

Enabling Secure and Effective Biomedical Data Sharing through Cyberinfrastructure Gateways

Shreya Goyal*, Saptarshi Purkayastha*, Tyler Phillips†,
Robert Quick‡, Alexis Britt*

*Department of BioHealth Informatics, †Department of Computer Science, ‡Pervasive Technology Institute

‡ Indiana University, *† Indiana University-Purdue University Indianapolis

Indianapolis, Indiana 46202, USA.

{shregoya, saptpurk, phillity, rquick, almbritt}@iupui.edu

Abstract—Scientific cyberinfrastructures promise solutions to computational challenges with rich resources; they embrace collaborative workflows in which users can access and share scientific data and computing resources to perform research and education tasks, which catalyze scientific discovery. The Dynaswap project reports on developing a coherently integrated and trustworthy holistic secure workflow protection architecture for cyberinfrastructures which can be used on virtual machines deployed through cyberinfrastructure (CI) services such as JetStream. This service creates a user-friendly cloud environment designed to give researchers access to interactive computing and data analysis resources on demand. The Dynaswap cybersecurity architecture supports roles, role hierarchies, and data hierarchies, as well as dynamic changes of roles and hierarchical relations within the scientific infrastructure. Dynaswap combines existing cutting-edge security frameworks (including an Authentication Authorization-Accounting framework, Multi-Factor Authentication, Secure Digital Provenance, and Blockchain) with advanced security tools (e.g., Biometric-Capsule, Cryptography-based Hierarchical Access Control, and Dual-level Key Management). The security technologies have been developed and integrated with the Open Medical Record System for enhanced security purposes and delivering a secure scientific infrastructure, which allows researchers, educators, practitioners, and students to remotely access and share sensitive data, computing resources, and workflows with flexibility and convenience while also having the highest security and privacy protection. The CI is being validated in life-science research environments and in the education settings of Health Informatics.

I. INTRODUCTION

The Dynaswap project has developed an integrated and trustworthy holistic secure workflow protection architecture for cyberinfrastructures. The project is a CI gateway that allows different users to access and share proprietary data and computing resources to perform research and education tasks, and to bring about scientific discovery. Dynaswap is developed as an OpenStack Image on the JetStream Cloud Infrastructure [1] supported by the National Science Foundation (NSF) for science and engineering research. This service creates a user-friendly cloud environment designed to give researchers access to interactive computing and data analysis resources on demand. The operational software environment that powers

JetStream is OpenStack [2]. JetStream uses the OpenStack platform [2] to deploy the virtual machines as Infrastructure as a Service (IaaS) on the Internet2 and XSEDE networks. JetStream also uses the Globus Online system for secure file transfer between VMs such that SSH, X.509, and OpenID security protocols are available to all collaborating researchers of JetStream CI.

However, although widely used by biosciences researchers, the JetStream security features do not satisfy the strict requirements involving sensitive healthcare data and protected health information [3]. Gaps include a lack of preconfigured setup for privacy-preserving workflows, accessible deidentification, role-based differential access, flexible authentication, and authorization schemes. This results in substantial resistance from data owners and clinical collaborators to share data over JetStream VMs due to fear of data security violations [4]. With a rise in precision health research, use of big data in healthcare, and demand for high-performance computing resources for data analytics and artificial intelligence, there is an urgent need for securing the CI through an easy-to-use, extensible, and robust cybersecurity architecture that can protect sensitive information being shared for scientific collaborations. The goal of the Dynaswap project is to incorporate existing cutting-edge security frameworks (including AAA, MFA, SDP, Blockchain) with advanced security tools (such as Biometric-Capsule, Cryptography-Based Hierarchical Access Control (CHAC), dual-level key management) to remotely access and share scientific data, computing resources, and workflows with flexibility and convenience while maintaining security and privacy.

The Dynaswap system architecture supports roles, role hierarchies, and data hierarchies, while ensuring protection with the latest security frameworks and tools. It also enables system administrators to dynamically change roles and hierarchical relationships within the cyberinfrastructure, allowing users to easily adjust data access in accordance with their needs. With these features, the gateway allows a wide range of users to collaborate by remotely sharing data, resources, and workflows in a flexible and secure environment. This architecture supports dynamic user (role) hierarchies and flexible associations between users and roles efficiently.

II. BACKGROUND

The users of the source data have different roles, as illustrated in Figure 1 for a subset of the users. There are also different role levels for each type of user [5]. For example, the nurses have multiple levels, including Certified Nursing Assistant, Registered Nurse, Licensed Practical Nurse, Nurse Practitioner, etc. In the same units, it is common for multiple nurses of different levels to work together for patient care. Cybersecurity challenges arise due to the complex access rights of these users, particularly when users may use medical records systems for clinical practice, research and training [6]. For instance, a physician should have unlimited access to all of the information of her/his own patients, but not full access to the information of other patients. A nurse should have only limited access rights to the information for those patients under her/his care according to the nurses' level. Clinical researchers can access information regarding their projects, while outside researchers can only access the information for their collaboration projects.

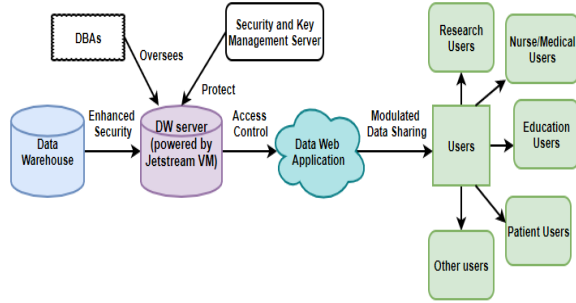


Fig. 1. The research infrastructure on which the Dynaswap security architecture is applied and validated

The project categorize the user roles and allows fine-tuned data access using Cryptography-Based Hierarchical Access Control (CHAC) as shown in Figure 2.

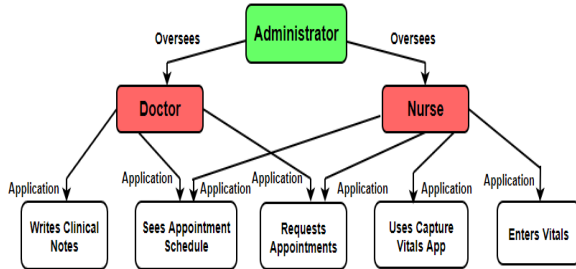


Fig. 2. Cryptography-Based Hierarchical Access Control

It is assumed that the users can be grouped in the CI by roles and each group has multiple users. The role groups can be hierarchically related. Similarly, data can also be

classified into different levels according to their sensitivities, privacy requirements, and/or structural composition relations and organized into hierarchies. Roles are associated with data, and users in a role group can access the corresponding data and all lower-level data in the hierarchy whose sensitivity is less than this data's sensitivity. Moreover, due to the role hierarchy, a user was automatically granted all the sub-roles of the role in the role hierarchy.

III. RELATED WORK

CILogon [7] is a federated X.509 certification authority that has been in use for secure access to cyberinfrastructures such as the Extreme Science and Engineering Discovery Environment (XSEDE). CILogon supports multiple levels of assurance and customizable interfaces for user communities. The benefit of CILogon is that it relies on federated authentication for determining user identities when issuing certificates, hence enabling users to obtain certificates using their existing university or google identities. This eliminates the need/burden of the users to have to manage different identities and credentials for different services [7]. CILogon relies on the InCommon identity federation when issuing certificates [7].

The use of a Credential Store to manage the diverse resource credentials, map the end-user identities with community accounts, and enable the credential delegation without human interaction while creating a trusted multi-hop authorization is fairly common in many Science Gateways projects. The Credential Store is a secure database for storing authentication data with added utilities for performing delegation and key generation [8]. Delegation is defined as the mechanism of transferring end-user resource credentials to gateway middleware. If a delegation mechanism is available, the Credential Store uses it to transfer the credentials. Where there is no such mechanism, the Credential Store uses other approaches such as SSH key generation and manual persistence of credentials [8].

The Science Gateway Platform (SciGaP) is an important CI gateway to advance scientific discovery that minimizes the net operating cost of Science Gateways [9]. The SciGaP project's goal is to develop a robust platform with the ability to provide generic middleware functionalities and features required by the science gateways. The SciGaP infrastructure promotes sustainability through scaling. The paper provides various examples of Science Gateways being operated through multi-tenanted SciGaP Infrastructure for different field of sciences. Some of them includes SEAGrid, Ultrascan, PGA, InterACTWEL, Next GEN Thermo DB, Prostate Cancer Prediction [9]. The Dynaswap project uses similar ideas for multi-tenanted infrastructure for deploying new VMs and still allowing users to share data, workflow and resources between each other.

Neuroscience Gateway (NSG) is an example of another science gateway infrastructure that makes neuroscience-specific computational tools conveniently available to the researchers and students [10]. This portal

allows the users to use the High Performance Computing (HPC) resources for modelling tasks with convenient web based interface that hides the technical details, enabling the users to focus on their research [10]. In the Dynaswap project, we are able to provide a host of security tools to researchers, educators and clinical users of the OpenMRS medical records platform, such that those from low- and middle-income countries [11] are able to access HPC resources through our gateway.

One of most popular gateways in the XSEDE community is the CIPRES (CyberInfrastructure for Phylogenetic REsearch) Science Gateway, designed to make the access easier and simpler to supercomputing resources [12]. This gateway allowed the researchers to get easy access to tools and resources for inferring evolutionary relationships. The availability of large and complex datasets have made it impossible to perform the analysis in local laptops. The CIPRES Science Gateway allows the to explore evolutionary relationships between species using supercomputers [12]. The iPlant Collaborative, another National Science Foundation (NSF) funded cyberinfrastructure (CI) project was launched in 2008 that allows the researchers to share their data, software tools, analysis pipelines, and best practices with their collaborators [13]. While originally developed for the plant science research community, iPlant has evolved to CyVerse and expanded to provide CI support across the life sciences [13].

IV. ARCHITECTURE

An overview of the Dynaswap's architecture is shown in Figure 3. Many security, data processing, and data transfer features were employed to ensure data protection and support collaborative and dynamic scientific research and training.

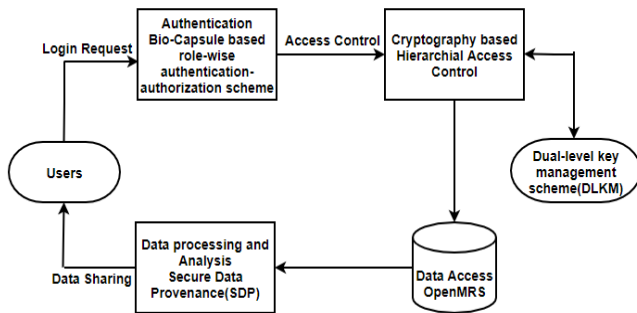


Fig. 3. Dynaswap secure scientific infrastructure for sensitive data access and sharing

A. Login Control

The first step in a user attempting to gain access to the project architecture is for that user to login and be authenticated by the system. The method of confirming a user's identity is a password-less Authentication-Authorization scheme (AuthN-AuthZ) based on a novel Bio-Capsule authentication model [14] [15].

1) *Bio-Capsule Authentication Model*: Using the new Bio-Capsule (BC) authentication model [14] [15], BioCapsule-based role-wise authentication-authorization scheme was developed to realize role-based access control and data sharing [16]. The BC generation process is through key extraction from a Reference Subject (RS - a random physical object or image) and the user's biometric data. The key is then transformed and user and RS attributes were fused to generate the Bio-Capsule [16]. This user-friendly method of authentication allows the user to use face identification to login and meet rigorous security standards [17].

2) *Authentication-Authorization Protocol*: Authentication verifies a user's identity and authorization gives that user access to specific accounts. The Bio-capsule model lent some unique advantages to Authentication-Authorization (AuthN-AuthZ) protocol. One advantage of the new BC model is that the RS (Reference Subject) can be shared by all users but each user receives a unique BC from his/her biometrics. Also, for two RSs, a user gets two totally different BCs using the same biometrics. This unique property enabled us to combine authentication and authorization together into one step.

B. Access Control

Once the user is authenticated by the system and authorized to login to a certain role, there were controls in place to determine what data the user can access based on the particular role that the user is authorized to perform in the system. Access to data within project architecture was defined by cryptography-based hierarchical access control (CHAC) [18]–[22] and the system was enhanced by Dual-Level Key Management Scheme (DLKM), which enabled the system administrators to dynamically change roles in the hierarchical model to suit their needs.

1) *Cryptography-Based Hierarchical Access Control (CHAC)*: Both roles and scientific data was organized in a hierarchical manner within the cryptography-based hierarchical access control scheme. If a user U with role R can access data D, U is automatically able to access all the data that are the children or descendants of D and in addition, to access all the data associated with the children or descendants of R. CHAC was used in this project to enable this kind of extended and fine-tuned access control. One important feature of our CHAC scheme is its ability to support dynamics, i.e., changes of a user's role or changes to the hierarchy.

2) *Dual-Level Key Management Scheme (DLKM)*: An effective key management scheme was designed that supports dynamic changes of hierarchies and user groups at both the node level and user levels. Key management allows to effectively grant users access to certain data, according to the user's role. The strength of this system includes its ability to dynamically change both the relationships between roles and users assigned to certain roles in the system. Also, the keys in the hierarchy have the property that any high-level key can derive its low-level keys but the reverse was not allowed.

C. Data Access and Computation

Once the user is authenticated and authorized to access a certain role, he or she must be able to view the appropriate data stored on the life sciences research server. Currently, Dynaswap works with the OpenMRS database, an open-source electronic medical records system [23]. However, the architecture could be applied to other databases where sensitive health information is stored. The project runs on an OpenMRS image (virtual machine) and allows users to securely access information stored in an OpenMRS database.

Once the data has been accessed researchers using this system would want to use a variety of computational methods to analyze and process the data. To protect the data, Secure Data Provenance (SDP) techniques were used, which aim to secure the data through its entire life-cycle [24]–[26].

V. VALIDATION AND TESTING

To make the secure workflow architecture a widely-used cyberinfrastructure, it needs to undergo rigorous testing and evaluation during its development and after completion. The project uses Common Vulnerability Scoring System (CVSS) v3 [27] as the vulnerability metrics for security assessment. The secured JetStream infrastructure is being assessed with rigorous testing by researchers and students in a semester-long graduate course in which a team of students simulate different user roles, are assigned various access privileges, and perform a variety of attacks against the cyberinfrastructure. Currently, the Usability testing for the BioCapsule authentication system is underway, and a survey has been sent to the students of the Privacy and Security course for the usability, security, and ease of use of the BioCapsule Authentication system implemented within the Dynaswap environment. Further, we are planning to include the testing and evaluation of the developed secure workflow and various clinical workflows. The evaluation design as a course assignment is shown in Tables 1 and 2 which includes the modules and CVSS metrics to be used for testing purposes. Table 2 provides an example of the evaluation of three types of workflows: login and access request workflow, data processing/analysis workflow, and data transfer among entities over internet workflow.

The CVSS score calculation is organized in two conceptually different groups, Exploitability metrics and Impact metrics. The specification for all CVSS metrics and possible values can be found in [27]. The results from the test will be analyzed and security measures will be improved as required.

VI. CONCLUSION

This paper discusses building a secure, holistic, fine-tuned, and multi-layered cybersecurity architecture that can be ported to multiple virtual machines on a popular CI like JetStream. This allows for scientific data/workflow so that researchers, educators, practitioners, and students can remotely access and share protected scientific data/workflow in a convenient yet finely-controlled manner.

TABLE I. Example course lab modules to test & validate the secure workflows

Module lab name	Secure Workflow
Module 1: Patient records management	1. Login to OpenMRS. Create patient cohort based on observations and diagnosis with a history of inflammatory bowel disease and ulcerative colitis. 2. Create a report and then chart the distribution of patient age and gender. 3. Create an aggregation report based on the procedure code for the cohort.
Module 2: Principles of Security and Privacy: CAINA principles and Parkarian Hexad	1. Login as a provider to get patient records for a patient who have not been seen by the provider (Confidentiality). 2. Modify a patient record and try to change the data auditing tables to reflect change by another user (Integrity). 3. Use an existing botnet, and target a server resource such that a provider from the network cannot access the patient record (availability).
Module 3: Data sharing and Protected Health Information	Create a cohort report based on patient symptom or condition and export C-CDA such that all HIPAA mandated 18 identifiers are removed from each patient record.

TABLE II. Example of evaluation of the security-enhanced JetStream infrastructure

Steps	Components tested	CVSS metric
Login	BC-based AuthN-AuthZ validation	Attack Vector [N,A,L,P]
Access control	Role-wise AuthN-AuthZ and CHAC	Privilege Required [N,L,H]
Access data	Role-wise AuthN-AuthZ, CHAC	Privilege Required [N,L,H]
Create/Compute	CHAC	Privacy-preservation, Integrity [H,L,N]
Update	SDP (Signature)	Report Confidence [X,C,R,U]
Upload	BC-based AuthN-AuthZ and CHAC	Privilege Requires [N,L,H]
Network transfer	SDP (Enc., MAC)	Availability Req. [X,H,M,L]

Notes: (1) Enc.: Encryption, MAC: Message Authentication Code. (2) For metric values in [H (high) and L (low)], each has its own meaning and empirical score.. E.g., for attack vector: N: Network (0.85), A: Adjacent Network (0.62), L: Local (0.55) and P: Physical (0.2). The specification for all CVSS metric and possible values can be found in [27].

The novelty of our Dynaswap approach is the robust and secure workflow it offers using advanced security technologies. For securing the login process, a role-wise password-less Authentication-Authorization scheme (AuthN-AuthZ) based on a novel Biometric-Capsule is used. Also, the fine-tuned cryptography-based hierarchical access control scheme (CHAC) was developed and integrated with the OpenMRS. This ensures that only the authenticated legitimate users are able to login into the infrastructures and access the data. For securing the data processing and analysis, use of SDP (secure digital provenance) is made to protect the scientific data and its provenance. The overall architecture makes sure that the stringent security measures are present at each step of the workflow. Another unique feature of the Dynaswap project is that it can accept and efficiently deal with user-role dynamics: the relationships among roles and among data can change and a user can join/be granted to, leave/be removed from, and change a role group according to changing application requirements and security needs.

The implemented secure architecture, when applied to healthcare and life-science infrastructure, will enable valuable clinical data to be broadly used by various users. Dynaswap will also help scale and protect NSF investments in national

cyberinfrastructure such that they can be useful to healthcare projects. It also promotes the use of the rich, real clinical data in education. It would be interesting to see and evaluate, in the future, if the usability and performance of the Dynaswap architecture will be acceptable practically for the researchers in different context and countries. For example, forgetting/deleting ones own data as a patient or revoking the key if someone leaves the system ("right to erasure" in EU law) needs to be implemented in the future.

VII. ACKNOWLEDGEMENT

The U.S. National Science Foundation supported this work under grant OAC-1839746. This work was made possible through a research allocation on the JetStream [1] public cloud infrastructure and XSEDE resources.

REFERENCES

- [1] C. A. Stewart, T. M. Cockerill, I. Foster, D. Hancock, N. Merchant, E. Skidmore, D. Stanzione, J. Taylor, S. Tuecke, G. Turner *et al.*, "Jetstream: a self-provisioned, scalable science and engineering cloud environment," in *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, 2015, pp. 1–8.
- [2] O. Sefraoui, M. Aissaoui, and M. Eleuldi, "Openstack: toward an open-source solution for cloud computing," *International Journal of Computer Applications*, vol. 55, no. 3, pp. 38–42, 2012.
- [3] C. A. Stewart, "Big data, big red ii, data capacitor ii, wrangler, jetstream, and globus online," *The Science of Cyberinfrastructure*, 2015.
- [4] C. Farcas, N. Balac, and L. Ohno-Machado, "Biomedical cyberinfrastructure challenges," in *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, 2013, pp. 1–4.
- [5] S. Purkayastha, A. Price, R. Biswas, A. J. Ganesh, and P. Otero, "From dyadic ties to information infrastructures: care-coordination between patients, providers, students and researchers," *Yearbook of medical informatics*, vol. 24, no. 01, pp. 68–74, 2015.
- [6] S. Purkayastha, J. W. Gichoya, and S. A. Addepally, "Implementation of a single sign-on system between practice, research and learning systems," *Applied clinical informatics*, vol. 8, no. 1, p. 306, 2017.
- [7] J. Basney, T. Fleury, and J. Gaynor, "Cilogon: A federated x. 509 certification authority for cyberinfrastructure logon," *Concurrency and Computation: Practice and Experience*, vol. 26, no. 13, pp. 2225–2239, 2014.
- [8] T. A. Kanewala, S. Marru, J. Basney, and M. Pierce, "A credential store for multi-tenant science gateways," in *2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE, 2014, pp. 445–454.
- [9] S. Marru, M. Piece, E. Abeysinghe, S. Pamidighantam, M. Christie, and D. Wannipurage, "Experiences from scaling scale science gateway operations," in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)*, 2019, pp. 1–4.
- [10] T. Carnevale, A. Majumdar, S. Sivagnanam, K. Yoshimoto, V. Astakhov, A. Bandrowski, and M. Martone, "The neuroscience gateway portal: high performance computing made easy," *BMC neuroscience*, vol. 15, no. S1, p. P101, 2014.
- [11] S. Purkayastha, T. D. Manda, and T. A. Sanner, "A post-development perspective on mhealth—an implementation initiative in malawi," in *2013 46th Hawaii International Conference on System Sciences*. IEEE, 2013, pp. 4217–4225.
- [12] M. A. Miller, W. Pfeiffer, and T. Schwartz, "Creating the cipres science gateway for inference of large phylogenetic trees," in *2010 gateway computing environments workshop (GCE)*. Ieee, 2010, pp. 1–8.
- [13] N. Merchant, E. Lyons, S. Goff, M. Vaughn, D. Ware, D. Micklos, and P. Antin, "The implant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences," *PLoS biology*, vol. 14, no. 1, p. e1002342, 2016.
- [14] Y. Sui, X. Zou, E. Y. Du, and F. Li, "Design and analysis of a highly user-friendly, secure, privacy-preserving, and revocable authentication method," *IEEE Transactions on Computers*, vol. 63, no. 4, pp. 902–916, 2013.
- [15] Y. Sui, X. Zou, and E. Y. Du, "Biometrics-based authentication: A new approach," in *2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2011, pp. 1–6.
- [16] T. Phillips, X. Zou, and F. Li, "A cancellable and privacy-preserving facial biometric authentication scheme," in *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2017, pp. 545–549.
- [17] A. K. Jain, K. Nandakumar, and A. Nagar, "Biometric template security," *EURASIP Journal on advances in signal processing*, vol. 2008, pp. 1–17, 2008.
- [18] M. J. Atallah, M. Blanton, N. Fazio, and K. B. Frikken, "Dynamic and efficient key management for access hierarchies," *ACM Transactions on Information and System Security (TISSEC)*, vol. 12, no. 3, pp. 1–43, 2009.
- [19] H.-Y. Chen, "Efficient time-bound hierarchical key assignment scheme," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 10, pp. 1301–1304, 2004.
- [20] I. Ray, I. Ray, and N. Narasimhamurthi, "A cryptographic solution to implement access control in a hierarchy and more," in *Proceedings of the seventh ACM symposium on Access control models and technologies*, 2002, pp. 65–73.
- [21] W.-G. Tzeng, "A time-bound cryptographic key assignment scheme for access control in a hierarchy," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 182–188, 2002.
- [22] S. Zhong, "A practical key management scheme for access control in a user hierarchy," *Computers & Security*, vol. 21, no. 8, pp. 750–759, 2002.
- [23] S. Purkayastha, R. Allam, P. Maity, and J. W. Gichoya, "Comparison of open-source electronic health record systems based on functional and user performance criteria," *Healthcare Informatics Research*, vol. 25, no. 2, pp. 89–98, 2019.
- [24] K.-K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. I. Seltzer, "Provenance-aware storage systems," in *Usenix annual technical conference, general track*, 2006, pp. 43–56.
- [25] S. Sultana, M. Shehab, and E. Bertino, "Secure provenance transmission for streaming data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1890–1903, 2012.
- [26] X. Wang, K. Zeng, K. Govindan, and P. Mohapatra, "Chaining for securing data provenance in distributed information networks," in *MILCOM 2012-2012 IEEE Military Communications Conference*. IEEE, 2012, pp. 1–6.
- [27] S. Hanford, "Common vulnerability scoring system, v3 development update," in *Technical report, Forum of Incident Response and Security Teams (FIRST)*, 2013.