Dynamic Watermarking for Security of Cyberphysical Systems

P. R. Kumar
Dept. of Electrical and Computer Engineering
Texas A&M University

With Bharadwaj Satchidanandan, Woo Hyun Ko, Tong Huang, Lantian Shangguan, Kenny Chour, Gopal Kamath, Le Xie, and Swaminathan Gopalswamy

Email: prk.tamu@gmail.com
Web: http://cesg.tamu.edu/faculty/p-r-kumar/
Cyberphysical systems

- Next generation of engineered systems in which computing, communication, and control technologies are tightly integrated
- Many societally important future applications
  - Smart grid
  - Automated transportation
  - Unmanned Air Vehicle Transportation System
  - Water treatment facilities
  - Telesurgery systems
  - ...
- Safety critical
  - Malfunctioning causes physical harm
- Critical infrastructure
  - Important to functioning of economy and society
Vulnerability of cyberphysical systems to attacks

- Hackers hitherto could tamper only with information or bits in cyber layer

- CPS tightly couples cyber and physical worlds
  - Actions in physical world taken based on information from cyber layer

- CPS, therefore, gives hackers ability to cause damage in physical world
Several attacks on critical infrastructure systems

- Maroochy-Shire, Australia, 2003, attack on sewage treatment system, commands issued which led to a series of faults in the system
- Attack on computers controlling Davis-Besse nuclear power plant in Ohio, 2003, Slammer worm disabled the safety monitoring system
- Stuxnet worm, 2010, exploited Microsoft Windows vulnerability to subvert critical computers controlling centrifuges in Iran uranium enrichment facility
- Attacks on Supervisory Control and Data Acquisition system, natural gas pipeline systems, trams, power utilities, and water systems, etc.

- And many more instances of attacks
  - Ukraine power grid
  - Water filtering plant in Pennsylvania
  - Demonstrations of cyber attacks in automated cars
  - ...
Isn't network security enough for CPS security?

- Network and information security implemented through periodic patching
  - CPS has a dynamic system in the loop, and may not admit controllers going online for patching

- Traditional notion of “Confidentiality, Integrity and Availability” in network and information security does not address real-time availability, which is critical for control system security

- Network or information security fundamentally cannot address physical layer attacks such as in Maroochy-Shire incident
Many societally critical systems are being networked
- Smart grid, Automated transportation, Unmanned Air Vehicle Transportation System, Process Control Systems

They are safety and economy critical
- Malfunctioning causes economic physical harm

There have been many attacks
- Stuxnet worm, 2010
- Ukraine power grid, 2015

Even after many decades we still cannot secure
- Operating Systems
- Internet

How can we possibly secure CPSs?
Abstraction of cyberphysical systems

- Overall system has
  - Physical plant
  - Actuators
  - Sensors
  - Routers
  - Computational nodes
  - Network

- But some of the routers, computation nodes, sensors, actuators may be compromised

- How do we secure the overall cyberphysical system?
Abstraction of security problem

- Some sensors, actuators may be compromised
- If information from a sensor is compromised, we say sensor is compromised
- It does not matter whether sensor is compromised or its information is compromised downstream

How do we secure the overall cyberphysical system when some sensors and actuators may be compromised?
Dynamic watermarking

- Actuator node superimposes a private excitation whose realization is unknown to other nodes.
Random noise $e_i(t)$ is privately added to the signal

- Private excitation “comes back” transformed from sensors
- Controller can check if reported measurements are appropriately correlated with $e(t)$
Illustration on simple first order SISO system

- \( x(t + 1) = ax(t) + bu(t) + w(t + 1) \) where \( w(t) \sim N(0, \sigma_w^2) \), i.i.d.
- Dynamic watermarking: \( u(t) = u^g(t) + e(t) \) with \( e(t) \sim N(0, \sigma_e^2) \),
- Closed-loop system is: \( x(t + 1) = ax(t) + bu^g(t) + be(t) + w(t + 1) \)
- So \( x(t + 1) - ax(t) - bu^g(t) - be(t) = w(t + 1) \sim N(0, \sigma_w^2) \)
- \( x(t + 1) - ax(t) - bu^g(t) = be(t) + w(t + 1) \sim N(0, \sigma_w^2 + b^2 \sigma_e^2) \)
- Two tests are conducted by actuator
  \[
  \lim \frac{1}{T} \sum_{t=0}^{T-1} \left( z(t + 1) - az(t) - bu^g(t) - be(t) \right)^2 = \sigma_w^2
  \]
  \[
  \lim \frac{1}{T} \sum_{t=0}^{T-1} \left( z(t + 1) - az(t) - bu^g(t) \right)^2 = b^2 \sigma_e^2 + \sigma_w^2
  \]
- If either test fails, then there is malicious sensor information
  - System goes into safety mode: Halted, rebooted, manual operation, etc.
Fundamental Guarantee provided by Dynamic Watermarking

- **Theorem**
  - Let \( v(t + 1) := z(t + 1) - az(t) - bu^g(t) - be(t) - w(t + 1) \)

- Then \( \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} v^2(t) = 0 \)

- **Interpretation:**
  \[
  z(t + 1) - az(t) - bu^g(t) - be(t) = w(t + 1) + v(t + 1)
  \]

- So reported sensor measurements can distort actual noise \( w(t) \) only by zero power signal \( v(t) \)
Proof - 1

- From the first test: \( \lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} (v[k] + w[k])^2 = \sigma_w^2 \)

- So \( \lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} v^2[k] + \lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} 2v[k]w[k] = 0 \)

- Goal is to show: \( \lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} w[k]v[k] = 0 \)

- From second test: \( \lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} (v[k] + be[k - 1] + w[k])^2 = b^2 \sigma_e^2 + \sigma_w^2 \)

- So \( \lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} e[k - 1]v[k] = 0 \)

- Would like to replace \( e[k-1] \) by \( w[k] \)

Need to use the fact that the attacker cannot distinguish between watermark \( e(k-1) \) and noise \( w(k) \)
Proof - 2

Since \[ w[k] = x[k] - ax[k-1] - bg_{k-1}(z^{k-1}) - be[k-1] \]

\((x^{k-2}, e^{k-2}) \rightarrow (x[k-1], x[k], z^k) \rightarrow w[k] + be[k-1] \rightarrow w[k]\) Markov Chain

Define the \(\sigma\)-algebra \(S_k := \sigma(x^k, z^k, e^{k-2})\)

Then \(S_k = \sigma(x^{k-2}, e^{k-2}, x[k-1], x[k], z^k, w[k] + be[k-1])\)

So \(\hat{w}[k] := E[w[k] | S_k]\)

\[
\begin{align*}
\hat{w}[k] &= E[w[k] | w[k] + e[k-1]] \\
&= \frac{\sigma_w^2}{b^2 \sigma_e^2 + \sigma_w^2} (be[k-1] + w[k]) = \beta (be[k-1] + w[k])
\end{align*}
\]
Proof - 3

- Let \( \tilde{w}[k] := w[k] - \hat{w}[k] \)

- Then \( \tilde{w}[k] \in S_k \) and \((\tilde{w}[k], S_k)\) is a Martingale difference sequence

- Also \( v[k] \in S_k \), in fact \( v[k] \in \sigma(x^k, z^k) \)

- By Martingale Stability Theorem

\[
\sum_{k=1}^{T} v[k] \tilde{w}[k] = o\left(\sum_{k=1}^{T} v^2[k]\right) + O(1)
\]

- Now

\[
\sum_{k=1}^{T} v[k] w[k] = \sum_{k=1}^{T} v[k](\hat{w}[k] + \tilde{w}[k]) = \sum_{k=1}^{T} v[k] \hat{w}[k] + o\left(\sum_{k=1}^{T} v^2[k]\right) + O(1)
\]
Proof - 4

- So \( \sum_{k=1}^{T} v[k] w[k] = \beta b \sum_{k=1}^{T} v[k] e[k-1] + \beta \sum_{k=1}^{T} v[k] w[k] + o(\sum_{k=1}^{T} v^2[k]) + O(1) \)

- Hence \( \sum_{k=1}^{T} v[k] w[k] = \frac{\beta b}{1 - \beta} \sum_{k=1}^{T} v[k] e[k-1] + o(\sum_{k=1}^{T} v^2[k]) + O(1) \)

- Now \( \sum_{k=1}^{T} v[k] e[k-1] = o(T) \)

- So it follows that \( \sum_{k=1}^{T} v[k] w[k] = o(\sum_{k=1}^{T} v^2[k]) + o(T) + O(1) \)

- So \( \sum_{k=1}^{T} v^2[k] + \sum_{k=1}^{T} 2v[k] w[k] = (1 + o(1))(\sum_{k=1}^{T} v^2[k]) + o(T) \)

- Taking limits and dividing by \( T \) gives the result
Stability consequences of Fundamental Guarantee of Dynamic Watermarking

**Theorem:**

- Suppose $|a| < 1$, i.e., system is open-loop stable,

- Then distortion $d[t] := z[t] - x[t]$ is zero power: 
  $$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} d^2[k] = 0$$

- Mean-square performance is same as reported performance
  $$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} x^2[k] = \lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} z^2[k]$$

- Suppose $u^g(t) = fx(t)$ with $|a + bf| < 1$

- Then mean square performance is optimal
  $$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} x^2[k] = \frac{\sigma_w^2 + b^2 \sigma_e^2}{1 - |a + bf|^2}$$
More general results

- Results extend to

- ARMAX Systems used in process control:

\[
y[t] = -\sum_{k=1}^{p} a_k y[t - k] + \sum_{k=0}^{h} b_k u[t - l - k] + \sum_{k=0}^{r} c_k w[t - k]
\]

- MIMO partially observed Gaussian systems

\[
x[t + 1] = Ax[t] + Bu[t] + w[t + 1]
y[t + 1] = Cx[t + 1] + n[t + 1]
\]
Example

- **System:**
  \[ y(t + 1) + 0.7y(t) - 0.2y(t - 1) = u(t) + 0.5u(t - 1) + w(t) \]
  \[ w(t) \sim N(0,1), \text{ i.i.d.} \]

- **Actuator applies**
  \[ u(t) = -0.7z(t) - 0.2z(t - 1) - 0.5u(t - 1) + e(t) \]
  \[ e(t) \sim N(0,1), \text{ i.i.d.} \]

- **Closed-loop system:**
  \[ y[t + 1] = 0.7(y[t] - z[t]) + 0.3(y[t - 1] - z[t - 1]) + e[t] + w[t + 1] \]

- **Sensor estimates process noise by**
  \[ \hat{w}[t + 1] := \frac{1}{2} (y[t + 1] - 0.7(y[t] - z[t]) - 0.3(y[t - 1] - z[t - 1])) \]
Example

- Simulates a fake system with a fake noise $n(t) - \hat{w}(t)$
  $n(t) \sim N(0,1), \text{i.i.d.}$

- Reports output of fake simulated system

- In absence of watermarking, actuator would not suspect any malicious measurements

- Sensor attack begins at time 4500
Do we really need both tests?
An attack which passes Test 1 (shows necessity of Test 2)

◆ Consider a system:
\[
x(t + 1) = ax(t) + bu(t) + w(t + 1)
\]
\[
u(t) = g_t(z^t) + e(t)
\]

◆ Consider the attack with spurious measurements
\[
z(t + 1) = az(t) + bg_t(z^t) + \left( \frac{b^2 \sigma_e^2 - \sigma_w^2}{b^2 \sigma_e^2 + \sigma_w^2} \right) \left( x(t + 1) - ax(t) - bg_t(z^t) \right)
\]

◆ This passes Test 1, but fails Test 2
An attack which passes Test 2 (shows necessity of Test 1)

- Consider a system:

\[ x(t+1) = ax(t) + bu(t) + w(t+1) \]
\[ u(t) = g_i(z^t) + e(t) \]

- Consider the attack with spurious measurements

\[ z(t+1) = az(t) + bg_i(z^t) + \left( x(t+1) - ax(t) - bg_i(z^t) + \xi(t+1) \right) \]

where \( \xi(t) \sim N(0, \sigma^2_\xi) \), i.i.d

- Passes Test 2 for any \( \sigma^2_\xi \), but fails Test 1
Non-Gaussian Noise

Consider the system with non-Gaussian noise

\[ x(t+1) = Ax(t) + Bu(t) + w(t+1) \]

\[ w(t) \sim P_w, \text{i.i.d.} \]

Then the watermark guarantee continues to hold if

\[ u(t) = g_t(z^t) + e(t) \text{ with } e(t) \sim P_e, \text{ and } P_e \equiv P_w \]

But then the watermark cannot be made negligible
What is the best we can do?
Characterization of when watermarking guarantee is feasible

**Theorem**

- Consider the system

\[ x(t + 1) = Ax(t) + Bu(t) + w(t + 1) \quad \text{with} \quad w(t) \sim P_w, \text{i.i.d.} \]

and the DW

\[ u(t) = g_t(z^t) + e(t) \quad \text{with} \quad e(t) \sim P_e \]

Then

\[ \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} v^2(t) = 0 \]

\[ E[e \mid e + w] = GE[w \mid e + w] + g \]
Implementations
Defending against attacks on autonomous transportation systems
Test of autonomous transportation system in CPS lab
System model for automatic vehicles

- Plant model for vehicle $i$ given by its kinematic equations

  
  $x_i[t + 1] = x_i[t] + h \cos(\theta_i[t])v_i[t] + h \cos(\theta_i[t])w_{ix}[t]$
  
  $y_i[t + 1] = y_i[t] + h \sin(\theta_i[t])v_i[t] + h \sin(\theta_i[t])w_{iy}[t]$
  
  $\theta_i[t + 1] = \theta_i[t] + h\omega_i[t] + h\omega_i\theta[t]$

- $h$ is the sampling period (100ms)
- $v_i[t]$ a control input, denoting speed
- $\omega_i[t]$ a control input, denoting angular
- $w_{ix}[t], w_{iy}[t], w_i\theta[t]$ all $N(0,2)$, i.i.d.

- Non-linear system
Watermarked system’s performance in absence of attack

- **Watermarked system**

  \[
  x_i[t+1] = x_i[t] + h \cos(\theta_i[t])u_i^s(z_1^t, z_2^t) + h \cos(\theta_i[t])e_{iv}[t] + h \cos(\theta_i[t])w_{ix}[t]
  \]

  \[
  y_i[t+1] = y_i[t] + h \sin(\theta_i[t])u_i^s(z_1^t, z_2^t) + h \sin(\theta_i[t])e_{iv}[t] + h \sin(\theta_i[t])w_{iy}[t]
  \]

  \[
  \theta_i[t+1] = \theta_i[t] + h \omega_i[t] + h e_{i\theta}[t] + h w_{i\theta}[t]
  \]

- **Performance with and without watermarking**

- **Watermarks do not result in any added penalty on performance**
Sensor attack

Sensor attack

\[ z_{2x}[t_A] = x_2[t_A] + \tau, \text{ where } \tau = \text{bias} \]

\[ z_{2x}[t + 1] = z_{2x}[t] + h \cos(\theta_2[t]) u_2^g(z_1^t, z_2^t) + \cos(\theta_2[t]) n[t] \]

\[ n[t] \sim \mathcal{N}(0, \sigma_x^2) \]

This attack passes Test 2, but fails Test 1
Test Statistics

- Fails Test 1

- Passes Test 2
Defending against arbitrary sensor attacks on autonomous vehicles

Does it actually work in practice?
Important question because simulation models do not capture noise

Typically vehicle simulation models only model the deterministic part

However our theory of defense relies crucially on stochastic considerations
  – Inability of attacker to separate system noise from watermark
  – Ability of DW Tests to detect correlations with watermark in sensed signals

So road noise plays a crucial role

So the test of the pudding is only in the tasting
Implementation on Lincoln MKZ and Road Test of DW Defense

With Lantian Shangguan, Kenny Chour, and Swaminathan Gopalswamy, and Gopal Kamath, Jaewon Kim, Woo Hyun Ko, Bharadwaj Satchidanandan, and P. R. Kumar
Architecture of Autonomous Vehicle

Drive By Wire System (DBWS): Steering, Brakes, Acceleration -> Autonomous Research Vehicle (ARV)

Exteroceptive Sensors (e.g. LIDARs, Cameras, RADAR) -> Local Sensor Processing

Proprioreceptive Sensors (e.g. Vehicle Speed, IMU, GPS)

CAN/Ethernet Interface

Autonomous Vehicle Controls on the ECU (ROS + Linux Based System) -> CAN/Ethernet Interface
Implemented system

- Attack on yaw rate sensor of Lincoln MKZ with programmable Drive-By-Wire System
- CAN interface allows access to OEM measured or calculated information, e.g., vehicle speed, yaw rate
- Steering, brakes and acceleration can also be controlled programmatically, with well-defined CAN interface
- Intel processor-based computer running Linux is primary controller – the Electronic control Unit (ECU)
- Controllers are built using the messaging system provided by ROS (Robot Operating System)
- Implementation at RELLIS Campus Proving Grounds
Constant bias attack on Yaw Rate Sensor

• Before 10 sec: Nominal behavior of autonomous circling at desired linear and angular velocities
• Attack begins at 10 sec: Positive constant bias of 2.5 rad/s leads controller to steer less
Observations

- Nearly instantaneous detection
- Watermark variances tested: 0.01, 0.001, 0.0001
- Watermark magnitude does not matter

Attack 2: Constant bias attack 0.5 rad/sec on Yaw Rate Sensor

Attack begins and is instantly detected
Attack on Process Control Systems: Bedrock of industry
Attack on Process Control system

◆ Two tank system

\[ u(t) \]  
Electric Motor

\[ y_1(t) = \text{Level of Tank 1} \]
\[ z_1(t) = \text{Reported Level of Tank 1} \]

Flow in

\[ y_2(t) = \text{Level of Tank 2} \]
DW detection of attack on level sensor of Tank 1
Noise in system

- Froth
  - Noise in the system
- Watermark has to survive this noise
- Dynamic Watermarking is fundamentally a stochastic approach
- Typically, simulation models only model the deterministic part of the system
  - Not the noise
- Hence simulation is not enough
- Experimentation is essential to establishing whether the method works
Defending against attacks on Automatic Generation Control (AGC) in Power Systems
Automatic generation Control

- AGC monitors frequency deviations and tie line flows across multiple “areas” and closes the loop with generators.
- Controls both frequency deviations and tie line power flows.
- So each area properly responds to its own load changes.
- Operates in the 30s to 20min time scale.
- Can be attacked by underreporting/overreporting frequency.

Sensed variables

Defense by DW
Four area synthetic power system with 10 generators
Inserted watermarking does not affect performance of AGC in normal operation
Destabilization Attack on AGC

- Frequency and tie-line under destabilization attack that starts at time 10 min
- Operator realizes attack only at 18 min
- But watermarking allows detection after 11 mins
Some attacks may not be destabilizing but watermarking can still detect

- Attack is not severe

- But watermarking still detects attack at time 11 mins
Remarks

- CPS is important for society and economy
- Lot of future infrastructure may be CPS
- Societally and economically important
- Security of CPS is a very rapidly emerging area
- Critical for safety of future infrastructure
- Lots of attacks have already been demonstrated
Concluding remarks

- A possible general purpose defense of cyberphysical systems?
- Many interesting problems


Thank you