Information Security and Digital Forensics in the world of Cyber Physical Systems

Andrew Jones
Cyber Security Centre,
University of Hertfordshire, Hatfield, UK
Security Research Institute, Edith Cowan University,
Perth, Australia
andy1.jones@btinternet.com

Stilianos Vidalis
Cyber Security Centre,
University of Hertfordshire, Hatfield, UK
s.vidalis@herts.ac.uk

Nasser Abouzakhar
Cyber Security Centre,
University of Hertfordshire, Hatfield, UK
n.abouzakhar@herts.ac.uk

***Abstract*— The security of Cyber Physical Systems and any digital forensic investigations into them will be highly dependent on data that is stored and processed in the Cloud. This paper looks at a number of the issues that will need to be addressed if this environment is to be trusted to securely hold both system critical and personal information and to enable investigations into incidents to be undertaken.**

Keywords*— *Digital Forensics, InformationSecurity, Cyber Physical Systems, Big Data

I. INTRODUCTION

As computing technology is incorporated into an ever widening range of applications that affect our everyday lives (Instagram, Google, Spotify, Uber, Seamless…), users are increasingly being asked to trust that it (the technology) will function correctly and that information that they provide, or that is collected about them, will be adequately protected. Menou (1995), as cited in [1], described information as "a product, which encompasses information as thing, as object, as resource, as commodity, what is carried in a channel (including the channel itself), the contents." Given today's socially-driven knowledge-centric virtual-computing era specific attributes such us interconnectivity, information exchange speed, and social impact, coupled with the lack of cyber-ethics, there is a need to expand the definition to include the concepts of 'community' and 'environment', addressing the different types of computing devices (e.g. smart phones, smart embedded devices, game consoles, laptops, computers, etc.) and a domain that goes beyond the concept of the term "cyber-domain". For the purposes of this paper we will use the term Information Environment (IE). The U.S. Department of Defence (DoD) has defined the Information Environment (IE) in [2], as "*... the information environment is the aggregate of individuals, organizations and systems (resources) that collect, process, disseminate, or act on information*." This definition can also be used for describing the Internet of Things (IoT), which consists of everyday objects that have uniquely identifiable embedded processors that are connected to the Internet.

The average person typically currently thinks of the types of objects that will be connected to the IoT as the fridge, the coffee maker or the washing machine. In reality there are a huge number of devices and applications that are already in use (to do things such as monitoring babies and toddlers, managing medicine usage, tracking personal activity levels, monitoring aging family members, e-Health and remote doctors, controlling kitchen appliances, controlling smart home sensors, controlling smart cities, navigation, tracking assets and may others). In the near future, the interconnection of such embedded devices is expected to enable automation in nearly all fields, including applications such as Smart Grids, integrated transport systems and Vehicular Ad-Hoc Networks (VANETs).

In this paper we look at the range of issues that must be considered when securing data used in these systems and the issues that will be faced when attempting to carry out a digital forensic analysis either as part of a criminal investigation or as part of an audit.

II. CYBER PHYSICAL SYSTEMS

A cyber physical system (CPS) is a system of collaborating digital systems that are controlling physical entities. A range of cyber physical systems already exist in areas such as aerospace, the automotive industry, chemical processes, the civil infrastructure, energy, healthcare, manufacturing, transportation, entertainment and consumer appliances. The current generation of devices normally consists of embedded systems, communications links and computers that are used to coordinate the activities of the individual entities. At the current time, two of the more obvious and visible manifestations of this in the UK are the Docklands light railway in London and the pod system at Heathrow airport terminal 5.

In the UK many of the major cities, including Milton Keynes, Birmingham and Glasgow are looking to develop intelligent transport systems in order to address the ever increasing problem of congestion in the existing transport systems. Overseas, places such as Abu Dhabi in the UAE have plans for a 'smart city' by 2030, which will see the introduction of a number of new transport modes in order to reduce the Emirate's reliance on the car. This will be achieved by implementing a network of public transport systems, including high speed rail and rapid transit options, such as trams and buses, as well as initiatives for walking and cycling.

Cyber physical systems, by definition, are real-time, intelligent, adaptive and predictive networked or distributed systems that produce and use a range of data inputs, with or without human interaction/intervention, to enable them to operate, and while most of the data will belong to the individual objects, there will often be links back to a person/user/customer in some form or another. Whether data refers to an 'object' or a 'person' they will overlap, interconnect and much of their value will be derived from these connections and interrelationships. As a result, issues that affect personal data such as trust, security and privacy are just as important in the CPS as they are in other aspects of computing. This in itself is an issue for concern. As reported in [3], there is still no singular privacy law (The European Union General Data Protection Regulation was not enacted at the time of writing). Coupling this with the attributes of the IoT as an Information Environment for CPSs, one can argue that there is a very real problem with a potentially serious impact towards the prevention and prosecution of cyber-crime.

It is increasingly clear that the security of Cyber Physical Systems (CPS) and Big Data must be dealt with in tandem as the one relies heavily on the other. The issues that both the CPS themselves, and the Big Data that they rely on, also have to be addressed together when any attempt is made to put in place an effective and appropriate level of security or carry out a forensic investigation of an information environment such as the IoT.

In addition to the IoT devices and CPS systems collecting, storing and processing data about entities and their environment, they are also doing the same with data about users/customers. This in itself is not a major problem, as there are many other systems that are doing the same thing, from banks and financial institutions to supermarkets and social networking websites. However, with each of the aforementioned examples, the user provides information to a known entity for a specific purpose and will sign an agreement (even if they don't read it) with regard to what that data can be used for.

When we start to look at integrated transport systems that manage perhaps road, rail and air transport, the issue starts to become more complex. In order for these integrated systems to operate efficiently, to give user satisfaction and to allow the service to be personalised to the user, they will have to collect varying levels of information about the user. For example, in an integrated transport system, it may be necessary for the system to know who the customer is, for card payment or so that their preferences can be used to personalise their journey. The systems will capture details of the customer's journey, perhaps across several modes of transport, and may use this information to enhance the experience of the service provided by adjusting the environment of the vehicle (perhaps the temperature of the pod or by playing music that the user has previously indicated that they like). The information may also be used to adjust the speed or route of the vehicle to ensure that it reaches its destination to enable an easy and timely connection to another mode of transport.

The personal information will be used together with data from a wealth of other sources such as traffic and environmental sensors, power monitors and vehicle operating sensors in order to make the system work efficiently and safely. The issue relating to trust, security and privacy is not one single system managing a dataset of personal data, but a number of systems working collaboratively, combining a number of different and disparate datasets in order to use extracted knowledge in non-authorised ways.

III. SECURITY

Naturally, in order for integrated transport systems to operate effectively, there will be a need to collect and process vast amounts of data from a large number of sources, some of which will belong to the entity that is operating the system and many more from, and owned by, external agencies. The systems will have a high level of complexity and it is only when the data is fused that it will fully serve the requirements. According to [4], "*a system is defined as a regularly interacting or interdependent group of items forming a unified whole*".

One of the security issues that will have to be considered is how the personally identifiable information (PII) will be protected. While individual systems might be adequately securing PII, this might not be the case when they get connected to a system of systems. Historically, security controls have been failing were there were interconnections or system merges. The issue is twofold: functionality and boundaries changing type. A boundary of a system is the point where the system is receiving or sending information to processes outside its control.

Another security issue is that of data aggregation. While an individual may be happy to provide information for use on the individual elements of the integrated system, the majority of them will be unaware of the potential effect of the aggregation of this data over the different elements of the system and over time. From the point of view of the owners of the integrated transport system, the personal information is important for use in billing, service optimisation and service delivery. The user has no option but to provide some information, as payment in cash may be problematic or not possible, and in order to get the best experience and personalisation from the service. The individual elements of information provided may have little

value, but when they are combined, over a period of time, they could allow for a significant profile of an individual to be created, such as their travelling patterns and the locations they visit, their current location, their preferences, the people they travel with etc. to be revealed.

From the early planning stages of this type of integrated system, the ownership and protection of this type of information needs to be considered, not only in the area of the storage and processing of the information, but also in the communication of it.

To understand the full extent of the security issues we will use Samsung Electronics UK as an example. Going back to the integrated and intelligent semi-autonomous and personalised transport system that we described in the previous section, customers may user their smart devices (smart-phones) for their interaction with the system. Activities such as authentication, purchase of tickets/goods, personalisation of experience, evaluation and feedback, itinerary management will be conducted using the phone, on the go, when and where the customer is, around the clock, without the need of a human operator on the other side. According to [5], Samsung processes customer and supplier information relating to:

- Personal details,
- Family details,
- Lifestyle and social circumstances,
- Education and employment details,
- Financial details,
- Goods and services
- Furthermore, Samsung states they process sensitive classes of information that includes:
- Racial and ethnic origin,
- Religious and other beliefs,
- Trade union membership,
- Physical and mental health details,
- Offences and alleged offences,
- Visual images, personal appearance and behaviour,
- Criminal proceedings and behaviour.

One could question the ethical reasoning behind the data Samsung collects and analyses, but this would be beyond the scope of this paper. The aforementioned personal data is a rather significant element of the total amount of data collected within the Samsung systems, and if not adequately protected, could allow for an individual to be tracked, or for their personal information to be stolen or modified. Fusing and datamining the mobile phone dataset (which is part of the data that Samsung is in control of) with the aforementioned transport and travel dataset discussed in the previous section and one could argue that every single person travelling in London could be extensively and continuously tracked and profiled, both in the virtual world and the physical world.

IV. BIG DATA

The volumes of data that will be produced by a range of sources; the integrated, processed and stored data to enable a CPS to work effectively, will be huge. A term which is increasingly being used to describe the large volume of data - both structured and unstructured – is 'Big Data'. Big data can have a significant value in itself and can also be analysed for insights that allow for better situational awareness and lead to better strategic business decisions.

The whole field of 'big data' and big data analytics and data mining is developing at a rate to meet the needs of large and complex systems. Big data has three main characteristics, known as the 3 'v's: Velocity, which describes the speed with which data comes in and out; Volume, which describes the ever increasing quantity of data; and Variety, which describes the range of data sources and types.

An article in [6] in March of 2014 gave an insight into the value of big data when it stated: '*that data analysis produces uncannily accurate results; that every single data point can be captured, making old statistical sampling techniques obsolete; that it is passé to fret about what causes what, because statistical correlation tells us what we need to know; and that scientific or statistical models aren't needed because, ….with enough data, the numbers speak for themselves'.*

According to another report from [7], *'there are five main ways in which using big data can create value. First, big data can unlock significant value by making information transparent and usable at much higher frequency. Second, as organizations create and store more transactional data in digital form, they can collect more accurate and detailed performance information on everything from product inventories to sick days, and therefore expose variability and boost performance. Leading companies are using data collection and analysis to conduct controlled experiments to make better management decisions; others are using data for basic low-frequency forecasting to high-frequency nowcasting to adjust their business levers just in time. Third, big data allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services. Fourth, sophisticated analytics can substantially improve decision-making. Finally, big data can be used to improve the development of the next generation of products and services. For instance, manufacturers are using data obtained from sensors embedded in products to create innovative after-sales service offerings such as proactive maintenance (preventive measures that take place before a failure occurs or is even noticed).*'

However, 'big data' can bring with it its own problems and, as with many other uses of technologies, big data solutions are being used in ways that were never intended by their developers. By its very nature, big data tends to exist in systems with a distributed architecture. Because most of the data that is used is unstructured and security is not inherent in many of the data sources (as already discussed in a previous section), both organisations and vendors have to retrofit security

into the systems that they use, and historically, vendors did not design security controls for distributed knowledge-based computing architectures.

The handling and protection of those elements of personal data, which will undoubtedly only form a very small subset of the overall picture, will need to be adequately addressed throughout their whole lifecycle. The designers of such systems, on top of everything else (scalability and complexity management, modularity and synthesis, interfacing with legacy systems, time synchronisation, validation and verification) will need to consider how this will be achieved and who will be responsible for these elements of data in a hugely complex system that is highly interconnected across the Internet.

A growing number of organisations are now using the concept of big data to store and analyse petabytes of data in order to gain better insights into their customers and also their own business in order to optimise the services and products that they offer and to ensure that they operate as efficiently as possible. As a result, the classification of the information has become essential. Apropos, in order to carry out any reasonable classification of the information, its ownership must be known and appropriate metadata must be collected. For most organisations, the ability to achieve sensible classification of data has either not been a priority or the ability to do so has eluded them to date. In 2011, the authors conducted data classification operations under a UK Government funded project. As an indication, we could acquire data at a rate of 150.37 MB/min. The de-duplication operation required 5 hours for 211.9GB and the indexing operation required 5 days for 149.4GB. It is understandable as to why the majority of the organisations have not as yet developed the capability for classifying data. Furthermore, if data classification is to be achieved, the ownership of both the raw data that is the input, as well as the outputs, must be known.

This will be essential in order to adequately secure both the business critical data of the organisation and the personal data of the customers. It is only when you can identify all of your operation–critical assets and understand the interrelationships of their vulnerabilities that you can develop and deploy adequate security measures to protect them. All of this will almost certainly have to be outsourced and take place in a cloud environment, as very few organisations will have the ability or desire to develop their own virtualised or physical infrastructure to deal with big data.

V. Digital Forensics of Cyber Physical Systems

In a previous section we have argued that the PII that is used within any CPS will be a small part of a much larger dataset. Digital forensic practitioners and academics have, over the last few years, developed procedures and toolkits for recovering data related to their investigations from large data stores such as web farms and the cloud (see [8], [9], [10] and [11]). While much of this data is unstructured, such as email or documents, it is contained in a structured architecture. This means that it is possible to identify things such as an individual or group of webmail accounts or the cloud storage space used by a specific user. In a CPS system, much of the PII may not be so easily isolated.

Some (but by no means all) of the issues that will need to be addressed for a digital forensic investigation on the big data that will be part of a CPS include: the capture of the relevant elements of structured data sources, unstructured data sources, real time data and time sensitive data (that which only exists for a short period of time) and the relevant meta-data about the data. Once the investigator has managed to do this, the next hurdle that they will need to overcome is that of correlating all of the disparate elements of information that they have gathered. There are currently very few tools that are available to the investigator and as yet, this issue has not been the subject of any real level of research. The reality is that there is going to have to be a rethink of what we consider to be digital forensics. The scientific basics of digital forensics were given in a definition by the Digital Forensic Research Workshop in 2001 as *"the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations."* The current reality is that in forensic investigations of big data, the much sought after standards that was set out in Daubert v. Merrell Dow (1993), which included evidentiary reliability, testing, error rate (is there a known error rate of the procedure?), publication (has the procedure been published and subject to peer review?); and acceptance (is the procedure generally accepted in the relevant scientific community?) is not currently achievable and is not likely to be so in the foreseeable future.

Even in this difficult environment, organisations can put in place measures that would assist an investigation. More than a decade ago, [12] produced a ten step process for forensic readiness and while the environment has changed, the steps outlined are still valid.

VI. Conclusions

The use of CPS is increasing and will affect an increasing number of people. We accept that fusion and transportation of data is an essential element in these systems. We also accept that the majority of the people in modern societies are not particularly concerned about their data as long as they can happily and securely use technologies and commercial products to enhancing their social lives. In order to adequately protect the information that is used on these systems it is essential that security measures are considered from the design phase onwards. The ownership of both PII and business critical information must be determined and correctly classified at each stage of the data lifecycle so that it can be properly protected.

We will turn the clock back to the 1990s when academics and practitioners alike were discussing the integrated supply chains and identified that the weakest link is the actor that will create a detrimental impact to the chain and its environment. Any system consists of a number of subsystems. Security standards must be adhered to by every subsystem. Boundaries must be clearly defined, associated stakeholders must be identified, and security controls must be implemented (and appropriately managed) throughout the lifetime of the systems and of the datasets. We are not suggesting that we reinvent the wheel. We are suggesting we should adopt best practice developed in other application domains into the CPS domain.

Finally, any digital forensic investigation is likely to be time consuming and complex and will continue to be hampered by the current lack of effective tools to deal with these complex and distributed environments unless considerable research into the discussed issues is carried out. There will also need to be significant effort made in the development of new regulations for governing CPS environment and regulating CPS stakeholders.

REFERENCES

[1] Chowdhury, T., and Vidalis, S. (2013) Proactively defending computing infrastructures through the implementation of live forensics and website capture in corporate network security. Third International Conference on Cybercrime, Security and Digital Forensics, Cardiff University, Cardiff, UK, June.

[2] U.S. Department of Defense. (2012) Information operations. Joint Publication: 3-13. Retrieved from http://www.dtic.mil/doctrine/new_pubs/jointpub_operations.htm, last access: 17.05.2016

[3] Duffy, D. (2015) Wearable Devices: how they affect their users privacy. Staffordshire University: Report. April 2015.

[4] Herrmann, D. S. (2002). "The first step in security engineering and information assurance: defining the system boundaries." EDPACS XXIX(10): 1-14.

[5] ICO. (2015) Data Protection Public Register. [Online] Available at: https://ico.org.uk/ESDWebPages/Entry/Z9408010. [Accessed 16.05.2016]

[6] Harford T., Big data: are we making a big mistake?, Financial Times, 28 march 2014,

[7] Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Hung Byers A., Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, May 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[8] Belorkar, A., & Geethakumari, G. (2011, December). Regeneration of events using system snapshots for cloud forensic analysis. In India Conference (INDICON), 2011 Annual IEEE (pp. 1-4). IEEE.

[9] Patrascu, A., & Patriciu, V. V. (2014). Logging System for Cloud Computing Forensic Environments. Journal of Control Engineering and Applied Informatics,16 (1), 80-88.

[10] Sibiya, G., Venter, H. S., & Fogwill, T. (2012). Digital forensic framework for a cloud environment.

[11] Zachary Reichert, Katarina Richards, Kenji Yoshigoe, Automated Forensic Data Acquisition in the Cloud, 2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems

[12] Rowlingson R., A Ten Step Process for Forensic Readiness, International Journal of Digital Evidence, Volume 2, Issue 3 http://www.utica.edu/academic/institutes/ecii/publications/articles/A0B13342-B4E0-1F6A-156F501C49CF5F51.pdf