Critical Infrastructure Protection: Requirements and Challenges for the 21st Century

Cristina Alcaraz¹, and Sherali Zeadally²

¹Computer Science Department, University of Malaga, Campus de Teatinos s/n, 29071, Malaga, Spain
alcaraz@lcc.uma.es

²College of Communication and Information, University of Kentucky, Lexington, KY 40506-0224, USA
szeadally@uky.edu

August 1, 2015

Abstract

Today, critical infrastructures have become an integral part of cyberspace and they play a vital role in supporting many of our daily activities (including travel, water and power usage, financial transactions, telecommunications, and so on). Today, the reliability, high performance, continuous operation, safety, maintenance and protection of these critical infrastructures are national priorities for many countries around the world. We explore the various vulnerabilities and threats currently present in critical infrastructures and describe protection measures that can be deployed to mitigate those threats. We highlight and discuss some of the challenging areas such as governance and security management, network design and secure communication channels, self-healing, modeling and simulation, wide-area situational awareness, forensic, and finally, trust management and privacy that must be considered to further enhance the protection of critical infrastructures in the future.

Keywords: Critical infrastructure, Protection, Risk, SCADA, Security, Situational Awareness

1 INTRODUCTION

A Critical Infrastructure (CI) consists a set of systems and assets, whether physical or virtual, so essential to the nation that any disruption of their services could have a serious impact on national security, economic well-being, public health or safety, or any combination of these [36]. The European Union (EU), through its European Programme for Critical Infrastructure Protection (EPCIP), also stresses the importance of CI protection to all its Member States and their citizens. To address CI Protection (CIP), an EPCIP communication, COM(2006) 786 final [33], was developed to establish a legislative framework on CIP to transparently operate and enable cooperation across different borders. According to the EPCIP, CIs can be classified as follows:
• Energy: energy production sources, storage and distribution (oil, gas, electricity).

• Information, Communication Technology (ICT): information system and network protection (e.g., the Internet); provision of fixed telecommunications; provision of mobile telecommunication; radio communication and navigation; satellite communication; broadcasting.

• Water: Provision of water (e.g., dams); control of quality; stemming and control of water quantity.

• Food and agriculture: Food provision, safety and security.

• Health care and public health: Medical and hospital care; medicines, serums, vaccines, and pharmaceuticals; bio-laboratories and bio-agents.

• Financial systems: banking, payment services and government financial assignment.

• Civil administration: government facilities and functions; armed forces; civil administration services; emergency services; postal and courier services.

• Public, legal order and safety: maintaining public and legal order, safety and security; administration of justice and detention.

• Transportation systems: road transport, rail transport, air traffic; border surveillance; inland waterways transport; ocean and short-sea shipping.

• Chemical industry: production and storage of dangerous substances; pipelines of dangerous goods.

• Nuclear industry: production and storage of nuclear substances.

• Space: Communication and research.

• Research facilities.

However, the National Infrastructure Protection Plan (NIPP) [28], defined by the United States Department of Homeland Security (DHS), also considers other critical sectors such as:

• National Monuments and Icons: monuments, physical structures, objects or geographical places that are acknowledged as representing national culture, or have a religious or historical importance.

• Commercial Facilities: commercial centers, office buildings, sports stadiums, any other place that can accommodate a large number of people.

• Critical Manufacturing: Transformation of materials into goods. This includes all the processes involved in manufacturing and transportation equipment.
Figure 1: A general architecture of a SCADA network based on remote substations

- Defense Industry Base: production facilities of military resources (e.g., weapons, aircraft or ships) and maintenance of essential services (e.g., communication) to protect a nation.

The aforementioned sectors together with their CIs are somehow connected to each other, creating a special interdependence relationship. This relationship means that a CI could require and depend on the services from another CI to work properly, and the latter might also need the output from the first infrastructure. This interdependence relationship could trigger a cascading effect when disruptions of services and functionalities appear within a CI [5]. Rinaldi et al. [71] concretely identified and analyzed up to four types of relationships: physical, geographic, cyber and logical. A physical interdependency refers to a dependency on receiving resources or raw material from other infrastructures. A geographic interdependency exists when multiple infrastructures share a close spatial proximity, and any problem located in one of them can reach the other CIs. A cyber interdependency is attributed to the existing dependencies in communication systems and their information. Logical corresponds to those systems, actions or decisions that connect an agent of one infrastructure to another agent belonging to another infrastructure which does not have a direct link through physical, geographic and cyber connections (e.g., bureaucratic or political decisions) [84].

Given the influence of information systems on the good performance of other CIs, this paper focuses mainly on Critical Information Infrastructures (CIIs) and their se-
curity issues. A CII basically consists of those “information processes supported by Information and Communication Technology which form CIs for themselves or that are critical for the operation of other critical infrastructures” [22]. This means that the vast majority of critical infrastructures currently deployed in our society are dependent on information systems to manage sensitive information associated with such infrastructures. This means that an unplanned event on cyber infrastructures may lead to serious consequences that may affect the performance, reliability, security and safety-critical of the underlying system. Therefore, protection measures are also needed, opening a new research area known as Critical Information Infrastructure Protection (CIIP).

In particular, CIIP comprises those “programs and activities of infrastructure owners, manufacturers, users, operators, R&D (Research and Development) institutions, governments, and regulatory authorities which aim at keeping the performance of critical (information) infrastructures in case of failures, attacks, or accidents above a defined minimum level of service and aim at minimizing the recovery time and damage. CIIP should therefore be viewed as a cross-sector phenomenon rather than being limited to specific sectors. CIIP should be closely coordinated with CIP from a holistic perspective” [32]. This importance is not only considered by the European Union, but also by the United States Government in its Law 107-296 [27], which states that: “protection of critical information infrastructures is important to the national defence and economic security of the nation that are designed, built, and operated by the private sector” (Law 107 296, Section 1001 titled Information Security). This law considers that CIs are CIs by themselves because their information is not normally in the public domain and it is related to the security of CIs or protected systems. In fact, ICTs can be considered as the backbone of a CI, which is composed of multiple communication links, network topologies and interfaces to manage and transmit sensitive data in a reliable and timely manner.

Typical CIs are, for example, Industrial control Systems (ICS), which are in charge of controlling and supervising services of industrial infrastructures, such as energy bulk generation systems, electrical distribution and transmission lines, water treatment systems, and oil and gas pipelines/refineries [13]. Their communication architectures include a set of communication links, topologies and technologies in order to receive and process information from their remote substations located close to the infrastructure being supervised (as shown in Figure 1). These substations are automated systems composed of a set of industrial engineering devices (e.g., Remote Terminal Units (RTUs), sensors and actuators) in charge of collecting and sending status related to the controlled infrastructure (e.g., levels of pressure, temperature or voltage). The supervision is not unidirectional. A great part of the information is sent from the SCADA center to remote substations through commands (i.e., actions in the field), which should be executed through actuators.

An ICS can be for example a SCADA system or a Distributed Control System (DCS). Although these two systems share common goals, there are slight differences between them. A SCADA system basically refers to an event-driven centralized network (i.e., wait for an event or a change within the system to carry out an action) for supervision and data acquisition of substations located over large and distant geographic areas. Figure 1 illustrates a SCADA network architecture. We note three main networks in Figure 1: the control center, substations and corporate network. The con-
trol center is responsible for controlling the overall performance of the entire system and managing its sensitive information. This management depends on the functionality of a set of servers, such as SCADA servers and historical servers. External accesses to these resources must be properly addressed and restricted through security mechanisms, such as firewalls, Demilitarized Zones (DMZ), Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS) or antivirus. Some of these accesses may come from corporate networks in charge of reporting statistical evaluations or updating strategic plans to increase productivity and business output. In contrast, a DCS is a process-oriented system, the control of which is limited in terms of dimension and geographic distribution.

This work focuses on SCADA systems for two main reasons. First, SCADA systems are, from a security point of view, one of the most widely researched systems in the literature [7, 57]. Second, they can be viewed as the main core and backbone of the next generation of electrical production and distribution systems of the 21st century, (also known as the Smart Grid) [65]. A Smart Grid is a system composed of a set of main sub-domains, such as energy bulk generation systems (either renewable or non-renewable), transmission and distributions lines, customers, providers, market and control systems. Each domain comprises various stakeholders and resources, all of them interconnected to each other to efficiently manage demand load and reduce unneeded power generation.

Unfortunately, the nature of SCADA systems means that they are exposed to numerous threats, which may be caused by (Hardware (HW)/Software (SW)) errors, mistakes caused by people (i.e., operational errors) or deliberate actions (i.e., malicious actions). These aspects, which can put the security of the control system and its CIs at risk, require that protection must be carefully provided which is also another focus of this chapter.

The rest of this paper is organized as follows. Section 2 analyzes the current security problems associated to the new technologies and information systems for control and automation of other CIs. This analysis includes the role of particular technologies, such as the Internet, wireless communication systems and embedded systems. Considering the relevance of these technologies, Section 3 addresses the security requirements and services that should be considered and needed to protect the critical information infrastructures in charge of protecting our CIs. Section 4 identifies and outlines a few challenging areas for protection that should be carefully addressed in the future, whereas Section 5 concludes the paper.

2 VULNERABILITIES AND THREATS INVOLVE PROTECTION MEASURES

Threats in CIs are mainly caused by existing vulnerabilities in HW/SW resources that can be exploited to produce unplanned changes in the service offered and a deviation from their normal behavior. These faults can be classified into two categories: internal and external faults. An internal fault corresponds to anomalous changes originating within the system. An external fault is related to those interactions that originate from
outside the system, such as natural phenomena, malicious actions or accidents. Irre-
spective of the cause, any fault within the system can then create an internal effect that
can collapse essential services and activities for the control. For example, an attack
on a sensor node may cause (HW/SW) errors that may affect the operations of other
essential resources for the control, such as RTUs. If this occurs, the central system will
then be unable to receive sensitive information from substations, becoming a system
blind to the real states of the system under control. This situation can also occur when
communication links stop functioning or are compromised by malicious entities, leav-
ing critical areas and their services unprotected. Therefore, ICSs are also themselves
considered CIs in [5].

These kinds of faults were also analyzed by Harrison et al. in [38] who proposed
a taxonomy based on the concept of cause-effect. This taxonomy consists of defining
a set of event vectors and effect vectors so as to define the motivation of a threat,
the methodology applied for each threat and their effects. An event vector describes
the threat agent, the motivation, the objective to be compromised, the method and the
technique applied to carry out such an objective. In contrast, an effect vector refers to
the impact of the affected infrastructure, its services and the sector, in addition to the
cause of the effect. In either case, these vectors are rather dependent on the type of
threat vector that can be exploited within the system in order to compromise its secu-
rity which generally depends on the characteristics and conditions of the system. For
example, SCADA systems normally follow a particular order of security requirement:
availability, integrity, and confidentiality [86]. A threat to the availability consists of
leaving unavailable essential resources for control and performance at all times, as well
as critical information (e.g., alarms, measurements or commands). A threat to the in-
tegrity aims to manipulate the critical information of resources, whereas a threat to the
confidentiality consists of eavesdropping on critical information.

It is also important to consider the level of dependence among resources or compo-
nents of a system, the segregation of functionalities and services [5]. This means that
when one component presents a particular anomaly, the result may take on a progres-
sive effect that may change its normal behavior, which may result in a crisis situation.
When the effect enters into cascading mode, the entire system and its services may
also become affected, with a high probability of reaching other CIs and their services.
When these adverse situations occur, it is of paramount importance to be aware of four
main factors: the scope of the effect, its magnitude, propagation and recovery [31]. The
first factor contributes to the loss/unavailability of an element and its impact within so-
ciety which could be rated according to the geographic coverage; i.e., international,
national, provincial/territorial, or local. The magnitude of the effect is related to the
degree (minor, moderate or major) of the loss according to the public, economic, en-
vironmental, interdependency, and political impact. For the latter two factors, time is
an essential parameter to measure the criticality of a situation because it determines at
what point the loss of an element could have a serious effect; and at what point it would
be possible to recover the functionality of the entire affected system.

Many of these threats are caused by the adaptation of the current ICTs for control
tasks and operations of critical services, such as the Internet and wireless communi-
cation technologies. The reason is quite simple. The technological introduction is
simultaneously increasing architectural complexities, and adding vulnerabilities, se-
curity risks and interoperability issues [5]. All of these aspects will be thoroughly discussed in the following section.

2.1 Technological Trends Bring New Security Issues

Nowadays, ICTs play a crucial role in the control and connectivity between critical entities. This is the case of SCADA systems where its control and supervision mainly depend on the reliability and security of the communication channels and information systems to send, compute or storage commands, alarms or measurements. We identify three main technological entities in this section: communication systems for large distances, wireless communication systems, and the influence of embedded systems for control operations.

2.1.1 Communication Systems for Covering Long Distances

Within this category it is important to highlight the operational benefits that the Internet can bring to enhance data acquisition and supervision activities. These benefits include global connectivity, flexibility, and data dissemination from anywhere and anytime through web interfaces and IP-based communication protocols (e.g., Hypertext Transfer Protocol Secure (HTTPS)). To extend functionalities and reduce maintenance and installation costs, remote substations also had to migrate to TCP/IP, opening their connections to ensure coexistence with other technologies and concurrency in order to balance work load and activities in the field. Moreover the use of standard TCP/IP protocol stack also led to the specification and standardization of SCADA communication protocols based on the concept of server-client communication. Some of these SCADA protocols are well-known within the research area and they include: Modicon Communication Bus (Modbus/TCP), IEEE Std 1815 (Distributed Network Protocol (DNP3)), IEC 60870-5-104, or Inter Control Center Protocol/Telecontrol Application Service Element-2 (ICCP-TASE2.0, IEC60870-6). The first two are dedicated to automation and control, and the latter ones focus on the interconnection between telemetry control systems (i.e., between SCADA systems).

This step towards the modernization of control systems has also encouraged to researchers and companies to analyze, design and provide HW and SW solutions for global collaborations and connectivity. For example, K. Suresh et al. designed a web-based SCADA prototype for dissemination of information using an Extensible Markup Language (XML) encoding [78]. This technique provides a virtual approach for experimentation with the possibility for including General Packet Radio Service (GPRS) and Wireless Application Protocol (WAP) connections. Adnan et al. designed in [1] a web-based multi layered distributed SCADA system to supervise terminals for truck loading and oil products pipeline shipping at refinery plants. Jain et al. presented in [50] a web-based expert system for smart and automatic diagnosis and control of power systems. From a commercial perspective, Exemys company currently has at its disposition, web and mobile cellular telemetry solutions, and hardware protocol converters to translate serial Modbus communication to TCP/IP SCADA communication [34]. Similarly, Yokowaga [81], and WebSCADA [80] also have web solutions for automation.
Unfortunately, we also cannot ignore the fact that the Internet, as a communication infrastructure itself, is also exposed to multiple threats, the majority of which come from traditional TCP/IP vulnerabilities. Typical threats might be, for example, to continuously resend control messages or request essential resources to exhaust computational or communication capabilities, as well as to eavesdrop on critical information through Man-in-The-Middle (MTM) techniques, or inject false messages to perform unsuitable actions or show false monitored values. Through these threats, any adversary might bypass the security of the system and once inside, tries to carry out other types of attacks such as the reading/modifying of files, memory dump or launching operational functional services through false commands. In addition to these vulnerabilities, it is also necessary to consider the existing security gaps of the vast majority of SCADA protocols. For example, the communication of Modbus/TCP is done in clear text, where a large part of the payload (e.g., sensitive information and network addresses) can be captured, manipulated and/or eavesdropped. This protocol also lacks authentication because Modbus sessions only verify the validity of specific parts of a message, such as the address and the function code. DNP3 also suffers from security deficiencies in spite of the fact that it was designed to carry out frequent Cyclic Redundancy Checks (CRC), data synchronization and uses several data formats. Nonetheless, there is a variant of DNP3, known as Secure DNP3, which implements an authentication system based on challenge-response together with a session key to verify the source node. Similarly, ICCP also suffers from certain limitations and security issues to ensure authentication and encryption [54].

Many of the threats are managed by public databases, such as the British Columbia Institute of Technology (BCIT), Industrial Security Incident Database (ISID), the Computer Emergency Response Teams (CERT) or Industrial Control System Cyber Emergency Response Team (ICS-CERT). These databases maintain a common threat repository so as to efficiently respond to security incidents. According to the last report published by the ICS-CERT, the number of incidents in the different critical sectors have significantly increased over the past few years (from 9 incidents in 2009 to 198 in 2011), and more so in the energy sector and its control systems. A large number of these incidents are normally caused by viruses, trojans and worms (such as the Stuxnet worm in 2010 and the Flame virus in 2012 [25]) that try to compromise the integrity of parts of the system or the entire system. Given the importance of these databases for protection, the European Council has recently approved the development of an European project called Testbed Framework to Exercise Critical Infrastructure Protection (CloudCERT). This project began in 2012 and aims to develop a complex technological system based on the paradigm cloud computing so as to facilitate information exchange (e.g., incidents) between CIs [24].

In this context, it is also necessary to highlight the role of cloud computing for CIP [4]. The cloud computing infrastructure provides a set of operational benefits for CIP, such as data redundancy, data availability, as well as survivability when essential parts of the system remain isolated or lost. For example, if a SCADA Center (temporarily or permanently) loses control of its operational services, another SCADA Center could retake control by using the ICCP protocol and the cloud infrastructure enables queries about the critical information (e.g., alarms, processes, measurements, etc.). The adaptation of this new paradigm for data redundancy and its importance for data recovery
introduces additional interesting benefits. Some of them are, for example, virtualization of assets and management of (private, public or hybrid) service-oriented architectures where the services are managed on-demand over the Internet [70]. Virtualization is based on the creation of a virtual platform of HW resources (storage, servers and network devices) and operating systems to reduce costs, information sharing, manageability and isolation. In addition, a cloud computing infrastructure can influence the development of industrial applications to ensure interoperability and cooperation among different organizations or entities.

However, the information registered inside the cloud is computed within the infrastructure which is shared among diverse providers and subscribers. Under these circumstances, security and privacy aspects should be considered to protect any sensitive data and its upload/download within the cloud through cryptographic schemes, mechanisms of authentication and identity management, access control and accounting, as well as trust management, governance, policies and regulation between providers [79]. This security has to be properly addressed because in the hypothetical case that the cloud infrastructure stores information related to incidents, the infrastructure might be targeted by adversaries who could trace and find existing vulnerabilities and weaknesses of CIs, and later carry out other malicious actions against these systems later on. In addition, aspects of maintenance should be equally considered to avoid loss of information. Data redundancy at different locations within the cloud should be taken into consideration along with the use of alarm management systems with the ability to respond and automatically return to previous states.

Other technologies designed for long distances such as mobile cellular technologies (e.g., 3G/4G, Universal Mobile Telecommunications System (UMTS), GPRS, GSM or Terrestrial Trunked Radio (TETRA)), Satellite, GPS, World inter-operability for Microwave Access (WiMAX, IEEE 802.16), Mobile Broadband Wireless Access (MBWA, IEEE 802.20)), or microwave systems also enable automation and control tasks at low costs remotely, in addition to guaranteeing mobility, collaboration, reliability and coexistence with other technologies. Field operators can directly interact with industrial devices (e.g., an RTU with a wireless transmitter) through their handheld device interfaces by sending commands and receiving information such as status, measurements or alarms. Similarly, there is also a suite of technologies medium and small control control applications such as Bluetooth (IEEE 802.15.1), Wireless Local Area Networks (WLANs, IEEE 802.11), Mobile Ad-hoc NETworks (MANETs) or Wireless Sensor Networks (WSNs) [67]. Moreover, cellular technology can be a quite cheap alternative to connect small groups of field devices and send non-critical information to the SCADA Center. Nonetheless, this technological option is not recommended given that a high number of cellular nodes for supervision can increase delays in the communication and costs associated with data transfers [67].

The adoption of these technologies have encouraged International organizations to specify their communication standards such as ZigBee [88], ISA100.11a [44], WirelessHART [39]. Moreover, a large part of the communication of an Advanced Metering Infrastructure (AMI) of a Smart Grid could depend heavily on these technologies to transfer significant information associated with customers or business utilities such as SCADA systems. For example, WBWA operates at 3.5GHz with a data rate of up to 1Mpbs-20Mpbs; and WiMAX operates at 2.3, 2.5 and 3.5GHz with a data rate up...
to 70Mbps. These transmission capacities could allow the AMI to send data streams to business utilities from smart meters. An AMI consists of a set of HW/SW technologies to measure, collect, analyze and show the level of energy usage, and connect metering devices (associated with electricity, gas, heat and water) to utility business systems. This connectivity is absolutely bidirectional where information is distributed to/from customers and other entities. These entities can be, for example, the control systems in charge of supervising energy substations.

The aforementioned networks are also becoming increasingly important for CIP. These networks allow human operators to establish in-situ local connectivity without going through the SCADA center, and mobility within the area. This is the case of the MANET networks, which help human operators gain authorized access to parts of the system (e.g., sensors, actuators or RTUs) and carry out operational activities such as data dissemination and management, response to incidents, configuration of parameters and maintenance [11]. Within this classification, it is worth highlighting the role of Wireless Personal Area Networks (WPANs, IEEE 802.15) for networks with small coverage where the control is limited to a reduced number of nodes, such as Bluetooth, Z-Wave or ZigBee. A variant of WPANs are the Low-Rate WPANs (LR-WPANs, IEEE 802.15.4 [43]), such as ZigBee, ISA100.11a, WirelessHART T M , and MiWi/MiWi Peer-to-Peer (P2P) networks [58].

The use of these wireless technologies also adds certain inconveniences such as operational delays, latencies, electro-magnetic or radio frequency interferences, and security issues. Indeed the abuse of repeaters and routers to intensify the signal might significantly increase end-to-end delays. In addition to slowing down the data transfer, altering the integrity of the information caused by interferences and obstacles is also possible. The result may be a variation from the Quality of Service (QoS) which can affect the integrity of the information and/or on the time required to process operational activities.

In addition, it is of paramount importance to address some aspects related to the coexistence and reliability of the communication between heterogeneous technologies. Some of these aspects are associated with authentication and authorization between devices, information security as well as interoperability to manage different types of messages with different formats. A threat to reliability of the communication may be, for example, a jamming attack, which consists of altering the radio frequency channel and this attack works regardless of the use of methods of frequency hopping offered by some existing standards, such as WirelessHART [39] and ISA100.11a [44]. Other threats might also seriously put the availability, integrity and confidentiality of the data and resources at risk. For example, threats to availability can be driven by Denial of Service (DoS) attacks related to flooding attacks (overloading of communication channels), selective forwarding attacks (selectively sending information to the next hop), sybil attacks (impersonating several identities), blackhole attacks (drop messages), sinkhole/wormhole attacks (directing information to particular nodes) and jamming attacks [9].

Threats to integrity are linked to those attacks that involve route falsification and sybil attacks. In contrast, threats to confidentiality can be launched through deliberate exposure attacks (intentionally revealing critical information), sniffing attacks (eavesdropping communication channels), traffic analysis, and physical attacks (steal/break
the hardware architecture to extract information from memory). Many of these threats and their countermeasures are thoroughly discussed in [9].

2.1.2 Embedded Systems and Cooperation between Objects

Embedded systems are based on constrained devices (also known as objects or things) with the capability to dynamically and autonomously work and actively interact with other devices by themselves. For collaboration, such objects have to be readable, recognizable, locatable, addressable, and/or controllable [74]. In this context, it is important to highlight the role of those industrial objects in charge of controlling and managing the energy generation and distribution systems that comprise a Smart Grid, such as RTUs, sensors, actuators, smart meters, phaser measurement units, mobile robots, vehicular nodes, storage devices, RFID tags, or even human operators hand-held interfaces. The cooperation of all these energy elements under a common communication infrastructure, via the Internet, leads to a new concept known as the Internet of Energy (IoE). IoE is an infrastructure based on standards and interoperable communication technologies and protocols that interconnect the energy network with the Internet, enabling the availability of power units when they are needed [53, 18].

WSNs are also playing a crucial role in CIP. According to [73, 37], this technology and its smart objects (i.e., sensor nodes) are able to maintain by themselves a continued control of a context and its surroundings, as well as the ability to detect, track and warn of threatening situations. Their collaborative capacities make them one of the most sought technologies currently being deployed in diverse applications and in different sectors. In addition, WSNs standalone and smart capacities enable them to adapt to the environmental conditions without losing functionality. Conventional sensors can work with 4MHz, 1KB of RAM and 4KB-16KB of ROM; whereas typical industrial nodes are configured with 4MHz-32MHz, 8KB-128KB of RAM, 128KB-192KM of ROM. The deployment of such devices depends on several factors, such as the criticality of the application context and the organizations protection needs. Moreover, these smart devices can also offer a suitable support for the construction of prevention and response tools, such as Early Warning Systems (EWSs) or IDSs [8, 6]. They can form part of the observation system in charge of perception, tracking, detection and alerting of anomalies or anomalous events. This also means that a critical system should require a special commitment to invest not only in new technologies for control at (almost) real-time but also in technological components for security and protection.

These attractive features for protection have recently encouraged governments to invest in dynamic and automated substations. This is the case of the American Recovery and Reinvestment Act (ARPA) of 2009, which invested around one hundred automated substations with thousands of sensor nodes to detect changes and prevent local or regional power blackouts [17]. In fact, it is expected that the vast majority of them will be connected with the Smart Grid using the 6LowPAN standard [59], interacting with other system objects compatible with the new version of the Internet Protocol IPv6 [19], known as Internet Protocols for the Smart Grid. However, interactivity from anywhere and anytime via the Internet will not be a trivial task. Studies made in [12] indicate that the full integration of sensors with the Internet using TCP/IP connections still is a open research field that need to be addressed to raise the con-
nectivity capacities. For example, if sensors are fully integrated into the Internet, any maintenance procedure of firmware might leave parts of the system not operational. On the other hand, the heterogeneity of objects and their arbitrary connections could also increase incompatibilities [74], and result in an infrastructure possibly more susceptible to vulnerabilities and threats.

3 PROTECTION TO ADDRESS BUSINESS CONTINUITY IN CRITICAL CONTEXTS

We have identified security weaknesses associated with the use of ICTs in control and automation in CIIs in the previous sections. In this section, we focus on the protection requirements expected from CIIs (e.g., SCADA systems) to protect CIs (e.g., energy substations) and the necessary requirements to protect the communication infrastructure itself (as shown in Figure 2). It should be also noted that a major part of these analyses are based on the studies made in [5], where a set of operational and security requirements for control systems are formally analyzed using dependency relationships and dependency theory [72].

When CIs trust the good performance and reliability of ICTs to manage its sensitive information, a set of functional services and requirements is needed from critical information infrastructures. These requirements include: performability 24/7, interop-
erability, scalability, extensibility, availability, reliability, resilience, safety-critical, autonomy and self-healing, usability, trust and collaboration between heterogeneous objects to jointly address anomalous/threatening situations, in addition to fault-tolerance and security.

The three first requirements are needed to adapt new HW and SW resources with existing ones. In other words, interoperability is concerned with the ability of the system, systems and organizations to work together to carry out a common goal. This concept is normally applied to those engineering systems composed of a set of social, technical, political, and organizational aspects all of which play an essential role for business continuity [5]. This also means that the control systems should ensure that both existing resources and new resources can cooperate and interact with each other without affecting the functionality of the system, its communication protocols, industrial devices, SW-based control components, and security services. To this end, it is advisable that parts of the system are responsible for the definition and maintenance of the governance of the entire system. This governance includes the specification and development of security policies and technical specifications, as well as the access and availability of technical and legal reports such as standards.

Similarly, scalability refers to the upgradability capabilities of the system to add or remove its HW resources; extensibility is related to the ability of the system to extend/modify its SW resources (e.g., security services, control applications) [20, 87]. The adoption of new resources should not trigger changes on the final service of a critical system. In particular, this adaptation should be the only means to provide the system with technological support to ensure: modernization, management of its services and performance in an attractive way. Both scalability and extensibility do not necessarily guarantee absolute interoperability and compatibility with the existing resources. Thus it is necessary to specify and comply with a set of technical and legal assets such as policies, standards, recommendations and good practices.

Availability and reliability are two tightly related concepts. Availability is associated with the probability that a system delivers its services when they are really required at an instant of time $T_z$. In contrast, reliability corresponds to the probability that a system is able to deliver its services properly and it does not lose its availability during a particular time period $[T_x,T_y]$, such that $T_z \in [T_x,T_y]$ [41]. This relationship between both properties means that if a control system needs to execute a set of operations to perform command (e.g., close/open valve), both the information infrastructure and intermediary objects should not disrupt or delay the normal sequence of execution. Otherwise, services delivered by the underlying system will not be available when they are needed, and therefore the system will not be reliable.

QoS is also a relevant property in CIP because a disruption or alteration of the system (either due to faults, incidents, errors or threats) could put the normal performance of the entire system at risk. To this end, Zheng et al. in [85] designed an adaptive QoS-aware fault tolerance strategy for web services based on a Service-Oriented Architecture (SOA) in order to dynamically adjust the system parameters to their optimal fault tolerant configurations. To develop a suitable QoS strategy for critical systems it is advisable to consider a set of additional parameters such as the level of heterogeneity, variable nature and interactivity of the environment, network topologies, weaknesses associated to objects, as well as interdependencies between nodes and systems. In this
way, it is possible to adjust essential parameters and design robust infrastructures with the ability to control faults or incidents.

On the other hand, to face adverse or threatening situations resilience and robustness should be properly addressed. In a nutshell, a system under threat should guarantee functionality all the time though certain parts of the system are being seriously compromised. In addition, and according to the studies made in [5], any fault could trigger a cascading effect due to the internal dependency relationships between resources and elements of the system itself. For example, a software error located at a network resource (e.g., a gateway) may generate a progressive effect that might delay/interrupt essential operations, or even isolate critical parts of the system such as substations. If such an effect is not controlled properly, it may reach the border of a critical system and ultimately affecting the business continuity of the rest of CIs.

To control the cascading effect, it is necessary to consider safety-critical aspects [5]. This property is in charge of avoiding or mitigating the propagation of the effect between CIs, which could result in human deaths, injuries, or physical or physiological damages. To prevent these situations, control networks should be mainly based on autonomous, dynamic and intelligent approaches to enable the system to make decisions by itself, and ensure prevention and response in an efficient and timely manner. Some of these solutions could be, for example, the design of dynamic alarm management solutions to locate the most suitable human operators with the best experience to deal with critical incidents at any time [8]. Currently, there are not enough related works in this area and lightweight solutions should be carefully developed.

Additionally, topics related to usability should be considered. Any user (expert or not) must be able to interact with the system through intuitive interfaces. This means that the design of these interfaces must not make the information (such as alarms or readings come from sensors) received from the system difficult to understand, and they must facilitate the options to speed up critical operations (i.e., management of actions in the field). In addition, when multiple heterogeneous objects are involved, the interfaces must map and manage their information properly without delaying operational tasks; and in the worst scenario (under a threatening situation) to dynamically locate the exact position of the affected system so that an immediate response can be generated. Finally, the heterogeneity of the environment and the implication of different networks, topologies and objects, as well as the deployment of services and applications must not reduce the business continuity and operational activity, which are normally measured in terms of computation or communication.

Following with the topic of heterogeneity, it is worth considering the collaboration between objects. For example, any active object deployed in the system has to know how to collaborate with other objects in a secure and transparent way, and how to proceed quickly without losing control of the system under observation at all times. In addition, all the objects have to be trustworthy nodes and they should trust information exchange to ensure a rapid response in the worst case scenario. Trust services can be also extended to other application contexts where new infrastructures play an essential role in the modernization of critical systems (such as the cloud computing). If a system depends on a cloud computing infrastructure to store backup instances, then CIs have to be able to trust the cloud and its elements (i.e., providers) for its management operations.
The network architecture also has to able to control any occurrence of unplanned events and guarantee an acceptable level of reliability and performance of its services. Therefore, fault-tolerance is a requirement that should be considered in any critical environment to ensure business continuity in spite of the presence of HW/SW faults within system components. One way to control these kinds of faults is through strict security policies, maintainability and testability based on validation and verification processes, in addition to considering aspects associated with redundancy and dynamic solutions related to fault-detection, fault-restoration, fault-removal and fault-location. These solutions make more sense when the context is mainly formed by different types of networks (e.g., MANETs-WSNs-the Internet) and objects actively interacting each other [74]. Lastly, security aspects should be extensively addressed in the entire SCADA architecture to ensure availability, integrity and confidentiality of resources and information.

4 CHALLENGING AREAS FOR THE PROTECTION FOR CRITICAL CONTEXTS

In this section we highlight the current highest priority security areas that should be properly addressed to build a secure and sustainable future. In particular, we discuss the following challenging areas of security for protecting critical environments: governance and security management, robust network design and secure communication channels, self-healing; modeling and simulation, Wide-Area Situational Awareness (WASA), forensic and learning, trust management and privacy. These areas are illustrated in Figure 3 where the bottom of the pyramid shows the protection foundation for any CII (e.g., SCADA systems).
4.1 Governance and Security Management

To use resources and assets efficiently, the system must be under the control of a suitable governance and security management. Governance is concerned with the set of security controls (i.e., actions) used to govern an organization. These controls are defined within security policies, standards, best practices, or recommendations. In particular, a security policy contributes to a set of action plans agreed or chosen by an organization, and it is the means by which security requirements must be properly specified in order to enforce security controls and management.

Security controls and their abstractions are in charge of regulating the overall behavior of the entire system made up of physical and virtual entities. These entities can be human entities (e.g., staff members, providers, customers, etc.) or HW/SW entities (e.g., applications, services, resources, objects, etc.). For interoperability between entities, a set of behaviors needs to be specified according to the type of application domain and its criticality, the existing interdependencies between organizations and resources, the information architecture and its coexistence with engineering systems, information management, associated risks. Important issues that must be addressed by security controls include: where, what, how, and when an action can change the functionality of a part of the system, and who should do it.

According to the latest report on security for control systems published by the DHS [29], controls can be categorized into a set of sub-controls which are described below. As these sub-controls are rather general, we further classify them into two categories: organizational security sub-control and operational sub-control. Both categories are defined as follows:

- **Organizational security sub-controls**: this category refers to all those security sub-controls related to the organizational management (both physical and cyber) of the entire system. These sub-controls include security policy, organizational security, personnel security, physical and environmental security, strategic planning, security awareness and training, monitoring and reviewing control system security policy (review security compliance according to the security policies), risk management and assessment, and security program management.

- **Operational sub-control**: this category comprises all those security sub-controls that allow a system to perform a set of activities (e.g., operational control or sensitive information management) securely. Within this classification, we include system and services acquisition (e.g., allocation or acquisition of control system assets, software, and services), configuration management, information and document management, system development and maintenance, system and communication protection, incident management and response, system and information integrity, access control, audit and accountability, and media protection.
Table 1: Compliance with organizational and operational standards for critical control systems.

<table>
<thead>
<tr>
<th>Organizational and Operational Standards</th>
<th>NIST 800-55</th>
<th>NISTIR 7628</th>
<th>ISA 99-1</th>
<th>ISA 99-2</th>
<th>ISO 17799</th>
<th>ISO 27001</th>
<th>ISO 27002</th>
<th>ISO 19791</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Organizational Security Sub-Controls</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Security Policies</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Personnel Security</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Physical and Environmental Security</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Strategic Planning</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Security Awareness and Training</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Monitoring and Reviewing Sec. Policy</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Risk Management and Assessment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Security Program Management</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Operational Security Sub-Controls</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System and Services Acquisition</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Configuration Management</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>System and Communications Protection</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Information and Document Management</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>System Development and Maintenance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Incident Management and Response</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>System and Information Integrity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Access Control</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Audit and Accountability</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Media Protection</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
A selective set of standards and recommendations is also currently available that deals with organizational and technical aspects. Some of the most representative standards on information systems and SCADA communication systems are, for example: NIST SP800-53 (Rev. 3) [62], NISTIR 7628 (for Smart Grids) [63], NIST, 2010b, IEC 62351 [40], WirelessHART or ISA100.11a. However, these standards are not the only applicable ones for critical environments. Traditional standards can also be useful and they include: ISA 99-1, ISA 99-2 [45], ISO 17799 [48], ISO 27001 [47], ISO 27002 [46], ISO 19791 [49], amongst others. Together with these standards, SCADA organizations could also use recommendations and guidelines for critical control systems to align their business models with an effective protection framework such as: NERC CIP-2 [60], GAO-04-140T [35], IEEE 1402 (physical security of energy substations) [42] or API 1164[16]. Both Table 1 and Table 2 summarize the security sub-controls described above. Moreover, it is possible to deduce from these two tables that the majority of current standards and recommendations aim to cover both organizational and operational aspects.

Security assessment is another aspect that should be considered as part of the governance of a critical system and its security management. This procedure consists of reviewing whether the current architecture, its interconnected objects and entities, as well as the information system, comply with the security policies and their prerequisites, and the business model. To facilitate this procedure it is essential to perform activities associated with maintenance and auditing. A large part of the maintenance task focuses on not only offering support to modify or repair faults but also on satisfying new requirements, improving performance, reducing costs by making future maintenance easier, or enabling adaptation to a changing context. Similarly, the auditing task consists of checking whether the current architecture complies with the prerequisites of the system or requires some type of improvement using activity registers (e.g., logs). To properly address accountability aspects, topics of responsibility and activity (e.g., storage, access and format) should also be addressed within the security policies.
Table 2: Compliance with technical standards, recommendations and guidelines for critical control systems.

<table>
<thead>
<tr>
<th>Technical Standards</th>
<th>Recommendations and Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEC 62351</td>
<td></td>
</tr>
<tr>
<td>FIPS 140-2</td>
<td></td>
</tr>
<tr>
<td>WirelessHART</td>
<td></td>
</tr>
<tr>
<td>ISA-100.11a</td>
<td></td>
</tr>
<tr>
<td>ZigBee</td>
<td></td>
</tr>
<tr>
<td>AGA 12-1</td>
<td></td>
</tr>
<tr>
<td>AGA 12-2</td>
<td></td>
</tr>
<tr>
<td>NERC CIP</td>
<td></td>
</tr>
<tr>
<td>GAO-04-140T</td>
<td></td>
</tr>
<tr>
<td>IEEE 1402</td>
<td></td>
</tr>
<tr>
<td>API Sec</td>
<td></td>
</tr>
</tbody>
</table>

Security Control

<table>
<thead>
<tr>
<th>Organizational Security Sub-Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security Policies</td>
</tr>
<tr>
<td>Organizational Security</td>
</tr>
<tr>
<td>Personnel Security</td>
</tr>
<tr>
<td>Physical and Environmental Security</td>
</tr>
<tr>
<td>Strategic Planning</td>
</tr>
<tr>
<td>Security Awareness and Training</td>
</tr>
<tr>
<td>Monitoring and Reviewing Sec. Policy</td>
</tr>
<tr>
<td>Risk Management and Assessment</td>
</tr>
<tr>
<td>Security Program Management</td>
</tr>
</tbody>
</table>

Operational Security Sub-Controls

| System and Services Acquisition    | ✓ ✓ ✓ ✓ ✓ ✓ ✓ |
| Configuration Management           | ✓ ✓ ✓ ✓ ✓ ✓ |
| System and Communications Protection | ✓ ✓ ✓ ✓ ✓ ✓ |
| Information and Document Management | ✓ ✓ ✓ ✓ ✓ ✓ |
| System Development and Maintenance  | ✓ ✓ ✓ ✓ ✓ ✓ |
| Incident Management and Response    | ✓ ✓ ✓ ✓ ✓ ✓ |
| System and Information Integrity    | ✓ ✓ ✓ ✓ ✓ ✓ |
| Access Control                      | ✓ ✓ ✓ ✓ ✓ ✓ |
| Audit and Accountability            | ✓ ✓ ✓ ✓ ✓ ✓ |
| Media Protection                    | ✓ ✓ ✓ ✓ ✓ |


19
4.2 Secure Design of Network Architectures and Control

As noted earlier, behind a SCADA network architecture there is an important engineering deployment based on HW/SW resources (cf. 2.1.1) which include the Internet together with its web-based SCADA interfaces, wireless communication systems and other technologies for the control and automation tasks. Emerging technologies should help human operators and engineering experts take over the control of the system from anywhere and at anytime. However, the operators need to know the level of security of the communication channels and those security mechanisms, services and approaches used (and where) to protect their control activities at all times. For example, the National Infrastructure Security Coordination Centre (NISCC) and its good practices on firewalls state that a critical network configuration should be based on a division into three main zones: firewalls, IDSs and DMZs. A (HW/SW) firewall is a component in charge of delimiting the boundaries of a system. It analyses incoming and outgoing network traffic, and determines whether this traffic should be allowed based on predetermined rules (e.g., IDs of messages must be unique and identifiable). An IDS/IPS consists of a HW/SW-based mechanism in charge of monitoring network traffic or system processes to detect those activities that seriously violate the security policies established (e.g., abuse of a particular resource or service). It is possible, in a SCADA system, to configure two types of IDSs; Network Intrusion Detection Systems (NIDSs) or Host Intrusion Detection Systems (HIDSs). A NIDS is an independent platform that identifies intrusions by examining network traffic and monitors multiple hosts. Their sensors, strategically deployed at vulnerable locations, are able to capture all network traffic and analyze the content of individual packets for malicious traffic. A HIDS does the same as a NIDS but concentrates on hosts. A DMZ corresponds to a physical or logical sub-network that exposes limited services to untrusted networks, usually the Internet or corporate networks. Within these networks critical servers are generally configured and they generally maintain historical or sensitive information (e.g., alarms, measurements, executed processes, etc.). In addition, the content of these servers could be further protected through encryption methods.

Although the configuration of these three zones is such that they would correspond to the first line of defense for control systems where access to critical servers can be reduced to a defense-in-depth [61] approach, the proprietary nature of SCADA protocols makes the use of security mechanisms for construction of these zones difficult. The vast majority of traditional security mechanisms do not always fit in well with the SCADA requirements and policies. Conventional IDS/IPS systems have to understand the inherent characteristics of SCADA protocols to define very specific rules, which would hinder the building of a scalable and extensible system. In addition, the rules attributed to IPSs have to be well-defined so that their actions should not put the security of the SCADA system at risk. An incorrect action may trigger an effect that may change the normal sequence of the system. For the integration of SW applications and information systems also needs to be protected. To this end, it is necessary to locate existing dependencies between services and applications so that it becomes possible to segment, isolate and protect those critical areas (e.g., applications for alarm management). One way to achieve this would be to reduce their visibility with appropriate access control mechanisms, privileges or roles, or even define restrictions through firewalls or IDSs.
These roles and privileges are assigned according to responsibility areas, functionality and trust, experience and knowledge. They must address the following issues through access control policies: identifying the users that are authenticated and authorized to carry out an action, where and when they may work, and what they could do within the system. These policies are supported by security mechanisms, specialized software and electronic devices (e.g., biometric systems, smart cards, electronic keys, etc.). Nonetheless, we cannot ignore the fact that some current SCADA architectures still rely on simple authentication mechanisms based on user/password where responsibilities are basically constructed with permissions that limit actions. This dependence forces the system to frequently update security credentials through variable patterns; check and close inactive accounts; limit the number of active/inoperative sessions; and automatically block those accounts that present a high rate of failed attempts.

4.2.1 Secure Communication between Control Devices

Apart from the network architecture, it is also necessary to protect the communication channels from external accesses. Confidentiality aspects can be addressed by using cryptographic services (e.g., Advanced Encryption Standard (AES) or Elliptic Curve Cryptography (ECC)) or additional security mechanisms that facilitate the encryption processes. These can be tunneling mechanisms to provide secure virtual connectivity between networks (e.g., Virtual Private Networks (VPNs)) or Bump-in-the-Wire devices which are positioned between the RS/EIA-232 port of the RTU and the modem. One of the first organizations to work on SCADA cryptography was the American Gas Association (AGA), which published two relevant reports: AGA-12 Part 1 [2] and AGA-12 Part 2 [56]. Both of these reports deal with the use and implementation of cryptographic services in serial channels and protocols based on sessions, using authentication services and symmetric keys generated by AES and SHA-1.

Integrity and authentication aspects should also be taken into consideration by using additional security mechanisms (e.g., challenge-response) or services provided or recommended by existing SCADA security standards such as the IEC-62351 [40]. This standard particularly recommends the use of Transport Layer Security/Secure Sockets Layer (TLS/SSL) protocols and digital certificates, message authentication code, key interchange (at least 1024 bits), and the use of cryptographic services such as Rivest, Shamir, Adleman (RSA) and Digital Signature Standard (DSS). It is worth mentioning that the TCP/IP security services offered by the RFC-6272 for the new version of the IPv6 was recently designed for Smart Grids networks [19]. When the environment is composed of objects such as industrial sensors, it is also necessary to consider security services for their communication protocols. Each communication protocol defines its own security restrictions according to its protocol stack. For example, the security in the PHY and MAC layers of ZigBee, WirelessHART and ISA100.11a depend on the IEEE 802.15.4 standard, which offers hardware support for AES-128 bits and the use of a Message Integrity Code (MIC)/Message Authentication Code (MAC) with 32/64/128 bits. This MIC consists of three main fields: a frame control (that includes the security mode (Cipher Block Chaining Message Authentication Code (CBC-MAC), Counter Mode (CTR) or Counter with CBC-MAC (CCM)), a unique counter for relay and key identifier), a security control and the data payload. In addition, the IEEE 802.15.4
standard provides sensor nodes with an Access Control List (ACL) with trustworthy neighbor nodes to authenticate the rest of the peers involved in the communication by themselves.

Depending on the application context (e.g., an open/closed environment) and the computational capabilities of embedded technologies (e.g., sensor nodes) network designers must define the security and network parameters that are necessary to protect the system at all times. An ideal tool to help in these types of configuration tasks is: SenseKey. This tool helps in the selection of the most suitable key management schemes according to the critical conditions of the context and the requirements (such as performance, resilience, scalability, extensibility or global/local communication) of the organization to protect its CIs. Sensekey is an extensible and scalable tool that enables the system to update its database with new protocols, including those key management schemes defined for SCADA communication protocols, the most important among them being ZigBee, ISA100.11a and WirelessHART [10].

On the other hand, network and routing configurations must not cause changes that can lead to congestions at specific points in the network or change its QoS. The protocol knowledge used and its requirements are ideal to try to adapt the object deployment to the network conditions, since each protocol defines its own protocol stack architecture, requirements and conditions for connectivity. For example, most wireless communication standards applicable to WSNs, such as ISA100.11a, WirelessHART and Zigbee, are based on the IEEE 802.15.4 standard and which specifies their PHY and MAC layers. The remaining layers are implemented above this standard, and they depend on the protocol features. Zigbee, for example, defines its own network layer where the nodes follow a network topology based on many-to-one, and the application layer. This layer also includes two important sub-layers; ZigBee Device Object (ZDO) and Application Framework.

4.3 Self-healing: Redundancy, Coordination and Self-Stabilization

Research into self-healing has its origin in fault-tolerance researches. The idea is that the system alone is able to handle either transient or permanent faults through local and individual actions in order to reach its acceptable states in spite of small disturbances. To address aspects associated with fault-tolerance, topics related to redundancy, coordination and self-stabilization should then be addressed as well [68]. Redundancy allows the underlying system to be able to recover its control, by itself, in adverse situations. This is due to the redundant capacities for maintaining backup copies at strategic locations or duplicating functionalities in the entire system (e.g., to configure both primary and standby resources, or specify secondary communication channels using store and forward protocols). In contrast, coordination enables the system to control, by itself, aspects associated with concurrency and consistency in distributed environments when different entities (e.g., software processes, human operators, control resources) significantly interact with each other. To achieve this, the coordination must also be supported by synchronization mechanisms based on actuation policies that regulate all those actions between the implied entities. These policies should be well-defined in order to delimit unauthorized accesses and avoid unsuitable actions that produce damages, interruptions or alterations during control activities in the field.
Moreover, the redundancy should also consider the topic of synchronization since parts of the system can often have a tendency to transient faults because of a problem of coordination. For example, redundant (HW/SW) resources can fall into unplanned states due to discordant executions of processes or inconsistent parameters or variables. One way of mitigating unforeseen states would be to use the concept of self-stabilization. This technique, initially introduced by Dijkstra in 1974 in [30], focuses on offering a support to dynamically control arbitrary transient faults by converging to normal states in a finite number of steps. These steps may be, from the diagnosis and location of a fault within a component to its removal, reparation or re-initialization.

Although self-stabilizing research still requires more attention from the scientific community to try to protect critical and complex systems against unplanned faults, there are traditional approaches in the current literature that should be also considered so that they can equally be applied to critical contexts. For example, Datta et al. proposed in [26] an approach to dynamically control the topic of mutual exclusion in distributed networks, where critical sections between successive executions may remain dependent on the ability of an arbitrary and distributed scheduler. Cheng et. al. similarly proposed in [21] a self-stabilizing approach to control mutual exclusion using token ad-hoc networks in which the resources present a mobile and arbitrary nature. These approaches, amongst others, could be quite useful for those critical contexts where there is no fair and unique scheduler, and/or the communication may depend on wireless infrastructures.

4.4 Modeling and Simulation

Modeling and simulation of multiple infrastructures are two of the most challenging research areas within CIIP. They aim to model and simulate normal or anomalous behaviors in order to analyze complexities, resilience of the infrastructure and correct functionality of fault-tolerance mechanisms to mitigate future risks. This implies a study into the causes, risks, consequences and effect by mapping and visualizing a global representation of existing entities, objects and resources, as well as their interconnections, irregular behaviors (e.g., abuse of resources) and the interdependent relationships between nodes.

Modeling and simulation aim to represent states and situations not only to improve the correctness, but also to understand current behaviors associated with the infrastructure itself and their interrelationships to improve the business continuity from a socio-economic point of view. For example, the organization could (i) calculate the technical/economic situation (in the long term) according to current interdependencies and their impact on a set of economic, social and legal aspects; (ii) locate weak spots in existing infrastructures to optimize future investments; (iii) study the social impact of large scale disruptions/threats; (iv) optimize the deployment of resources/objects; or (v) define strategic plans for mitigation and preparedness. Moreover, modeling and simulation could be used to offer the input to define a suitable governance and security management with new policies or strategic plans, such as business and market plans, contingency and emergency plans, or recovery and mitigation plans (as we mentioned previously).

According to Rinaldi in [72] there are up to six possible ways of modeling and
simulating systems: (i) aggregate supply and demand approaches in charge of analyzing the loss of infrastructure assets; (ii) physics based models using standard engineering techniques (e.g., engineering safety techniques) to evaluate physical aspects of infrastructures; (iii) agent-based models to model operational functionalities and physical states of infrastructures; (iv) population mobility model analysis of how mobile consumers affect the integrity of an infrastructure; (v) Leontief input-output models (economic models) to analyze the spreading of a risk between interdependent infrastructures under time dependencies; and (vi) dynamic simulation approaches to visually represent infrastructure operations, and the effects and consequences resulting from a disruption. Many of these techniques are also analyzed in depth in [69], where various approaches and methods are surveyed, and the challenges associated with modeling and simulation are discussed. Among the most important challenges, it is worth mentioning the complexity of modeling interdependent infrastructures, the time frame for simulations, the type and number of samples and events to be used for observations, and the collection of data.

4.5 Wide-Area Situational Awareness: Prevention, Detection and Response

One of the future challenges that must be addressed in the future is how to avoid or mitigate, as much as possible, collateral effects caused by faults or threats. According to the National Institute of Standards and Technology (NIST) and Federal Energy Regulatory Commission (FERC), this challenge should be considered as one of the priority research areas that should be addressed to protect CIs such as the Smart Grid systems. This priority research area, known as WASA [14, 65], focuses on monitoring critical systems located over large geographic distributions in (almost) real-time, ensuring prevention, detection and response to problems before serious disruptions appear within the system.

Prevention and detection focus on anticipating and/or detecting any internal or external faults that can provoke a deviation from normal state. This protection depends on the effectiveness of proactive tools which are normally supported by high-level security services. In particular, an anomaly prevention service aims to recognize any possible anomalous occurrences at any given moment (i.e., a fault-detection) or in advance (i.e., a fault-prevention). The former (fault-detection) corresponds to those systems (e.g., IDSs) in charge of monitoring the network traffic through patterns, rules and a knowledge source based on events that have occurred in the past. The latter type (fault-prevention) is based on automatic and dynamic tools with the capability to predict the presence of faults, such as EWSs. Similarly, a boundary service is based on the isolation concept where the system has to be able to defend its perimeter against external entities so as to establish a first line of defense through firewalls, IDSs and/or DMZs (cf. Section 4.3). Through cryptographic services it is possible to ensure information protection by using encryption and key management, as well as the prevention of suspicious actions by managing authentication, authorization and non-repudiation.

Response makes use of reactive and recovery tools that help the system to automatically and dynamically address threatening situations, and in a worst case scenario
tries to restore/recover lost states, parameters or variables. Reactive tools must rapidly react to any adverse situation given the criticality of the environment. They must be well-configured to ensure business continuity. Some of their configurations could be, for example, to activate boundary services to isolate compromised components, prioritize activities and services, or limit privileges and accesses so as to reduce any negative impact on legitimate components. A good example of this type of tool is the distributed tool known as Event Monitoring Enabling Responses to Anomalous Live Disturbances (EMERALD), which can trace malicious activities through and across large networks, guaranteeing attack isolation and automated response [77]. Recovery tools are in charge of removing faults within a component as well as being in charge of recovering operational normal states according to security policies (i.e., self-restoration or self-healing). A good example of this kind of tool is one developed by the SELFMAN European project [75] which focused on implementing self-healing in large-scale distributed systems with the capability to automatically and dynamically handle, reconfigure and remove anomalous states. Generally, these protection services for situational awareness can depend on existing technologies (such as the Internet, geographic positioning systems, MANETs, or WSNs) to work [11]. For example, the Internet and geographic position systems offers the capability of global connectivity, location and visualization of the most sensitive physical points that require particular attention from the system at any given moment. MANETs provides local connectivity to attend to a situation in-situ and in real-time whereas WSNs can offer capacities for continued monitoring, detecting, tracking and alerting to threatening situations (cf. Section 2.1.1). However, and regardless of these advances in CIIP, there is a lack of research in this area. Existing situational awareness approaches do not ensure complete protection based on prevention, detection and response as a whole [65], in addition to considering that such a protection should comply with the conditions and prerequisites of the control context (e.g., efficient and rapid operational performance 24/7) [5].

4.6 Forensics and Learning

When anomalous symptoms or faults or threats arise within a critical system, it is recommended to study promptly the sequence of evidence, the modes of operations and the identities that have triggered such situations. Unfortunately, this procedure still depends on conventional forensic techniques, methodologies and guidelines. For this reason, the Colloquium for Information Systems Security Education (CISSE) [23] created a working group in 2008 to implement dynamic techniques suitable for critical contexts, adjusting them to the four basic stages attributed in any forensic methodology which include: information recollection, information analysis, data analysis and report. In addition, and given large scale dimensions of some CIs, forensic techniques should be designed in such a way that they can be applied from anywhere, anytime, and in any mode (either at on-line and off-line mode (i.e., alive or dead nodes), or in-situ), and without putting at risk the performance of the system and its continuity. This means that the techniques applied should not cause the critical system or its sub-systems to fall into anomalous or transient states that can cause isolations or problems in the delivery of services. One way to mitigate this situation is to make use of redundant systems during the four basic forensic phases (cf. Section 4.3).
For the implementation of forensic techniques and methodologies it is also advisable to take into account the restrictions of the context. These constraints can be associated with architectural complexities and interdependencies, existence of dependencies on large ICTs and components developed by third parties (i.e., Commercial Off-The-Shelf (COTS) components); coexistence of heterogeneous technologies and so on. Moreover, computational and storage differences between devices could limit the capacities of some forensic stages mentioned above, such as the gathering and analysis of evidences. We need to develop lightweight mechanisms that can be supported by constrained field devices (e.g., sensors, hand-held interfaces, smart meters, RTUs, or any limited industrial object involved in monitoring tasks) to efficiently analyze and correlate situations (i.e., incidents). A possible solution for evidence collection would be to use external and powerful storage devices that can capture traffic without affecting the overall performance of the system. The additional support from intrusion detection systems/agents and network sensors could help tasks to monitor traffic.

Correlation and analysis of information could result in a valuable input parameter for learning techniques. These techniques would enable the architecture to acquire certain dynamic and autonomous capabilities for decision making. In a nutshell, the system could, for example, learn (by itself) from sequences of anomalous events and automatically generate new patterns/rules to offer a rapid response. In fact, data-mining [82] techniques can be leveraged to predict and discover new behavior patterns through specific techniques, such as sequential patterns [76] or time series and statistical analysis [15]. The selection and implementation of lightweight learning mechanisms should constitute a priority research area, where a set of sensitive variables and conditions (such as the criticality of the environment) have to be properly addressed as well as the existing interdependencies among entities.

4.7 Trust Management and Privacy

As mentioned previously, a simple way to compute trust is through reputation (i.e., a mathematical concept that enables a system to increase not only its decision-making processes but also its ability to compute the level of reliability of observed entities). Taking advantage of these benefits, an initial incident management framework for control systems is proposed in [3]. This framework assigns and monitors alarms according to the severity of the incident together with the staffs’ work availability according to their contract, workload to assist an emergency situation and their experience. The computation of the concept of trust and reputation can be carried out through different approaches where their models can be based on logic, graph theory, bayesian networks, or on particular centralized architectures [51]. Beyond that, most of the research issues in this area, such as lightweight solutions, are still needed.

A recent issue, particularly related to advances in Smart Grid technologies and their relationships to various domains is the topic of privacy. In this context the communication with electrical utilities can be carried out through bidirectional infrastructures (i.e., AMI communication) where the transferred information, obtained from smart meter devices, enables utilities to improve the efficiency of the energy distribution processes. Indeed, control utilities use the information from these bidirectional networks to control the demand load during peak times or reduce unnecessary power generation.
Keeping the privacy in this context involves protecting the communication channels (either via the Internet or wireless networks), and any activity associated to lifestyle routines. Privacy becomes an important issue when the activity pattern may be deduced by analyzing the signals received from home appliances, known as load signatures or power fingerprints [83]. Kalogridis et al. proposed in [52] a power management model based on the use of batteries in appliances in order to register different load signatures inside the smart meter thereby hiding real electricity usage values. More research in this area is still needed to develop cost-effective, robust solutions that can be easily supported by appliances devices and smart meters.

Research on location-based privacy should also be addressed when the location of industrial devices or objects (e.g., sensors) are exposed, putting the resilience of the system at risk [86]. Most of the existing approaches proposed that aim to protect the location information are based on the intrinsic features of exchanged signals (e.g., strength, coverage, etc.), the observed network traffic, the number of hops or the coverage area [55]. It is also imperative to prevent external entities from inferring the location of devices by analyzing, for example, the network traffic [66]. However, these solutions usually require synchronization techniques, additional hardware, computational resources that could become excessive for limited devices (e.g., sensors, smart meters). Therefore, additional research is needed and new lightweight solutions should be developed in the future.

5 CONCLUSION

Protection measures should be considered when new, but also conventional, technologies and information systems are being adapted to control the vast majority of our critical infrastructures. These measures should not only include traditional security mechanisms to detect and react against potential threats, but the system should be also based on intelligent mechanisms with the capability to identify vulnerabilities and faults that can be exploited by intruders.

The goal of the attacker is basically to try to bypass the security mechanisms so that once inside the system others types of threats can be launched, such as memory dump, execution of false commands to activate/deactivate critical assets, modification of state values or critical processes, etc. This paper has analyzed, on the one hand, the relevance of new technologies in control and automation tasks, and on the other hand, the motivations for the need to protect control systems when using these technologies and their information systems. Moreover, these analyses also include a study of security requirements we need to protect the control systems and the protection of the controlled critical infrastructures themselves.

We have identified several research areas that should be explored further in the future to enhance the protection of critical infrastructures. These priority research areas include governance and security management, robust network design and secure communication channels, self-healing, modeling and simulation, wide-area awareness situational, forensic and learning, trust management and privacy.
6 ACKNOWLEDGMENTS

The research leading to these results has received funding from the MARIE-CURIE COFUND programme U-Mobility, co-financed by the University of Málaga, the EC FP7 under GA No. 246550 and the Ministerio de Economía y Competitividad (COFUND2013-40259).

References


