Secure Control: Intrusion Detection and Identification

Yilin Mo

July 2, 2015

1 Fault Detection and Identification

1.1 Detection

Consider the following system:

$$x(k+1) = Ax(k) + Bu(k),\ y(k) = Cx(k). \tag{1}$$

We assume that (A, C) is observable, B is full column rank. Suppose that $u(k)$ is the fault signal. We will say that a fault occurs when $u(k) \neq 0$ for some k.

Define $\mathcal{Y} = (y(0), y(1), \dots)$ and $\mathcal{U} = (u(0), u(1), \dots)$. Clearly, $\mathcal{Y}$ is a function of $x(0)$ and $\mathcal{U}$. Thus, we will write

$$y = f(x(0), \mathcal{U}).$$

Fact: f is a linear operator.

Question: Can we know whether a fault occurs from $y(k)$?

There are two cases depending whether we know $x(0)$ or not.

Suppose that $x(0)$ is known. Then the nominal trajectory is given by

$$\mathcal{Y}^* = f(x(0), 0).$$

On the other hand, if there exists a $\mathcal{U} \neq 0$, such that

$$\mathcal{Y}^* = f(x(0), \mathcal{U}),$$

then there is no way for us to know whether there is a fault or not given y^*. Notice that

$$\mathcal{Y}^* = f(x(0), 0) = f(x(0), \mathcal{U}) \Rightarrow f(0, \mathcal{U}) = 0,$$

which gives the following theorem:

Theorem 1. *The following statements are equivalent:*

1. *The fault is detectable with known initial conditions.*

2. *The following implication holds:*

$$f(0,\mathcal{U}) = 0 \implies \mathcal{U} = 0$$

3. *The system is left-invertible, i.e., the mapping from $\mathcal{U}$ to $\mathcal{Y}$ defined by $\mathcal{Y} = f(0,\mathcal{U})$ is one to one.*

If the initial condition is unknown, then the nominal trajectory will be a set of

$$Y^* = \{\mathcal{Y} : \mathcal{Y} = f(x(0), 0) \text{ for some } x(0) \in \mathbb{R}^n\}.$$

By the similar argument, if there exists a $\mathcal{U}$ and $x(0)'$, such that

$$\mathcal{Y} = f(x(0)',\mathcal{U}) \in Y^*,$$

then there is no way to know whether a fault occurs or not given $\mathcal{Y}$. By linearity of $\mathcal{Y}$, we know that

$$\mathcal{Y} = f(x(0), 0) = f(x(0)',\mathcal{U}) \implies f(x(0)' - x(0),\mathcal{U}) = 0,$$

which leads to the following theorem:

Theorem 2. *The following statements are equivalent:*

1. *The fault is detectable with unknown initial conditions.*

2. *The following implication holds:*

$$f(x(0),\mathcal{U}) = 0 \implies x(0) = 0 \text{ and } \mathcal{U} = 0.$$

3. *The system has no non-trivial zero dynamics (strongly observable), i.e.,*

$$f(x(0),\mathcal{U}) = 0 \implies x(k) = 0, \forall k.$$

4. *The system does not have an invariant zero, i.e., there does not exists an $z \in \mathbb{C}$, and non-zero $x_0 \in \mathbb{R}^n$ and $u_0 \in \mathbb{R}^m$, such that*

$$Ax_0 + Bu_0 = zx_0, \text{ and } Cx_0 = 0. \tag{2}$$

Proof. We will only prove 3 $\implies$ 2. Suppose that there exists $x(0)$ and $\mathcal{U} \neq 0$, such that $f(x(0),\mathcal{U}) = 0$. Let us define the subspace $\mathcal{V} \in \mathbb{R}^n$ as

$$\mathcal{V} \triangleq \text{span}(x(0), x(1), \ldots,).$$

$\mathcal{V} \neq \{0\}$. Since

$$Ax(k) = x(k+1) - Bu(k),$$

we know that

$$A\mathcal{V} \subseteq \mathcal{V} + \mathrm{col}(B), \tag{3}$$

where $\mathrm{col}(B)$ is the column space of B. Furthermore,

$$C\mathcal{V} = 0 \implies \mathcal{V} \subseteq \ker(C),$$

where $\ker(C)$ is the null space of C. By (3), we know that there exists an K, such that

$$(A + BK)\mathcal{V} \subseteq \mathcal{V}.$$

Hence, there exists $x_0 \in \mathcal{V}$, which is an eigenvector of $A+BK$ with corresponding eigenvalue z. Define $u_0 = Kx_0$, then z, x_0, u_0 satisfies (2). □

Remark 1. *The system is called strongly detectable if the following implication holds:*

$$f(x(0), \mathcal{U}) = 0 \implies x(k) \to 0.$$

This implies that even there might exists an undetectable attack, the effect of the attack on the state is decaying over time. One can prove that a system is strongly detectable if and only if all the invariant zeros of the system are stable.

1.2 Identification

Consider the following system:

$$x(k+1) = Ax(k) + \sum_{i\in\mathcal{I}} B_i u_i(k),\; y(k) = Cx(k), \tag{4}$$

where $u_i(k)$ denotes the ith fault and we say it occurs if $u_i(k) \neq 0$ for some k. We assume that at most one fault occurs and we want to identify which one.

Suppose that $x(0)$ is known, then all possible trajectories generated by the ith fault can be written as

$$Y_i = \{\mathcal{Y} : \mathcal{Y} = f(x(0), B_i\mathcal{U}_i)\},$$

where $B_i\mathcal{U}_i = (B_i u_i(0), B_i u_i(1), \dots)$. We claim that we can distinguish the ith fault and the jth fault if

$$\mathcal{Y} = f(x(0), B_i\mathcal{U}_i) = f(x(0), B_j\mathcal{U}_j) \implies \mathcal{U}_i = \mathcal{U}_j = 0.$$

Notice that

$$f(x(0), B_i\mathcal{U}_i) = f(x(0), B_j\mathcal{U}_j) \Leftrightarrow f\left(0, \begin{bmatrix} B_i & B_j \end{bmatrix} \begin{bmatrix} \mathcal{U}_i \\ -\mathcal{U}_j \end{bmatrix}\right) = 0.$$

Therefore, the fault is identifiable if and only if for any $i \neq j$, $(A, \begin{bmatrix} B_i & B_j \end{bmatrix}, C)$ is left invertible.

Similarly, with unknown initial conditions, the fault is identifiable if and only if for any $i \neq j$, $(A, \begin{bmatrix} B_i & B_j \end{bmatrix}, C)$ has no invariant zeros.

2 Generic Detectability

We model a network composed of m agents as a graph $G = \{V, E\}$. $V = \{1, 2, \ldots, m\}$ is the set of vertices representing the agents. $E \subseteq V \times V$ is the set of edges. $(i, j) \in E$ if and only if j can send information to i. The graph can be directed.

Define the neighbors $\mathcal{N}_i$ of agent i as the set of agents who can send information to i, i.e.,

$$\mathcal{N}_i \triangleq \{j : (i, j) \in E,\, j \neq i\}.$$

Suppose each agent has a state $x_i(t)$. The agent update the state based on the following update equation:

$$x_i(k+1) = a_{ii}x_i(k) + \sum_{j \in \mathcal{N}_i} a_{ij}x_j(k) + u_i(k),$$

where $u_i(k)$ is a malicious input. A node is benign if $u_i(k) = 0$ for all k. It is malicious if $u_i(k) \neq 0$ for some k.

We can write the above equation in matrix form as:

$$x(k+1) = Ax(k) + Bu(k),$$

where $B = \begin{bmatrix} e_{i_1}, \ldots, e_{i_f} \end{bmatrix}$, where $\{i_1, \ldots, i_f\}$ are the set of malicious node. Furthermore, for node i, we can define

$$C_i = \begin{bmatrix} e_{i_1} \\ \vdots \\ e_{i_l} \end{bmatrix},$$

where $\mathcal{N}_i \bigcup \{i\} = \{i_1, \ldots, i_l\}$. As a result, for a benign node i, it observes

$$y(k) = C_i x(k).$$

One can see that there is a straight forward connection between the topology of the network and the graph associated with linear structured system (A, B, C_i)

Theorem 3. *(Assuming unknown initial condition:) If the graph G has connectivity $k > f$, and i be a benign node. Then for almost any A, B matrices, node i can detect the existence of a malicious behavior. On the other hand, if $k \leq f$, then there exists a set of malicious node $\{i_1, \ldots, i_f\}$, such that no node can detect the malicious behavior for any A, B.*