TRUSTED COMPUTING IN DATA SCIENCE: VIABLE COUNTERMEASURE IN RISK MANAGEMENT PLAN

Uchechukwu Emejeamara[1], Udochukwu Nwoduh[2] and Andrew Madu[2]

[1]IEEE Computer Society, Connecticut Section, USA
[2]Department of Computer Science, Federal Polytechnic Nekede, Nigeria.

ABSTRACT

The need for secure data systems has prompted, the constant reinforcement of security systems in the attempt to prevent and mitigate risks associated with information security. The purpose of this paper is to examine the effectiveness of trusted computing in data science as a countermeasure in risk management planning. In the information age, it is evident that companies cannot ignore the impact of data, specifically big data, in the decision making processes. It promotes not only the proactive capacity to prevent unwarranted situations while exploiting opportunities but also the keeping up of the pace of market competition. However, since the overreliance on data exposes the company, trusted computing components are necessary to guarantee that data acquired, stored, and processed remains secure from internal and external malice. Numerous measures can be adopted to counter the risks associated with data exploitation and exposure due to data science practices. Nonetheless, trusted computing is a reasonable point to begin with, in the aim to protect provenance systems and big data systems through the establishment of a 'chain of trust' among the various computing components and platforms. The research reveals that trusted computing is most effective when combined with other hardware-based security solutions since attack vectors can follow diverse paths. The results demonstrate the potential that the technology provides for application in risk management.

KEYWORDS

Trusted Computing, Security, Data, Data Science, Provenance, Risk Management, Big Data, Trusted Platform Module, Platform Computation Register

1. INTRODUCTION

The constantly growing need for information has prompted data scientists to create more advanced models that can improve the gathering and analysis of big data. However, as stated by FOX [10], most technological developments come along with numerous risks that impact significantly on the information systems of companies. As a result, not only may the companies fail to achieve the intended results but also risk losing significant information including its big data. Faced with these increasing threats, cyber security teams and data scientists have sought refuge in trusted computing that creates a 'chain of trust' among the various computing platforms. The trusted computing block is enforced with numerous chained signatures and encryptions that attempt to prevent the successful malice that would appear to be successful in their absence [5]. Since data is such a critical resource, companies lack the freedom to expose it recklessly to agents of malice due to the costs that would be incurred. These costs can come in the form of customer loyalty, lawsuits, loss of competitive edge, and loss of revenue among

David C. Wyld et al. (Eds): SPTM, IPPR, AIS, CSIT, DaMi, AMLA - 2020
pp. 17-26, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.100602

others. The costs are adequate to hamper the growth of the company. This paper proposes a feasible risk management plan that utilizes trusted computing technologies to protect the company information during the utility of data through various data science approaches. This paper also reveals pertinent features of TC that makes it a viable countermeasure for inclusion in Risk Management Plans for organizations. It helps in the generation of cryptographic keys that prevent the modification of the software or the software data where it is applied. TC can be used as a component of other recognized security countermeasures, including authentication, IDS, firewalls, access controls, and VPN through software modifications and enhancements to the hardware.

2. TECHNICAL SURVEY

2.1. Trusted Provenance

Data integrity is a primary goal for information security systems. To achieve this objective, one of the primary approaches is to guarantee the reliability of provenance systems. Provenance systems are centered on the concept that the original data used in the formation of certain systems should be safeguarded for auditing and reference. When faced with security challenges, system auditors can always retreat to the original framework that created a system with code and infrastructure to understand the actual source of the problem. This concept brings to light the necessity of trusted computing in provenance systems. Since provenance systems have such value when addressing security challenges, Lyle and Martin [6] explain that failure to have 'trusted pervasive hardware infrastructure' would only lead to increased susceptibility of the information systems to attackers. Lyle and Martin [6] adds that malicious agents, target the provenance system since they can best inform them of the details of the specific infrastructural system. Hence, maintaining secure provenance systems is critical towards the protection of the data systems within any organization.

Companies are over-relying on external data for their information. This trend is evident from the unmatched utility of big data and analytics tools in the management's decision-making engagements. However, cyber threats evolve with evolving technologies and business needs. In this way, cyber-attacks target company information either in its storage form, which Bao, Chen, and Obaidat [9] refer to as 'data at rest', data being processed, or data at transit. While these threats are evident, data scientists and security teams have the task to ensure that their data is safe from attacks. One of the simplest ways for attackers to successfully infiltrate information systems, according to Hu et al. [4], is to gain access to provenance information. This information, as explained by Hu et al. [4], gives hackers the root resources about the target system paralleling their knowledge of their target system with that of the system owners. Problematically, the stakes that hackers with provenance information will not just steal information but overrun a system or obliterate its resources, are rather high [6]. In this respect, the risk factors for data loss or obliteration when provenance information is stolen are significantly critical that the best or possibly the only solution would be to prevent successful infiltration.

2.2. Security Risks in Data Science

Security in data science can take various forms. The basic security approaches include security on the software resources, hardware or infrastructural resources, data protection as a security concept independent from the former two, and data anonymization. Gordo [2] explains that security in data science would also include information warfare due to the increasingly high rates of availability of strategically deceptive data meant to promote misinformed moves by various corporations and institutions. It has taken new approaches that include significant turns of events

since the birth of psychologically deceptive approaches such as social engineering. In this context, information warfare has been one of the most significant challenges in data science leading data scientists, to get misinformed patterns about the data available. Moreover, the correctness of data can also render it obsolete when its usability is compromised by hackers. Hence, having the right information must always be paired with reliably secure systems to make viable use of any type of data.

The challenges and solutions to data security revolve around the physical infrastructure, software infrastructure, the data, and people within and around an organization. Physical infrastructure security protects data from technical failures and physical and virtual malice agents. While data scientists have the task to remain vigilant, infrastructure resources are managed by the information security teams. The security of the infrastructure is a primary factor to guarantee the security of the data within its systems. Software infrastructure focuses on the vulnerabilities that could be exploited by hackers. It supplements hardware infrastructure by creating frameworks using which virtual attacks can be detected, prevented, and repelled. In other cases, the software can take the usual task of informing the computer security incident response team to include human decisions in the process. Data security is another issue outside the ordinary context of information security fostered by infrastructural components. However, data scientists with security skills can identify anomalies in data that alert them of falsifications [3]. Such trends can protect data scientists from the risk of using false information that should also alert them of a security challenge in their information systems. Moreover, encryption promotes data anonymity regardless of the security standards applied by an organization. Finally, the people in any organization determine significantly the security of data within the data systems. Essentially, the organizational culture based on its dedication to follow the set policies explains the accountability to protect data resources and prevent insider threats. Nonetheless, all these components of information security must be effective to establish the chain of trust among them and achieve the necessary security standards.

2.3. Specific Data Science Challenges

Privacy and security remain the topmost challenges in data science. Cai and Zhu [7] explains that the bigger the data a company holds, the more susceptible it is to attackers. This susceptibility raises the question of the privacy of information stored within the information systems. Notably, while confidential information often revolves around customer information, data science confidentiality may take a slightly different perspective and focus on the quality of information achieved from interpreting the data. Most decisions are made based on the information available which they would like to conceal. Notably, this information also forms part of their competitive advantage, which, if adequately sustainable, would lead other companies to target competitors with better selling propositions for the market. Regardless of the type of data used, privacy must be guaranteed to prevent lawsuits which can lead to significant losses in revenue and tainting of the brand image.

The increasing significance of data in corporate decision-making processes has made the use of data compulsory. According to Ingrams [1], big data utility is not only fostered by the need for competitive advantage but also the opportunity costs of ignoring its utility. In explanation, improved security has been achieved through the analysis of big data that provides proactive insights into the security issues that could be affecting other companies and could happen at a specific company. Ignoring such aspects keeps the company away from information leading them to make uninformed security decisions that could risk the data of the company. However, while this opportunity cost is irresponsible for any security teams to incur, data science practices such as data collection, evaluation of its validity, data analysis, and decision making are, by itself, a complex and costly engagement. The resources required are overly expensive, and the

information may not necessarily be useful as to cover the costs if the process is not well-informed. For such reasons, Data Scientists employed must be well trained and experienced professionals, to keep cost within considerable limits.

By contrast, increasing challenges in the area of data science have been caused by the nature of data used, the tools, or the skills. Essentially, before data scientists can assume that the data

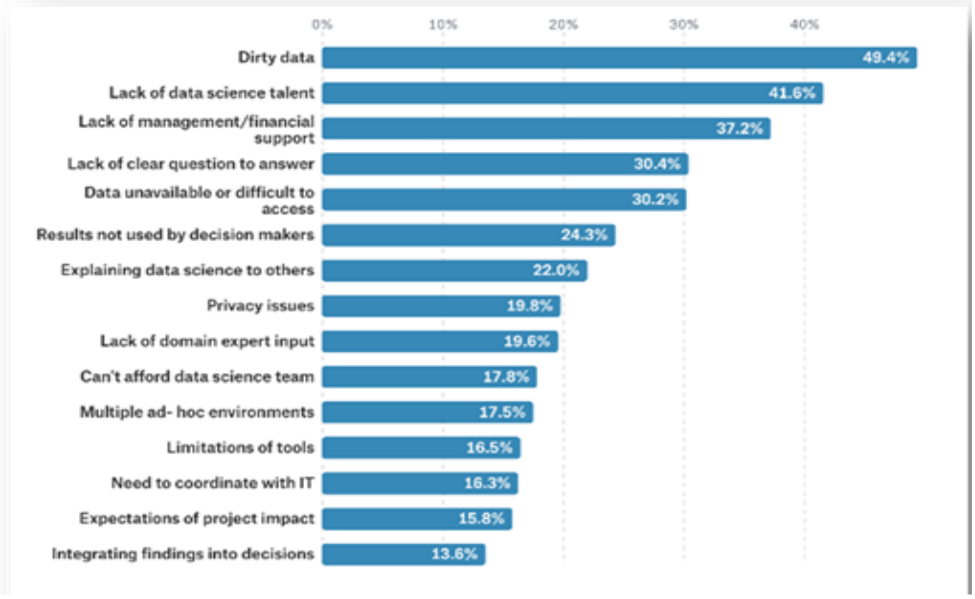

Fig. 1 Significance of Challenges in Data Science [11]

will give the right results, they need to ensure the data itself is the right one. Fig 1, highlights the danger posed by disorganized, inaccurate, and incoherent data. Bad data not only costs the company the resources consumed but also the misinformed decisions that will be made. Moreover, the company is more likely to fall behind competition due to the failure to draw correct business insights [7]. Bad analytics, on the other hand, results in the misinterpretation of data. Bad analytics can take the form of using the wrong tools or using the right tools without the foreknowledge of the best data science practices. The consequence is the derivation of wrong patterns or their misinterpretation. Similarly, the risk, in this case, is still the lagging behind the competition.

3. METHODS

The Risk Management Plan (RMP) will be implemented similarly to any other security approach. As trusted computing pertains more to the individual components, scholarly works will be used to assess the effectiveness of the approaches.

3.1. Risk Identification

The first process in risk planning is the identification of the risk aspect. The risk factor associated with data science with regards to computing has been identified by numerous scholars to be highly significant [7]. The infrastructural, data, and people-related security aspects, for instance, would have a considerable impact on the risk positioning of the company when considered. If ignored, data science practices could be rendered not only obsolete but also expensive. Additionally, issues with big data pose a significant risk since they can foster bad decisions when

analyzed incorrectly or when the data is not right. Such instances can pose a dangerous cost risk should the misinformed decisions affect crucial operations of the company. Moreover, according to Bilić [8], data science is at such risks from the falsification of information by intentional agents to mislead their competition into arriving at the wrong conclusions. Hence, the identification of these risks is the first step in creating sufficient countermeasures in trusted computing with regards to data science.

3.2. Risk Assessment

The assessment of a risk factor helps teams to substantiate its threat and determine the weight of the associated risk. Three common risk factors are assessed that include threats, malware, and anomalies. In the context of data science, some aspects of assessing anomalies use trusted computing principles that seek to find consistency between data sets such as information stored in the PCRs [6]. By contrast, anomalies can arise from the data itself due to falsifications or decreased security. When the software components identify such anomalies, trusted computing is compromised since the chain of trust no longer exists. Similarly, malware risks can be assessed to confirm the risk factors they pose. Notably, while machine learning has been improving the utility of big data, cyber threats have increased as they attempt to use the same principles such as dynamic code analysis to compromise the effectiveness of the data science machine learning algorithms. The risk, in this case is so significant that it can threaten the company's sustainability as it lags behind competition due to ill-informed decisions. Trusted computing also aims at identifying threats, such as logs of failed and successful authentications and be able to explain the incidents. The assessment of such incidents can explain how targeted the company is and the risk posed by any threats.

3.3. Risk Control

The final process in the risk management plan is the control of the identified risks. The risks are diverse and numerous approaches must be applied to fully address the challenges. Provenance systems can expose the company considerably. The chain of trust between the computing systems in the data science computing environment would require subtle implementations, such as the use of endorsement keys stored in the TPM, memory curtaining that isolates critical memory components from the less sensitive ones, sealed storage that binds information to its infrastructural components, a remote attestation that connects with legitimate parties remotely, and trusted third party that is based on the security measures of a remote computing base [5]. However, other than having these security solutions in place, ethical data scientists with adequate skills and resources are also at the bottom line of controlling the risks, otherwise, the costs would be dire. Overall, the risk control process is a collaborative effort between all stakeholders to guarantee the achievement of the desired outcomes. Fig. 2 depicts the process flow in RMP.

Fig. 2 RMP Flow

3.4. Trusted Computing

Trusted Computing (TC) is a hardware-based method that enforces the integrity of software or platforms. It protects a system from software-level attacks [19]. It operates on two principle ideas, namely remote attestation and sealed memory [12]. Remote attestation is concerned with the identification and declaration of the software that runs on a remote computer while sealed memory authorizes specific software stacks to access stored secrets. Maene et al. [12] explained that TC provides both local and remote attestation. Through sealing, TC wraps data in a manner that prevents unwrapping without the decrypting key. From these two foundational ideas, TC enables the development of security protocols, including authentication, encryption, and the management of digital rights. The simple framework for achieving trust using TC is as depicted in Fig. 3.

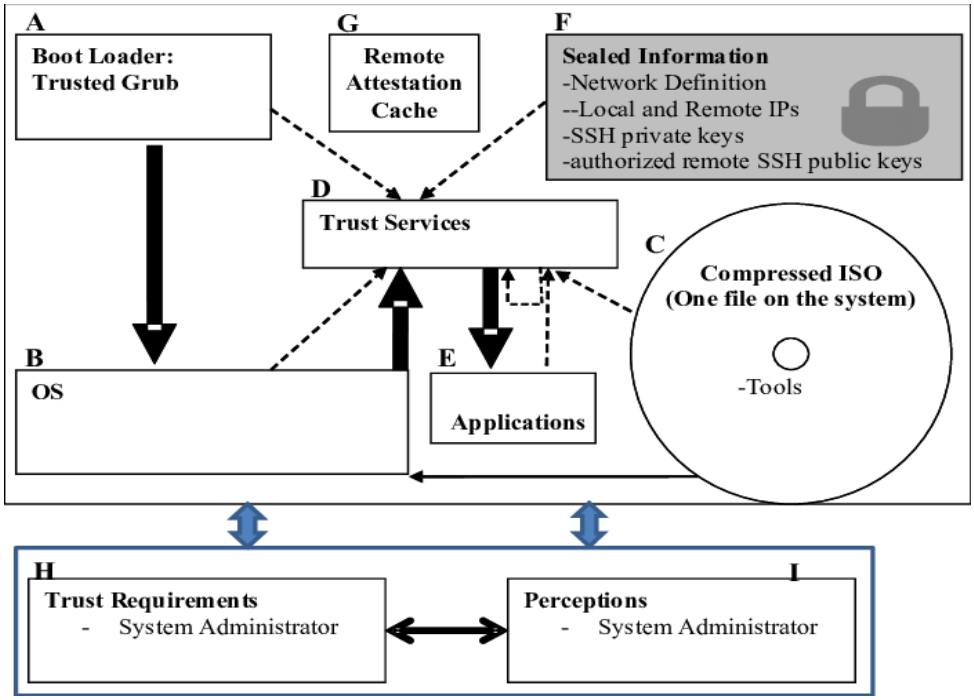

Fig. 3 Trust development framework based on TC

With the proliferation of mobile devices that rely heavily on digital data, TC provides an effective scheme for securing the data in applications installed on various smartphones. Using a Trusted Execution Environment (TEE), Fan et al. [13] applied the TC framework to minimize data losses from mobile devices. The TEE enabled the authors to divide sensitive files into file slices, encrypt the individual file slices/parts, and map individual file parts to whole files in cloud computing [13]. They further reported that the scheme thus developed was highly efficient and secure. According to Maene et al. [12], the trusted computing architecture guarantees users the protection against software-level attackers, thereby acting as a countermeasure against the exploitation of software vulnerabilities. The most practical use of TC is found when it is integrated with other security protocols.

4. DISCUSSION

Risk management teams have no option but to include IT security comprehensively into their plans. The risk factor with data science is rather critical that the failure to promote trusted

computing among data-handling components can render the data science operations obsolete. In explanation, trustworthy behaviour is enforced through the use of standards that ensure a 'chain of trust' between the hardware and software components that have security roles to play. The Trusted Platform Module (TPM), for instance, stores isolated encryption keys that are used by the users to authenticate various processes and platform computation registers which hold integrity registers [6]. The TPM, essentially, guarantees that records are maintained regarding the various computing processes that are necessary for auditing to confirm that the normal processes have not been compromised. However, trusted computing uses backup values about processes to validate the processes such as that any inconsistencies between the data would require an explanation. Hence, trusted computing can ensure that anomalies are addressed when components do not function as expected.

4.1. Efficacy of Trusted Computing

Trusted Computing, in the context of data science, can be used as a countermeasure in Risk Management. Trusted Computing (TC) is effective in securing data, making it an instrumental technology in data science. In mobile technologies and cloud computing, Trusted Computing enables the minimization of data losses through file slicing and slice encryption and allows for the mapping of file slices to cloud computing environments [13]. The technology is also effective in securing applications that reside on mobile devices from unauthorized access. Trusted Computing also provides an effective solution for promoting the privacy and security of data stored in cloud infrastructure through the use of technologies such as Intel Software Guard Extensions (SGX) [15]. Thus, organizations that provide Software-as-a-Service (SaaS) cloud solutions can effectively implement TC-based privacy and security policies. When used along with hardware-based solutions such as TEE and other security systems, Trusted Computing becomes an effective security solution. Its effectiveness has been proved for protection against unauthorized access, man-in-the-middle attacks, and password guessing attacks [14].

The practicable solutions that TC offers, based on its integration with other security technologies, makes it useful in RMP countermeasures in the contexts of data science, cloud computing, Internet of Things (IoT), and blockchain applications. In IoT solutions, TC provides a platform for addressing the challenge of users' loss of control over data, as evidenced by the HyperNet framework [16]. In blockchain applications, Trusted Computing can be used to promote user control over their data through the Proof-of-Credibility option that it offers [17]. Furthermore, Trusted Computing principles allow for the development of various trust enhancement frameworks that are a necessity in blockchain applications [18]. Thus, Trusted Computing finds applications in technologies that inform the future of the data science landscape.

While it protects specific applications from attacks, Trusted Computing is not an effective countermeasure against attacks that target other sections of a computing system. The attacker model it adopts assumes the ability of the cybercriminal to tamper with the OS, launch malicious software, sniff the network, perform MITM attacks, modify traffic, break network-based cryptographic primitives, and launch denial-of-service (DoS) attacks [12]. Consequently, Trusted Computing is not an effective countermeasure against the threats facing an information system unless it is used along with other hardware-based security solutions that protect the OS, network, and traffic.

4.2. Trusted Computing Viability in Risk Management

From literary research, trusted computing is a significant aspect of IT security. It promotes the availability of mutually secure systems due to the increased security in each specific component. One of the most significant benefits of trusted computing is the capacity to protect data systems

through the presence of built-in processes. These processes revolve around encryptions and hashing preventing successful attacks. Notably, data encryption and hashing is an added layer of security that ensures that information remains confidential even in the event of successful infiltration. This way, the data is protected from compromise.

Being an approach for establishing trust between components, trusted computing also guarantees that other systems are secure. In explanation, trust can only be established between the components when each of them is secure from malice. In this way, the approach seems to play the same role in promoting the security of all infrastructural components in a distributed computing environment. This aspect is a limitation since it seems to give trusted computing a broad connotation as an umbrella term, which it is not. Moreover, trusted computing acts as a facilitator for creating a safer environment even when resources are decentralized from the main systems.

4.3. Challenges

Trusted computing works at a rather low level of computing. Although it is one of the best ways of protecting data systems, trusted computing on its own may not be able to guarantee that security will be achieved. For instance, trusted computing can perform the best task to promote the security of provenance systems. However, attack vectors do not necessarily follow that path. Other approaches such as social engineering have been used extensively that organizations can only focus on trusted computing alone as the viable countermeasure to mitigate the risks associated with threats in data science.

4.4. Verdict

The presence of a plethora of vectors posits that companies cannot focus on one of them. However, the significance of any of them should not be overlooked since they add up to the overall security of the data systems within any organization. Trusted computing offers significant advantages for its users concerning data science. With the basic security challenges having been solved, the additional approaches can guarantee a comprehensive and adequate addressing the security concerns that revolve around the use of data within and external to a company. For this reason, trusted computing is a viable countermeasure in risk management planning but can only be applied as a complementary strategy reinforcing other security mechanisms.

5. CONCLUSION

Trusted computing is one of the primary approaches that use security methods outside the common software-related mechanisms. The understanding of the significance of trust in security begins with the foreknowledge of the risk impacts of not improving the effectiveness of the various security and protection of information systems. Trust is established when all components can guarantee the security that the data being stored, processed, or in transit in these systems is secure from both internal and external threats. With this security guaranteed, the remaining risk mitigation measure would be the assurance that the data scientists perform the right tasks correctly and with the right tools. In this way, the countermeasures for the risk management plan would be viable.

REFERENCES

[1] A. Ingrams, "Public Values in the Age of Big Data: A Public Information Perspective", *Policy & Internet*, vol. 11, no. 2, pp. 128-148, 2018. Available: 10.1002/poi3.193.

[2] B. Gordo, "“Big Data” in the Information Age", *City & Community*, vol. 16, no. 1, pp. 16-19, 2017. Available: 10.1111/cico.12219.

[3] B. Tellenbach, M. Rennhard and R. Schweizer, "Security of Data Science and Data Science for Security", *Applied Data Science*, pp. 265-288, 2019. Available: 10.1007/978-3-030-11821-1_15.

[4] D. Hu, D. Feng, Y. Xie, G. Xu, X. Gu and D. Long, "Efficient Provenance Management via Clustering and Hybrid Storage in Big Data Environments", *IEEE Transactions on Big Data*, pp. 1-1, 2019. Available: 10.1109/tbdata.2019.2907116.

[5] E. Padma, "Trusted Attestation System for Cloud Computing Environment Using Trusted Platform Module", *Internet of Things and Cloud Computing*, vol. 5, no. 3, p. 38, 2017. Available: 10.11648/j.iotcc.20170503.11.

[6] J. Lyle and A. Martin, *Trusted Computing and Provenance: Better Together*. Oxford: Oxford University Computing Laboratory, 2010.

[7] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era", *Data Science Journal*, vol. 14, no. 0, p. 2, 2015. Available: 10.5334/dsj-2015-002.

[8] P. Bilić, "Search algorithms, hidden labour and information control", *Big Data & Society*, vol. 3, no. 1, p. 205395171665215, 2016. Available: 10.1177/2053951716652159.

[9] R. Bao, Z. Chen and M. Obaidat, "Challenges and techniques in Big data security and privacy: A review", *Security and Privacy*, vol. 1, no. 4, p. e13, 2018. Available: 10.1002/spy2.13.

[10] S. FOX, "Policing - The technological revolution: Opportunities & challenges!", *Technology in Society*, vol. 56, pp. 69-78, 2019. Available: 10.1016/j.techsoc.2018.09.006.

[11] N. Thabet and T. Soomro, "Big Data Challenges", *Journal on Computer Engineering and Information Technology*, vol. 4, no. 3, 2015. Available: 10.4172/2324-9307.1000133.

[12] P. Maene, J. Götzfried, R. d. Clercq, T. Müller, F. Freiling and I. Verbauwhede, "Hardware-based trusted computing architectures for isolation and attestation," *IEEE Transactions on Computers*, vol. 67, no. 3, pp. 361-374, 2018.

[13] Y. Fan, S. Liu, G. Tan, X. Lin, G. Zhao and J. Bai, "One secure access scheme based on trusted execution environment," in *12th IEEE International Conference On Big Data Science And Engineering*, New York, NY, 2018.

[14] E. F. Cahyadi, Y.-C. Chou, C.-Y. Yang and M.-S. Hwang, "An improved mutual authentication scheme with smart cards and password under trusted computing," in *2018 the 2nd annual International Conference on Cloud Technology and Communication Engineering*, Nanjing, China, 2018.

[15] A. T. Gjerdrum, R. Pettersen, H. D. Johansen and D. Johansen, "Performance principles for trusted computing with Intel SGX," in *International Conference on Cloud Computing and Services Science*, Porto, Portugal, 2017.

[16] H. Yin, D. Guo, K. Wang, Z. Jiang, Y. Lyu and J. Xing, "Hyperconnected network: A decentralized trusted computing and networking paradigm," *IEEE Network*, vol. 32, no. 1, pp. 112-117, 2018.

[17] D. Fu and L. Fang, "Blockchain-based trusted computing in social network," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, China, 2016.

[18] Y. Wu, Y. Qiao, Y. Ye and B. Lee, "Towards improved trust in threat intelligence sharing using blockchain and trusted computing," in *2019 Sixth International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, Granada, Spain, 2019.

[19] M. Alhaidary, S. M. M. Rahman, M. Zakariah, M. S. Hossain and A. Alamri, "Vulnerability analysis for the authentication protocols in trusted computing platforms and a proposed enhancement of the OffPAD protocol," *IEEE Access*, vol. 6, pp. 6071-6081, 2018.