Enhancing Big Data Security with Collaborative Intrusion Detection

Zhiyuan Tan, University of Twente
Upasana T. Nagar, Xiangjian He, and Priyadarsi Nanda, University of Technology Sydney
Ren Ping Liu, Commonwealth Scientific and Industrial Research Organisation (CSIRO)
Song Wang, La Trobe University
Jiankun Hu, University of New South Wales

A collaborative intrusion detection system (CIDS) plays an important role in providing comprehensive security for data residing on cloud networks, from attack prevention to attack detection.

Cloud computing delivers a flexible network computing model that allows organizations to adjust their IT capabilities on the fly with minimal investment in IT infrastructure and maintenance. Because an organization need only pay for the services it uses, it can focus on its core business instead of handling technical issues.

In the cloud computing context, network-accessible resources are defined as services. These services are typically delivered via one of three cloud computing service models:

- Infrastructure as a service (IaaS) offers storage, computation, and network capabilities to service subscribers through virtual machines (VMs).
- Platform as a service (PaaS) provides an environment for software application development and hosts a client's applications in a PaaS provider's computing infrastructure.
- Software as a service (SaaS) delivers on-demand software services via a computer network, eliminating the cost of purchasing and maintaining software.

2325-6095/14/$31.00

These technical and business advantages, however, don't come without cost. The security vulnerabilities inherited from the underlying technologies (that is, virtualization, IP, APIs, and datacenter) prevent organizations from adopting the cloud in many critical business applications.[1] Generally speaking, cloud computing is a service-oriented architecture (SOA). Earlier work gives a comprehensive dependability and security taxonomy framework revealing the complex security cause-implication relations in this architecture.[2] We summarize cloud computing vulnerabilities by underlying technology in the sidebar.

These vulnerabilities leave loopholes, allowing cyberintruders to exploit cloud computing services and threatening the security and privacy of big data. Various security schemes, such as encryption, authentication, access control, firewalls, intrusion detection system (IDSs), and data leak prevention systems (DLPSs), address these security issues. In this complex computing environment, however, no single scheme fits all cases. These schemes should thus be integrated and cooperate to provide a comprehensive line of defense.

Intrusion Detection for Securing Cloud Computing

IDSs aim to provide a layer of defense against malicious uses of computing systems by sensing attacks and alerting users. Because it's impossible to prevent all cyberattacks, IDSs have become essential to securing cloud computing environments.

IDSs are commonly categorized by the type of data source involved in detection. Host-based IDSs (HIDSs) detect malicious events on host machines. They handle insider attacks (which attempt to gain unauthorized privileges) and user-to-root attacks (which attempt to gain root privileges to VMs or the host). Network-based IDSs (NIDSs) monitor and flag traffic carrying malicious contents or presenting malicious patterns. This type of IDS can detect direct and indirect flooding attacks, port-scanning attacks, and so on.

Although to some extent, DLPSs can be considered a type of IDS, they're more tailored to data security. However, it's difficult to completely guarantee data security using DLPSs alone. Attackers who gain control of the host machines can modify the DLPS settings, thereby completely disclosing data to those attackers. Moreover, even though firewalls can block unwanted network traffic packets according to a predefined rule set, they can't detect sophisticated intrusive attempts such as flooding and insider attacks. IDSs, DLPSs, and firewalls are therefore not interchangeable security schemes but collaborative ones.

Conventional IDSs

Conventional IDSs are mostly standalone systems residing on computer networks or host machines. They can be categorized as misuse-based or anomaly-based IDSs, depending on the detection mechanism applied.

Misuse-based IDSs enjoy high detection accuracy but are vulnerable to all zero-day intrusions.[3] This is due to the underlying detection mechanism that checks for a match with existing attack signatures. Obviously, an IDS can't generate signatures for an unknown attack. Anomaly-based IDSs show promise for detecting zero-day intrusions,[4,5] but are prone to high false positives.

Current enterprise networks (such as cloud computing environments) typically have multiple entry points. This topology is intended to enhance a network's accessibility and availability, but it leaves security vulnerabilities that sophisticated attackers can exploit using advanced techniques, such as cooperative intrusions.

Unlike traditional attack mechanisms, cooperative attack mechanisms are launched simultaneously by slave machines within a botnet. Attackers organize instances of this attack type to penetrate an enterprise network through all its entry points. By evenly distributing the attack traffic volume to the different entry points, these cooperative intrusions can evade detection of traditional standalone IDSs set in front of the entry points. This is because network traffic behavior at each entry point doesn't significantly deviate from normal behavior. After traveling through the entry points, the attack instances are directed to a single targeted service within the enterprise network.

Moreover, many of the existing intrusions can occur collaboratively and simultaneously on nodes throughout a network. Attackers can initiate automated attacks targeting all vulnerable services within a network simultaneously,[6] rather than focusing on a specific service.

VULNERABILITIES IN UNDERLYING TECHNOLOGIES

Vulnerabilities in the cloud's underlying technologies allow cyberintruders to exploit cloud computing services and threaten the security and privacy of big data.

Virtualization

Virtualization facilitates multitenancy and resource sharing (such as physical machines and networks) and enables maximum utilization of available resources. Categories include full, OS-layer, and hardware-layer virtualizations.

Virtual machines (VMs) can gain full access to a host's resources if isolation between the host and the VMs isn't properly configured and maintained. (In this case, the VMs escape to the host and seize root privileges.) In addition, a VM's security can't be guaranteed if its host is compromised. Hosts and their VMs share networks via a virtual switch, which VMs could use as a channel to capture the packets transiting over the networks or to launch Address Resolution Protocol (ARP) poisoning attacks. Finally, because a host shares computing resources with its VMs, a guest could launch a denial-of-service (DoS) attack via a VM by taking up all the host's resources.

IP Suite

The IP suite, the core component of the Internet, ensures the functioning of internetworking systems and allows access to remote computing resources.

Defects in the implementation of the TCP/IP protocol suite can lead to a variety of attacks, including IP spoofing, ARP spoofing, DNS poisoning, Routing Information Protocol (RIP) attacks, flooding, HTTP session riding, and session hijacking.

Application Programming Interfaces

APIs provide interfaces for managing cloud services, including service provisioning, orchestration, and monitoring. Areas of vulnerability include weak credentials, authorization checks, and input-data validation, which could allow an attacker to seize root privileges. Developers might introduce defects during the design and implementation of cloud APIs or introduce new security vulnerabilities when fixing bugs.

Datacenter

Datacenter technologies allow administrators to manage and store data. Data is often stored, processed, and transferred in plaintext, which can be compromised, leading to the loss of confidentiality. Attackers might also find residual data from data that's been deleted. Finally, in a datacenter, data from different users (both legitimate users and intruders) is mixed together with weak separation, providing opportunities for an intruder to access the data of the legitimate users.

Need for Collaborative Intrusion Detection

Conventional standalone IDSs are susceptible to cooperative attacks, so are unsuitable for collaborative environments (such as a cloud computing environment). To defend against this type of attack, collaborative intrusion detection systems (CIDSs) correlate suspicious evidence between different IDSs to improve the intrusion detection efficiency. Unlike conventional standalone IDSs, a CIDS shares traffic information with the IDSs located at a local network's entry points.

In practice, we can organize IDSs within a CIDS in a decentralized[7] or hierarchical[8] manner over a large network. These IDSs communicate directly with each other or with a central coordinator, according to the applied mode of organization.

In a decentralized CIDS, each IDS can generate a complete attack diagram of the network by ag-

gregating network information received from other IDSs in the CIDS. Detection of malicious attempts is undertaken locally at each IDS. In a hierarchical CIDS, a coordinator is a central point responsible for information aggregation. The central coordinator, which analyzes the aggregated information, generates a complete attack diagram of the network.

Limitations of Current Collaborative IDSs

Collaborative IDSs seem promising for detecting cooperative intrusions. However, existing system architectures aren't without criticism. In CIDSs, network data summarization is an important precursor to reliable intrusion detection.[9] However, traditionally, network information is collected and processed by IDS software built on a single network device that only deals with the traffic flowing in and out of that device. It therefore has limited traffic information. In addition, the computation of network data summarization is proportional to the amount of traffic flow that single device experiences. Such an approach has drawbacks in terms of both accuracy and efficiency.

In terms of accuracy, without knowledge of the network data from other nodes, any summarization is specific to a partial and insignificant portion of all available data over the entire network. Exchanging and combining these summarizations later, without the actual data, provides a minimal information gain.

In terms of efficiency, nodes with denser traffic require additional computation to process summarization. Because summarization is a pure overhead operation, in an ideal environment, a node will have less traffic to process when performing summarization tasks.

Security is another concern for existing CIDSs. If a CIDS is compromised, the entire cloud computing environment is in danger. Conventional IDS software, installed on a single network device, analyzes and maintains network information on the device but doesn't include security properties that ensure confidentiality, authentication, and integrity. Thus, CIDSs that are designed simply by integrating conventional IDS software without proper security enhancements are vulnerable to attacks.

Collaborative Intrusion Detection Framework

Given the defects of existing CIDSs, a new sophisticated CIDS framework could strengthen the security of cloud computing systems. However, cloud computing presents unique issues. With a large, dense network of nodes forming a cloud environment, cloud computing offers us unprecedented opportunities for making available network data from all nodes. At the same time, it requires that we perform summarization and combine the results in a distributed and parallel manner. In addition, because we're now dealing with all the network data in the entire cloud, where an unknown number of categories can exist, the summarization algorithms will need to expand their categories on demand to automatically create new clusters when they discover new types of traffic emerging.

Given the characteristics of cloud computing, we must consider several desirable properties when designing a new CIDS framework. These properties include fast detection of various attacks with minimal false positive rates, scalability with the expansion of the cloud computing system, self-adaption to changes in the cloud computing environment, and resistance to compromise.[10] Figure 1 shows the framework of our proposed CIDS, which meets these requirements.

As Figure 1 shows, HIDSs and NIDSs cooperate to perform intrusion detection at the host and network levels, and each IDS in the network is equipped with signature- and anomaly-based detectors.[11] This tactic ensures better detection accuracy in both known and unknown attacks.

There are two categories of nodes in this framework—*cooperative agent* and *central coordinator*. These nodes form a collaborative system whose security is assured through the implementation of various security mechanisms.

Cooperative Agents

Cooperative agents stand at the front lines and detect misuses on host machines or malicious behavior on networks. These agents are equipped with HIDSs or NIDSs depending on their location—agents installed on a host machine to detect suspicious events are equipped with HIDSs, whereas agents monitoring traffic on a network are equipped with NIDSs.

In our framework, the cooperative agents located on host machines are a new type of HIDS, requiring no instrumentation within VMs ~~or models~~ processes at the VM granularity level (that is, treating VMs as individual processes and modeling VM behaviors accordingly). This scheme ensures that our detection system complies with service-level agreements (SLAs) and legal restrictions, which might not allow an IaaS provider to make amendments or perform intensive monitoring and surveillance on client VMs. It also alleviates the ineffectiveness of NIDSs on encrypted traffic. The host-based cooperative agents inform a central coordinator when they detect an intrusive behavior or activity.

Cooperative agents residing at the network level conduct first-tier detection, defending against

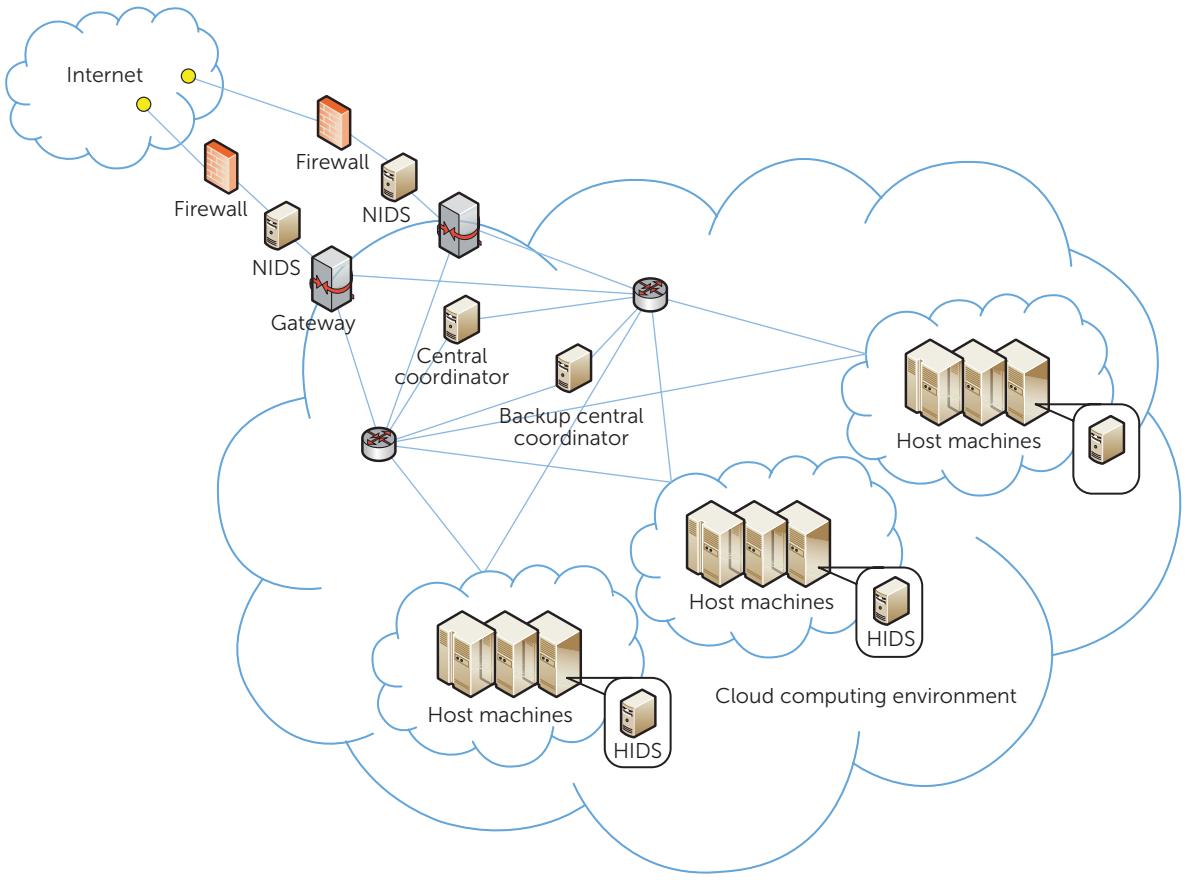

FIGURE 1. Framework of a collaborative intrusion detection system (CIDS). The figure illustrates how the different types of fellow IDSs are deployed in a cloud computing environment, and how they cooperate with each other and central coordinators in detecting intrusions. *(HIDS: host-based IDS, NIDS: network-based IDS)*

generic attacks that present abnormality within the network traffic and don't involve sophisticated cooperation. The network-based cooperative agents alert a central coordinator to any suspicious packets detected. Meanwhile, these agents summarize network traffic flowing through the network in a distributed and parallel manner. In network data summarization, the nonparametric Bayes could be a suitable machine learning approach for solving the challenges of cloud computing.[12] Network summarization is particularly important for detecting cooperative intrusions, such as distributed denial-of-service (DDoS) attacks. These summarizations are periodically sent to a central coordinator, as we discuss next.

This parallel summarization is empowered by cloud computing through the MapReduce framework.[13] The MapReduce framework provides seamless and effortless integration of our CIDS framework into a distributed and parallel architecture by treating the network-based cooperative agents as slave nodes and the central coordinator as a master node. The MapReduce framework manages all details, ranging from scheduling to information aggregation.

Central Coordinator

Finally, the network traffic aggregation is performed on the central coordinator, which generates a complete attack diagram of the entire network (that is, the cloud computing system). Based on this aggregation, the central coordinator is capable of capturing sophisticated cooperative intrusions that the individual network-based cooperative agents miss. When intrusive behaviors (including those identified by the cooperative agents and the central coordinator) are detected, the central coordinator raises an alert to a system administrator.

It's worth noting that a hybrid detector combining ~~misuse~~ and anomaly-based detection mechanisms can help reduce the time needed to detect and enhance the detection accuracy of known and unknown attacks.

Security Mechanisms

To ensure that the CIDS is resistant to compromise, we use authentication and encryption as well as an integrity check. Because the CIDS works 24/7, energy-efficient group key distribution schemes are preferable for secure key distribution and node authentication.[14,15] These schemes provide a strong, secure mechanism for updating group keys when nodes join in or leave the network or a node is being compromised. They're also resilient to collusion attacks, in which multiple nodes are compromised and coordinated for attack. Finally, a backup central coordinator runs alongside the main coordinator to prevent a single point of failure. The coordinators' roles can be exchanged depending on actual requirements and network conditions.

Future studies will explore the framework's implementation and application on different cloud computing systems. Focuses of our future studies will be casted on algorithms for distributed and parallel data summarization on cloud computing, and their implementation on the MapReduce framework, as well as new detection approaches for HIDSs.

Acknowledgments

The work described here was performed when Zhiyuan Tan was a research associate with the School of Computing and Communications at the University of Technology, Sydney.

References

1. C. Modi et al., "A Survey on Security Issues and Solutions at Different Layers of Cloud Computing," *J. Supercomputing*, vol. 63, no. 2, 2013, pp. 561–592.
2. J. Hu et al., "Seamless Integration of Dependability and Security Concepts in SOA: A Feedback Control System Based Framework and Taxonomy," *J. Network and Computer Applications*, vol. 34, no. 4, 2011, pp. 1150–1159.
3. Y. Meng, W. Li, and L.-F. Kwok, "Towards Adaptive Character Frequency-Based Exclusive Signature Matching Scheme and Its Applications in Distributed Intrusion Detection," *Computer Networks*, vol. 57, no. 17, 2013, pp. 3630–3640.
4. G. Creech and J. Hu, "A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguous and Discontiguous System Call Patterns," *IEEE Trans. Computers*, vol. 63, no. 4, 2014, pp. 807–819.
5. Z. Tan et al., "A System for Denial-of-Service Attack Detection Based on Multivariate Correlation Analysis," *IEEE Trans. Parallel and Distributed Systems*, vol. 25, no. 2, 2014, pp. 447–456.
6. S. Savage, "Internet Outbreaks: Epidemiology and Defenses," keynote address, Internet Soc. Symp. Network and Distributed System Security (NDSS 05), 2005; http://cseweb.ucsd.edu/~savage/papers/InternetOutbreak.NDSS05.pdf.
7. S. Ram, "Secure Cloud Computing Based on Mutual Intrusion Detection System," *Int'l J. Computer Application*, vol. 2, no. 1, 2012, pp. 57–67.
8. S.N. Dhage and B. Meshram, "Intrusion Detection System in Cloud Computing Environment," *Int'l J. Cloud Computing*, vol. 1, no. 2, 2012, pp. 261–282.
9. D. Hoplaros, Z. Tari, and I. Khalil, "Data Summarization for Network Traffic Monitoring," *J. Network and Computer Applications*, vol. 37, Jan. 2014, pp. 194–205.
10. A. Patel et al., "An Intrusion Detection and Prevention System in Cloud Computing: A Systematic Review," *J. Network and Computer Applications*, vol. 36, no. 1, 2013, pp. 25–41.
11. A.K. Jones and R.S. Sielken, *Computer System Intrusion Detection: A Survey*, tech. report, Dept. of Computer Science, Univ. of Virginia, 2000; http://atlas.cs.virginia.edu/~jones/IDS-research/Documents/jones-sielken-survey-v11.pdf.
12. N. L. Hjort et al., eds. *Bayesian Nonparametrics*, vol. 28, Cambridge Univ., 2010.
13. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Comm. ACM*, vol. 51, no. 1, 2008, pp. 107–113.
14. B. Tian et al., "A Mutual-Healing Key Distribution Scheme in Wireless Sensor Networks," *J. Network and Computer Applications*, vol. 34, no. 1, 2011, pp. 80–88.
15. B. Tian et al., "Self-Healing Key Distribution Schemes for Wireless Networks: A Survey," *Computer J.*, vol. 54, no. 4, 2011, pp. 549–569.

ZHIYUAN TAN *is a postdoctoral research fellow in the Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, Enschede, Netherlands. His research interests include network security, pattern recognition, machine learning, and distributed systems. Tan received a PhD from the University of Technology Sydney (UTS), Australia. He's an IEEE member. Contact him at z.tan@utwente.nl.*

UPASANA T. NAGAR *is a PhD student in the School of Computing and Communications at the University of Technology, Sydney (UTS), Australia,*

and a student member of the Research Centre for Innovation in IT Services and Applications (iNEXT) at UTS. Her research interests include network security, pattern recognition, and cloud computing. Nagar received a bachelor's degree in electronics from the National Institute of Technology, Surat. Contact her at Upasana.T.Nagar@student.uts.edu.au.

XIANGJIAN HE *is a professor of computer science in the School of Computing and Communications at the University of Technology, Sydney (UTS). He's also director of the Computer Vision and Recognition Laboratory, leader of the Network Security Research group, and a deputy director of the Research Centre for Innovation in IT Services and Applications (iNEXT) at UTS. His research interests include network security, image processing, pattern recognition, and computer vision. He received a PhD in computer science from the University of Technology Sydney (UTS), Australia. He's an IEEE senior member. Contact him at Xiangjian.He@uts.edu.au.*

PRIYADARSI NANDA *is a senior lecturer in the School of Computing and Communications at the University of Technology, Sydney (UTS), Australia. He's also a core research member at the Centre for Innovation in IT Services Applications (iNEXT) at UTS. His research interests include network security, network QoS, sensor networks, and wireless networks. Nanda received a PhD in computer science from the University of Technology Sydney (UTS), Australia. He's an IEEE senior member. Contact him at Priyadarsi .Nanda@uts.edu.au.*

REN PING LIU *is a principal scientist of networking technology at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and an adjunct professor at Macquarie University and the University of Technology, Sydney (UTS), Australia. His research interests include MAC protocol design, Markov analysis, quality-of-service scheduling, TCP/IP internetworking, and network security. Liu received a PhD in electrical and computer engineering from University of Newcastle, Australia. He's an IEEE senior member. Contact him at Ren.Liu@csiro.au.*

SONG WANG *is a senior lecturer with the Department of Electronic Engineering, La Trobe University, Melbourne, Australia. Her research interests include biometric security, blind system identification, and wireless communication. Wang received a PhD in electrical and electronic engineering from the University of Melbourne. Contact her at song.wang@latrobe.edu.au.*

JIANKUN HU *is a full professor and research director of the Cyber Security Lab, School of Engineering and IT, University of New South Wales at the Australian Defence Force Academy, Canberra, Australia. His research interests are in the field of cybersecurity including biometrics security. Hu received a PhD in control engineering from Harbin Institute of Technology, China. He's an IEEE member. Contact him at j.hu@adfa.edu.au.*