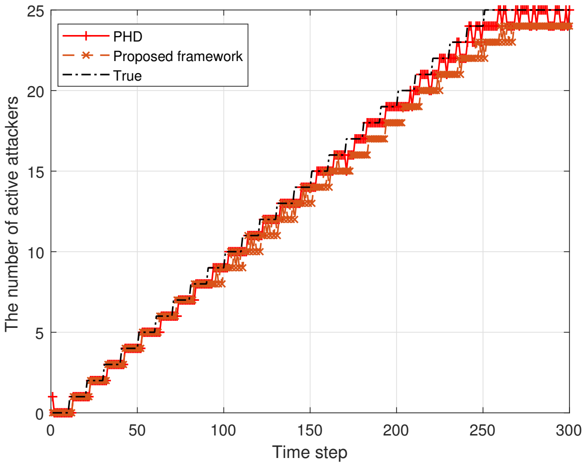

Fig. 8: The detection results of the number of attackers over 100 MC trials in case 2.

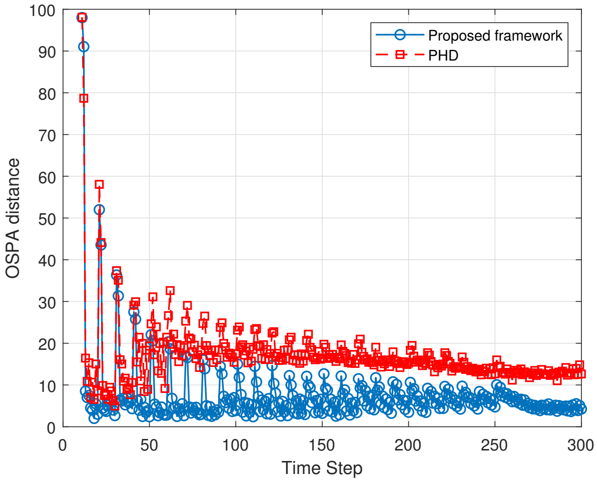

Fig. 9: Joint detection errors over 100 MC trials in case 2.

3) Results and Discussion: The performance is evaluated via the following metric,

$$Accuracy = \frac{TP}{TP + FP},$$

where TP and denote the truly detected attacks and FP denote the falsely detected attacks. It tells us about the ratio of correct classification with respect to all test samples.

The classification results are shown in TABLE I. It can be observed that both SVM and MLP classifiers achieve high accuracy (more than 97%) for the training samples. It can be also seen that, for the test samples, the MLP classifier is far superior to the SVM classifier, since it achieves 91.16% accuracy. However, it is worth noting that this accuracy is still slightly lower than the accuracy of the proposed method (97.49%).

To make the above conclusions more convincing, we investigate the accuracy of the MLP classifier with different deep learning configurations, varying hidden lay from 1 to 50. The results are presented in TABLE II. Generally speaking, for the MLP classifier with no more than 20 hidden layer, with the increasing of the number of hidden layer, the accuracy also monotonously increases. However, for the MLP classifier with more than 30 hidden layer, the accuracy dramatically becomes worse. That's reasonable, since too many layers may lead to some problems such as overfitting, making training more difficult and even affecting the expressiveness and generalization

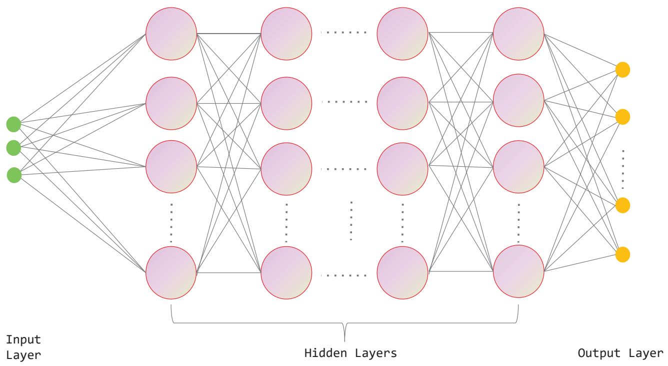

Fig. 10: The adopted MLP architecture.

ability of the trained model.

In summary, both the data-driven methods and the proposed method have their individual advantages. The data-driven methods also achieve good accuracy, and are easy to be deployed, since only the prior information including the training samples and the total number of attackers are needed. As a comparison, without the training step, the proposed framework achieves slightly higher accuracy. The reason can be attributed to the use of the prior statistical information about both the attackers' behaviors and false alarms. Overall, both the data-driven methods such as the MLP classifier and the knowledge-based methods like the proposed framework have their own advantages and the reader can have a choice in accordance with his needs.

TABLE I: The accuracy of different methods.

	SVM	**MLP**	**Proposed framework**
Training samples(%)	97.5	99.14	None
Test samples(%)	75.58	91.16	97.49

TABLE II: The accuracy of MLP with varying hidden layer.

Number of hidden layer	**1**	**2**	**3**	**5**	**10**	**20**	**30**	**50**
Training samples(%)	67.85	98.9	99.00	99.14	98.98	99.24	3.85	3.88
Test samples samples(%)	74.14	86.41	90.13	91.16	90.72	93.33	3.71	4.22

VII. Conclusion

This article proposes a labeled RFS-based framework to deal with the problem of multiple integrity attacks detection and identification, including the labeled RFS-based problem formulation and a solution based on the δ-GLMB filter. The proposed framework achieves the simultaneous detection of multiple integrity attacks and accurately identifies each attacker. This work further reveals the application prospect of coping with the problems of CPSs cybersecurity via the labeled RFS theory. In the future, we will continue to focus on the attempt to deal with some intractable CPSs security problems such as secure state estimation and secure control by exploiting the labeled RFS theory.

Appendix

TABLE III: The specific details of integrity attacks.

Attacker index	Launched time step	Persistent time step	First attacked sensor(case 1)	First attacked sensor(case 2)	Attack type
1	11	389	1	1	FDI
2	21	379	3	2	FDI
3	31	369	5	3	FDI
4	41	359	7	4	FDI
5	51	349	9	5	FDI
6	61	339	11	6	FDI
7	71	329	13	7	FDI
8	81	319	15	8	FDI
9	91	309	17	9	BI
10	101	299	19	10	BI
11	111	289	21	11	BI
12	121	279	23	12	BI
13	131	269	25	13	BI
14	141	259	27	14	BI
15	151	249	29	15	BI
16	161	239	31	16	BI
17	171	229	33	17	RE
18	181	219	35	18	RE
19	191	209	37	19	RE
20	201	199	39	20	RE
21	211	189	41	26	RE
22	221	179	43	32	RE
23	231	169	45	55	RE
24	241	159	47	60	RE
25	251	149	49	70	RE

References

[1] Z. Ji, C. Chen, J. He, S. Zhu, and X. Guan, "Edge sensing and control co-design for industrial cyber-physical systems: Observability guaranteed method," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13 350–13 362, Dec. 2022.

[2] Z. Zhang, R. Deng, P. Cheng, and Q. Wei, "On feasibility of coordinated time-delay and false data injection attacks on cyber-physical systems," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 8720–8736, June 2022.

[3] X. Cao, L. Liu, W. Shen, A. Laha, J. Tang, and Y. Cheng, "Real-time misbehavior detection and mitigation in cyber-physical systems over WLANs," *IEEE Trans. Industr. Inform.*, vol. 13, no. 1, pp. 186–197, Feb. 2017.

[4] F. Farivar, M. Haghighi, A. Jolfaei, and M. M. Alazab, "Artificial intelligence for detection, estimation, and compensation of malicious attacks in nonlinear cyber-physical systems and industrial IoT," *IEEE Trans. Industr. Inform.*, vol. 16, no. 4, pp. 2716–2725, April 2020.

[5] F. Zhang, H. Kodituwakku, J. W. Hines, and J. Coble, "Multilayer data-driven cyber-attack detection system for industrial control systems based on network, system, and process data," *IEEE Trans. Industr. Inform.*, vol. 15, no. 7, pp. 4362–4369, July 2019.

[6] H. Zhang, Y. Qi, J. Wu, L. Fu, and L. He, "DoS attack energy management against remote state estimation," *IEEE Trans. Control. Netw. Syst.*, vol. 5, no. 1, pp. 383–394, Mar. 2018.

[7] K. Zhang, C. Keliris, M. M. Polycarpou, and T. Parisini, "Detecting stealthy integrity attacks in a class of nonlinear cyber-physical systems: A backward-in-time approach," *Automatica*, vol. 141, pp. 1–14, July 2022.

[8] S. Kim, Y. Eun, and K. Park, "Stealthy sensor attack detection and real-time performance recovery for resilient CPS," *IEEE Trans. Industr. Inform.*, vol. 17, no. 11, pp. 7412–7422, Nov. 2021.

[9] Y. Luo, L. Cheng, Y. Liang, J. Fu, and G. Peng, "Deepnoise: Learning sensor and process noise to detect data integrity attacks in CPS," *China Commun.*, vol. 18, no. 9, pp. 192–209, Sept. 2021.

[10] Y. Mo, R. Chabukswar, and B. Sinopoli, “Detecting integrity attacks on SCADA systems,” *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 4, pp. 1396–1407, July 2014.

[11] M. Wixey, “Multiple attackers: A clear and present danger,” *A Sophos X-Ops Active Adversary Whitepaper*, pp. 1–23, Aug. 2022.

[12] K. Townsend, “Cyberattack victims often attacked by multiple adversaries: Research,” https://www.securityweek.com/cyberattack-victims-often-attacked-multiple-adversaries-research, accessed Aug. 2022.

[13] R. Katipally, L. Yang, and A. Liu, “Attacker behavior analysis in multi-stage attack detection system,” in *Proc. 7th CSIIRW Conf.*, Oak Ridge, USA, Oct. 2011, pp. 1–4.

[14] C. Yang, Z. Shi, H. Zhang, J. Wu, and X. Shi, “Multiple attacks detection in cyber-physical systems using random finite set theory,” *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 4066–4075, Sept. 2020.

[15] Y. Guan and X. Ge, “Distributed attack detection and secure estimation of networked cyber-physical systems against false data injection attacks and jamming attacks,” *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 4, no. 1, pp. 48–59, Mar. 2018.

[16] B.-T. Vo and B.-N. Vo, “Labeled random finite sets and multi-object conjugate priors,” *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3460–3475, July 2013.

[17] B. Hussain, Q. Du, B. Sun, and Z. Han, “Deep learning-based DDoS-attack detection for cyber-physical system over 5G network,” *IEEE Trans. Industr. Inform.*, vol. 17, no. 2, pp. 860–870, Feb. 2021.

[18] S. Tan, J. M. Guerrero, P. Xie, and R. Han, “Brief survey on attack detection methods for cyber-physical systems,” *IEEE Syst. J.*, vol. 14, no. 4, pp. 5329–5339, Dec. 2020.

[19] M. S. Mahmoud, M. M. Hamdan, and U. A. Baroudi, “Modeling and control of cyber-physical systems subject to cyber attacks: A survey of recent advances and challenges,” *Neurocomputing*, vol. 338, no. 21, pp. 101–115, April 2019.

[20] D. Ding, Q. Han, Y. Xiang, X. Ge, and X. Zhang, “A survey on security control and attack detection for industrial cyber-physical systems,” *Neurocomputing*, vol. 275, no. 31, pp. 1674–1683, Jan. 2018.

[21] J. Zhang, L. Pan, Q. Han, C. Chen, S. Wen, and Y. Xiang, “Deep learning based attack detection for cyber-physical system cybersecurity: A survey,” *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 3, pp. 377–391, Mar. 2022.

[22] A. J. Gallo, M. S. Turan, F. Boem, T. Parisini, and G. Ferrari-Trecate, “A distributed cyber-attack detection scheme with application to DC microgrids,” *IEEE Trans. Automat. Contr.*, vol. 65, no. 9, pp. 3800–3815, Sept. 2020.

[23] S. Kim, Y. Eun, and K.-J. Park, “Stealthy sensor attack detection and real-time performance recovery for resilient CPS,” *IEEE Trans. Industr. Inform.*, vol. 17, no. 11, pp. 7412–7422, Nov. 2021.

[24] W. Lucia, K. Gheitasi, and M. Ghaderi, “Setpoint attack detection in cyber-physical systems,” *IEEE Trans. Automat. Contr.*, vol. 66, no. 5, pp. 2332–2338, May 2021.

[25] W. Yan, L. K. Mestha, and M. Abbaszadeh, “Attack detection for securing cyber physical systems,” *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8471–8481, Oct. 2019.

[26] K. Huang, Z. Tao, Y. Liu, B. Sun, C. Yang, W. Gui, and S. Hu, “Adaptive multimode process monitoring based on mode-matching and similarity-preserving dictionary learning,” *IEEE Trans. Cybern.*, vol. 53, no. 6, pp. 3974–3987, June 2023.

[27] G. Intriago and Y. Zhang, “Online dictionary learning based fault and cyber attack detection for power systems,” in *Proc. IEEE PESGM*, Washington, USA, July 2021, pp. 1–5.

[28] Y. Zhang, C. Yang, K. Huang, and Y. Li, “Intrusion detection of industrial internet-of-things based on reconstructed graph neural networks,” *IEEE Trans. Netw. Sci. Eng.*, pp. 1–12, June 2022, to be published. doi:10.1109/TNSE.2022.3184975.

[29] A. Deng and B. Hooi, “Graph neural network-based anomaly detection in multivariate time series,” in *Proc. AAAI*, Virtual Conf., Feb. 2021, pp. 4027–4035.

[30] X. Ge, Q. Han, M. Zhong, and X. Zhang, “Distributed krein space-based attack detection over sensor networks under deception attacks,” *Automatica*, vol. 65, no. 9, pp. 3800–3815, Sep. 2019.

[31] F. Boem, A. J. Gallo, G. Ferrari-Trecate, and T. Parisini, “A distributed attack detection method for multi-agent systems governed by consensus-based control,” in *Proc. 56th IEEE CDC*, Melbourne, Australia, Dec. 2017, pp. 1–4.

[32] W. Duo, M. Zhou, and A. Abusorrah, “A survey of cyber attacks on cyber physical systems: Recent advances and challenges,” *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 5, pp. 784–800, May 2022.

[33] M. Pérez-Jiménez, B. B. Sánchez, A. Migliorini, and R. Alcarria, “Protecting private communications in cyber-physical systems through physical unclonable functions,” *Electronics*, vol. 8, no. 4, pp. 390–412, April 2019.

[34] J. Qin, M. Li, L. Shi, and X. Yu, “Optimal denial-of-service attack scheduling with energy constraint over packet-dropping networks,” *IEEE Trans. Automat. Contr.*, vol. 63, no. 6, pp. 1648–1663, June. 2018.

[35] H. Zhang, P. Cheng, L. Shi, and J. Chen, “Optimal denial-of-service attack scheduling with energy constraint,” *IEEE Trans. Automat. Contr.*, vol. 60, no. 11, pp. 3023–3028, Nov. 2015.

[36] C. Yang, F. Li, Z. Shi, R. Lu, and K. Choo, “A crowdsensing-based cyber-physical system for drone surveillance using random finite set theory,” *ACM Trans. CPS*, vol. 3, no. 4, pp. 1–22, Oct. 2019.

[37] M. Beard, B.-T. Vo, and B.-N. Vo, “A solution for large-scale multi-object tracking,” *IEEE Trans. Signal Process.*, vol. 68, pp. 2754–2769, April 2020.

[38] T. Nguyen, B.-N. Vo, B.-T. Vo, D. Kim, and Y. Choi, “Tracking cells and their lineages via labeled random finite sets,” *IEEE Trans. Signal Process.*, vol. 69, pp. 5611–5626, Sept. 2021.

[39] Y. Li, H. Voos, M. Darouach, and C. Hua, “An algebraic detection approach for control systems under multiple stochastic cyber-attacks,” *IEEE/CAA J. Autom. Sinica*, vol. 2, no. 3, pp. 258–266, July 2015.

[40] B.-N. Vo, B.-T. Vo, and D. Phung, “Labeled random finite sets and the Bayes multi-target tracking filter,” *IEEE Trans. Signal Process.*, vol. 62, no. 24, pp. 6554–6567, Dec. 2014.

[41] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, “A consistent metric for performance evaluation in multi-object filtering,” *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3447–3457, Aug. 2008.

[42] S. Reuter, B.-T. Vo, B.-N. Vo, and K. Dietmayer, “The labeled multi-Bernoulli filter,” *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3246–3260, June 2014.