Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Hierarchical Attack Identification for Distributed Robust Nonlinear Control

Sarah Braun [*,**] **Sebastian Albrecht** [*] **Sergio Lucia** [**,***]

[*] *Siemens Corporate Technology, 81739 Munich, Germany (e-mail: sarah.braun@siemens.com)*
[**] *Technische Universität Berlin, 10587 Berlin, Germany*
[***] *Einstein Center Digital Future, 10117 Berlin, Germany*

Abstract: Developing tools for attack identification in large-scale networked control systems is a research area of increasing significance for the secure and reliable operation of autonomous control systems. Due to scalability limits and privacy issues of individual subsystems, attack identification methods should not rely on global model knowledge. We address systems of interconnected nonlinear subsystems with coupled dynamics or constraints in a distributed control setup. The local controllers share information about the coupling variables of the subsystems and are designed to be robust towards attacks and uncertain influences through neighboring subsystems. We present a scalable hierarchical attack identification method which monitors the evolution of the coupling variables after an attack occurred in some unknown subsystem. Based on the mutual exchange of local sensitivity information among the subsystems, the propagation of the attack through the network is approximated. The propagation equations are used to formulate a quadratic program whose solution determines the attack signal that explains the observed network evolution best. The developed approach is applied to the IEEE 30 bus system to illustrate attack identification in power systems with faulty buses.

Keywords: attack identification, distributed and nonlinear control, robust model predictive control, cyber-physical systems, power systems

1. INTRODUCTION

Many relevant technologies such as energy grids, traffic networks or building systems involve the operation of large-scale networked control systems, which typically consist of several interacting subsystems. In safety-critical infrastructures, potential system faults and deliberate attacks have to be approached by suited prevention and defense mechanisms to ensure a secure operation. For this purpose, the controllers should be robust towards disturbances and supplemented with monitoring methods that reveal attacks. For both system control and attack identification, distributed approaches that require only partial model knowledge are more suitable than classical centralized approaches since they decrease the computational complexity and support privacy between the subsystems. To approach a secure operation of large cyber-physical systems in a scalable manner, we present a hierarchical method for attack identification based on partial model knowledge and limited information transmission in a distributed robust control setup.

Many distributed control strategies, in particular distributed model predictive control (MPC), have been developed for specific industrial applications such as power systems (Camponogara et al., 2002), multi-vehicle coordination (Dunbar and Murray, 2006) or transportation networks (Negenborn et al., 2008). Also algorithmic results analyzing fundamental theoretical properties like stability have received significant attention (e.g., Venkat et al., 2005). For reviews and classifications of existing distributed MPC approaches we refer to the surveys of Scattolini (2009) and Christofides et al. (2013). Distributed methods by definition include some exchange of information between the subsystems, e.g., about planned state trajectories, such that each subsystem has more information about its neighbors' influences than in a fully decentralized approach. This typically improves the performance since, to the eyes of a subsystem in a fully decentralized approach, the evolution of the neighboring subsystems and thus their influence on its own dynamics are unknown. If the subsystem obtains some information about the uncertainty range, it can design its local controller to be robust towards these uncertainties. This is the underlying idea of the distributed MPC scheme by Farina and Scattolini (2012), combining the exchange of corridors around reference trajectories with robust MPC. The future values of a subsystem's *coupling variables* are guaranteed to lie in these corridors. Several works follow similar ideas, e.g., Lucia et al. (2015) extend the approach to nonlinear MPC (NMPC) and present sufficient conditions for Input-to-State stability and recursive feasibility. They introduce the notion of *contracts* for the exchanged corridors and elaborate on how to obtain them by reachable set analysis.

Apart from neighboring coupling variables we consider a second, possibly far more significant source of uncertainty

[⋆] This work was supported by the German Federal Ministry of Education and Research (BMBF) via the funded research project AlgoRes (01S18066B).

for each subsystem, namely attacks. We model an *attack* as the malicious disturbance of any signal in the local closed-loop system, taking up the formulation of Pasqualetti et al. (2013) who define an attack as an unknown signal affecting a linear control system. They characterize the concepts of attack detection and identification as the tasks to reveal the existence and location of an attack, respectively. As a generic approach towards attack identification they suggest a sparse recovery problem revealing the input signal that explains the observed system evolution best.

While many of the available model-based approaches towards attack identification focus on centralized systems (see Ding, 2008), also distributed and hierarchical attack identification methods attain increasing interest. For linear systems, some existing methods based on local model information require special network structures. For instance, Pasqualetti et al. (2010) focus on networks with weakly coupled subsystems or leading subsystems with better communication capabilities. Others analyze observer-based techniques such as Shames et al. (2011) who consider second-order linear systems with consensus control laws, for which they construct a bank of unknown-input observers to detect and identify faults. Boem et al. (2018) extend this idea to nonlinear systems and design local estimators for identification purposes, assuming that each subsystem knows all possible faults that may occur.

We consider the problem class of distributed networked systems with nonlinear dynamics and several interacting subsystems, which are physically coupled through their local dynamics and in the form of constraints as formally described in Section 2. We present a hierarchical attack identification method in Section 4 which does not require global model knowledge. Instead, all subsystems exchange sensitivity information about their coupling variables, evaluated at the current iterate. While the local dynamics or objectives are not revealed such that privacy is maintained, the sensitivity information allows to approximate the propagation of the attack through the network. Based on this propagation, the proposed method computes the global disturbed input signal which explains the observed data best and thus indicates which subsystems were attacked and which only propagated their neighbors' errors. In contrast to Boem et al. (2018), we do not assume the set of possible attacks to be finite and known, which constitutes a restrictive assumption if one considers malicious attacks rather than system faults. Instead, we solve a continuous optimization problem for sparse signal recovery that provides a systematic approach to reveal any possible attack. The method is strongly connected to the contract-based distributed NMPC setup in Section 3 since the exchanged sensitivities indicate how a subsystem's coupling variables are influenced by a deviation of the neighbors' couplings from the previously stated nominal contract values. To illustrate the results, we apply attack identification to the IEEE 30 bus power system in Section 5.

2. PROBLEM FORMULATION

We consider a nonlinear system of systems, exposed to the risk of attacks, with state $x \in \mathbb{X} \subseteq \mathbb{R}^{d_x}$, initial state $x^0 \in \mathbb{X}$, control $u \in \mathbb{U} \subseteq \mathbb{R}^{d_u}$ and discrete-time dynamics

$$x^+ = f\left(x, a(u)\right), \tag{1}$$

where the *attack function* $a : \mathbb{U} \to \mathbb{U}$ indicates how the input u is modified by the attack and constitutes an a priori unknown influence on the dynamics. If no attack is present, $a(\cdot)$ is the identity and u directly controls the system via $x^+ = f(x, u)$ in an undisturbed manner. In case of an attack, $a(u) \neq u$ holds and the set

$$I_a \coloneqq \{i \in \{1, \dots, d_u\} : (a(u))_i \neq u_i\}$$

contains the attacked components. If an attack disturbs all d_u control inputs, i.e., $I_a = \{1, \dots, d_u\}$, it might not be possible to guarantee a feasible state evolution even with a robust controller. Therefore, the problem class of attacked systems (1) will only be manageable for a subset $\mathcal{A} \subseteq \{F : \mathbb{U} \to \mathbb{U}\}$ of possible attack functions F. It is important to note that we do not make any structural assumptions on $\mathcal{A}$ nor do we assume that $\mathcal{A}$ is known to the identification method we propose. We will later comment on how $\mathcal{A}$ is approximated to design controllers which are robust towards attacks. We assume that modifications in the input u by an attack a cannot be measured directly but only their impact on the state is observable. Even though we refer to the input u as control, it may represent any signal such that various types of attacks are modeled. System (1) consists of a partition $\mathcal{P}$ into dynamically coupled subsystems, each of which is described by

$$\begin{aligned} x_I^+ &= f_I(x_I, a_I(u_I), z_{\mathcal{N}_I}), \\ z_I &= h_I(x_I), \end{aligned} \tag{2}$$

where x_I and u_I denote the local state and control in subsystem I, $a_I \in \mathcal{A}_I$ is the unknown local attack function and f_I the differentiable discrete-time dynamics. Interconnections between the subsystems are modeled via coupling variables $z_I \in \mathbb{R}^{d_{z_I}}$, which are related to the local states x_I by differentiable functions h_I. The set of all subsystems $J \in \mathcal{P} \setminus \{I\}$ influencing the dynamics of I via their couplings z_J is called neighborhood of I and denoted as $\mathcal{N}_I$. Each subsystem I is subject to nonlinear constraints

$$g_I\left(x_I, a_I(u_I), z_{\mathcal{N}_I}\right) \leq 0,$$

which may depend on the neighbors' couplings $z_{\mathcal{N}_I}$. For the control and identification methods in Section 3 and 4 to be efficient, the partition should be chosen such that each subsystem is of manageable size and for the total number $d_z = \sum_{I \in \mathcal{P}} d_{z_I}$ of coupling variables it holds $d_z \ll d_x$.

This setup naturally leads to a hierarchical structure; on a lower level a set of distributed controllers is employed. On a higher level, each subsystem I is considered one entity with associated coupling z_I. By observing the error propagation through the high-level network, we aim to identify the subsystem containing the attacker by computing a sparse attack signal which explains the observed behavior.

3. CONTRACT-BASED DISTRIBUTED MPC

In this section, we describe a distributed control setup to control system (1), taking into account the uncertainty that arises due to possible attacks. The local dynamics of each subsystem I are influenced by two types of uncertainty, namely the couplings $z_{\mathcal{N}_I}$ of the neighboring subsystems and their own potentially disturbed control $a_I(u_I)$, which differs from the controller input u_I if an attack occurs. In critical infrastructures with safety and reliability requirements, the local controllers have to provide *robust* control strategies such that the constraints are satisfied

no matter which attack a_I and coupling values $z_{\mathcal{N}_I}$ occur. We apply robust NMPC to achieve robust optimal control in real time. These approaches typically require the uncertainty to be represented by a set of possible scenarios such as the attack set $\mathcal{A}_I$ in which the uncertain attack a_I is contained. We will later elaborate on how we approximate the unknown, generally infinite set $\mathcal{A}_I$ for numerical computations. To provide a set of potentially occurring realizations for the coupling variables z_I, we consider a setup where each subsystem I at each sampling time t shares a *contract* $\mathcal{Z}_I$ with its neighbors, constituting a sequence of sets containing the future trajectory of its coupling variable z_I on the considered horizon $t, \dots, t+N$. A schematic representation of this setup is given in Fig. 1.

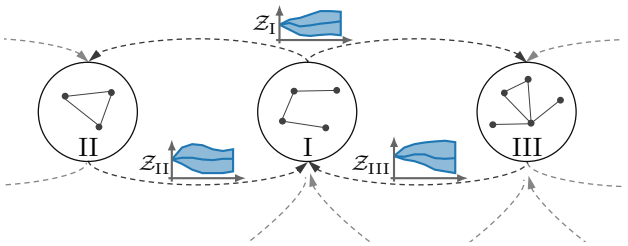

Fig. 1. Schematic overview of the underlying distributed control setup based on an illustration from Lucia et al. (2015). The subsystems are interconnected through physical couplings or coupled constraints (illustrated by dashed edges) and share contracts in which the future coupling values lie (depicted in blue).

A contract takes the uncertain influences of potentially occurring internal attacks a_I and its neighbors' couplings $z_{\mathcal{N}_I}$ into account. Additionally, each subsystem shares a *nominal* coupling trajectory $\bar{z}_I$ which is the predicted undisturbed trajectory that it will follow if no attack occurs in subsystem I and all neighbors also act according to their nominal strategies $\bar{z}_{\mathcal{N}_I}$. Closely following the concepts and notation of Lucia et al. (2015), the contract of subsystem I for time k calculated at time t is denoted as $\mathcal{Z}_I(k|t)$. It is defined considering the reachable set $\mathcal{X}_I(k|t)$ of state x_I at time k, calculated at time t, which is given as

$$\mathcal{X}_I(k|t) \coloneqq \{f_I\left(x_I, a_I(u_I), z_{\mathcal{N}_I}\right) : \\ x_I \in \mathcal{X}_I(k-1|t), a_I \in \mathcal{A}_I, z_{\mathcal{N}_I} \in \mathcal{Z}_{\mathcal{N}_I}(k|t)\}$$

for a given input u_I. The contract $\mathcal{Z}_I(k|t)$ is derived as

$$\mathcal{Z}_I(k|t) \coloneqq \{h_I(x_I) : x_I \in \mathcal{X}_I(k|t)\}.$$

The aforementioned nominal state and coupling values $\bar{x}_I$ and $\bar{z}_I$ are computed as

$$\bar{x}_I(k|t) \coloneqq f_I\big(\bar{x}_I(k-1|t), u_I(k-1|t), \bar{z}_{\mathcal{N}_I}(k|t)\big) \text{ and}$$
$$\bar{z}_I(k|t) \coloneqq h_I\big(\bar{x}_I(k|t)\big)$$

with undisturbed input $u_I(k-1|t)$. After the distributed computation of the control inputs at sampling time t, subsystem I receives the contract $\mathcal{Z}_J(k|t)$ and the nominal trajectory $\bar{z}_J(k|t)$ from each neighbor $J \in \mathcal{N}_I$ and locally aggregates the contracts and nominal values for the next sampling time $t+1$ as

$$\mathcal{Z}_{\mathcal{N}_I}(k|t+1) = \Pi_{J\in\mathcal{N}_I} \mathcal{Z}_J(k|t) \text{ and}$$
$$\bar{z}_{\mathcal{N}_I}(k|t+1) = \Pi_{J\in\mathcal{N}_I} \bar{z}_J(k|t),$$

extended by $\mathcal{Z}_{\mathcal{N}_I}(t+1+N|t+1) = \mathcal{Z}_{\mathcal{N}_I}(t+N|t+1)$ and $\bar{z}_{\mathcal{N}_I}(t+1+N|t+1) = \bar{z}_{\mathcal{N}_I}(t+N|t+1)$ to have contract values available at all time steps of the new horizon. Employing the neighbors' contracts $\mathcal{Z}_{\mathcal{N}_I}$ as an uncertainty interval for the unknown couplings $z_{\mathcal{N}_I}$, the local optimization problem under uncertainty for subsystem I at sampling time t with horizon N reads as follows:

$$\begin{aligned} \min_{x_I(\cdot), u_I(\cdot)} \quad & \sum_{k=t}^{t+N-1} l_I\big(x_I(k), a_I(u_I(k)), z_{\mathcal{N}_I}(k)\big) \\ \text{s.t.} \quad & x_I(k+1) = f_I\big(x_I(k), a_I(u_I(k)), z_{\mathcal{N}_I}(k+1)\big), \\ & x_I(t) = x_I^t, \\ & g_I(x_I(k), a_I(u_I(k)), z_{\mathcal{N}_I}(k)) \le 0, \\ & u_I(k) \in \mathbb{U}_I, x_I(k) \in \mathbb{X}_I, \\ & \text{for all } z_{\mathcal{N}_I}(k+1) \in \mathcal{Z}_{\mathcal{N}_I}(k+1|t), a_I \in \mathcal{A}_I, \\ & \text{for } k = t, \dots, t+N-1, \end{aligned} \tag{3}$$

where x_I^t is the initial state at time t and l_I the local cost function. As long as the attack sets $\mathcal{A}_I$ are unknown, so are the contracts $\mathcal{Z}_I$ and approximations $\widetilde{\mathcal{A}}_I$ of the attack sets are required. Apart from that, for nonlinear functions f_I it is nontrivial to determine the reachable sets $\mathcal{X}_I(k|t)$ (e.g., Sahlodin and Chachuat, 2011). Since they depend on the choice of the input u_I, their computation should be combined with solving the optimization problem (3). To compute robust solutions of (3), considering the uncertain influence through $a_I(u_I)$ and $z_{\mathcal{N}_I}$, we apply the multi-stage scheme by Lucia et al. (2013), which offers two advantages in our setup. First, it is less conservative than classical min-max-approaches for robust MPC (Campo and Morari, 1987) because it takes into account that future inputs can be adapted once new measurements will be available. Second, multi-stage NMPC provides a simple tool to approximate the reachable sets $\mathcal{X}_I(k|t)$. It requires the uncertainty range to be approximated by a given discrete set Σ_0 of sampling values and models the uncertain state evolution by a scenario tree like the one shown in Fig. 3. The branches at time k represent all realizations $\sigma \in \Sigma_k$, where Σ_k is obtained by considering all possible combinations of samples in Σ_0 up to stage k or some predetermined earlier stage N_r, the so-called "robust horizon". Depending

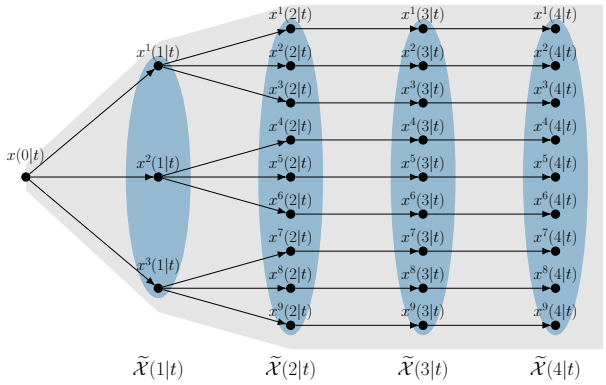

Fig. 2. A scenario tree for the uncertain state evolution on $t, \dots, t+N$ with $N = 4$ according to Lucia et al. (2013). The gray shadow covers the reachable sets $\mathcal{X}(k|t)$ explored by the tree and the ellipses show the approximations $\widetilde{\mathcal{X}}(k|t)$ based on different state realizations $x^\sigma(k|t)$ corresponding to the tree's nodes.

on the realization $\sigma \in \Sigma_k$, different states $x_I^\sigma(k|t)$ are obtained for $k = t+1, \dots, t+N$ at sampling time t by propagating the parent node with a given control input

and the realization $\sigma \in \Sigma_k$. These state samples provide an approximation of the reachable set $\mathcal{X}_I(k|t)$ by

$$\{x_I^\sigma(k|t) : \text{for all } \sigma \in \Sigma_k\} =: \widetilde{\mathcal{X}}_I(k|t) \subseteq \mathcal{X}_I(k|t),$$

indicated with ellipses in Fig. 3. A corresponding approximation of the contracts $\mathcal{Z}_I(k|t)$ can easily be derived as

$$\{h_I\left(x_I^\sigma(k|t)\right) : \text{for all } \sigma \in \Sigma_k\} =: \widetilde{\mathcal{Z}}_I(k|t) \subseteq \mathcal{Z}_I(k|t).$$

To guarantee recursive feasibility (Lucia et al., 2015), it is necessary to ensure that the information exchanged at each sampling time is consistent with the information exchanged in the past. This can be achieved using consistency constraints requiring

$$\mathcal{Z}_I(k|t) \subseteq \mathcal{Z}_I(k|t-1). \tag{4}$$

The corresponding constraints on each element $h_I(x_I^\sigma(k|t))$ of the approximated contracts $\widetilde{\mathcal{Z}}_I(k|t)$ read as

$$\min \widetilde{\mathcal{Z}}_I(k|t-1) \leq h_I\left(x_I^\sigma(k|t)\right) \leq \max \widetilde{\mathcal{Z}}_I(k|t-1) \tag{5}$$

and must be enforced as additional constraints when solving problem (3) with multi-stage NMPC. For nonlinear dynamics, the approximated contracts are not guaranteed to contain the real coupling values in cases where the actual realization of the uncertainty is not considered in the scenario tree. Lucia et al. (2014) approach this problem by a rigorous computation of reachable sets that allows to give guarantees. In practice, classical multi-stage NMPC has proven to provide promising results such that the simpler reachable set computation is typically sufficient.

Since the contracts $\mathcal{Z}_I(k|t)$ at time t by definition depend on $\mathcal{Z}_{\mathcal{N}_I}(k|t)$, which are computed at the previous time step $t-1$, the existence of a set of initial contracts is a critical requirement (see Lucia et al., 2015).

Assumption 1. (Initial contracts). We assume that there are initial contracts $\mathcal{Z}_I(k|0)$ and nominal trajectories $\bar{z}_I(k|0)$ for each $I \in \mathcal{P}$, $k = 0, \dots, N$, such that for every subsystem I all predicted coupling values are within $\mathcal{Z}_I(k|0)$ as long as the same holds for all its neighbors.

Given such a set of initial contracts, we assume that at each time t a feasible solution of problem (3) together with a new contract satisfying the consistency constraints (4) can be computed for each I. Then it holds by construction of the contracts that the coupling variables at time $t+1$ lie in the previous contracts $\mathcal{Z}_I(t+1|t)$ and attain the nominal value $\bar{z}_I(t+1|t)$ if no disturbances in u_I or $z_{\mathcal{N}_I}$ occur. In a distributed control setup one should not assume the availability of a centralized initial solution and we thus propose the distributed Algorithm 1 to compute initial contracts in an iterative offline fashion. It terminates when the computed contracts for all subsystems are contained in the ones from the previous iteration.

Remark 1. For convenience, each subsystem locally applies centralized control to solve problem (3). Also multiple hierarchical levels are possible in the sense that the subsystems are again subdivided and apply distributed control.

4. HIERARCHICAL ATTACK IDENTIFICATION

While attack detectors monitor the system to recognize the presence of an attack, attack identification is used to localize it. Local detectors typically compute an estimate $\widehat{x}_I$ of the state x_I exploiting local model information and compare it to a state measurement $\widetilde{x}_I$ (e.g., Boem et al., 2018). If the estimation error $\|\widetilde{x}_I - \widehat{x}_I\|$ exceeds a given threshold δ_I, the subsystem reports that an attack has probably occurred. In this paper, we focus on attack identification and propose a novel method that can be combined with any state-of-the-art attack detector.

Algorithm 1 Computation of initial contracts

1: **For each** subsystem $I \in \mathcal{P}$:
  1.1: Compute initial coupling value $z_I^0 := h_I(x_I^0)$
  1.2: Set $\widetilde{\mathcal{Z}}_I^0(k|0) = \{z_I^0\}$, $\bar{z}_I^0(k) = z_I^0 \ \ \forall k = 0:N$ and communicate this first contract to neighbors
  1.3: Locally build $\widetilde{\mathcal{Z}}_{\mathcal{N}_I}^1, \bar{z}_{\mathcal{N}_I}^1$
2: **For** $j = 1, \dots,$ `MAX_ITER`:
  2.1: **For each** subsystem $I \in \mathcal{P}$:
    2.1.1: Build scenario tree by branching on different realizations in $\mathcal{A}_I$ and $\widetilde{\mathcal{Z}}_{\mathcal{N}_I}^j$
    2.1.2: Solve problem (3) for $t = 0$
    2.1.3: Derive contract $\widetilde{\mathcal{Z}}_I^j(k|0)$ and nominal $\bar{z}_I^j(k|0)$ $\forall k = 0:N$ and exchange with neighbors
    2.1.4: Locally build $\widetilde{\mathcal{Z}}_{\mathcal{N}_I}^{j+1}, \bar{z}_{\mathcal{N}_I}^{j+1}$
  2.2: **If** $\widetilde{\mathcal{Z}}_I^j(k|0) \subseteq \text{conv}(\widetilde{\mathcal{Z}}_I^{j-1}(k|0))$ $\forall I \in \mathcal{P}, k = 0:N$, **Return** $\widetilde{\mathcal{Z}}_I(\cdot|0) := \widetilde{\mathcal{Z}}_I^j(\cdot|0) \ \ \forall I \in \mathcal{P}$
  2.3: Increment j, $j := j+1$.

Since the subsystems are interconnected through the coupling variables z_I, the cause for a deviation $\|\widetilde{x}_I - \widehat{x}_I\| > \delta_I$ does not necessarily have to be an attack $a_I(u_I) \neq u_I$ in I but can be located anywhere else in the network. Therefore, a suited identification scheme should not be fully decentralized but take the mutual interference of the subsystems and a potential spread of the attack into account. The basic idea of our attack identification is as follows. At each subsystem I, we monitor the evolution of the coupling variables z_I and compute the deviation $\widetilde{z}_I - \bar{z}_I$ of the measured value from the nominal value. By solving a suitably designed signal recovery problem, we want to find out at which subsystem the observed deviation is only due to the propagation of its neighbors' errors and which subsystem actively disturbs the network. For this purpose, we derive linear approximations of the attack propagation through the network at the current point in time.

According to (2), z_I is a function of x_I, which in turn depends on $a_I(u_I)$ and $z_{\mathcal{N}_I}$, such that z_I implicitly depends on $x_I, a_I(u_I)$ and $z_{\mathcal{N}_I}$ through some function ζ_I. It holds

$$z_I = \zeta_I(x_I, a_I(u_I), z_{\mathcal{N}_I}).$$

We consider the coupling values $z_I(t)$ at one fixed sampling time t with given nominal contract values $\bar{z}_I(t|t-1)$. For the sake of brevity, the time index is dropped in the following. We want to analyze the behavior of z_I if small deviations in $a_I(u_I)$ and $z_{\mathcal{N}_I}$ from the corresponding nominal values $\bar{a}_I(u_I)$ and $\bar{z}_{\mathcal{N}_I}$ occur, denoting by $\bar{a}_I := \text{id}$ the identity function modeling the nominal case of no attack. For this purpose, we compute a linear approximation of $\zeta_I(x_I, \cdot, \cdot)$ in a neighborhood around the nominal value $\zeta_I\big(x_I, \bar{a}_I(u_I), \bar{z}_{\mathcal{N}_I}\big)$. Recall that the contracts are designed such that ζ_I attains the nominal value $\bar{z}_I$ if $a_I = \bar{a}_I$ and $z_{\mathcal{N}_I} = \bar{z}_{\mathcal{N}_I}$, i.e., $\zeta_I\big(x_I, \bar{a}_I(u_I), \bar{z}_{\mathcal{N}_I}\big) = \bar{z}_I$. Denoting the deviation $z_I - \bar{z}_I$ of the actual coupling value z_I from the nominal value $\bar{z}_I$ as Δz_I and the deviation $a_I(u_I) - \bar{a}_I(u_I)$ of the potentially disturbed control input $a_I(u_I)$ from the undisturbed input $\bar{a}_I(u_I)$ as Δa_I, it holds

$$\Delta z_I = \frac{\partial \zeta_I}{\partial a_I}\big(x_I, \bar{a}_I(u_I), \bar{z}_{\mathcal{N}_I}\big)\Delta a_I + \frac{\partial \zeta_I}{\partial z_{\mathcal{N}_I}}\big(x_I, \bar{a}_I(u_I), \bar{z}_{\mathcal{N}_I}\big)\Delta z_{\mathcal{N}_I} + o\left(\left\|\begin{pmatrix}\Delta a_I \\ \Delta z_{\mathcal{N}_I}\end{pmatrix}\right\|\right) \tag{6}$$

for $\Delta a_I, \Delta z_{\mathcal{N}_I} \to 0$ due to Taylor's theorem. The Jacobian $\frac{\partial \zeta_I}{\partial a_I}$ evaluated at $(x_I, \bar{a}(u_I), \bar{z}_{\mathcal{N}_I})$ can be computed by applying the chain rule as follows:

$$\frac{\partial \zeta_I}{\partial a_I}\big(x_I, \bar{a}_I(u_I), \bar{z}_{\mathcal{N}_I}\big) = \frac{\partial h_I}{\partial x_I}(x_I)\frac{\partial f_I}{\partial a_I}\big(x_I, \bar{a}_I(u_I), \bar{z}_{\mathcal{N}_I}\big)$$

and $\frac{\partial \zeta_I}{\partial z_{\mathcal{N}_I}}$ analogously. The matrices $\frac{\partial f_I}{\partial a_I}$, $\frac{\partial f_I}{\partial z_{\mathcal{N}_I}}$ indicate how the optimal solution x_I depends on small perturbations in $a_I(u_I)$ or $z_{\mathcal{N}_I}$ (see (2)) and thus represent sensitivity matrices. If all subsystems share sensitivity information in the form of $\frac{\partial \zeta_I}{\partial a_I}$, $\frac{\partial \zeta_I}{\partial z_{\mathcal{N}_I}}$ evaluated at $(x_I, \bar{a}(u_I), \bar{z}_{\mathcal{N}_I})$, a linear approximation for the propagation of the attack through the network can be set up based on equations (6), omitting the remainder term. If additionally the deviations $\Delta\widetilde{z}_I \coloneqq \widetilde{z}_I - \bar{z}_I$ of the measured coupling values from the nominal values are shared, the problem of attack identification can be formulated using the following sparse signal recovery problem:

$$\begin{aligned} \min_{\Delta a, \Delta z} \quad & \alpha_1 \|\Delta z - \Delta\widetilde{z}\|_2^2 + \alpha_2 \|\Delta a\|_1 \\ \text{s.t.} \quad & \Delta z_I = \frac{\partial \zeta_I}{\partial a_I}\left(x_I, \bar{a}_I(u_I), \bar{z}_{\mathcal{N}_I}\right)\Delta a_I \\ & \qquad + \frac{\partial \zeta_I}{\partial z_{\mathcal{N}_I}}\left(x_I, \bar{a}_I(u_I), \bar{z}_{\mathcal{N}_I}\right)\Delta z_{\mathcal{N}_I} \quad \forall I \in \mathcal{P}, \end{aligned} \tag{7}$$

where $\alpha_1, \alpha_2 \in \mathbb{R}_{\geq 0}$ are weighting factors for the two cost components. Note that the deviations Δa in the control inputs caused by the attack are unknown, whereas $\Delta\widetilde{z}$ is assumed to be measurable. A solution of problem (7) tracks the measured deviations $\Delta\widetilde{z}$ as close as possible while obeying the approximated error propagation and taking a minimum possible attack as a basis. This sparsity assumption is common in attack identification (e.g., Pasqualetti et al., 2013; Liu et al., 2014). To penalize the number of attacked subsystems, one would replace the ℓ_1-norm $\|\Delta a\|_1$ in (7) by the ℓ_0-"norm" $\|\Delta a\|_0$, counting the number of non-zero elements. This would, however, transform problem (7) into a mixed-integer problem, which is why it is typically relaxed to the ℓ_1-norm that can be expressed by linear constraints. Candès and Tao (2005) proved sufficient conditions under which the approximation of the ℓ_0-"norm" with the ℓ_1-norm is exact.

The identification problem (7) is a quadratic program with linear constraints and thus computationally easy to solve. Given an optimal solution $(\Delta a^*, \Delta z^*)$, the nonzero components in Δa^* indicate those subsystems I for which $a_I^*(u_I^*) \neq u_I^*$, i.e., those subsystems in which an attack occurred. By defining a suitable threshold ε, the identification method identifies all subsystems I with $\|\Delta a_I^*\| > \varepsilon$ as attacked. Note that problem (7) is a global identification problem describing the entire network and not only referring to one subsystem. Depending on the network structure, we consider different procedures for conducting the identification process. For instance, the network may be equipped with some central superior instance which collects all sensitivity information, solves problem (7) and possibly also has the power to exclude suspicious subsystems from the network. As an alternative, all subsystems could exchange the sensitivity information among each other such that each can locally solve the identification problem (7) and share its suspicions with the others. This requires a decision rule like a majority vote to finally accuse and possibly exclude some subsystem, which is a common necessity in distributed approaches towards attack identification.

The proposed identification method is a hierarchical approach since no information about the local dynamics f_I or costs l_I is used. We think that requiring the subsystems to publish local evaluations of their solutions' sensitivities is within the privacy limits since they do not allow other subsystems to easily draw conclusions about the local objectives. Additionally, the approach scales significantly better than a fully centralized nonlinear method because typically the number of variables in (7) is considerably smaller than the number $d_x + d_u$ of variables affecting the global dynamics. First, this is due to the fact that in many applications one can model the subsystems and their couplings such that $d_z \ll d_x$. Second, if one only wants to figure out which subsystem I disturbed the network with an attack $a_I(u_I) \neq u_I$ but is not interested in a specific component $(u_I)_i$, the number of variables in (7) can be further reduced: Instead of $\frac{\partial \zeta_I}{\partial a_I}$, the subsystems may publish a submatrix of full rank by omitting dependent columns.

5. ATTACK IDENTIFICATION IN POWER SYSTEMS

We consider the problem of attack identification in power systems and apply our hierarchical method from Section 4 to the IEEE 30 bus system shown in Fig. 3. The contract-based distributed NMPC approach is implemented based on the do-mpc environment for multi-stage NMPC (Lucia et al., 2017), which uses CasADi for automatic differentiation and optimization (Andersson et al., 2019) and Ipopt for solving nonlinear problems (Wächter and Biegler, 2006). The reachable sets and resulting contracts are approximated with multi-stage NMPC as described in Section 3. To deal with the size of the scenario tree, multi-stage NMPC does not use branching on the entire horizon N but only up to the robust horizon $N_r \leq N$ (Lucia et al., 2013). Additionally, we only enforce the consistency constraints (5) on the approximated contracts $\widetilde{\mathcal{Z}}(k|t)$ for $k = 0, \ldots, N_r$ but not for $k = N_r + 1, \ldots, N$. This is not sufficient for recursive feasibility but already yields reliable contracts in our numerical experiments. We compute the sensitivity matrices from Section 4 based on finite differences. For each subsystem we use a piecewise constant discretization to embed the neighbors' coupling variables $z_{\mathcal{N}_I}$ into the optimal control problem.

The multi-stage scheme requires the user to provide sampling sets of the uncertainty ranges. The contract approximations $\widetilde{\mathcal{Z}}_I$ are represented by all combinations of the componentwise upper and lower bounds and the nominal coupling values. To obtain discrete representations of the attack spaces $\mathcal{A}_I$ for $I \in \mathcal{P}$, we choose $(s_I)^3$ sampling elements from $\mathcal{A}_I$ for some small number $s_I \leq d_{u_I}$ as follows. We arbitrarily pick s_I out of the d_{u_I} input components $(u_I)_i$ in subsystem I and consider three sample attacks on $(u_I)_i$, namely one attack which does not modify $(u_I)_i$ and one attack disturbing $(u_I)_i$ by decreasing or increasing

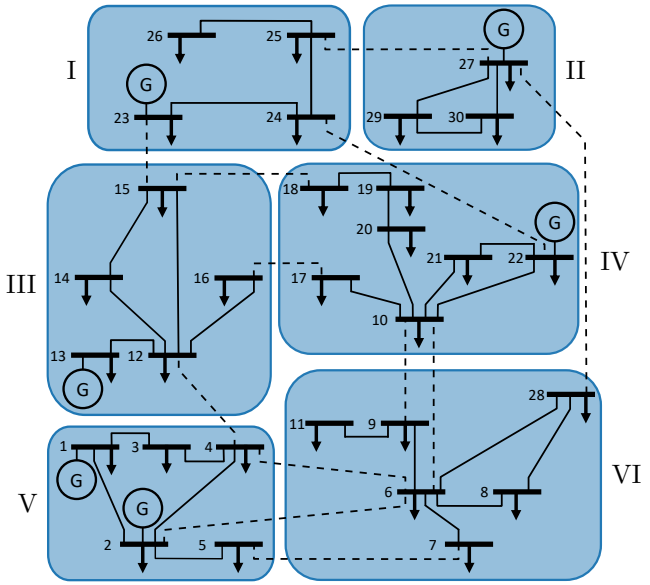

Fig. 3. Partition of IEEE 30 bus system with six generators into six dynamically coupled subsystems I–VI. Dashed edges show transmission lines connecting two subsystems and thus represent physical couplings.

it by 10% towards its lower or upper bound, respectively. Choosing $s_I = d_{u_I}$ and thus branching on potential attacks on *all* inputs entails too large scenario trees such that we choose $s_I = 1, 2$ or 3 for each I.

The considered IEEE 30 bus system consists of 30 buses and 41 transmission lines. It contains six generator units located at the buses 1, 2, 13, 22, 23 and 27 and six constant loads at the buses 3, 7, 14, 19, 26 and 30. The neighborhood of bus i is denoted by N_i and consists of all nodes that are connected with i by a transmission line. We assume that all buses are connected to synchronous machines and model the behavior of the machine in bus i using the so-called swing equation, for a detailed derivation of which we refer to Kundur et al. (1994):

$$m_i\ddot{\theta}_i + d_i\dot{\theta}_i = P_{\text{in},i} - \sum_{j\in N_i} P_{ij},$$

where θ_i is the phase angle of bus i, m_i and d_i describe inertia and damping coefficients of the machine, respectively, $P_{\text{in},i}$ is the power infeed at bus i and P_{ij} is the active power flow from bus i to bus j. Neglecting the dynamics of the transmission lines, the power flow P_{ij} between bus i and bus j can be modeled as

$$P_{ij} = |V_i|\,|V_j|\,b_{ij}\sin(\theta_i - \theta_j),$$

where $|V_i|$ is the voltage magnitude at bus i and b_{ij} is the susceptance of the power line between buses i and j. Realistic values for these parameters as well as a steady state of the system, providing initial values for θ and P_{in}, are based on Matpower (Zimmerman et al., 2010). The dynamic coefficients d_i and m_i for the generator buses are taken following De Tuglie et al. (2008) and Kundur et al. (1994), while the coefficients for the load buses are arbitrarily chosen from a range of realistic values.

With $k_{ij} := |V_i|\,|V_j|\,b_{ij}$, the following optimal control problem with states θ_i, $\omega_i := \dot{\theta}_i$ and inputs $P_{\text{in},i}$ for $i = 1, \dots, 30$ describes the problem of optimal frequency control:

$$\begin{aligned}\min_{\theta,\omega,P_{\text{in}}} \quad & \|\omega\|_2^2 \\ \text{s.t.} \quad & \dot{\theta}_i = \omega_i, \\ & \dot{\omega}_i = \frac{1}{m_i}\left(P_{\text{in},i} - d_i\omega_i - \sum_{j\in N_i} k_{ij}\sin(\theta_i - \theta_j)\right),\end{aligned}$$

for all $i = 1, \dots, 30$. As all parameters are chosen in a per-unit system with a 200kV base, ω describes the deviation from the nominal frequency of 60Hz and should thus be minimized. To solve the problem in a distributed manner, we define six subsystems I–VI consisting of three to seven buses as indicated in Fig. 3. Transmission lines running between two subsystems are shown as dashed lines in Fig. 3 and indicate the subsystem neighborhoods $\mathcal{N}_I$. For every subsystem, those phase angles θ_i are defined as coupling variables that have at least one incident edge leaving the subsystem. The couplings of subsystem III, for instance, are given as $z_{\text{III}} = (\theta_{12}, \theta_{15}, \theta_{16})$, influencing the neighbored subsystems I, IV and V. Initial contracts for the distributed control setup from Section 3 are computed offline with Algorithm 1, which converges after four iterations. We consider a time horizon of 20s and discretize it with time steps of length $\Delta t = 0.05$s. The system is attacked simultaneously at the load buses 4 and 15 during the five time steps $2\text{s}, \dots, 2.2\text{s}$. The attackers modify the inputs $P_{\text{in},4}$, $P_{\text{in},15}$ to the maximum or minimum possible load of 0.4 and 0 p.u. We denote the first time step of the attack as $t_a := 2$s. The resulting disturbed trajectories for ω_i on $[0, 20\text{s}]$ for all buses i are shown in Fig. 4, where the same color is used for all buses that belong to the same subsystem. Even though only two nodes are attacked,

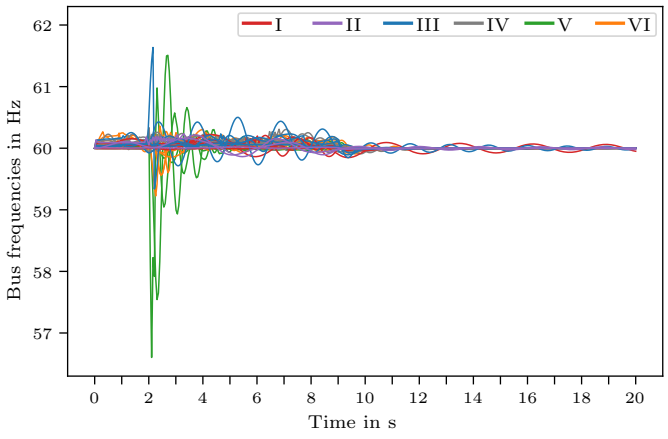

Fig. 4. Disturbed trajectories for all 30 bus frequencies showing the impact of an attack during $2\text{s}, \dots, 2.2\text{s}$. Due to dynamic couplings, not only the frequencies of the two attacked buses are affected by the attack but several ones, also from not attacked subsystems.

Fig. 4 reveals that several ω_i of multiple subsystems are disturbed and thus illustrates how the attack spreads through the network due to the internal coupling of the system. The conservativeness of the robust NMPC scheme prevents the nominal frequency of 60Hz to be reached exactly, see for example the time $[0, 2\text{s}]$ before the attack. To increase setpoint accuracy, we gradually reduce the conservativeness of multi-stage NMPC from $t = 5$s on by branching on decreasing disturbances, starting at 10% around the nominal value, until finally non-robust NMPC

is applied from $t = 9$s on. Due to the large impact of the attack, some states still oscillate at the end of the considered simulation time.

The effects of the attack also clearly show when comparing measurements $\widetilde{z}_I$ of the coupling values at each subsystem with the corresponding nominal values $\bar{z}_I$ computed at the previous time step. We assume perfect measurements $\widetilde{z}_I = z_I$ and do not introduce an additional source of noise in problem (7) apart from the linearization error in the attack propagation. Fig. 5 illustrates a snapshot of the high-level network, in which each subsystem is represented by one node, at the last time step $t_a + 4\Delta t$ at which the attackers are active. The color of each node indicates the total absolute deviation $\|\Delta\widetilde{z}_I\|_1$ of the subsystem's coupling variables from the respective nominal values in a logarithmic scale. Even though only subsystems III and V are attacked, the coupling values deviate significantly for multiple subsystems.

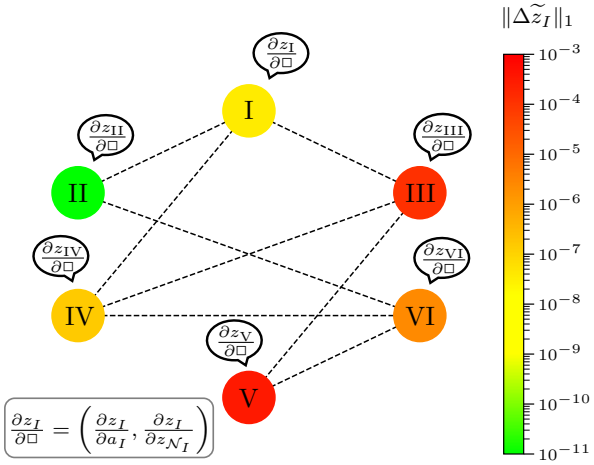

Fig. 5. If some subsystems are attacked, the disturbance propagates through the network such that the measured coupling values $\widetilde{z}$ deviate from the expected nominal values $\bar{z}$. The figure shows a snapshot of a partition of the IEEE 30 bus system at one point in time, where the node colors indicate the total coupling deviation in each subsystem on a logarithmic scale.

The deviations $\Delta\widetilde{z}$ in the coupling variables together with the exchanged sensitivity matrices $\frac{\partial \zeta_I}{\partial a_I}$, $\frac{\partial \zeta_I}{\partial z_{\mathcal{N}_I}}$ are the basis for the attack identification method from Section 4, which solves the signal recovery problem (7). While the global model contains $d_x = 60$ states and $d_u = 30$ control inputs, the considered partition from Fig. 3 yields only $d_z = 18$ coupling variables and the subsystems are of sufficiently small size to be handled by robust NMPC. Assuming that instead of the sensitivity matrices $\frac{\partial \zeta_I}{\partial a_I}$ only full-rank submatrices are transmitted as explained at the end of Section 4, the identification problem (7) contains at most $d_z + \sum_{I\in\mathcal{P}} \min\{d_{u_I}, d_{z_I}\} = 2d_z = 36$ variables instead of $d_x + d_u = 90$, which underlines one of the benefits of our hierarchical method. A second advantage is that it is not based on testing a finite number of known possible attacks, but provides a systematic model-based approach to identify any possible attack. The identification problem is solved at each time step, given the respective coupling deviations and sensitivity matrices evaluated at this time step. For weighting the tracking cost and the sparsity cost in (7), we use $\alpha_1 = 10^8$ and $\alpha_2 = 1$. In order to evaluate the proposed identification method, we compare the computed solution $\Delta a^*(t)$ at each time step t with the actual deviation $\Delta a(t) = a(u(t)) - u(t)$ caused by the really occurring, unknown attack a. For all time steps $t = t_a, \ldots, t_a + 4\Delta t$ at which the attackers are active, Fig. 6 shows the actual total deviation $\|\Delta a_I(t)\|_1$ for each subsystem as a bar in the upper half of the figure, and the presumed deviation $\|\Delta a_I^*(t)\|_1$ according to the solution of (7) with a lighter bar in the lower half. A logarithmic scale is chosen to visualize also small values in Δa^*. The solutions $\Delta a^*(t)$

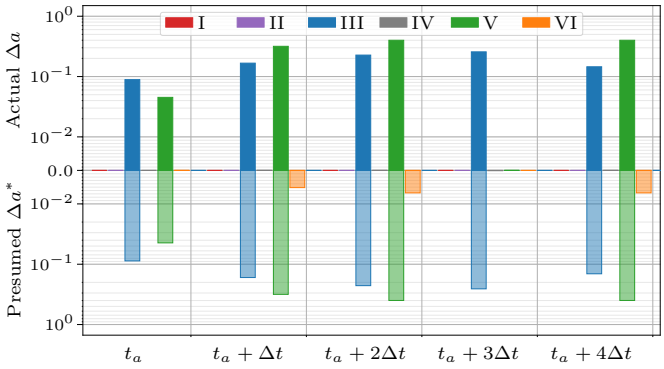

Fig. 6. Comparison of the actual, but unknown deviations $\Delta a(t)$ and the computed presumed deviations $\Delta a^*(t)$ in per unit on a logarithmic scale for five time steps during which the system is attacked. While there are small differences between reality and presumption, partly suspecting the benign subsystem VI, the proposed method clearly identifies the disturbing subsystems III and V with presumed deviation values Δa^* that are very close to the actual deviations Δa.

serve as a basis for suspecting individual subsystems at each point t in time. As explained in Section 4, we introduce a threshold ε, e.g., we choose $\varepsilon = 0.01$, and declare each subsystem I as suspicious at time t for which $\|\Delta a_I^*(t)\|_1$ exceeds ε. Hence, it is crucial that $\Delta a_I^* \approx \Delta a_I$ not only for the attacked subsystems (being subsystems III and V in the considered example), but also for the remaining subsystems. If the former is not true for some attacked subsystem, the attack might not be identified at time step t. If the latter is not true, some subsystem which actually does not actively disturb the network might be suspected. Fig. 6 shows that the computed solutions $\Delta a^*(t)$ for each $t = t_a, \ldots, t_a + 4\Delta t$ are very similar to the actual, unknown deviations $\Delta a(t)$. All differences between Δa and Δa^* are in the order of 10^{-3}. Small discrepancies can for example be seen at time steps $t_a + \Delta t$, $t_a + 2\Delta t$ and $t_a + 4\Delta t$ when the benign subsystem VI gets assigned deviation values $\Delta a_{\mathrm{VI}}^*(t) \in [0.005, 0.007]$. Despite these small inaccuracies, Fig. 6 clearly reveals that the deviations Δa_{VI}^* assigned to subsystem VI are significantly smaller compared to those of the attacked subsystems III and V. With the chosen threshold $\varepsilon = 0.01$, in all time steps the attacked subsystems are identified while none of the benign subsystems is suspected.

Remark 2. It should be noted that the computation of the sensitivities in the propagation equations (6) only requires the exchange of nominal trajectories $\bar{z}_I$ but not necessarily

full contracts $\mathcal{Z}_I$. As a consequence, the proposed attack identification can also be applied with a non-robust NMPC scheme as long as nominal trajectories are exchanged and thus constitutes a powerful method towards system resilience in its own right. However, we think that in many safety critical applications it is reasonable to apply robust schemes for guaranteed constraint satisfaction. To decrease the entailed conservativeness and speed up computation times, one can scale "how robust" the controllers are designed, for example by choosing more or less conservative approximations of the attack set $\mathcal{A}$ or even by applying non-robust MPC at time periods during which attacks are considered highly unlikely.

6. CONCLUSION

We have designed a novel attack identification algorithm for systems of systems which is based on the mutual exchange of sensitivity matrices that allow to approximate the propagation of an attack through the network. It is combined with a distributed predictive control scheme which leads to a scalable control method. The hierarchical attack identification, while requiring information from all subsystems, is formulated as a smaller-size convex quadratic program that can be solved efficiently. An important advantage of the proposed scheme is that it is not based on testing a finite possible number of attacks but provides a systematic, model-based approach towards attack identification. By successfully identifying attacks in the IEEE 30 bus system, we verified the potential of the proposed method for a practically relevant scenario.

REFERENCES

Andersson, J., Gillis, J., Horn, G., Rawlings, J., and Diehl, M. (2019). CasADi: a software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11, 1–36.

Boem, F., Riverso, S., Ferrari-Trecate, G., and Parisini, T. (2018). Plug-and-play fault detection and isolation for large-scale nonlinear systems with stochastic uncertainties. *IEEE Transactions on Automatic Control*, 64, 4–19.

Campo, P. and Morari, M. (1987). Robust model predictive control. In *IEEE American Control Conference*, 1021–1026.

Camponogara, E., Jia, D., Krogh, B., and Talukdar, S. (2002). Distributed model predictive control. *IEEE Control Systems Magazine*, 22, 44–52.

Candès, E. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51, 4203–4215.

Christofides, P., Scattolini, R., de la Pena, D., and Liu, J. (2013). Distributed model predictive control: A tutorial review and future research directions. *Computers & Chemical Engineering*, 51, 21–41.

De Tuglie, E., Iannone, S., and Torelli, F. (2008). A coherency-based method to increase dynamic security in power systems. *Electric Power Systems Research*, 78, 1425–1436.

Ding, S. (2008). *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer.

Dunbar, W. and Murray, R. (2006). Distributed receding horizon control for multi-vehicle formation stabilization. *Automatica*, 42, 549–558.

Farina, M. and Scattolini, R. (2012). Distributed predictive control: A non-cooperative algorithm with neighbor-to-neighbor communication for linear systems. *Automatica*, 48, 1088–1096.

Kundur, P., Balu, N., and Lauby, M. (1994). *Power system stability and control*. McGraw-Hill New York.

Liu, L., Esmalifalak, M., Ding, Q., Emesih, V., and Han, Z. (2014). Detecting false data injection attacks on power grid by sparse optimization. *IEEE Transactions on Smart Grid*, 5, 612–621.

Lucia, S., Finkler, T., and Engell, S. (2013). Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty. *Journal of Process Control*, 23, 1306–1319.

Lucia, S., Kögel, M., and Findeisen, R. (2015). Contract-based predictive control of distributed systems with plug and play capabilities. *IFAC-PapersOnLine*, 48, 205–211.

Lucia, S., Paulen, R., and Engell, S. (2014). Multi-stage nonlinear model predictive control with verified robust constraint satisfaction. In *IEEE Conference on Decision and Control*, 2816–2821.

Lucia, S., Tătulea-Codrean, A., Schoppmeyer, C., and Engell, S. (2017). Rapid development of modular and sustainable nonlinear model predictive control solutions. *Control Engineering Practice*, 60, 51–62.

Negenborn, R., De Schutter, B., and Hellendoorn, H. (2008). Multi-agent model predictive control for transportation networks: Serial versus parallel schemes. *Engineering Applications of Artificial Intelligence*, 21, 353–366.

Pasqualetti, F., Carli, R., Bicchi, A., and Bullo, F. (2010). Identifying cyber attacks via local model information. In *IEEE Conference on Decision and Control*, 5961–5966.

Pasqualetti, F., Dörfler, F., and Bullo, F. (2013). Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58, 2715–2729.

Sahlodin, A. and Chachuat, B. (2011). Convex/concave relaxations of parametric ODEs using Taylor models. *Computers & Chemical Engineering*, 35, 844–857.

Scattolini, R. (2009). Architectures for distributed and hierarchical model predictive control – a review. *Journal of Process Control*, 19, 723–731.

Shames, I., Teixeira, A., Sandberg, H., and Johansson, K. (2011). Distributed fault detection for interconnected second-order systems. *Automatica*, 47, 2757–2764.

Venkat, A., Rawlings, J., and Wright, S. (2005). Stability and optimality of distributed model predictive control. In *IEEE Conference on Decision and Control*, 6680–6685.

Wächter, A. and Biegler, L. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106, 25–57.

Zimmerman, R., Murillo-Sánchez, C., and Thomas, R. (2010). Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26, 12–19.