Differentially-Private Distributed Fault Diagnosis for Large-Scale Nonlinear Uncertain Systems *

Vahab Rostampour * Riccardo Ferrari * André M.H. Teixeira ** Tamás Keviczky *

* Delft Center for Systems and Control, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands.
(e-mail: {v.rostampour, r.ferrari, t.keviczky}@tudelft.nl)
** Department of Engineering Sciences, Uppsala University, PO Box 534, SE-75121, Uppsala, Sweden.
(e-mail: andre.teixeira@angstrom.uu.se)

Abstract: Distributed fault diagnosis has been proposed as an effective technique for monitoring large scale, nonlinear and uncertain systems. It is based on the decomposition of the large scale system into a number of interconnected subsystems, each one monitored by a dedicated Local Fault Detector (LFD). Neighboring LFDs, in order to successfully account for subsystems interconnection, are thus required to communicate with each other some of the measurements from their subsystems. Anyway, such communication may expose private information of a given subsystem, such as its local input. To avoid this problem, we propose here to use differential privacy to pre-process data before transmission.

© 2018, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.

Keywords: Privacy Preserving, Differential Privacy, Distributed Fault Diagnosis, Uncertain Network of Nonlinear Systems.

1. INTRODUCTION

The problem of fault diagnosis and security for large scale nonlinear systems such as critical infrastructures or interconnected Cyber Physical Systems (CPS) have received increasing attention in the recent years (Kyrriakides and Polycarpou (2014)). Indeed, one way to increase the resiliency of such systems to faults or deliberate cyber attacks is to endow them with architectures capable of monitoring, detecting, isolating and counteracting such anomalies and threats. Such systems being large scale, centralized monitoring and diagnosing architectures are rarely feasible, thus favoring distributed or decentralized ones. While decentralized solutions do not require communication between diagnosis nodes, they are not able to account for interconnection effects between different parts, or subsystems, of the large scale system being monitored. As this may lead to unacceptable performances, distributed methods, which instead do require communication, are thus preferable (Ferrari et al. (2012); Zhang and Zhang (2013); Zhang et al. (2013); Ge and Han (2014); Riverso et al. (2016); Noursadeghi and Raptis (2017)). One unexplored issue about the implementation of such distributed schemes, regards indeed the necessity of communication between neighbouring nodes. In the case where such nodes may be operated by different, possibly competing entities, mutual communication may be opposed as it may lead to leaking privacy-sensitive information. We may consider as an example a smart grid where neighbouring diagnosis nodes are each monitoring different subgrids with distributed energy sources and each is managed by its own grid operator. The two grid operators must exchange data about nodes on their respective boundaries in order to allow for grid balancing, but they would rather keep private the way that they are allocating energy supply to their different energy sources and satisfying their energy demand (Han et al. (2014b); Sankar et al. (2011)). A powerful and mathematically rigorous concept for dealing with privacy problems is differential privacy. This concept emerged in the Computer Science community (Dwork et al. (2006, 2014)), but recently found applications in Control Systems as well (see for instance Han et al. (2014b,a, 2017); Le Ny and Pappas (2014); Mo and Murray (2017)). It assumes that each piece of user data whose privacy must be protected is contained in a separate record in a database. A trusted party, called curator, maintains such database and answers queries posed by possibly adversarial, external parties. Differential privacy aims at modifying the query output to guarantee that no adversarial can guess whether a single record is present or has been altered, either by combining the results from several queries, or using side-channel information. In the previous example, the role of user data is taken by the local input applied to a subgrid, while the query corresponds to the communication of a subsystem boundary values to adversarial neighbours, such values being dependent on the subsystem local input...
2. PROBLEM STATEMENT

In this paper we will consider the case of a large-scale dynamical system $\mathcal{S}$, originating from the interconnection of $N$ smaller subsystems $\mathcal{S}_I$, $I = 1, \ldots, N$. Following Ferrari et al. (2012), we will allow each subsystem to be monitored by a dedicated agent $\mathcal{L}_I$, called Local Fault Detector (LFD), having access to locally available information, coming from measurements on its subsystem, and information from neighboring agents (see Fig. 1).

2.1 Large-scale System Dynamics

We will assume $\mathcal{S}$ to be described by the following nonlinear uncertain discrete time system

$$
\begin{align*}
    x_{k+1} &= g(x_k, u_k, w_k, f_k) \\
    y_k &= x_k + v_k
\end{align*}
$$

(1)

where $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$ and $y_k \in \mathbb{R}^n$ are the state, the input and the output of $\mathcal{S}$ at discrete time index $k \in \mathbb{N}$, respectively, while $g : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}^n$ models the state dynamics. The variable $w_k \in \mathbb{R}^p$, instead, represents unavoidable modeling uncertainties affecting eq. (1), while $f_k \in \mathcal{F} \subseteq \mathbb{R}^q$ represents a parametrization of the whole class of faults that can affect $\mathcal{S}$. Such formulation is purposely as general as possible, and comprises the cases where $w_k$ and $f_k$ affects the dynamics $g$ as additive or multiplicative terms, or where they affect one or more parameters that appear in the definition of $g$: we conventionally assume anyway that for $w_k = 0$ and $f_k = 0$ the nominal and healthy behavior of $\mathcal{S}$, that is in the absence of uncertainties and faults, is obtained. Furthermore, $g$ will be assumed to be differentiable and Lipschitz with respect to $u$, as detailed in Ass. 3. Finally, for the sake of simplicity here the full state is assumed to be available, up to a measurement uncertainty $v_k \in \mathbb{R}^m$: the extension to general input-output systems could be addressed similarly to Ferrari et al. (2008).

Assumption 1. No faults act on the system, that is $f_k = 0$, for $0 \leq k < k_f$, with $k_f$ being the anomaly occurrence time. Moreover, the variables $x_k$ and $u_k$ remain bounded before and after the occurrence of an anomaly, i.e., there exist some stability regions $\delta = \delta^+ \times \delta^- \subset \mathbb{R}^n \times \mathbb{R}^m$, such that $(x_k, u_k) \in \delta$ for all $k$.

Assumption 2. $w_k$ and $v_k$ are random variables defined on some probability spaces $(\mathcal{W}, \mathcal{B}(\mathcal{W}), \mathcal{P}_W)$, and $(\mathcal{V}, \mathcal{B}(\mathcal{V}), \mathcal{P}_V)$, respectively, where $\mathcal{W} \subseteq \mathbb{R}^p$, $\mathcal{V} \subseteq \mathbb{R}^m$, $\mathcal{B}(\cdot)$ denotes a Borel $\sigma$-algebra, and $\mathcal{P}_W, \mathcal{P}_V$ are probability measures defined over $\mathcal{W}$, $\mathcal{V}$, respectively. Furthermore, $w_k$ and $v_k$ are not correlated and are independent from $x_k$, $u_k$ and $f_k$, $\forall k$.

2.2 Sub-systems Dynamics

We assume that $\mathcal{S}$ can be described through a non-overlapping decomposition $\mathcal{D}$ into $N$ subsystems $\mathcal{S}_I$, with $I \in \{1, \ldots, N\}$, each defined via an extraction index $n_I$-tuple $I$ (see Ferrari et al. (2012)). It is then possible to define a local state $x_{I,k} \in \mathbb{R}^{n_I}$, where $x_{I,k} := \text{col}(x_{I}^{(i)} : i = I^{(j)}, j = 1, \ldots, n_I)$, and similarly a local output $y_{I,k}$ and a local measurement uncertainty $v_{I,k}$. The local input $u_{I,k}$ is instead built with all the components of $u_k$ that structurally affect at least one component of $x_{I,k+1}$, and similarly for the local $w_{I,k}$ and $f_{I,k}$.

Definition 1. A variable $c$ structurally affects a variable $a = b(c, d)$ through a multi-input function $b$, and is written $c \rightarrow a$, if there exists at least a pair of distinct values $\bar{c}$ and $\bar{c}'$ and a value $\bar{d}$ such that $\bar{a} = b(\bar{c}, \bar{d})$ is distinct from $\bar{a}' = b(\bar{c}', \bar{d})$.

Remark 1. It is important to stress that here we are only assuming that we have a structural knowledge of the effect of $w_k$ and $f_k$ on each component of $g$. This does not preclude the capability for our problem formulation to capture the case where the uncertainty, or the anomaly, are not parametric and arbitrary. For instance, we could assume in this case the dynamics to be decomposable as $g(x_k, u_k, w_k, f_k) = g^*(x_k, u_k) + w_k + f_k$, where $g^*$ represents the nominal dynamics, and $w_k$ and $f_k$ are arbitrarily varying signals, but respecting Assumptions 1 and 2.

We can proceed further and describe the dynamics of the generic subsystem $\mathcal{S}_I$ as
we can introduce a stability region as Assumption 1 holds for the original system.

Remark 2. The role of the local dynamics $g$ in the right order the components of $g$ that are contained in the index tuple $I_i$. In general we cannot assume that all the resulting subsystems $S_i$ are decentralized, i.e. their dynamics depend only on the local state $x_i$, therefore we introduced the interconnection variable $x_{NI,i}$ as in Ferrari et al. (2012).

Definition 2. The interconnection variable $x_{NI,i} \in \mathbb{R}^{n_{NI}}$ of the subsystem $S_i$ is the vector $x_{NI,i} := \text{col}(x_{i,:}^{(j)} \ni x_{I,i+1}, i \in \{1, \ldots, n_I\})$.

The role of $x_{NI,i}$ is to describe the functional dependence of the local dynamics $g_i$ on state components from other subsystems, which we will call neighboring subsystems or simply neighbors. The set of all the neighbors of $S_i$ will be denoted by $N_I$.

Remark 2. As Assumption 1 holds for the original system $S$, then it will continue to do so for every subsystem and we can introduce a stability region $S_i$ for each one, where the local state $x_i$ and input $u_i$ are assumed to always belong. Similarly, we can easily build the domains $\mathcal{V}_I$, $\mathcal{W}_I$, $\mathcal{F}_I$ and $\mathcal{V}_{NI}$ of, respectively: the local measurement and modeling uncertainties, the local fault parameters, and the measurement uncertainties of the interconnection variable.

2.3 Residual Generator

For fault detection purpose each LFD $\mathcal{L}_I$ shall compute a residual $r_{I,k} := y_{I,k} - \hat{y}_{I,k}$ and compare it to a dynamic detection threshold. In this subsection the residual will be addressed. As a direct extension of Rostampour et al. (2017), it shall be obtained as the output estimation error of the following nonlinear estimator

$$
\hat{x}_{I,k+1} = g_I(x_{I,k}, u_{I,k}, \Delta x_{NI,k}, 0, 0) + \Lambda (y_{I,k} - \hat{y}_{I,k}) ,
\hat{y}_{I,k} = \hat{x}_{I,k} \tag{3}
$$

where $\hat{x}_I, \hat{y}_I \in \mathbb{R}^{n_I}$ are, respectively, the local state and output estimates, $y_{NI,k} \in \mathbb{R}^{n_{NI}}$ are the measurements of the interconnection variables $x_{NI,k}$, $\Lambda \triangleq \text{diag}(\lambda^i, i = 1 \ldots n_I)$ is a diagonal matrix, and $\lambda^i \in (0,1)$ denotes some filtering parameters chosen to guarantee the stability of the estimator.

By using eqs. (1) and (3), we can then write the residual dynamics as

$$
r_{I,k+1} = \Lambda r_{I,k} + \delta_{I,k} , \tag{4}
$$

where the total uncertainty $\delta_{I,k}$ is a stochastic process representing the uncertain part of the residual dynamics: $\delta_{I,k} := g_I(x_{I,k}, u_{I,k}, \Delta x_{NI,k}, w_{I,k}, f_{I,k}) - g_I(y_{I,k} - v_{I,k}, u_{I,k}, \Delta y_{NI,k}, v_{NI,k}, w_{I,k}, f_{I,k}) + g_I(y_{I,k} - v_{I,k}, u_{I,k}, \Delta y_{NI,k}, v_{NI,k}, w_{I,k}, \hat{f}_{I,k}) - g_I(y_{I,k} - v_{I,k}, u_{I,k}, \Delta y_{NI,k}, v_{NI,k}, w_{I,k}, f_{I,k})$.

(5)

Thanks to Ass. 1, 2 and eq. (5), it follows that, given $y_{I,k}$, $u_{I,k}$, and $y_{NI,k}$, $\delta_{I,k}$ is a conditioned random variable on a probability space $(\Delta_{I,k}, \mathfrak{B}(\Delta_{I,k}), \mathbb{P}_{\Delta_{I,k}})$, where $\Delta_{I,k}$ is a time varying set defined as follows.

Definition 3. The time varying total uncertainty set $\Delta_{I,k} \subset \mathbb{R}^{n_I}$ at time index $k$ is defined as

$$\Delta_{I,k} := \{ \delta_{I,k} \mid r_{I,k} = y_{I,k} - \hat{y}_{I,k}, u_{I,k}, \Delta y_{NI,k}, v_{NI,k}, w_{I,k} \in \mathcal{W}_I, f_{I,k} \in \mathcal{F}_I, v_{I,k} \in \mathcal{V}_I, v_{I,k+1} \in \mathcal{V}_I, v_{NI,k} \in \mathcal{V}_{NI} \} ,$$

with $\delta_{I,k}$ being computed according to (5).

As a special case of Definition 3, we introduce the uncertainty set corresponding to a healthy plant as follows.

Definition 4. The time varying healthy total uncertainty set $\Delta_{I,k}^0 \subset \mathbb{R}^{n_I}$ at time index $k$ is defined as

$$\Delta_{I,k}^0 := \{ \delta_{I,k} \mid r_{I,k} = y_{I,k} - \hat{y}_{I,k}, u_{I,k}, \Delta y_{NI,k}, v_{NI,k} \in \mathcal{W}_I, v_{I,k+1} \in \mathcal{V}_I, u_{I,k}, v_{NI,k} \in \mathcal{V}_{NI} \} ,$$

where $\delta_{I,k}$ is computed according to (5).

Remark 3. The role of $\Delta_{I,k}$ and $\Delta_{I,k}^0$ is to quantify the range of possible values that $\delta_{I,k}$ can take, respectively, corresponding to situations when a fault may be present, and in a healthy situations where a fault is absent. Apart from simple cases, no closed form is available for computing such sets, and numerical approximations techniques such as those in Dabbene et al. (2015) may be used.

We can now introduce a compact notation for the residual generator described by eqs. (3),(4),(5), through a mapping function $\Sigma_I : \mathbb{R}^{n_I} \times \mathbb{R}^{n_I} \rightarrow \mathbb{R}^{n_I}$ defined as

$$r_{I,k+1} = \Sigma_I(r_{I,k}, \delta_{I,k}) := \Lambda r_{I,k} + \delta_{I,k} . \tag{6}$$

Remark 4. While at time index $k$ the residual $r_{I,k}$ can be computed from $y_I$ and $\hat{y}_I$ and is thus a deterministic quantity, from (4), (6) it follows that the next value $r_{I,k+1}$ is a random variable on the same probability space as $\delta_{I,k}$.

Given these preliminaries, it is now possible to write the following two fundamental definitions (see Fig. 2).

Definition 5. The time varying residual set $\mathcal{R}_{I,k+1}$ at index $k + 1$ is defined as the image of the set $\Delta_{I,k}$ through $\Sigma_I$, that is

$$\mathcal{R}_{I,k+1} := \Sigma_I(\delta_{I,k}, \Delta_{I,k}) = \{ r_{I,k+1} \mid r_{I,k+1} = \Sigma_I(\delta_{I,k}, \Delta_{I,k}) \} ,$$

Definition 6. The time varying healthy residual set $\mathcal{R}_{I,k+1}^0$ at index $k + 1$ is defined as the image of the set $\Delta_{I,k}^0$ through $\Sigma_I$, that is

$$\mathcal{R}_{I,k+1}^0 := \Sigma_I(\delta_{I,k}, \Delta_{I,k}^0) = \{ r_{I,k+1} \mid r_{I,k+1} = \Sigma_I(\delta_{I,k}, \Delta_{I,k}^0) \} .$$

For ease of notation, when there is no ambiguity, in the rest of the paper we will drop the index $I$ to denote that a quantity refers to the generic subsystem $S_I$ or the generic LD agent $\mathcal{L}_I$. The index $N$ will be retained to indicate the neighbor set of the generic subsystem or agent.

2.4 Fault Detection Threshold Design Problem

In order to reduce the detrimental effects on fault detectability of deterministic thresholds, which in practice can be overly conservative, in this paper we will seek a probabilistically robust threshold instead (see Rostampour et al. (2017)). In particular, by extending Boem et al.
978

the desired probabilistic robustness of the threshold $\bar{\delta}_r$.

Remark 5. Whether $\delta_{r, k}$ vary over its domain $\Delta_{r, k}$ and fixing the residual $r_{I, k}$ to its actual value. The domain $\Delta_{r, k}$ in turn is computed through eq. (5) by letting $v_{I, k}$, $w_{I, k}$, $f_{I, k}$ and $v_{N_{I}, k}$ vary over their respective domains, and fixing the local output and input $y_{I, k}$ and $u_{I, k}$, as well as the interconnection variables measurement $y_{N_{I}, k}$, to their actual values. The healthy residual set $\mathcal{R}_{I, k+1}$ can be obtained similarly, but by fixing the value $f_{I, k} \in \{0\}$.

(2015) and Ferrari et al. (2017), we propose the following residual evaluation logic and threshold for fault detection

$$d_M(r_{I, k+1}) \leq \bar{d}_M \triangleq \frac{n_f}{\alpha} \Rightarrow \mathcal{S}_I \text{ is healthy}$$

(7)

based on the Mahalanobis distance of the residual

$$d_M(r) \triangleq \sqrt{(r - \mu_r)^T C_r^{-1} (r - \mu_r)}$$

(8)

where $\mu_r \triangleq \mathbb{E}[r] \in \mathbb{R}^{n_I}$ and $C_r \triangleq \text{Cov}[r] \in \mathbb{R}^{n_I \times n_I}$ are the expected value and covariance matrix of the random variable $r_I$. Indeed, thanks to the Multivariate Chebyshev Inequality (see Chen (2007)), we can bound the probability of false positives during healthy conditions as

$$\mathbb{P} \left[ d_M(r_I) > \bar{d}_M \right] < 1 - \alpha,$$

(9)

where $\alpha \in (0, 1)$ is a user defined constant representing the desired probabilistic robustness of the threshold $\bar{d}_M$ and $\mathbb{P}$ denotes probability.

Remark 5. While the detection logic (7) employs a static threshold, it must be noted that it is equivalent to testing whether $r_{I, k+1}$ belongs to a time varying ellipsoid centered in $\mu_{r_{I, k+1}}$ and whose orientation and size are described by $C_{r_{I, k+1}}$. Indeed the moments of $r_{I, k+1}$ depend via the mapping $\Sigma_I$ on those of $\delta_{r, k}$, which are not assumed to be time invariant unless $g_I$ is a linear function and $v$ and $w$ are stationary processes. In such case and also when $g_I$ is bilinear the techniques proposed in Ferrari et al. (2017) can be employed.

Remark 6. For the general nonlinear case we will assume that the moments can be approximated by their corresponding sample moments by generating a number $N$ of samples $r_{I, k+1}^j$, with $j = 1 \ldots N$, as in Boem et al. (2015) and Rostampour et al. (2017).

3. DIFFERENTIALLY PRIVATE FRAMEWORK

The distributed fault detection scheme outlined in the previous section requires every agent $\mathcal{L}_I$ to have the following quantities communicated by its neighbours: their output components appearing in $y_{N_I}$, which are needed to update the nonlinear estimator (3) and thus generate the residual $r_I$; and $N$ samples of the measurement uncertainties appearing in $v_{N_{I}, k}$, such that the sample moments of $r_I$ can be computed by repeated evaluation of eqs. (4) and (5), and used to evaluate the residual according to (7).

The goal of this section is indeed to show how, relying on concepts from Differential Privacy (DP), an agent $\mathcal{L}_I$ and its neighbours may communicate with each other without exposing private information on their local input. The next subsection will introduce the basics of DP.

3.1 The Concept of Differential Privacy

Differential privacy “addresses the paradox of learning nothing about an individual while learning useful information about a population” (Dwork et al. (2014)). The initial concern that drove its development is in fact protecting the privacy of human individuals, for instance when personal health data is collected and used in medical studies.

As a preliminary notion, we need to introduce the concepts of database and of query.

Definition 7. A database $D$ of length $n$ is a set $D = \{d_1, d_2, \ldots, d_n\}$ taking values in $\mathbb{D}$, where $\mathbb{D}$ is the universe of all possible databases.

Definition 8. A query $q$ is a mapping $q : \mathbb{D} \rightarrow \mathbb{R}^{n_q}$, where $n_q$ is the size of the result provided by the query.

In DP it is assumed that data contained in a database $D$ can be accessed only through the results of queries, which are answered by the subject holding $D$, called curator. Protecting the privacy of an element $d_i$ in $D$ can thus be obtained by making the results of any query run on $D$ insensitive enough to the single $d_i$. This can also be expressed by ensuring that two adjacent databases are nearly indistinguishable from the answers to a query.

Definition 9. (Han et al. (2017)) Two databases $D = \{d_1, \ldots, d_n\}$ and $D' = \{d'_1, \ldots, d'_n\}$ are said to be adjacent, and it is written as $\text{adj}(D, D')$, if there exists $i \in \{1, \ldots, n\}$ such that $d_j = d'_j$ for all $j \neq i$.

This is enforced by introducing so-called mechanisms, which are randomized mappings from the universe $\mathbb{D}$ to some subset in $\mathbb{R}^{n_q}$, and letting the curator use the mechanism in lieu of the query. A mechanism that acts on a database is said to be differentially private if it complies with the following definition from Dwork et al. (2006).

Definition 10. Given $\epsilon \geq 0$ as the desired level of privacy, a mechanism $M$ preserves $\epsilon$–differential privacy if for all $\mathbb{R} \subset \text{range}(M)$ and all adjacent databases $D$ and $D'$ in $\mathbb{D}$, it holds that

$$\mathbb{P} \left[ M(D) \in \mathbb{R} \right] \leq e^\epsilon \mathbb{P} \left[ M(D') \in \mathbb{R} \right].$$

(10)

Remark 7. A smaller $\epsilon$ implies higher level of privacy. By using differential privacy, one can hide information at the individual level, no matter what side information others may have. Definition 10 shows that DP is based on randomization, but is independent on the contents of databases, as long as they belong to $\mathbb{D}$ and are adjacent.
A popular mechanism in the DP literature is the so-called Laplace mechanism, that introduces a Laplacian additive noise dependent on the query \( \ell_p \)-sensitivity

**Definition 11.** (Han et al., 2017, Definition 10) For any query \( q : D \rightarrow \mathbb{R}^n \), the \( \ell_p \)-sensitivity of \( q \) under the adjacency relation, adj, is defined as

\[
\sigma := \max \{|q(D) - q(D')|_p : D, D' \in D \text{ s.t. } \text{adj}(D, D')\}.
\]

It is worth mentioning that \( \ell_p \)-sensitivity of \( q \) does not depend on a specific database \( D \). We now recall the following results from (Han et al., 2014a, Theorem 9).

**Proposition 1.** Consider a query \( q : D \rightarrow \mathbb{R}^m \) whose \( \ell_2 \)-sensitivity is \( \sigma \). Define the mechanism \( M \) as \( M(D) = q(D) + \nu \), where \( \nu \in \mathbb{R}^m \) is a random vector whose probability density function is given by \( p_\nu(\nu) \propto \exp(-\epsilon \|\nu\|/\sigma) \). Then the mechanism \( M \) preserves \( \epsilon \)-differential privacy.

### 3.2 Privacy-Preserving Mechanism

The proposed privacy-preserving framework for distributed fault detection will be now presented. To simplify the notation and formulation, we will assume without loss of generality the case of a given agent \( L \) having a single neighbor \( L_N \), connected through an interconnection variable \( x_N \in \mathbb{R}^n \). We will also drop the time indexes to simplify our notation whenever possible. As said previously, \( L_N \) will send to \( L \) at each time indexes its last interconnection variable measurement \( y_N \). From the point of view of the DP formulation, agent \( L_N \) is the curator of a database that contains the local input \( u_{N,k-1} \) and that at time \( k \) is answering a query from \( L \) by providing the measurement \( y_N \), which depends on the previous local state of \( L_N \) and on \( u_{N,k-1} \) via its dynamics (2). In general it does not hold that \( u_N \) can be reconstructed from values of \( y_N \). Anyway, in the DP setting a privacy breach does not require the capability of fully reconstructing a piece of information, but only the capability of determining whether it will cause the query result to belong or not to a given set (Def. 10). This, in turns, depends on the query sensitivity (Def. 11).

For these reasons, \( L_N \) does want to replace such answer with a mechanism that guarantees the privacy of \( u_N \).

Before proceeding further, we need an extended definition of adjacency.

**Definition 12.** Two control actions \( u_N, u_N' \in U \subseteq \mathbb{R}^{m_N} \) are two adjacent control inputs at time step \( k - 1 \) if and only if \( \|u_N - u_N'\|_0 \leq 1 \), and it is written \( \text{adj}(u_N, u_N') \).

Such a distance between databases is referred to as the Hamming distance, i.e., the number of rows on which they differ. The set \( U \) is a compact set over which the input sequence \( \{u_{N,k}\}_{k=0}^\infty \) can take values.

**Remark 8.** Following Defin. 12, we can say that two adjacent control inputs belong to a bounded set \( U \) such that:

\[
\max_{i \in \{1, \ldots, m_N\}} |(u_N^{(i)} - u_N'^{(i)})| \leq 2\zeta,
\]

where \( \zeta \geq 0 \) is a positive constant number which depends on the set \( U \).

Since the query \( q(\cdot) \) answered by \( L_N \) is actually the output of the generic subsystem \( S_N \), the constant \( \sigma \) that appears in Definition 11 can be computed as

\[
\sigma_{N} = \max_{u_N, u_N' \in U} \|g_N(\psi_N, u_N) - g_N(\psi_N, u_N')\|_{\psi_N \in \Psi} \tag{11}
\]

where \( g_N(\psi_N, k-1, u_{N,k-1}) := y_{N,k} \) represents a compact notation for \( S_N \) dynamics in (2). The new quantity \( \psi_N \in \Psi \) represents the other variables, apart from the input \( u_N \), which influence \( S_N \), and is defined as \( \psi_N := (x_N, u_N, f_N, w_N, f) \), with \( \Psi := S^x \times S_N^x \times W \times F \). The bound \( \sigma_{N} \) can be seen as a bound on the global \( \ell_p \)-sensitivity of the mapping function \( g_N(\psi_N, u_N) \) with respect to the control input \( u_N \) at each time step \( k \) for all \( p \geq 1 \). The following assumption is needed to compute \( \sigma_{N} \).

**Assumption 3.** The nonlinear dynamics function \( g_N(\psi_N, u_N) \) of the generic subsystem \( S_N \) is measurable and differentiable in \( u_N \) such that at each sampling time \( k \)

\[
\frac{\partial g_N(\psi_N, u_N)}{\partial u_N} \neq 0, \forall u_N \in U, \psi_N \in \Psi,
\]

and there exists a constant \( L \) for all time step \( k \), \( u_N, u_N' \in U \) and \( \psi_N \in \Psi \) such that:

\[
\|g_N(\psi_N, u_N) - g_N(\psi_N, u_N')\|, \|\psi_N - \psi_N'\| \leq L\|u_N - u_N'\| \tag{12}
\]

where \( \psi_N \) and \( \psi_N' \) are two vectors obtained by concatenating \( \psi_N \) with \( u_N \) and \( u_N' \), respectively. We refer to \( L \) as the Lipschitz constant of the nonlinear function \( g_N(\psi_N, u_N) \) of the generic subsystem \( S_N \).

**Remark 9.** An essential factor is the differentiability of \( g_N(\psi_N, u_N) \) in order to derive the sensitivity of the output signal with respect to small variations (adjacent relations) of input control signals. The key assumption is the Lipschitz condition (12). An approximation of the Lipschitz constant \( L \) at time step \( k \) can be calculated from eq. (2) using the available values of \( \psi_N \in \Psi \) and drawing a sufficiently high number of samples of the uncertainties \( v_N \) and \( u_N \), following a Monte Carlo approach.

**Proposition 2.** The global \( \ell_2 \)-sensitivity of the output of the generic subsystem \( S_N \) is bounded by \( \sigma_{N} = 2\zeta L \).

**Proof.** Following Defin. 12 and Rem. 8 together with Ass. 3, the proof is straightforward by making use of eqs. (11) and (12), from which we can derive the inequality

\[
\max_{u_N, u_N' \in U} \|u_N - u_N'\| = \max_{u_N, u_N' \in U} \max_{i \in \{1, \ldots, m_N\}} \|u_N^{(i)} - u_N'^{(i)}\| \leq 2\zeta L .
\]

The proof is completed. \( \square \)

We are now ready to state the problem that we are going to address in the present section.

**Problem 1.** Find a randomized mechanism \( M_u \) such that it preserves \( \epsilon_u \)-differential privacy for the neighboring agent \( L_N \) under the adjacency relation described in Definition 12.

**Proposition 3.** The mechanism \( M_u(u_N) = g_N(\psi_N, u_N) + \nu_{u_N} \), where \( u_N \) is the control input signal and \( \nu_{u_N} \in \mathbb{R}^{m_N} \) is a noisy vector drawn from a probability density function that is proportional to \( \exp(-\epsilon_u \|\nu_{u_N}\|/2\zeta L) \), is \( \epsilon_u \)-differentially private.

**Proof.** The proof is the direct result of combining Proposition 2 with Proposition 1. \( \square \)
Output Signal as Database. By looking at mechanism $M_u$ in Proposition 3 it can be seen that it is equivalent to a mechanism $M_y$ acting on a database containing $y_N$, where the query is an identity. Indeed this equivalence can be easily shown. We first introduce the following

Definition 13. Two output signals $y_N$, $y'_N \in \mathcal{Y} \subseteq \mathbb{R}^{n_N}$ are two adjacent output signals if and only if $\|y_N - y'_N\|_0 \leq 1$, and it is written as $\text{adj}(y_N, y'_N)$. The set $\mathcal{Y}$ is a compact set over which the output sequence $\{y_N, k\}_{k=0}^\infty$ can take values, and since two output signals belong to $\mathcal{Y}$, we can have:

$$\max_{z \in \{1, \ldots, n_N\}} |(y_N)^{(i)} - (y'_N)^{(i)}| \leq 2\xi,$$

where $\xi \geq 0$ is a positive constant number which depends on the set $\mathcal{Y}$.

Since the query is an identity mapping, a bound $\sigma_{\mathcal{Y}_N}$ on the global $\ell_2$-sensitivity of such a query can be obtained from:

$$\sigma_{\mathcal{Y}_N} = \max_{y_N, y'_N \in \mathcal{Y}} \|y_N - y'_N\| \leq 2\xi. \quad (13)$$

The following proposition provides a randomized mechanism $M_y$ such that it preserves $\epsilon_y$-differential privacy for the agent $L_N$ under the adjacency relation of Definition 13.

Proposition 4. The mechanism $M_y(y_N) = y_N + \nu_{y_N}$, where $y_N$ is the output signal and $\nu_{y_N} \in \mathbb{R}^{n_N}$ is a noisy vector drawn from a probability density function that is proportional to $\exp(-\epsilon_y \|\nu_{y_N}\|/2\xi)$, is $\epsilon_y$-differentially private.

Proof. It directly results from combining Proposition 2 with Proposition 1.

We next provide a theoretical connection between $M_u(u_N)$ and $M_y(y_N)$.

Theorem 1. Let $M_u(u_N)$ and $M_y(y_N)$ be the two randomized mechanisms introduced in Propositions 3 and Proposition 4 for a generic nonlinear system dynamics $S_N$ such that they preserve $\epsilon_u$ and $\epsilon_y$ level of differential privacy with $\epsilon_y = \epsilon_u \frac{\xi}{\epsilon_L}$, respectively. Given $\xi$ in Remark 8 and $\xi$ in Definition 13 with $L$ in Assumption 3, if $\xi \leq \epsilon_L$, then,

$$P[M_u(u_N) \in \mathcal{R}_u] = P[M_y(y_N) \in \mathcal{R}_y] = e^{\epsilon_y} \leq e^{\epsilon_u}.$$

Proof. Following Propos. 3 together with Propos. 4, let $p_{u_N}$ and $p_{u'_N}$ denote the probability density function of $M_u(u_N)$ and $M_u(u'_N)$, respectively, and let $p_{y_N}$ and $p_{y'_N}$ denote the probability density function of $M_y(y_N)$ and $M_y(y'_N)$, respectively. We now compare $p_{u_N}$ and $p_{u'_N}$. At some arbitrary point $z \in \mathbb{R}^{n_N}$ in order to show the first inequality in the above assertion as follows:

$$\frac{p_{u_N}(z)}{p_{u'_N}(z)} = \frac{\exp\left(-\epsilon_u \|y_N(x_N, y_{N_k}) - z\|_2\right)}{\exp\left(-\epsilon_u \|y'_N(x'_N, y'_{N_k}) - z\|_2\right)} = \frac{\exp\left(-\epsilon_y \|y_N - z\|_2\right)}{\exp\left(-\epsilon_y \|y'_N - z\|_2\right)} = \frac{p_{y_N}(z)}{p_{y'_N}(z)}$$

where the second equality follows from choosing $\epsilon_y = \frac{\epsilon_u \xi}{\epsilon_L}$. Observe that $\epsilon_y \leq \epsilon_u$ holds for $\xi \leq \epsilon_L$.

It is important to highlight that Theorem 1 is the first result, to the best of our knowledge, toward privatizing a desired database, e.g. the control input actions, using another database, e.g. the output signals of a generic nonlinear system dynamics $S_N$. Theorem 1 provides a theoretical link between two randomized mechanisms $M_u(u_N)$ in Proposition 3 and $M_y(y_N)$ in Proposition 4. Strictly speaking, one can consider the output signals of a generic dynamical system $S_N$ as a database to develop a randomized mechanism $M_y(y_N)$ such that it preserves $\epsilon_y$-differential privacy together with achieving the $\epsilon_u$-differential privacy of the input control signals as the main desired privacy goal by considering that $\epsilon_y = \frac{\epsilon_u \xi}{\epsilon_L}$ and $\xi \leq \epsilon_L$.

4. NUMERICAL STUDY

In this section we are going to present the results of a numerical study, in order to illustrate the effectiveness of the proposed approach. The system under study will be a multi-tank system (see Ferrari et al. (2012) for details on modeling such a system), whose structural graph contains 22 nodes, each representing a state variable corresponding to the level of a tank, while edges represent pipes interconnecting such tanks (Fig. 3). The graph has been obtained by application of the Barabási-Albert model Albert and Barabási (2002), which as known leads to scale-free networks. After labeling the nodes according to their degree, in descending order, two sub-systems have been obtained by defining the extraction index tuples $\mathcal{I}_1 = \{1, 2, 6, 8, 9, 13, 16, 17, 19, 21, 22\}$ and $\mathcal{I}_2 = \{3, 4, 5, 7, 10, 11, 12, 14, 15, 18, 20\}$ of 11 elements each. Finally, in order to make the interconnection between
the two resulting subsystems asymmetric and thus more interesting, an edge between nodes 1 and 3 has been added, on top of the edges produced by the Barabási-Albert algorithm. The actual tanks’ cross section have been chosen on the basis of the data provided by the manufacturer. Drains with the same section as interconnecting pipes have been assumed to be connected to terminal nodes (i.e. nodes with unitary degree). A single source pump, with a sinusoidal time profile with a frequency of 0.1 Hz, has been connected to tank no. 1. All tank levels are assumed to be measured, with a gaussian measurement uncertainty with zero mean and a standard deviation equal to 0.15 m. When building the LFD estimators, a gaussian parametric uncertainty is introduced, having zero mean and a variance equal to 5% and 1.5%, respectively, of the tanks and pipes cross sections. The privacy mechanism $M_u$ will be used, with the value $\zeta = 0.01$: a number $N = N_u = 512$ of samples is used by each LFD to compute the moments $\mu_r$ and $C_r$ appearing in the Mahalanobis distance definition (8) and for generating the set $X_N$ that is communicated to neighbouring LFDs.

The fault that is presented in the current study represents a clogging in the pipe between tanks 1 and 3, reducing its flow to 50% of its nominal value. The reason we have chosen this kind of fault is that it affects exactly each subsystem interconnection variable, and as such may be hidden, that is made undetectable, by the introduction of the privacy mechanism. The following figures present the results obtained by simulating such fault occurring at time $T_f = 125$s. In particular, the effect on detection performance of varying privacy levels, that is of a varying parameter $\epsilon$, were analyzed.

Fig. 4 shows the time behaviour of tank no. 3, which is allocated to LFD no. 2 and is an interconnection variable for LFD no. 1. It can be seen how, for the smallest value of $\epsilon$ considered in this study ($\epsilon = 0.04$), the privatized version of its level can be dramatically different from the real one. Nevertheless, its estimated value by LFD no. 2 can still be relatively close to the real one, and clearly show its sensitivity to the fault happening at 125 s. In particular, Fig. 5 shows how the residual computed by LFD no. 2 is still sensitive to the fault and able to cross the threshold at about 148 s. In this figure the normalized Mahalanobis distance of the residual, that is $\frac{d_M(r_2)}{d_M}$, is plotted: a successful detection occurs when this quantity gets larger than 1. Finally, Fig. 6 presents a boxplot analysis of the effect of varying $\epsilon$ on the detection time of LFD no. 2. In order to produce such plot, a Monte Carlo approach has been used. For each value of $\epsilon$ 64 rounds were simulated, where every source of uncertainty in the system (that is $v$ and $w$) and in the mechanism $M_u$ have been implemented thanks to the default Matlab/Simulink random numbers generators. The seeds of the generators have been independently initialized before each round according to the system clock time, to avoid undesired repetitions and correlations. In particular, a formula of the type $seed = round(s(clock) \times mult + off)$ has been used, where $s$ is an arbitrary monotone function, and mult and off are real numbers that are unique for each source of uncertainty. The end results, albeit a bit unexpected, shows how, for the chosen range of values $^2$ of $\epsilon$, the median of the detection time is not significantly affected by the addition of the privacy mechanism. This first result implies that indeed the proposed privacy-preserving mechanism is feasible and will not hamper fault detection performances.

---

1 Matlab/Simulink R2016a on Mac Os X 10.11.6

2 No successful detection was obtained for values smaller than 0.04
properties of the original diagnosis scheme. Influence of the privacy mechanism on the detectability where a Monte Carlo analysis was used to show the negligible effect of the privacy mechanism applied to the subsystem output, was presented in Theorem 1. Simulation results were included, preserving the local input of each subsystem control algorithms and policies. A novel theoretical result linking the privacy level of the local input to the privacy level corresponding to a given privacy mechanism applied to the subsystem output, was presented in Theorem 1. Simulation results were included, where a Monte Carlo analysis was used to show the negligible influence of the privacy mechanism on the detectability properties of the original diagnosis scheme.

5. CONCLUSIONS

This paper presented for the first time, to the best of the authors’ knowledge, a differentially private approach to distributed fault diagnosis of large scale, nonlinear and uncertain systems. The data whose privacy must be preserved was considered to be the local input of each subsystem, which corresponds to protecting the privacy of each subsystem control algorithms and policies. A novel theoretical result linking the privacy level of the local input, to the privacy level corresponding to a given privacy mechanism applied to the subsystem output, was presented in Theorem 1. Simulation results were included, where a Monte Carlo analysis was used to show the negligible influence of the privacy mechanism on the detectability properties of the original diagnosis scheme.

REFERENCES


