2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Undetectable Attacks on Boolean Networks

Shiyong Zhu, Jianquan Lu, Jinde Cao, Lin Lin, James Lam, Michael Ng, and Shun-ichi Azuma

Abstract— **In this paper, driven by the ever-increasing cybersecurity threats, we study the undetectable attack problems for Boolean networks (BNs), which model distributed systems with a limited capacity of storage and bandwidth of communication. Given a consistent monitor, undetectable attacks are formalized for BNs as those do not yield an output sequence out of the nominal output sequence set. By the graph-theoretic approach, undetectable attacks are characterized by a reachability problem of a directed cycle in the augmented transition graph. On the other hand, the algebraic approach also derives a necessary and sufficient criterion for undetectable attacks by testing the existence of the nonzero elements in the constructed matrix. While all these derived results are only computationally efficient for relatively small-size BNs. The detection of attack signals is indeed NP-hard. In other words, there is no polynomial-time algorithm to check the detectability of an attack signal or an attack node set unless NP=P.**

Index Terms—**Boolean networks, Security, NP-hardness, Undetectable attacks, Algebraic state space representation.**

I. Introduction

With the ever-expanding distributed monitoring, control, and communication in the Cyber-Physical Systems (CPSs), investigations on cybersecurity have gradually become one prevalent research stream that is of both theoretical and practical significance. As hinted by a variety of recent frequent industrial security incidents, e.g., Stuxnet malware in 2010, the distributed configurations in "networked" systems inevitably bring a certain level of vulnerability from the hidden malicious adversaries. In general, CPSs can be regarded as a combination of two different functional worlds, where the behaviors of the physical layer have some dynamic features with the exploration of external environments while the basic mechanisms of the cyber layer are implemented for the targets of perception, computation, decision, and control. By their nature, the dominating threat of cybersecurity focuses on the physical layers of CPSs with dynamic features, especially inspiring a large amount of recent research interest in the security problems of dynamic systems.

Among the diverse types of models for the physical layer of CPSs, one prevalent system model is the classical linear time-invariant (LTI) system. According to different attack schemes and targets, a variety of attacks has been proposed to explore the security of CPSs modeled by such LTI systems including, but not limited to, Denial-of-Service (DoS) attacks [1], undetectable attacks [2], covert attacks [3], as well as false data injection attacks [4]. With the full knowledge of system parameters, one threatening category of attack, also of particular concern, is the undetectable attacks, which target the transmission zeros of LTI systems and, thus, hide the internal state divergence of the considered system in the null space of outputs, see [2], for example. Such an attack is also referred to as a zero-dynamics attack. In consideration of their high destructiveness, undetectable attacks have already been explored and disserted in the multiagent systems [5], and distributed optimization [6], apart from LTI systems.

Different from traditional LTI systems with accurate quantization for system states, it is impractical for such a setup because communications over networks always have limited capacities for storage and bandwidth. To overcome such issues, it is of both theoretical and practical significance to quantize the system states into finite discrete values. It reminds us of logical networks with evolving system dynamics along with a series of logical rules. Without loss of any generality in the techniques and applications, the logical networks can be reduced to those with binary variables, namely, Boolean networks (BNs), which have already been recognized to possess high potential in gene regulatory networks [7]–[10], multiagent systems [11], smart homes [12], transportation [13], and robotics [14]. To our knowledge, there are still no available results concerned with the cyber security of BNs under undetectable attacks. This partly inspires our interest in defining and developing the related results on the undetectable attacks in BNs to mitigate the security of systems in imperfect cases caused by the limited capacities of storage and communication.

To date, plentiful research on BNs mainly focuses on their dynamics and control synthesis, e.g., controllability [15], [16], observability [17]–[20], stabilization [21]–[24], decoupling problem [25], and optimal control [26], [27].

This work is supported in part by the National Natural Science Foundation of China under Grant Nos. 62273286 and 61833005, in part by the General Research Fund under Grant 17201820, in part by the Research Foundation of Yunnan Province No. 202105AF150011, and in part by the Grant-in-Aid for Transformative Research Areas (A) # 20H05969 (Molecular Cybernetics-Development of Minimal Artificial Brain by the Power of Chemistry) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Shiyong Zhu is with the Department of Systems Science, the School of Mathematics, Southeast University, Nanjing 210096, China (corresponding author, email: zhusy0904@gmail.com).

Jianquan Lu is with the Department of Systems Science, the School of Mathematics, Southeast University, Nanjing 210096, China (email: jqluma@seu.edu.cn).

Jinde Cao is with the School of Mathematics, Southeast University, Nanjing 210096, China, with the Yonsei Frontier Lab, Yonsei University, Seoul 03722, South Korea, and also with Department of Computer Science and Engineering, Yunnan University, Kunming 650091, China (e-mail: jdcao@seu.edu.cn).

Lin Lin and James Lam are with the Department of Mechanical Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong (email: linlin00wa@gmail.com; james.lam@hku.hk).

Michael Ng is with the Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong (email: mng@maths.hku.hk).

Shun-ichi Azuma is with the Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan (e-mail: sazuma@i.kyoto-u.ac.jp).

Such research enthusiasm undeniably has led to a new peak by applying the algebraic state space representation (ASSR) approach based on the semi-tensor product (STP) of matrices, see monograph [28] and the references therein. By writing each binary variable into its canonical form, the ASSR approach provides an equivalent transformation to represent the nonlinear logical systems in a linear form so that several concepts and results in classical control theory have been developed in this area. Noting that the resulting equivalent systems do not have the same topological structures, it does not possess the transmission zeros as LTI systems. This renders the existing results on the undetectable attacks in the LTI systems not applicable to this field (cf., e.g., [2], [29], [30]).

In this paper, we define the undetectable attacks on the BNs and seek to develop a series of necessary and sufficient criteria. The detailed contributions of this paper are summarized as follows:

(a) First of all, undetectable attacks are proposed for BNs to capture the dynamics of physical systems with limitations on the both capacities of storage and communications. Different from the traditional framework to address the undetectable attacks in [2], [29], and [30], several necessary and sufficient criteria are derived to check if an attacked node set is undetectable by resorting to the ASSR approach.

(b) From the graph-theoretic and matrix algebraic standpoints, respectively, these derived conditions are fairly efficient for relatively small-scale BNs.

(c) The computational complexity of judging if a BN is attacked is elaborated, for which we also verify that this problem is NP-hard. Thus, there is no algorithm in a polynomial-time amount of time subject to node number unless NP=P. We remove the detailed proofs of the main theoretic results in this version because of the page limitations and refer interested readers to our subsequent journal version.

The remainder of this paper is organized as follows. Section II defines the undetectable attacks on BNs, where several lemmas are given to serve as the cornerstone of the following research. In Section III, with the help of the ASSR technique, we capture the undetectable attacks by graph-theoretic and matrix algebraic approaches, respectively. Here, the fundamental time complexity to test our criteria is also analyzed. Finally, Section IV gives a brief concluding remark.

We end this section by defining several necessary mathematical notations. Given two integers a and b with $a<b$, $[a,b]_{\mathbb{N}}$ denotes the integers between a and b. Let $\mathscr{D}:=\{1,0\}$ and $\mathscr{D}^n:=\mathscr{D}\times\mathscr{D}\times\cdots\times\mathscr{D}$. Given a matrix $A\in\mathbb{R}^{m\times n}$, $[A]_{ij}$ is its (i,j)-th element, $\mathsf{col}_i(A)$ is the i-th column of matrix A, and $\mathsf{col}(A)$ is the set composed of all its columns. Given an integer $i\in[1,n]_{\mathbb{N}}$, δ_n^i stands for the i-th column of identity matrix I_n and Δ_n is the set $\left\{\delta_n^i|i\in[1,n]_{\mathbb{N}}\right\}$. $|S|$ denotes the number of elements in the set S. "$*$" represents the Khatri-Rao product of matrices $A\in\mathbb{R}^{m\times n}$ and $B\in\mathbb{R}^{m\times n}$, defined by

$$A*B:=\left(\mathsf{col}_1(A)\otimes\mathsf{col}_1(B),\ldots,\mathsf{col}_n(A)\otimes\mathsf{col}_n(B)\right).$$

"$\top$" is the matrix transport. The matrix $A\in\mathbb{R}^{m\times n}$ is called a logical one if it holds that $\mathsf{col}(A)\subseteq\Delta_m$, for which expression $A=[\delta_m^{i_1},\delta_m^{i_2},\ldots,\delta_m^{i_n}]$ can be simplified as $A=\delta_m[i_1,i_2,\ldots,i_n]$. Denote by $\mathscr{B}\sum$ the Boolean sum of binary variables. $\lceil\frac{m}{n}\rceil$ and $\lfloor\frac{m}{n}\rfloor$ stand for the remainder and modular of m divided by n, respectively.

II. Formulation of Undetectable Attacks

In this section, we define and study the undetectable attacks on BNs by using a consistent external monitor.

In general, the standard mathematical model of a BN with n state nodes and p output measures can be described by

$$\begin{cases}\mathsf{x}_i(t+1)=\mathsf{f}_i([\mathsf{x}_j(t)]_{j\in\mathsf{N}_i}),\ i\in[1,n]_{\mathbb{N}},\\ \mathsf{y}_k(t)=\mathsf{h}_k([\mathsf{x}_j(t)]_{j\in\mathsf{L}_k}),\ k\in[1,p]_{\mathbb{N}}\end{cases}\tag{1}$$

where $\mathsf{f}_i:\mathscr{D}^{|\mathsf{N}_i|}\to\mathscr{D}$ and $\mathsf{h}_k:\mathscr{D}^{|\mathsf{L}_k|}\to\mathscr{D}$ are Boolean functions with system states x_i, $i\in[1,n]_{\mathbb{N}}$ as their variables. $y_k\in\mathscr{D}$, $k\in[1,p]_{\mathbb{N}}$ are output measures. $\mathsf{x}(t):=[\mathsf{x}_1(t),\mathsf{x}_2(t),\ldots,\mathsf{x}_n(t)]^\top\in\mathscr{D}^n$ and $\mathsf{y}(t):=[\mathsf{y}_1(t),\mathsf{y}_2(t),\ldots,\mathsf{y}_p(t)]^\top\in\mathscr{D}^p$ stand for the compact representation of system state and output measures, respectively. Noting that BN (1) does not suffer from malicious attacks, it is referred to as the *nominal* one.

For the BN (1) attacked by malicious adversaries, we use $\mathfrak{A}\subseteq[1,n]_{\mathbb{N}}$ to represent the set of attacked nodes, i.e., the targeted nodes suffering from attacks, $\mathfrak{a}_i:t\mapsto\mathscr{D}$, $i\in\mathfrak{A}$ to describe the false data signals, and $\boxtimes_i$, $i\in\mathfrak{A}$ to denote the operators for attackers to inject the false data signals. The *attacked* BN is written by

$$\begin{cases}\mathfrak{x}_i(t+1)=\mathsf{f}_i([\mathfrak{x}_j(t)]_{j\in\mathsf{N}_i})\boxtimes_i\mathfrak{a}_i(t),\ i\in\mathfrak{A},\\ \mathfrak{x}_i(t+1)=\mathsf{f}_i([\mathfrak{x}_j(t)]_{j\in\mathsf{N}_i}),\ i\in[1,n]_{\mathbb{N}}\backslash\mathfrak{A},\\ \mathfrak{y}_k(t)=\mathsf{h}_k([\mathfrak{x}_j(t)]_{j\in\mathsf{L}_k}),\ k\in[1,p]_{\mathbb{N}}\end{cases}\tag{2}$$

wherein the attack data signals toward BN (1) is denoted by $(\boxtimes_i,\mathfrak{a}_i(t))|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$. The state trajectory of nominal BN (1) starting from initial state $\mathsf{x}(0)$ is denoted by

$$\mathsf{x}(\mathsf{x}(0),t)|_{t\in\mathbb{N}^+}:=\{\mathsf{x}(0),\ldots,\mathsf{x}(n),\ldots\}$$

and that of BN (2) starting from initial state $\mathfrak{x}(0)$ under attack is

$$\mathfrak{x}(\mathfrak{x}(0),\mathfrak{a}_i,t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}:=\{\mathfrak{x}(0),\ldots,\mathfrak{x}(n,\mathfrak{a}(n-1)),\ldots\}$$

where $\mathfrak{a}(t):=\left[\mathfrak{a}_1(t),\mathfrak{a}_2(t),\ldots,\mathfrak{a}_{|\mathfrak{A}|}(t)\right]^\top$.

Following the monitors for LTI systems proposed in [2], we define the monitors in a similar manner to judge if the considered BN is attacked via model dynamics and outputs. Formally, a monitor, denoted by χ, is a deterministic algorithm with the input data $F=(\mathsf{f}_i|_{i\in[1,n]_{\mathbb{N}}},\mathsf{h}_i|_{i\in[1,p]_{\mathbb{N}}},y_i(t)|_{i\in[1,p]_{\mathbb{N}},t\in\mathbb{N}^+})$ to output "True" and "False", which indicate that the BN is attacked and unattacked, respectively.

Definition 2.1 (see [2]): Consider BN (2) with attack node set $\mathfrak{A}$. A monitor, denoted by $\chi(F):=(\chi_1(F),\chi_2(F))$,

is composed of two parts, wherein $\chi_1(F)$ captures the existence of attackers and $\chi_2(F)$ identifies the attack set $\mathfrak{A}$:

(a) $\chi_1(F) = \mathsf{True}$ only if $\mathfrak{A} \neq \emptyset$ and $(\boxtimes_i, \mathfrak{a}_i)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ changes the dynamics of BN (1); and $\chi_1(F) = \mathsf{False}$, otherwise.

(b) $\chi_1(F) = \mathsf{False}$ if and only if $\chi_2(F) = \emptyset$.

(c) $\chi_2(F) = \mathfrak{A}$ only if the set $\mathfrak{A}$ is the unique and smallest set $S \subseteq [1,n]_{\mathbb{N}}$ such that $y_i(t)|_{i\in[1,p]_{\mathbb{N}},t\in\mathbb{N}^+} = \mathsf{y}(\mathsf{x}', \mathfrak{a}_i, t)|_{i\in S,t\in\mathbb{N}^+}$ for a certain state $\mathsf{x}' \in \mathscr{D}^n$.

In this paper, we are only interested in the "consistent" monitor χ_1, i.e., for inputs $F_1 = F_2$, it holds that $\chi_1(F_1) = \chi_1(F_2)$.

Different from the conditions for monitors in the work of [2], we also require in condition (a) of Definition 2.1 that attackers endowed with $(\boxtimes_i, \mathfrak{a}_i(t))|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ modify the state trajectories of BN (1), i.e., the considered attack is *meaningful*; otherwise, *meaningless*. This excludes the situation where the state trajectories of attacked BN (2) starting from all initial states are not affected by attack signals, due to the specialty of logical operators.

Remark 2.1: We can simplify the attack operator "$\boxtimes_i$" to be XOR without loss of any generality, as operator XOR is enough to steer the arbitrary state trajectory of attacked nodes in BN (2) to any desired one by injecting the elaborate attack signals $\mathfrak{a}_i(t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$. Therefore, in the following analysis part, we can also stipulate $\boxtimes_i$, $i \in \mathfrak{A}$ are all XOR operators.

Definition 2.2: Consider a BN (2) with an attack node set $\mathfrak{A}$. The attack signal $(\boxtimes_i, \mathfrak{a}_i(t))|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ is said to be detectable (resp., identifiable) if $\chi_1(F) = \mathsf{True}$ (resp., $\chi_2(F) = \mathfrak{A}$). Furthermore, the attack set $\mathfrak{A}$ is said to be detectable if all attack signals $(\boxtimes_i, \mathfrak{a}_i(t))_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ are detectable.

Lemma 2.1: The attack signal $(\boxtimes_i, \mathfrak{a}_i(t))|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ on BN (2) is undetectable if and only if $\mathsf{x}(\mathfrak{x}_0, \tau) \neq \mathfrak{x}(\mathfrak{x}_0, \mathfrak{a}, \tau)$ and $\mathsf{y}(\mathsf{x}_0, t) = \mathfrak{y}(\mathfrak{x}_0, \mathfrak{a}, t)$ hold for certain initial states $\mathsf{x}_0, \mathfrak{x}_0 \in \mathscr{D}^n$, certain $\tau \in \mathbb{N}^+$, and all $t \in \mathbb{N}^+$.

Proof: (If) For the inputs

$$F_1 = \left\{\mathsf{f}_i|_{i\in[1,n]_{\mathbb{N}}}, \mathsf{h}_i|_{i\in[1,p]_{\mathbb{N}}}, \mathsf{y}(\mathsf{x}_0,t)|_{t\in\mathbb{N}^+}\right\}$$

and

$$F_2 = \left\{\mathsf{f}_i|_{i\in[1,n]_{\mathbb{N}}}, \mathsf{h}_i|_{i\in[1,p]_{\mathbb{N}}}, \mathfrak{y}(\mathfrak{x}_0, \mathfrak{a}_i, t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}\right\}$$

with certain initial states x_0 and $\mathfrak{x}_0$, one has that $\chi_1(F_1) = \chi_1(F_2) = \mathsf{False}$, because the considered monitor is a consistent one. Moreover, the condition $\mathsf{x}(\mathsf{x}_0, \tau) \neq \mathfrak{x}(\mathfrak{x}_0, \mathfrak{a}, \tau)$ implies that the attack signal $(\boxtimes_i, \mathfrak{a}_i(t))|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ is undetectable.

(Only if) Assuming that $\mathsf{x}(\mathfrak{x}_0,t)|_{t\in\mathbb{N}^+} = \mathfrak{x}(\mathfrak{x}_0, \mathfrak{a}_i, t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ holds for all initial states $\mathfrak{x}_0 \in \mathscr{D}^n$ under the attack signal $\mathfrak{a}_i(t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$, then this attack signal $\mathfrak{a}_i(t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ is meaningless. On the other hand, if $\mathsf{y}(\mathsf{x}_0,t) \neq \mathfrak{y}(\mathfrak{x}_0, \mathfrak{a}, t)$ for any x_0 and $\mathfrak{x}_0$, then this attack signal is detectable as it contradicts with conditions (b) and (c) in Definition 2.1 no matter whether the attack signal $\mathfrak{a}_i(t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ is meaningful. ■

Remark 2.2: The most essential difference of Lemma 2.1 from Lemma 3.1 in the work of [2] is the condition $\mathsf{x}(\mathfrak{x}_0, \tau) \neq \mathfrak{x}(\mathfrak{x}_0, \mathfrak{a}_i, \tau)|_{i\in\mathfrak{A}}$ for a certain initial state $\mathfrak{x}_0$ and a certain τ. To be specific, by the nature of BN (1), if $\mathfrak{A} \neq \emptyset$ and $\boxtimes = \otimes$, injecting attack signal $\mathfrak{a}_i(t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+} = \mathsf{x}_i(t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$, will not modify the state trajectories of BN (1) from any initial state. Such an attack signal is meaningless as it cannot modify the attacked system dynamics albeit the attack node set $\mathfrak{A} \neq \emptyset$. Such attack signals are excluded in this paper.

Remark 2.3: Although the attack detection in Definition 2.1 is dependent on the consistent monitors, it is obvious from Lemma 2.1 that the detectability of attack signal $(\boxtimes_i, \mathfrak{a}_i(t))|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ only relies on the attack signals and networks.

Remark 2.4: The concept of attack detection is related to fault detection in [31]. While the faults in BNs occur by accident, the attacks here are deliberate and designable.

The detection of attack signals does not rely on the conventional observability or detectability of BN (1), which requires the distinguishability of different initial or current states for BN (1). This claim is evident from the following BN counter with three nodes.

Example 2.1: Consider the three-node BN counter with output measure $\mathsf{y}_1(t) = \mathsf{x}_1(t) \wedge \mathsf{x}_2(t) \wedge \mathsf{x}_3(t)$. By the graph construction introduced in Section III, the state transition graph of this counter is depicted in Fig. 1, where the green vertex stands for the state outputting variable 1 and the white vertices describe the ones outputting variables 0. By Theorem 1 in [15], one can readily conclude that this is an observable BN and, therefore, is also detectable.

Subsequently, we demonstrate that the attacked node set $\mathfrak{A} = \{1\}$ is undetectable. We can set $\mathsf{x}_0 = \mathfrak{x}_0 = (1,1,1)^\top$ with attack signal $\mathfrak{a}_1(t)|_{t\in\mathbb{N}^+} = \{1,0,0,0,0,0,1,1,1,1,0,\dots\}$, in which case $\mathsf{y}(\mathsf{x}_0,t) = \mathfrak{y}(\mathfrak{x}_0, \mathfrak{a}(t), t) = \{1,0,0,0,0,0,0,0,1,\dots\}$.

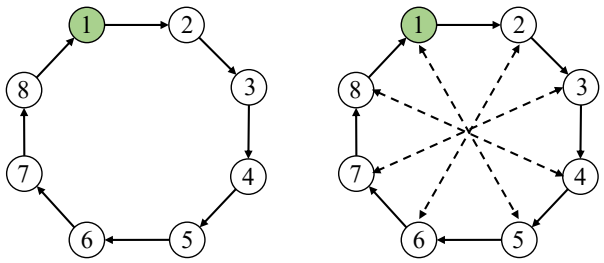

Fig. 1. (a) State transition graph of an observable and detectable BN counter. (b) State transition graph of BN counter with an attacked node set $\mathfrak{A} = \{1\}$. Vertices labeled i, $i \in [1,8]_{\mathbb{N}}$ stand for the state (i_1, i_2, i_3) with $(1-i_1)2^2 + (1-i_2)2^1 + (1-i_3)2^0 + 1 = i$. The solid and dashed edges, respectively, denote transitions corresponding to meaningful attack signals and meaningless ones.

On the other hand, while the output matrix of the counter is $y_1(t) = x_2(t) \wedge x_3(t)$, this counter is unobservable. In Fig. 2, it claims that the attack node set $\mathfrak{A} = \{1\}$ is undetectable with $\mathsf{x}_0 = (1,1,0)^\top$ and $\mathfrak{x}_0 = (0,0,1)^\top$ attacked by signal $\mathfrak{a}_1(t)|_{t\in\mathbb{N}^+} = \{1,0,0,1,1,1,1,0,0,0,0,1,\dots\}$.

From this counter, we demonstrate that, no matter whether BN (1) is observable or unobservable and detectable or undetectable, the network is always risky to be attacked.

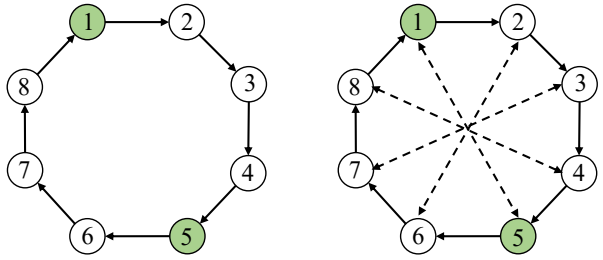

Fig. 2. (a) State transition graph of an unobservable and undetectable BN counter. (b) State transition graph of this BN counter with attack node set $\mathfrak{A}=\{1\}$.

III. Analysis of Undetectable Attacks

In this section, we shall develop a series of necessary and sufficient criteria for undetectable attacks with the help of the STP of matrices. To this end, the ASSR of BNs, developed in [28], is briefly reviewed here with the STP of matrices.

To begin with, as an extension of the normal matrix product to break through the requirement of dimensional matching in matrix product, the STP of matrices can be defined as follows:

Definition 3.1 (see [28]): Given two matrices $M\in\mathbb{R}^{p\times q}$ and $N\in\mathbb{R}^{s\times t}$, their STP is defined by

$$M\ltimes N:=(M\otimes I_{\iota/q})(N\otimes I_{\iota/s})$$

where ι is the least common multiple of integers q and s.

To cope with the nonlinear operators in logical dynamics, we define the bijective correspondence "$\sim$" between binary variables and its canonical form by $\iota\sim\delta_2^{2-\iota}$ with variable $\iota\in\mathscr{D}$. Such correspondence can be naturally extended to the mapping between $\mathscr{D}^n$ and Δ_{2^n} by $[\mathsf{x}_1,\mathsf{x}_2,\dots,\mathsf{x}_n]^\top\sim\delta_2^{2-\mathsf{x}_1}\ltimes\delta_2^{2-\mathsf{x}_2}\ltimes\cdots\ltimes\delta_2^{2-\mathsf{x}_n}$. By defining the canonical form of Boolean variables, we can convert arbitrary logical functions into their multilinear form by resorting to the STP of matrices.

Lemma 3.1 (see [28]): For arbitrary n-ary logical function $f(\mathsf{x}_1,\mathsf{x}_2,\dots,\mathsf{x}_n):\mathscr{D}^n\to\mathscr{D}$, there exists a logical matrix $M_f\in\mathscr{L}_{2\times 2^n}$, called the structure matrix of f, such that

$$f(\mathsf{x}_1,\mathsf{x}_2,\dots,\mathsf{x}_n)\sim M_f\ltimes_{i=1}^n x_i$$

where $\ltimes_{i=1}^n x_i:=x_1\ltimes x_2\ltimes\cdots\ltimes x_n$ with $x_i\sim\mathsf{x}_i$.

Other than breaking the dimensional compatibility rule in matrix multiplication, the STP of matrices also has some superior properties to the traditional matrix product. These properties are briefly provided as follows.

Lemma 3.2 (see [28]): The following calculation properties hold for the STP of matrices with $A\in\mathbb{R}^{m\times n}$, $v\in\Delta_n$ and $u\in\Delta_m$: (a) $vA=(I_n\otimes A)v$; (b) $uv=W_{[n,m]}vu$; (c) $v=M_d^l uv$; (d) $v\ltimes v=M_r\ltimes v$, where $W_{[n,m]}:=I_m\otimes I_n$, $M_d^l:=\mathbf{1}_m^\top\otimes I_n$, and $M_r:=\left[\delta_n^1\ltimes\delta_n^1,\delta_n^2\ltimes\delta_n^2,\dots,\delta_n^n\ltimes\delta_n^n\right]$.

By Lemma 3.1, we can write each Boolean variable into its canonical form. On the other hand, using the properties in Lemma 3.2 and defining $\mathsf{x}:=\ltimes_{i=1}^n x_i$, $\mathsf{y}:=\ltimes_{i=1}^p y_i$, $\mathfrak{x}:=\ltimes_{i=1}^n\mathfrak{x}_i$, $\mathfrak{y}:=\ltimes_{i=1}^p\mathfrak{y}_i$ and $\mathfrak{a}=\ltimes_{i\in\mathfrak{A}}\mathfrak{a}_i$, we can write the ASSR of BN (1) and attacked BN (2), respectively, as

$$\begin{cases}\mathsf{x}(t+1)=\mathsf{F}\ltimes\mathsf{x}(t),\\ \quad\mathsf{y}(t)=\mathsf{H}\ltimes\mathsf{x}(t)\end{cases}\tag{3}$$

with $\mathsf{F}:=\mathsf{F}_1*\mathsf{F}_2*\cdots*\mathsf{F}_n$, and

$$\begin{cases}\mathfrak{x}(t+1)=\mathfrak{F}_{\boxtimes}\ltimes\mathfrak{x}(t)\ltimes\mathfrak{a}(t),\\ \quad\mathfrak{y}(t)=\mathsf{H}\ltimes\mathfrak{x}(t)\end{cases}\tag{4}$$

with $\mathfrak{F}_{\boxtimes}:=\mathsf{F}_1^{\mathfrak{a}}*\mathsf{F}_2^{\mathfrak{a}}*\cdots*\mathsf{F}_n^{\mathfrak{a}}$ and $\mathsf{H}:=\mathsf{H}_1*\mathsf{H}_2*\cdots*\mathsf{H}_p$, where $\mathsf{x}\in\Delta_{2^n}$, $\mathfrak{x}\in\Delta_{2^n}$ and $\mathfrak{a}\in\Delta_{2^{|\mathfrak{A}|}}$, F_i, H_i and $\mathsf{F}^{\mathfrak{a}}$ are the structure matrices of f_i, h_i and $\mathsf{f}_i\boxtimes_i\mathfrak{a}_i$, respectively.

A. Graph-Theoretic Criteria

In this subsection, we first characterize the undetectable attack set $\mathfrak{A}$ on the basis of state transition graphs of BNs, which can be built from the network transition matrix F in BN (3). Notice that the system matrix F in BN (3) is a Boolean matrix, thus it is associated with a directed graph G, written by an ordered pair $G:=(V,E)$, where $V:=\{1,2,\dots,2^n\}$ is the vertex set of this constructed directed graph, and a solid directed edge (i,j) joining i to j exists in the directed graph G, i.e., $(i,j)\in E$ if and only if $\mathsf{F}_{ji}=1$. In a similar manner, the transition matrix $\mathfrak{F}_{\boxtimes}$ can be split into $2^{|\mathfrak{A}|}$ blocks as $\mathfrak{F}_{\boxtimes}=\left[(\mathfrak{F}_{\boxtimes})_1,(\mathfrak{F}_{\boxtimes})_2,\dots,(\mathfrak{F}_{\boxtimes})_{2^{|\mathfrak{A}|}}\right]$ in accordance with the different attack signals $\mathfrak{a}$, where each submatrix $(\mathfrak{F}_{\boxtimes})_i$, $i\in[1,2^{|\mathfrak{A}|}]$ is also Boolean. Therefore, we can build $2^{|\mathfrak{A}|}$ labeled state transition graphs $G_i^{\mathfrak{a}}=(V,E_i,l_i)$, which are associated with a dashed edge set in set $E_i\backslash E$ and solid edges in set $E_i\cap E$. In this way, the labeled state transition of attacked BN (2) is jointly represented by

$$G^{\mathfrak{a}}:=\bigcup_{i=1}^{2^{|\mathfrak{A}|}}G_i^{\mathfrak{a}}=\left(V,\bigcup_{i=1}^{2^{|\mathfrak{A}|}}E_i,\bigcup_{i=1}^{2^{|\mathfrak{A}|}}l_i\right).$$

As a consequence, the state transition information of BNs (3) and (4) are, equivalently, characterized by these two directed graphs G and $G^{\mathfrak{a}}$. Hereafter, they are referred to as the state transition graph of BN (3) and labeled state transition graph of BN (4), i.e., the graphic representation of BNs.

Finally, we color the vertices of these two graphs in accordance with the output measures satisfying that the colors of vertices i and j are the same if and only if $\mathsf{col}_i(\mathsf{H})=\mathsf{col}_j(\mathsf{H})$.

In what follows, we focus on their state transition graphs and proceed to explore the undetectable attacks of BNs.

Theorem 3.1: Attack signal $(\boxtimes_i,\mathfrak{a}_i(t))|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ in BN (2) is undetectable if and only if there exists a vertex v_o in the state transition graph $G^{\mathfrak{a}}$ such that the color sequence $\mathscr{C}$ of BN (2), driven by attack signal $\mathfrak{a}_i(t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$, satisfies that

(a) sequence $\mathscr{C}$ shapes a periodic sequence with length l, denoted by $\{c_1,c_2,\dots,c_l\}$, after T-step transitions;
(b) there is a directed cycle $C=\{q_1,q_2,\dots,q_l\}$ in the state transition graph G with the same color sequence $\{c_1,c_2,\dots,c_l\}$;
(c) for the prefix T of output color sequence, there is a vertex v_o' to q_1 in the state transition graph G with the

same color as the first T colors in the set $\mathscr{C}$;

(d) the directed path driven by signal $\mathfrak{a}_i(t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$ from the initial vertex v_0 contains a dashed arc.

Having roughly characterized the undetectable attack signal $\mathfrak{a}_i(t)|_{i\in\mathfrak{A},t\in\mathbb{N}^+}$, we proceed to concern the undetectable attack node set.

Theorem 3.2: Attack node set $\mathfrak{A}$ in BN (2) is undetectable if and only if state transition graphs G and $G^{\mathfrak{a}}$ contain the same colored cycles and the same colored prefix chain starting from a dashed edge.

Arguably, even if the above propositions capture the undetectable attacks necessarily and sufficiently, they are not easily checkable. This inspires us to follow the approach in [32] by augmenting the state transition graphs G and $G^{\mathfrak{a}}$ to develop a more computationally checkable algorithm.

Here, the augmented directed graph $\mathsf{G}=(\mathsf{V},\mathsf{E})$ combines the vertex set with its duplicate to get a vertex pair (i,j), i.e., $\mathsf{V}=V\times V$. As for the edge set E, $((i,j),(s,t))\in\mathsf{E}$ if and only if $(i,s)\in G$ and $(j,t)\in G^{\mathfrak{a}}$ hold and pair (i,j) and (s,t) are the same colored. Moreover, it is labeled if and only if $(j,t)\notin\mathsf{G}$.

Theorem 3.3: Attack node set $\mathfrak{A}$ in BN (2) is undetectable if and only if there is a directed cycle in the augmented graph G reachable from a vertex v at the tail of a labeled edge.

Remark 3.1: Finally, we analyze the total time complexity of the above graph-theoretic criteria. For BN (4) with 2^n state variables, $2^{|\mathfrak{A}|}$ attack variables, and 2^p output variables, it takes time $O\left(2^{2n+|\mathfrak{A}|+p}\right)$ to establish the corresponding augmented directed graph G. On the other hand, we also spend $O(2^{2n+|\mathfrak{A}|})$ time to check the directed cycles with the same color and time $O(2^{2n+|\mathfrak{A}|+p})$ on the depth-first search to find the reachability from a certain labeled edge. Therefore, the total time complexity is $O(2^{2n+|\mathfrak{A}|+p})$ for Theorem 3.3, which essentially can only be applied to relatively small-size BNs.

B. Algebraic Criteria

In the above subsection, we have established several necessary and sufficient conditions for BNs on the basis of their state transition graphs. In the sequel, we shall develop an algebraic approach to deal with the undetectable attack node set $\mathfrak{A}$ in BN (4) with relatively small sizes.

Towards this end, we first define the augmented variables $\mu(t)=\mathsf{x}(t)\ltimes\mathfrak{x}(t)$ and $\nu(t)=y(t)\ltimes\mathfrak{y}(t)$ with their dynamics, respectively, being governed by

$$\begin{aligned}\mu(t+1)&=\mathsf{x}(t+1)\ltimes\mathfrak{x}(t+1)\\&=\mathsf{F}\mathsf{x}(t)\mathfrak{F}_{\boxtimes}\mathfrak{x}(t)\mathfrak{a}(t)\\&=\mathsf{F}(I_{2^n}\otimes\mathfrak{F}_{\boxtimes})\mathsf{x}(t)\mathfrak{x}(t)\mathfrak{a}(t)\\&=\mathsf{F}(I_{2^n}\otimes\mathfrak{F}_{\boxtimes})\mu(t)\mathfrak{a}(t):=\mathfrak{L}\mu(t)\mathfrak{a}(t)\end{aligned}\tag{5}$$

and

$$\begin{aligned}\nu(t)&=y(t)\ltimes\mathfrak{y}(t)\\&=\mathsf{H}\mathsf{x}(t)\ltimes\mathsf{H}\mathfrak{x}(t)\\&=\mathsf{H}(I_{2^p}\otimes\mathsf{H})\mathsf{x}(t)\mathfrak{x}(t):=\mathfrak{K}\mu(t)\end{aligned}\tag{6}$$

where $\mathfrak{L}:=\mathsf{F}(I_{2^n}\otimes\mathfrak{F}_{\boxtimes})$ and $\mathfrak{K}:=\mathsf{H}(I_{2^p}\otimes\mathsf{H})$.

With this algebraic expression, the following two sets are defined as

$$\mathfrak{S}:=\left\{\delta_{2^{2n}}^{j}|_{\mathscr{B}}\sum_{i=1}^{2^p}\mathfrak{K}_{(i-1)2^p+i,j}=1\right\}$$

and

$$\mathfrak{T}:=\left\{\delta_{2^{2n}}^{(i-1)2^n+i}|i\in[1,2^n]_{\mathbb{N}}\right\}$$

where the above set $\mathfrak{S}$ denotes the multiply of states $\mathsf{x}(t)$ of BN (1) and $\mathfrak{x}(t)$ of BN (2) attacked by the node set $\mathfrak{A}$ with the same output measures, and set $\mathfrak{T}$ stands for the multiply of system state $\mathsf{x}(t)$ and its duplicate. Both sets play different roles in this part of the study. To be specific, the set $\mathfrak{S}$ is used for the undetectable outputs, and the set $\mathfrak{T}$ is imported for the meaningful attack signals.

According to condition $\mathsf{y}(\mathsf{x}_0,t)|_{t\in\mathbb{N}^+}=\mathfrak{y}(\mathfrak{x}_0,\mathfrak{a},t)|_{t\in\mathbb{N}^+}$ in Lemma 2.1, the invariance of state trajectory in set $\mathfrak{S}$ is necessary. Subsequently, the attack-invariant subset of set $\mathfrak{S}$ is defined and adapted from the control-invariant subset in the work of [33].

Definition 3.2: Given $S\subseteq\Delta_{2^{2n}}$, the subset $S^*\subseteq S$ is said to be attack-invariant with respect to (5) if for any $\delta_{2^{2n}}^{k}\in S^*$, there exists an attack signal $\delta_{2^{|\mathfrak{A}|}}^{a_k}$ such that $\mathfrak{L}\ltimes\delta_{2^{|\mathfrak{A}|}}^{a_k}\ltimes\delta_{2^{2n}}^{k}\in S^*$.

In the work of [33], the largest invariant subset of a given set S is calculated with a concise matrix deduction. We adopt this lemma to deal with the largest attack-invariant subset of BN (5).

Denote the indicator matrix $D_{\mathfrak{I}}$ subject to a given subset $\mathfrak{I}\subseteq\Delta_{2^{2n}}$ as follows:

$$\mathsf{col}_i(D_{\mathfrak{I}})=\begin{cases}\delta_{2^{2n}}^{i},\ \delta_{2^{2n}}^{i}\in\mathfrak{I},\\ \delta_{2^{2n}}^{0},\ \delta_{2^{2n}}^{i}\notin\mathfrak{I}.\end{cases}$$

Lemma 3.3: Given a set $\mathfrak{S}\subseteq\Delta_{2^{2n}}$, the largest attack-invariant subset of set $\mathfrak{S}$ can be calculated by

$$\mathfrak{S}^*=\{\beta\in\Delta_{2^{2n}}\mid\beta+_{\mathscr{B}}\alpha=\alpha\},$$

where

$$\alpha:=\mathbf{1}_{2^{2n}}^{\top}\ltimes_{\mathscr{B}}\left[\left(D_{\mathfrak{S}}\ltimes\mathfrak{L}\ltimes\mathbf{1}_{2^{|\mathfrak{A}|}}\right)^{|\mathfrak{S}|}D_{\mathfrak{S}}\right].$$

In the sequel, we are in a position to build a necessary and sufficient condition from an algebraic view of points for the undetectable attack node set $\mathfrak{A}$ in BN (2).

Theorem 3.4: The attack set $\mathfrak{A}$ in BN (2) is undetectable if and only if matrix

$$D_B\ltimes_{\mathscr{B}}\mathfrak{L}\left(I_{2^{2n+|\mathfrak{A}|}}+_{\mathscr{B}}\mathsf{W}_{\left[2^{|\mathfrak{A}|},2^n\right]}\left(I_{2^{|\mathfrak{A}|}}\otimes M_rM_d^l\right)\right)\ltimes_{\mathscr{B}}D_{\mathfrak{S}^*}\tag{7}$$

is not a zero matrix, where $B=\mathfrak{I}^c\cup\mathfrak{S}^*$ and

$$\mathfrak{S}^*=\{\beta\in\Delta_{2^{2n}}\mid\beta+_{\mathscr{B}}\alpha=\alpha\}.$$

C. Complexity of Attack Detection

As mentioned in the above subsections, although all the above results from the graph-theoretic approach or algebraic approach seem theoretically perfect as necessary and sufficient criteria, their time complexity would exponentially increase with respect to the number of nodes in BN (2).

Therefore, they are merely suitable for the detection of attacked nodes set on a relatively small-size BN. We seek to analyze the NP-hardness of detecting the attack set for BN (1), in order to conclude if there exists an algorithm with polynomial-time complexity to detect the attack set $\mathfrak{A}$ for general BNs. In this version, we directly conclude the following theorem.

Theorem 3.5: The problem of detecting if BN (2) is attacked is NP-hard.

IV. Conclusion

In this paper, we have investigated the undetectable attacks in BNs. By the ASSR approach, a series of necessary and sufficient criteria for the undetectable attack set has been developed for BNs from the graphic and algebraic views of points, respectively. However, noting that the developed criteria are endowed with exponentially increasing time complexity subject to node number, these results only suit relatively small-size BNs. Moreover, we also dissect the feasibility of building a polynomial-time algorithm for general large-scale BNs. However, the detectability of the attack node set in has been shown to be NP-hard. The theoretical results have been also validated in the simulating part to conduct the effectivity, while the detailed proofs will appear in our future journal version.

References

[1] S. Feng, A. Cetinkaya, H. Ishii, P. Tesi, and C. De Persis, "Networked control under dos attacks: Tradeoffs between resilience and data rate," *IEEE Transactions on Automatic Control*, vol. 66, no. 1, pp. 460–467, 2020.

[2] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[3] A. O. de Sá, L. F. R. da Costa Carmo, and R. C. Machado, "Covert attacks in cyber-physical control systems," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1641–1651, 2017.

[4] J. Miloševič, T. Tanaka, H. Sandberg, and K. H. Johansson, "Analysis and mitigation of bias injection attacks against a kalman filter," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 8393–8398, 2017.

[5] 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 *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 5961–5966, IEEE, 2017.

[6] L. An and G.-H. Yang, "Distributed sparse undetectable attacks against state estimation," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 1, pp. 463–473, 2021.

[7] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, 1969.

[8] S.-i. Azuma, T. Yoshida, and T. Sugie, "Structural monostability of activation-inhibition Boolean networks," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 2, pp. 179–190, 2017.

[9] S. Zhu, J. Cao, L. Lin, J. Lam, and S.-i. Azuma, "Toward stabilizable large-scale boolean networks by controlling the minimal set of nodes," *IEEE Transactions on Automatic Control*, to be published, doi: 10.1109/TAC.2023.3269321.

[10] S. Zhu, J. Lu, S.-i. Azuma, and W. X. Zheng, "Strong structural controllability of Boolean networks: Polynomial-time criteria, minimal node control, and distributed pinning strategies," *IEEE Transactions on Automatic Control*, vol. 68, no. 9, pp. 5461–5476, 2023.

[11] Y. Wang, C. Zhang, and Z. Liu, "A matrix approach to graph maximum stable set and coloring problems with application to multi-agent systems," *Automatica*, vol. 48, no. 7, pp. 1227–1236, 2012.

[12] M. H. Kabir, M. R. Hoque, B.-J. Koo, and S.-H. Yang, "Mathematical modelling of a context-aware system based on Boolean control networks for smart home," in *The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014)*, pp. 1–2, IEEE, 2014.

[13] K. E. Haynes, R. G. Kulkarni, L. A. Schintler, and R. R. Stough, "Intelligent transportation system (its) management using Boolean networks," *Spatial Dynamics, Networks and Modelling*, pp. 121–138, 2006.

[14] M. Braccini, A. Roli, E. Barbieri, and S. A. Kauffman, "On the criticality of adaptive Boolean network robots," *Entropy*, vol. 24, no. 10, p. 1368, 2022.

[15] D. Laschov and M. Margaliot, "Controllability of Boolean control networks via the Perron-Frobenius theory," *Automatica*, vol. 48, no. 6, pp. 1218–1223, 2012.

[16] E. Weiss, M. Margaliot, and G. Even, "Minimal controllability of conjunctive Boolean networks is NP-complete," *Automatica*, vol. 92, pp. 56–62, 2018.

[17] Y. Guo, "Observability of Boolean control networks using parallel extension and set reachability," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 12, pp. 6402–6408, 2018.

[18] K. Zhang and K. H. Johansson, "Efficient verification of observability and reconstructibility for large Boolean control networks with special structures," *IEEE Transactions on Automatic Control*, vol. 65, no. 12, pp. 5144–5158, 2020.

[19] E. Fornasini and M. E. Valcher, "Observability, reconstructibility and state observers of boolean control networks," *IEEE Transactions on Automatic Control*, vol. 58, no. 6, pp. 1390–1401, 2012.

[20] S. Zhu, J. Cao, L. Lin, L. Rutkowski, J. Lu, and G. Lu, "Observability and detectability of stochastic labeled graphs," *IEEE Transactions on Automatic Control*, to be published, doi: 10.1109/TAC.2023.3278797.

[21] R. Li, M. Yang, and T. Chu, "State feedback stabilization for Boolean control networks," *IEEE Transactions on Automatic Control*, vol. 58, no. 7, pp. 1853–1857, 2013.

[22] N. Bof, E. Fornasini, and M. E. Valcher, "Output feedback stabilization of Boolean control networks," *Automatica*, vol. 57, pp. 21–28, 2015.

[23] M. Meng, J. Lam, J.-E. Feng, and K. C. Cheung, "Stability and stabilization of Boolean networks with stochastic delays," *IEEE Transactions on Automatic Control*, vol. 64, no. 2, pp. 790–796, 2018.

[24] M. Meng, J. Lam, J.-E. Feng, and K. C. Cheung, "Stability and guaranteed cost analysis of time-triggered Boolean networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3893–3899, 2018.

[25] D. Cheng and H. Qi, "Controllability and observability of Boolean control networks," *Automatica*, vol. 45, no. 7, pp. 1659–1667, 2009.

[26] Y. Wu and T. Shen, "A finite convergence criterion for the discounted optimal control of stochastic logical networks," *IEEE Transactions on Automatic Control*, vol. 63, no. 1, pp. 262–268, 2017.

[27] D. Laschov and M. Margaliot, "Minimum-time control of Boolean networks," *SIAM Journal on Control and Optimization*, vol. 51, no. 4, pp. 2869–2892, 2013.

[28] D. Cheng, H. Qi, and Z. Li, *Analysis and Control of Boolean Networks: A Semi-Tensor Product Approach*. London, U.K.: Springer-Verlag, 2011.

[29] Y. Chen, S. Kar, and J. M. Moura, "Dynamic attack detection in cyber-physical systems with side initial state information," *IEEE Transactions on Automatic Control*, vol. 62, no. 9, pp. 4618–4624, 2016.

[30] J. Milošević, A. Teixeira, K. H. Johansson, and H. Sandberg, "Actuator security indices based on perfect undetectability: Computation, robustness, and sensor placement," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3816–3831, 2020.

[31] E. Fornasini and M. E. Valcher, "Fault detection analysis of Boolean control networks," *IEEE Transactions on Automatic Control*, vol. 60, no. 10, pp. 2734–2739, 2015.

[32] D. Laschov, M. Margaliot, and G. Even, "Observability of Boolean networks: A graph-theoretic approach," *Automatica*, vol. 49, no. 8, pp. 2351–2362, 2013.

[33] Y. Guo, P. Wang, W. Gui, and C. Yang, "Set stability and set stabilization of Boolean control networks based on invariant subsets," *Automatica*, vol. 61, pp. 106–112, 2015.