Securing Cyber-Physical Spaces with Hybrid Analytics: Vision and Reference Architecture

Daniel De Pascale[1](✉), Mirella Sangiovanni[1], Giuseppe Cascavilla[2], Damian A. Tamburri[2], and Willem-Jan Van Den Heuvel[1]

[1] Jheronimus Academy of Data Science, Tilburg University,
Tilburg, The Netherlands
`d.de.pascale@tue.nl`

[2] Jheronimus Academy of Data Science, TU/e - Eindhoven University of Technology,
Eindhoven, The Netherlands

Abstract. Considering the massive increase in the number of crimes in the last decade, as well as the outlook toward smarter cities and more sustainable urban living, the emerging *cyber-physical space* (CPS) obtained by the interaction of such physical spaces with the *cyber* elements around them (e.g., think of Internet-of-Things devices or hyperconnected mobility), plays a key role in the protection of urban social living, e.g., social events or daily routines. For example, the hyperconnectedness of a CPS to many networks can lead to potential vulnerability. This vision paper aims to outline a vision and reference architecture where CPS protection is center-stage and where CPS models as well as so-called hybrid analytics work jointly to help the Law Enforcement Agents (LEAs), e.g., in event monitoring and early detection of criticalities. As a part of validating said reference architecture, we implement a case study in the scope of VISOR, a Dutch government project aimed at improving CPS protection using hybrid analytics. We conduct a field experiment in the Paaspop social event and festival grounds to test and select the most appropriate device configuration. There we experiment with a CPS protection pipeline featuring several components reflected in the reference architecture, e.g., the KGen middleware, a prototype tool to anonymize structured big data using genetic algorithms, and SENSEI, a framework for dark web marketplace analytics. We conclude that hybrid analytics offer a considerable ground for more sustainable CPS.

Keywords: Cyber physical space (CPS) · Internet of Thing (IoT) · Law Enforcement Agents (LEA) · Genetic algorithm (GA)

1 Introduction

Cyber-physical systems (CPS) integrate physical processes with computational engineered systems [12,14] to solve a real-world problem. An example of CPS is

S. Katsikas et al. (Eds.): ESORICS 2022 Workshops, LNCS 13785, pp. 398–408, 2023.
https://doi.org/10.1007/978-3-031-25460-4_23

self-driving cars, where the physical part (the vehicle) meets the cyber part (the self-driving software) to solve the self-driving engineering problem.

Being center-stage in supporting smarter urban living, CPS' is increasingly becoming one of the main targets of all sorts of attacks. For example, cyberattacks are often discovered in many societal scenarios (e.g., value transactions, money exchange, post and stamp services). Similarly, CPS does not stand isolated from their operational context—a physical space with sensors that intercommunicate with the CPS itself—rather they are contiguous with such context and therefore are at risk from multiple engagements in such context.

For example, in 2016, a significant distributed denial of service (DDOS) attack interrupted many Internet services in the USA and Europe [17], causing delays in civil transport, as well as economic and financial meltdowns. Garroppo et al. in [8] investigated how to identify social events by analyzing anomalies in cellular traffic data. Similarly, Yuan et al. [23] did anomaly behavior detection in a crowd scene. They propose the analysis of anomaly motion behavior between individuals investigating the optical histogram flow between frames.

To address these kinds of safety issues, several works act under the physical domain [16] (i.e., analysis of anomaly behavior patterns through IoT devices). Other works use social and dark web analysis, providing models helping Law Enforcement Agents (LEAs) to analyze marketplaces and forums trending on the dark web, extracting useful insight for future investigations [10].

In this work, we present a proposed framework to introduce a reference architecture providing an alternative approach where physical space protection models—e.g., knowledge graphs [11] obtained on the physical spaces' components and connectors—work jointly with dark web analytics to improve the detection and response against critical actions. Starting from the models executed with IoT devices data (e.g., drug trading between persons), the idea is to use social and dark web analytics to improve these models in detecting, among others, meta-data describing the drugs or creating a more appropriate response through insights extracted from the Internet or other open-source intelligence databases.

The paper is organized as follows: in Sect. 2 we discuss the vision of our approach, explaining the architecture outlined in Sect. 1. Section 3 describes the data sources used as a baseline for the proposed architecture analytics. In Sect. 4 the data received from the previous step are anonymized before the analysis. Section 5 illustrates two approaches for countering illegal trafficking activities. Section 6 is the core of this paper and its vision. It explains the general idea behind the "hybrid analytic" concept and describes the case study we aim to set up. Lastly, Sects. 7 and 8 discuss, respectively, about the limitation, discussion and conclusion of this work.

2 Cyber-Physical Space Protection: A Hybrid Analytics Approach

One of the main strengths of a cyber-physical system (CPS) is the ability to monitor and control a physical environment—such as a social event, or a public

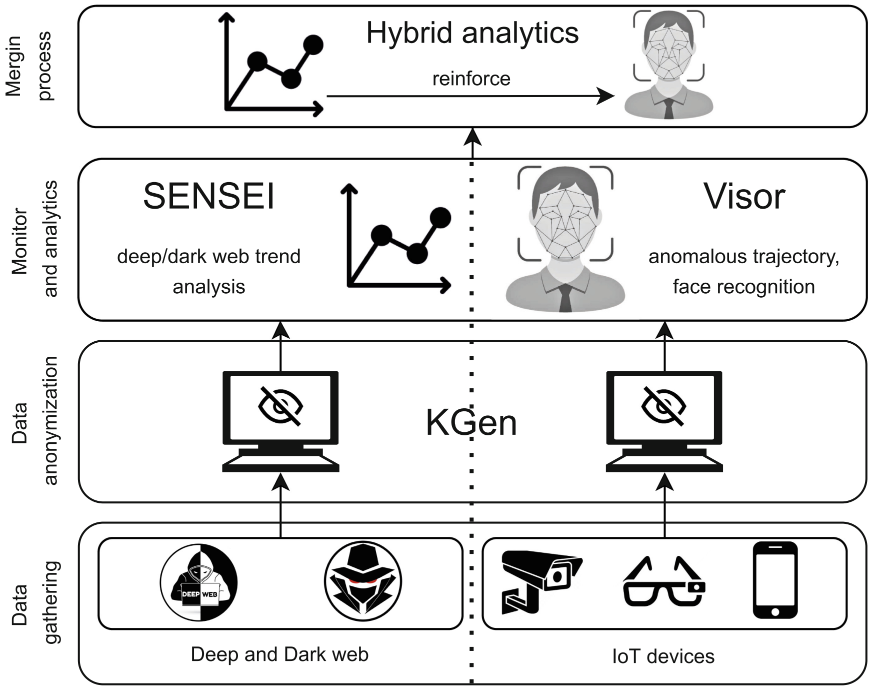

Fig. 1. Architecture of a CPS monitoring using both social event monitoring and deep/dark web analysis. It is divided into four layers: data gathering (1), data anonymization (2), monitor and analytics (3), merging process (4).

space [13]—using and interconnecting multiple sensor networks and data sources. In this type of CPS scenario, the design of models to predict fraudulent behavior plays a key role. However, the accuracy of CPS models needs to be continuously improved by using analytics that predicts critical events, e.g., terrorist attacks in a public space. Several methods protect CPS for the specific context in the state-of-the-art. For example, Nagarajan et al. [16] developed an alternative approach for anomaly detection in the CPS domain. The author proposed a deep learning technique, namely Convolutional Neural Network with Kalman Filter based Gaussian-Mixture Model, to identify anomalous behavior in the CPS context. Another CPS scenario has been provided by Du et al. [6]. Their work addresses the detection of pickpocket suspects in a large-scale public context. Studies in the literature focus on identifying anomalies in passengers' movement patterns [3]. The novelty of the work proposed in [3] is the usage of public transit records to recognize pickpocket actions.

This paper focuses on integrating information extracted from public spaces using analysis performed on the deep and dark web while blending this information with available models of the phenomenon under support, for example, the protection of the CPS and sustainable urban living within the spaces around it.

Our work provides a reference architecture to reinforce physical space protection through IoT devices augmented with deep and dark web analytics. For example, in a social event scenario, the recognition of drugs can be significantly improved if a dark web markets analysis results in massive trading of a specific drug in that area or country. This way, an analytic service model can be learned on that specific drug to improve its accuracy and deployed at the edge of the CPS. The following section generalizes this simplistic approach and outlines a reference architecture.

2.1 Reference Architecture

Figure 1 represents the multi-layer architecture offered in this work, with four different self-contained layers built on top of each other according to a layered software architecture style [2]. Each layer has a specific purpose following the typical data-intensive computing services architecture [1], from data gathering to a hybrid analytics service layer. More specifically, the hybrid analytics architecture encompasses the following four layers:

1. **Data gathering layer:** the first layer is in charge of downloading raw data from the web and social events. It includes IoT devices like smart glasses, smartphones, CCTV, and drones. Along with other devices, they collect real-time data to monitor and analyze social event scenarios. Data gathered include images, video, Global Positioning System (GPS), metadata, and sound signal, depending on the device. All the data are saved into a cloud or local storage and then passed to the next layer for the anonymization step.
2. **Data anonymization layer:** after gathering the raw data, they must be anonymized, ensuring the usability of the data. While data completely anonymized is useless, data partially anonymized can contribute to the analytics and preserve the anonymity. The data to anonymize must be structured. If data is unstructured (e.g., image, video), they bypass the anonymization layer to be directly analyzed in the monitor and analytic layer. Otherwise, all the structured data are anonymized before the following analysis. For example, item location distribution and price of a dark marketplace product must be anonymized using generalization rules [20]. Prices can be anonymized using range values criteria. Contrarily, the location anonymization process follows the generalization hierarchy criteria. For each anonymization layer, the level of abstraction is higher (e.g., Roma -> Lazio -> Italy -> Europe). The product anonymized is used in the monitoring and analytic layer (layer 3).
3. **Monitor and analytic layer:** this step of the architecture performs monitoring and analysis processes. The cyber and the physical part are analyzed separately with different results. Examples of analysis performed are 1) trend analysis for the dark web drugs monitoring, 2) anomalous trajectory and pickpocket detection for the social event monitoring. For example, coming back to the anonymized product of layer 2, using the item location, we can accomplish a trend analysis of the number of drugs in a specific country.

4. **Merging process layer:** at the end of the architecture, the analysis is joined together to check if there is an improvement in the results compared to the disjoint analysis. The main core of the hybrid analytic layer is to use the results of the dark web analysis to reinforce and improve the analysis of physical space protection models. For example, suppose the analytic on the dark web provides a massive usage of amphetamine in the Netherlands. In that case, we can set up a machine learning model trained with amphetamine to improve the accuracy of the detection process.

A more detailed description of each layer is available in the following sections.

3 Data Collection

In the first layer of the architecture in Fig. 1, the data are gathered for future analysis. The macro area sources involved in this project are twofold: a) cyber sources, with pages crawled from the deep and dark web and b) physical sources, with data gathered through IoT devices during social events.

3.1 Cyber Sources: Deep and Dark Web Pages

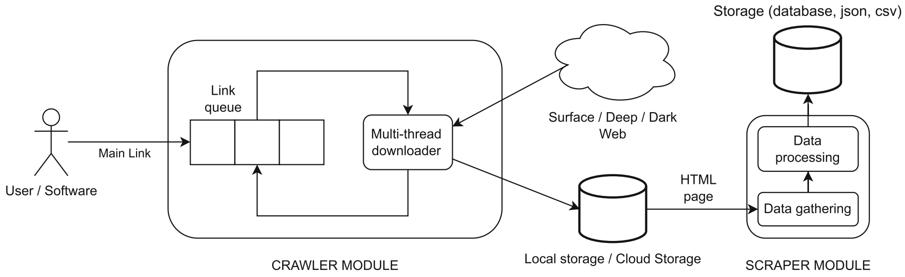

Fig. 2. Architecture of a multi-thread crawler. It represents the main workflow of the crawler, showing the iteration over a link queue list to download every web page found, and its integration with the scraper module, in charge to extract insights from the crawled web pages.

We implemented a crawler to download web pages automatically from the deep and dark web. The main goal is a solid dataset to analyze further and prevent crimes. Figure 2 shows a general architecture of the crawler. After feeding the crawler with a dark web onion link to crawl, it then downloads all the web pages, extracts information from them, and finds a new link to analyze. At first, the link queue contains only the main link. From the main link analysis, the tool extracts all the website links, puts them in the link queue, and iterates this process for each link extracted. To maximize the numbers of pages crawled, the main link might be the home page of a deep or dark website (e.g., onion links of a dark web marketplace found in the Hidden Wiki [22]). This process runs until all the links have been analyzed or the user stops the tool. Lastly, the pages downloaded and the information scraped are saved online (cloud-based solution) or locally.

3.2 Physical Sources: IoT Devices

IoT devices can be used to gather information from social events. For example, smart glasses and smartphones ease data gathering in a dynamic context. They allow the detection of entities with a low-range field of view (fov). On the other hand, drones have a high range of fov and can record video far from the action, keeping the camera focused on a specific area. In addition, CCTVs do not have any battery limitations, and the advantage is to focus on hotspot areas.

The data gathered are anonymized and stored in an encrypted hard drive to preserve their privacy, following the GDPR [21].

4 Data Anonymization

The main purpose of this architecture layer is, from one side, to anonymize the data crawled and gathered in the previous layer while keeping the main characteristics of the dataset unchanged and still useful for analysis and extraction of criminal behavior.

4.1 KGen: A Data Anonymization Tool for Structured Data

We developed KGen to ensure the right level of anonymization. KGen is a tool for data anonymization based on the k-anonymity property. As part of the project with the Dutch Tax Authority, our goal was to create a tool that could anonymize a big dataset in a reasonable time.

The main challenge is to provide a dataset anonymized while keeping the main properties of the data useful for future analysis. Moreover, it is crucial to ensure the k-anonymity property itself.

Several algorithms address the k-anonymity problem providing an optimal solution [7,15,18–20], but none among them handle big datasets. To account for the trade-off mentioned above, KGen develops an approach based on the Genetic Algorithm [9], providing a pseudo-optimal solution in a reasonable time for practical usage. In the KGen context, a solution represents a different level of generalization of each attribute to anonymize. For example, if an attribute to anonymize is a date with the format YYYY-MM-DD, there are at most two anonymization levels. In the first level, the day is obfuscated (YYYY-MM); in the second level, only the year is visible (YYYY). The scope of KGen is to minimize the level of generalization of each attribute while ensuring the k-anonymity principle.

The workflow of the approach is shown in Fig. 3. The tool generates a set of solutions containing random levels of anonymization of each attribute, starting from the dataset and a configuration file with the dataset's metadata. The

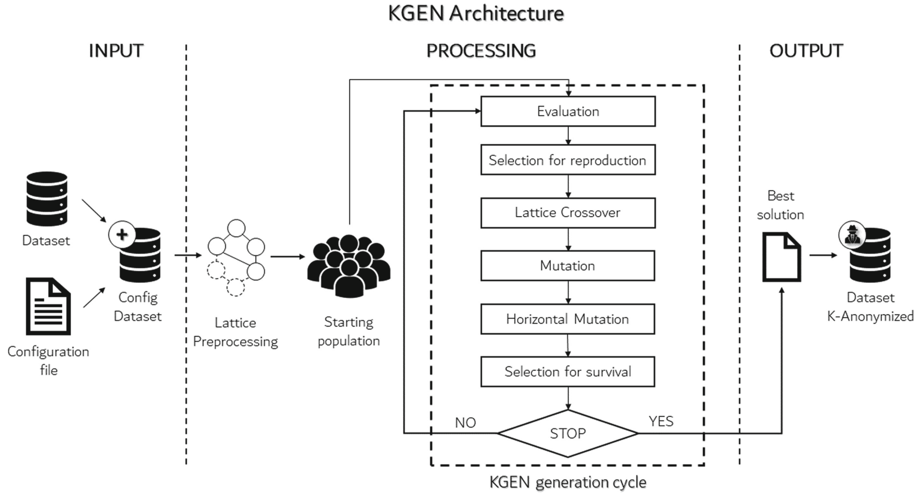

Fig. 3. KGEN Pipeline. It is divided into three steps (separated by dotted vertical lines), input, processing, and output; the KGEN-GA architecture is described in the processing step.

metadata config file has two dataset attribute information: 1) a boolean value used to determine whether a variable can be anonymized or not, and 2) the attribute type (e.g., numeric, string, data). Iterating multiple times, the genetic algorithm reaches the best pseudo-optimal solution that, in the end, is applied to the starting dataset to generate the anonymized version.

The main advantage of a genetic algorithm approach is the application of an anonymity tool even when a dataset contains lots of attributes to anonymize in a reasonable time. Nonetheless, its main disadvantage resides in the solution quality, expressed as the distance between a genetic algorithm solution and the best solution generated by a heuristic approach.

5 Data Analytics and Monitoring

After the data anonymization is performed in the previous step, the data is used for analysis and monitoring purposes. However, cyber data are different from physical data due to the data origin and the sources. Consequently, the process provides two methods: 1) analysis of cyber data crawled for the deep and dark web, and 2) analysis of data extracted from IoT devices placed in physical space. As a part of the architecture, we implement a case study to analyze the behavior of anomalous trajectories in a social event scenario, namely VISOR [4] and SENSEI [5], a web-based platform for dark web analytics.

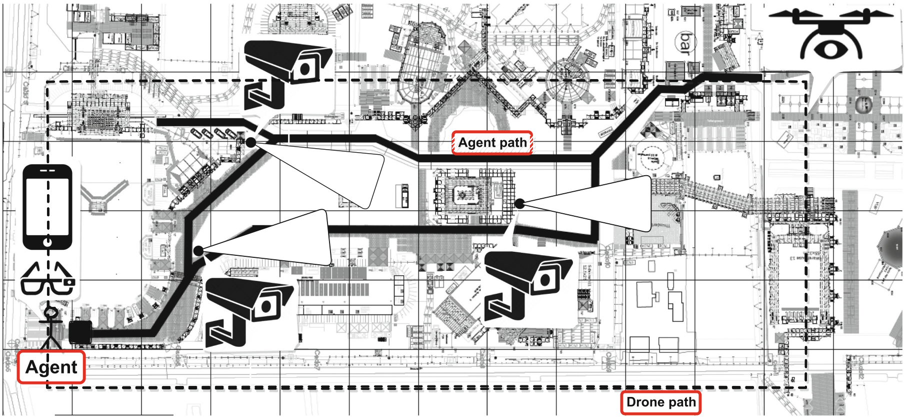

Fig. 4. Planimetry of survelliance scenario in physical event: (1) The dashed trajectory line is the path monitored by the drone, (2) the solid line has been followed by security staff, wearing smartglass and smartphone, (3) the circles, instead, are the cctv camera.

5.1 Monitor and Analytics of a Physical Environment

VISOR is a project born in collaboration with the Noord-Brabants Dutch Police and municipality stakeholders. Its primary purpose is to provide a platform for monitoring anomalous behavior in a social event (e.g., moshpit detection in a crowded concert). We led a case study focused on monitoring the social event Paaspop. The planimetry, shown in Fig. 4, represents a video surveillance scenario covered by the CCTV and the patrol done by agents and drones. In this scenario, three agents patrol the solid line in the planimetry, starting from the same point. The drone follows the dashed line, making a patrol of the event, covering the entire place. The three CCTV, represented by black circles, monitor a specific area, indicated by white cones. During the patrol, the information is sent in real-time and processed to detect illicit content (e.g., drug trading, moshpit). If the analysis recognizes an illegal event, agents on the field receive an alert notification with coordinates and images of such event.

The first study led in the VISOR project is an experience report done during the Paaspop event [4]. The next step is to experiment with the data gathered to evaluate different approaches helpful in recognizing and predicting illicit events.

5.2 Dark Web Analytics

The primary purpose of SENSEI is to build an investigation platform to help Law Enforcement Agencies (LEAs) analyze big data coming from the Dark Web. The framework provides a collection of tools for big data analysis to extract valuable insights for cybercrime investigations. The framework's features include but are not limited to trend analysis of specific temporal snapshots, network analysis of vendors, and comparison of trends between different countries to evaluate the movement of illicit goods across the world. Moreover, the platform allows narrowing the field, acting on a specific time range to provide more specific information.

6 Hybrid Analytic Services

The last layer of the architecture, namely Merging Process is the core of our research. Unlike layer 3 (Monitor and Analytic), the Hybrid Analytic layer is in charge of merging the two analyses to provide more fine-grained and detailed insights useful for an investigation.

The idea is to set up a case study with the Dutch Police to address the challenges inherited from hybrid analytics. Starting from a survey, we want to gather information regarding possible scenarios and data to build a prototype tool to help them in the monitoring process. For example, considering our study and analytics collected from the dark web marketplace accomplished by the SENSEI framework, a scenario could detect illegal trades in a Dutch Social event using the drug trend extracted from the dark web analysis as a baseline of a machine learner training process. The result of this experimentation allows us to estimate the effectiveness of our approach. For example, suppose the dark marketplace analytics show massive cocaine trades in the Netherlands, and the models used by the police show the same result. In that case, our web analytics is a good predictor for monitoring in the field. Hence, our analysis can be integrated into the real-time monitoring process to reinforce the analysis, lowering the effort of their models.

7 Limitation and Threat to Validity

This section outlines the major limitation we perceive in our work for each tool and case study.

The main limitation of KGen is the inability to anonymize unstructured data. It means that most data extracted by IoT devices, like video and images, can not be anonymized. Hence, following the GDPR principle, we need to integrate KGen with other anonymization tools if we are dealing with sensible data.

VISOR presents a threat in the monitoring process of the Paaspop event. Indeed, no specific illicit event happened during our gathering process. Hence, we could not have tested the notification alert of an illicit event on the field. A way to overcome this limitation is to lead a case study in a controlled environment where we set all the scenarios in advance.

Moreover, the crawler presents some limitations. For example, if a website is closed, obscured, or accessible through new captcha generation, the crawler can not download the website.

8 Discussion and Conclusion

In this vision paper, we introduce the architecture to facilitate the monitoring and analysis of CPS. Our architecture lays the foundation for further research using physical space protection models and social, deep, and dark web analysis. We have illustrated a preliminary study of the projects involved in this research work, passing from the KGen anonymization prototype tool to Visor and SENSEI projects.

The proposed architecture eases the monitoring and detection process in complex, critical, and dynamic scenarios. To validate our architecture, we set up a case study, namely VISOR, to monitor urban social living using hybrid analytics. We set up an experience report in the Paaspop event to validate the best IoT devices. We use SENSEI framework analytics to gather insights from the dark web and integrate them into the detection process. Then, we combine the KGen anonymization tool with the data extracted from web pages and IoT devices. Machine models could improve their efficiency in monitoring and detecting illicit goods by using information extracted from the dark web.

We conclude that the CPS might benefit from a hybrid analytic approach to improve the security level in a cyber-physical scenario.

8.1 Future Work

The following work of this project is to implement a case study with the Dutch police to evaluate the hybrid analytic approach in a different real scenario. The scope of this study is to find the best techniques for cyber-threat intelligence analysis to provide the best accuracy in a hybrid analytics context.

References

1. Casale, G., Li, C.: Enhancing big data application design with the DICE framework. In: Mann, Z.Á., Stolz, V. (eds.) ESOCC 2017. CCIS, vol. 824, pp. 164–168. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-79090-9_13
2. Cervantes, H., Kazman, R.: Designing Software Architectures: A Practical Approach. Addison-Wesley Professional, Boston (2016)
3. Da Silva, T.L.C., de Macêdo, J.A., Casanova, M.A.: Discovering frequent mobility patterns on moving object data. In: Proceedings of the Third ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, pp. 60–67 (2014)
4. De Pascale, D., Cascavilla, G., Sangiovanni, M., Tamburri, D.A., van den Heuvel, W.J.: Internet-of-things architectures for secure cyber-physical spaces: the visor experience report. arXiv preprint arXiv:2204.01531 (2022)
5. De Pascale, D., Cascavilla, G., Tamburri, D.A., Van Den Heuvel, W.J.: Sensei: scraper for enhanced analysis to evaluate illicit trends. SSRN 3976047 (2022)
6. Du, B., Liu, C., Zhou, W., Hou, Z., Xiong, H.: Detecting pickpocket suspects from large-scale public transit records. IEEE Trans. Knowl. Data Eng. **31**(3), 465–478 (2018)

7. El Emam, K., et al.: A globally optimal k-anonymity method for the de-identification of health data. J. Am. Med. Inform. Assoc. **16**(5), 670–682 (2009)
8. Garroppo, R.G., Niccolini, S.: Anomaly detection mechanisms to find social events using cellular traffic data. Comput. Commun. **116**, 240–252 (2018)
9. Goldberg, D.E., Holland, J.H.: Genetic algorithms and machine learning. Mach. Learn. **3**(2), 95–99 (1988)
10. Hayes, D.R., Cappa, F., Cardon, J.: A framework for more effective dark web marketplace investigations. Information **9**(8), 186 (2018)
11. Hogan, A., et al.: Knowledge graphs (2020)
12. Lee, E.A.: Cyber physical systems: design challenges. In: 2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC), pp. 363–369. IEEE (2008)
13. Lee, E.A.: CPS foundations. In: Design Automation Conference, pp. 737–742. IEEE (2010)
14. Lee, E.A.: The past, present and future of cyber-physical systems: a focus on models. Sensors **15**(3), 4837–4869 (2015)
15. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM (2005)
16. Nagarajan, S.M., Deverajan, G.G., Bashir, A.K., Mahapatra, R.P., Al-Numay, M.S.: IADF-CPS: intelligent anomaly detection framework towards cyber physical systems. Comput. Commun. (2022)
17. Perrone, G., Vecchio, M., Pecori, R., Giaffreda, R., et al.: The day after Mirai: a survey on MQTT security solutions after the largest cyber-attack carried out through an army of IoT devices. In: IoTBDS, pp. 246–253 (2017)
18. Samarati, P.: Protecting respondents identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
19. Sweeney, L.: Guaranteeing anonymity when sharing medical data, the Datafly system. In: Proceedings of the AMIA Annual Fall Symposium, p. 51. American Medical Informatics Association (1997)
20. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Internat. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**(05), 571–588 (2002)
21. Voigt, P., von dem Bussche, A.: The EU General Data Protection Regulation (GDPR). A Practical Guide, 1st edn. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57959-7
22. Wikipedia: The Hidden Wiki (2022). https://en.wikipedia.org/wiki/The_Hidden_Wiki
23. Yuan, Y., Fang, J., Wang, Q.: Online anomaly detection in crowd scenes via structure analysis. IEEE Trans. Cybern. **45**(3), 548–561 (2014)