

# Flurry: a Fast Framework for Reproducible Multi-layered Provenance Graph Representation Learning

Maya Kapoor  
Parsons Corporation  
UNC Charlotte  
mkapoor1@uncc.edu

Joshua Melton  
UNC Charlotte  
jmelto30@uncc.edu

Michael Ridenhour  
UNC Charlotte  
mridenh7@uncc.edu

Mahalavanya Sriram  
UNC Charlotte  
msriram2@uncc.edu

Thomas Moyer  
UNC Charlotte  
tmoyer2@uncc.edu

Siddharth Krishnan  
UNC Charlotte  
skrishnan@uncc.edu

## ABSTRACT

Complex heterogeneous dynamic networks like knowledge graphs are powerful constructs that can be used in modeling data provenance from computer systems. From a security perspective, these attributed graphs enable causality analysis and tracing for analyzing a myriad of cyberattacks. However, there is a paucity in systematic development of pipelines that transform system executions and provenance into usable graph representations for machine learning tasks. This lack of instrumentation severely inhibits scientific advancement in provenance graph machine learning by hindering reproducibility and limiting the availability of data that are critical for techniques like graph neural networks. To fulfill this need, we present FLURRY, an end-to-end data pipeline which simulates cyberattacks, captures provenance data from these attacks at multiple system and application layers, converts audit logs from these attacks into data provenance graphs, and incorporates this data with a framework for training deep neural models that supports preconfigured or custom-designed models for analysis in real-world resilient systems. We showcase this pipeline by processing data from multiple system attacks and performing anomaly detection via graph classification using current benchmark graph representational learning frameworks. FLURRY provides a fast, customizable, extensible, and transparent solution for providing this much needed data to cybersecurity professionals.

## CCS CONCEPTS

• Security and privacy → Intrusion/anomaly detection and malware mitigation; • Computing methodologies → Learning latent representations; Neural networks; • Computer systems organization;

## KEYWORDS

Graph Representation Learning, Data Provenance, Cyber Security, Complex Systems

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

KDD '22, August 14-18, 2022, Washington, DC, USA

© 2022 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## ACM Reference Format:

Maya Kapoor, Joshua Melton, Michael Ridenhour, Mahalavanya Sriram, Thomas Moyer, and Siddharth Krishnan. 2022. Flurry: a Fast Framework for Reproducible Multi-layered Provenance Graph Representation Learning. In *Proceedings of SIGKDD '22: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '22)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

The lineage or record of data modification, access, and usage, known as data provenance, has been a long-standing tool for resilient cybersecurity systems to use in attack tracing. Data may be sourced from multiple layers of a host, including the kernel or operating system level, the application or data layers, the browser layer, and the physical or virtualized file system. A single host may produce several gigabytes of log data a day even at a coarse level of provenance capture [21]. With multiple host systems generating intranet data over long periods of time, it is evident that data provenance management and analysis is a big data problem.

Resilient systems use graph representations of data provenance in intrusion and anomaly detection in order to perform attack tracing and to find causal dependencies to form mitigation strategies used against all kinds of attacks. Provenance graphs, often represented as heterogeneous networks, are comprised of system entities and objects interconnected by a variety of interactions generated from agents who affect changes in the system. Representing system activity using knowledge graphs effectively captures the complex and multi-faceted nature of system activity and lends itself to structural analysis by graph machine learning algorithms, causal inference, and time series analysis over dynamically generated networks. Unsupervised learning from the structure of the network representation has allowed security specialists to detect even zero-day attacks that have not been previously seen and therefore existing supervised machine learning techniques lack the training data to effectively identify such attack vectors [13, 14]. Additionally, the ability of graph machine learning techniques to effectively scale up to web-scale graphs, coupled with the use of graph summarization techniques [13, 18] has allowed for “low and slow” attack patterns like Advanced Persistent Threats (APTs) to be detected [15] despite the sheer amount of data gathered over long periods of time, which has thwarted traditional IDSes.

The advantages of graph machine learning applied to provenance graphs for anomaly detection have inspired a growing, intersectional research community of cybersecurity specialists, systems engineers, data scientists, and machine learning experts. Each member of this community has specialized knowledge of different parts of the non-trivial process of logging data, engineering the data into graphs, and learning from this graph-structured data. Currently, there is not a data pipeline for automatically producing data from cyberattacks, generating provenance graphs, and testing graph machine learning algorithms on the data; rather, this process must be done manually by experts from across the fields. This can be challenging, for example, for data scientists who may not have access to cyberattack data or the specialization to generate and capture this kind of data themselves. On the other hand, cybersecurity specialists may not have the graph machine learning expertise to effectively apply the latest machine learning approaches to the system provenance data they have on hand.

In order to bridge the gap between these communities, enable automatic data generation, and optimize the learning process, we propose FLURRY, a fast provenance framework for graph generation and analysis. FLURRY is a virtual machine environment equipped with provenance capture and graph generation tools. As opposed to learning on static datasets, FLURRY dynamically executes automated cyberattacks and converts these logs into provenance graphs ingestible by state-of-the-art graph machine learning tools. With the FLURRY system, researchers in provenance graph representation learning are no longer required to perform their work in separate silos. The process from attack simulation to anomaly detection is transparent, explainable, and readily extensible for use with current techniques and for future development of novel techniques for automated provenance analysis. Researchers may additionally use our provenance generation system with their own machine learning models for plug-and-play functionality to test new methods of representation learning.

The FLURRY system provides the following contributions:

- An end-to-end pipeline which can take a system execution, model it as a multi-layer provenance graph, and use graph representation learning to detect anomalies in that execution,
- Dynamic, reproducible data sets of web-based injection attacks and brute force password attacks,
- Visualized execution of cyberattacks at the push of a button for full clarity into our experiments and simplicity for future researchers,
- Concepts for multi-layered data provenance across kernel and application layers, bridging the semantic gap between user and kernel space,
- A plug-and-play framework for graph learning models to analyze complex provenance graphs. Furthermore, our framework provides a readily usable sandbox to test graph machine learning models designed to analyze whole system provenance graphs,
- A platform to bring together researchers from the machine learning community and the systems community.

## 2 PRELIMINARIES

### 2.1 Data Provenance

Provenance refers to the lineage of an object over time. Specifically, provenance is a record of data origination, modification, and interactions. Provenance originated in the context of art history, but has been widely adapted even to computer system data. This kind of data provenance records the history of system data over time, but also importantly captures relationships among data. These associations can provide credibility to the data and be used to establish trust. Furthermore, this trustworthiness makes data reusable for both scientific research and root cause analysis. Provenance records may be replayed or analyzed in a security context to determine potential cyber attacks and mitigate those threats as a response.

Due to its relational properties, provenance data can be well-represented as a provenance graph. The W3C Incubator Group has defined the PROV-DM model in order to represent provenance data as directed, acyclic graphs. For node types, data provenance can be recorded as entities, which are real, physical or digital concepts; activities, which are interactions between or among entities; or agents, which can act upon entities. PROV-DM also defines several relation types between these nodes [30]. Provenance graphs are *directed*, meaning that an entity for example acts upon an activity in a directional relationship. The graphs are also *acyclic* due to the time series of actions. The graphs which are produced by provenance data are by nature *heterogeneous*, as nodes will be different files, processes, inodes, and more diverse node types which will have their own unique features. Provenance produces *multigraphs* which have multiple relationships between nodes of varying edge types, making the graph problem both rich with information and uniquely challenging for analysis.

Figure 1 presents a high-level abstraction of provenance gathered from the Google Chrome web browser running on the FLURRY system. In the diagram, the Google Chrome web browsing service (task) was generated by the process memory which was derived from the binary path. This task is running on the host machine and receiving messages from a socket. Furthermore, the task is informing itself of versioning activity as it is running. This depiction represents only a small portion of the gigabytes of data that are generated as part of the provenance process over lengthy periods of time.

### 2.2 Graph Representation Learning

For data provenance, graph representation learning may be used to detect anomalies in execution graphs. Executions of anomalous behavior (for example, a cross-site scripting attack) will generate provenance subgraphs which differ structurally and in their hidden representations from benign execution graphs (for example, web requests from a browser). Graph machine learning can aid cyber security analysts in detecting and adapting to unforeseen attacks in resilient systems.

We represent this multi-relational provenance data as a heterogeneous graph  $G = (V, E, O, T, R)$ . This directed graph contains a node-set  $V$ , an edgeset  $E$ , a set of node types  $O$ , a set of edgetypes  $T$ , and a set of canonical relation types  $R$ . Each node  $v \in V$  maps to a node type  $o_v \in O$  and each edge  $(u, v) \in E$  maps to an edge type  $t_{u,v} \in T$ . Each canonical relation type  $r \in R$  maps to a tuple  $(o_u, t_{u,v}, o_v)$ .

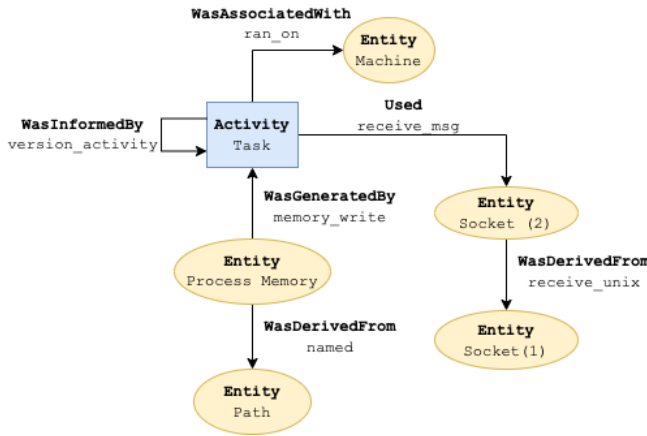


Figure 1: Sample provenance graph from Google Chrome web browsing activity.

The aim of the graph representation learning component of this framework is to learn a function which maps a provenance graph  $G$  to an expressive, low-dimensional vector representation, or embedding,  $z$ , where distance between embeddings correlates to similarity between the graphs. In this work graph embeddings are used to train a classifier to detect graphs containing anomalous behavior.

### 3 RELATED WORK

The collection, storage, and summarization of data provenance have been well-studied in the security research community, and representational learning on heterogeneous provenance graphs has recently become a prominent field in the machine learning community [1]. Despite long-standing research efforts in their respective communities, the intersection of data provenance and graph representational learning lacks many foundational frameworks that provide abundant public datasets, encourage reproducibility, and enable the use of graph neural networks for system provenance data. Recent efforts in developing graph-based techniques that employ neural networks for automated analysis of system provenance data make the need for such frameworks increasingly clear.

#### 3.1 Graph Representation Learning

Graph representation learning seeks to generate low-dimensional feature vectors (or embeddings) for nodes, edges, or graphs. Traditional methods for generating embeddings for homogeneous graphs, such as DeepWalk [22] and node2vec [9] used random walks on graphs to learn about the local and global neighborhoods in which nodes participate, allowing for generation of node representations that preserve graph structural information. Advances in deep graph neural networks have introduced the message-passing paradigm and convolutional graph embedding approaches (GCNs) [10, 16]. Such GCNs leverage spectral or spatial convolutional filters along with neighborhood aggregation to encode graph structural properties in node embeddings. Graph Attention Networks (GAT) [29] further improved on such graph convolutional approaches by introducing node-level attention mechanisms during neighbor aggregation.

Such convolutional approaches were originally limited to homogeneous networks, but in recent years a number of heterogeneous graph embedding frameworks have been proposed to extend graph neural networks to such multi-layered networks. Relational-GCN [26] extends the idea of GCNs by employing separate convolutional operations for each relation within a knowledge graph and then aggregating node representations across the multiple relations in which each node participates. Heterogeneous Attention Networks (HAN) similarly extend the idea of GAT to heterogeneous networks by employing multiple convolutional filters defined over specified metapaths in a graph. To consolidate the various representations of nodes from multiple metapaths, HAN employs semantic attention to learn an optimal aggregation of the multiple embeddings. Graph isomorphism network (GIN) [33] explores the expressive power of GCN models for graph classification. The authors prove that graph convolutional approaches can at most equal the Weisfeiler-Lehman algorithm for graph isomorphism and that sum aggregation is more expressive of the characteristic of multi-sets when compared with mean or max aggregators. In addition to these works, other recent frameworks extend the ideas of random walks on heterogeneous graphs [1, 6], and spectral approaches [17] for use with graph neural networks that are effective techniques for unsupervised learning on heterogeneous graphs.

#### 3.2 Deep Learning on Provenance Graphs

**3.2.1 StreamSpot.** Manzoor et al propose StreamSpot [18] as a clustering-based anomaly detection approach on heterogeneous streaming graphs. They analyze browsing data and downloading software in an effort to identify drive-by download attacks as anomalous behavior.

**3.2.2 SIGL.** SIGL [14] uses a word2vec-based node embedding approach in conjunction with a long short-term memory network to capture long-term dependencies in the provenance graph. Instances that fail to reach a minimum threshold of reconstruction loss in the decoder stage are determined to be anomalous software installations.

**3.2.3 Unicorn.** Similar to StreamSpot, Unicorn [13] uses a graph sketching technique based on similarity hashing to reduce graph size to a fixed space. For classification, the Unicorn system employs a version of the Weisfeiler-Lehman subtree graph kernel algorithm. They use clustering-based anomaly detection algorithms to aggregate similar graph hashes and identify anomalous provenance graph signatures.

**3.2.4 PROV-GE<sub>m</sub>.** In PROV-GE<sub>m</sub> [15], the authors propose a graph embedding framework based on graph convolutional networks coupled with relational self-attention to generate informative representations of provenance graphs. This combination allows PROV-GE<sub>m</sub> to encode the multi-faceted relationships and graph structural properties captured in heterogeneous data provenance graphs. PROV-GE<sub>m</sub> conducts anomaly detection experiments using the datasets published by the authors of StreamSpot and Unicorn, outperforming the aforementioned methods on graph anomaly detection and suggesting great potential for graph neural network approaches in automated analysis of data provenance graphs.

While the datasets in these related works are publicly available to researchers, they are static in nature. System provenance data cannot be easily reproduced with certainty since system specifications and attack scenarios may differ, and potential researchers must rely on the sparse documentation of previously published provenance data generation frameworks. In this work, we propose a paradigm shift to dynamic data collection for provenance graph analysis. With the FLURRY system, researchers can design their own benign and attack scenarios with support for a wider variety of attack types than in previous research. Raw system data is dynamically created, and the logs are converted to graphs using our extensible software module. This ensures that researchers also have access to the original system logs, providing a better representation of real-world attack scenarios than curated, pre-processed datasets [11]. FLURRY's intuitive GUI illustrates attack scenarios in the Chrome browser in real time, providing a transparent and easily understandable system for provenance data generation usable by both security experts and other non-security researchers from areas like machine learning and data science.

### 3.3 Automated Attack Systems

**3.3.1 Xanthus.** In conjunction with Unicorn [13], Han et al highlight the lack of publicly available provenance graph datasets and emphasize the deficits of those datasets that are made available to researchers. They propose a push-button orchestrated system [12] for recreating test and training data for provenance-based intrusion detection systems. Xanthus introduces the idea of a distributable virtual machine image with configurable jobs for automated recreation of attack scenarios, but this system requires the user to orchestrate the attack and to ensure that the VM and job scripts are correctly configured to mimic real-life scenarios. This requires extensive systems knowledge and makes the framework difficult to use for other researchers focused in data science and not security.

## 4 SYSTEM DESIGN

FLURRY is designed to work as both independent software modules or a joint pipeline to support provenance graph learning from data creation and storage to examination. In the following sections, we detail each stage of this process which can work jointly or independently at the discretion of the user. Throughout the system, we offer methods for importing and exporting data in multiple formats for maximum adaptability.

### 4.1 Cyber Attack Automation

For FLURRY v1.0, we focus on generating provenance data for web-based cyber attacks. In this architecture, provenance may be captured for data exchanged across sockets, server-client or peer-to-peer requests and responses, software or other attachments queried for and returned, file reads and writes and memory access on both server and client, and running processes on the server and client, to name a few points of interest. In order to simulate web-based cyber attacks on a web server, we run an instance of the Damn Vulnerable Web Application [5] on the Fedora virtual machine that is configured with a traditional XAMPP stack including Apache2 web server and MySQL database for storage. This PHP-based application that allows security students and professionals to practice skills and

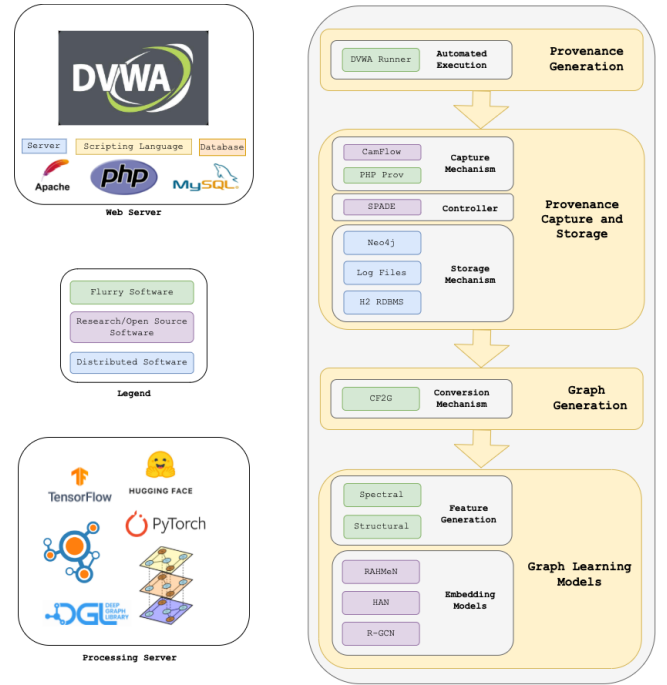


Figure 2: A system overview of the fast FLURRY framework.

tools in a legal environment. Because it is vulnerable, it is strongly recommended not to run the application on any web server with real user information; therefore, we also recommend not uploading proprietary or private data to FLURRY to prevent exfiltration.

In order to start the DVWA application and web services needed to run it, we provide a graphical user interface for researchers to start all dependencies with the push of a button. Once initiated, the user may select one of the attack scenarios to run and generate provenance, or a benign scenario which corresponds to one of the attacks from our **DVWA Runner** application. This level of abstraction over our automated scripts allows for rapid generation of gigabytes of data in mere seconds even by inexperienced users. More details about the attacks FLURRY currently supports and the benign behavior is provided in the datasets section.

As provenance is generated on the backend, the user will see our web driver service run the scenario in real time on the screen. This allows for full transparency of the attack process so that researchers are aware of how their data was made. In addition to the user interface, we provide the Python scripts themselves so that more experienced users may engineer new scenarios or make adjustments according to their needs.

### 4.2 Data Provenance Capture and Management

The capture of data provenance is an entire subset field of data engineering. At the systems level, there are multiple layers, or planes, at which data is exchanged. In the *kernel space*, systems like Hi-Fi [23], the Linux Provenance Modules (LPM) [2] framework, and CamFlow [21] capture IPC mechanisms, network activity, and access to the kernel itself from the perspective of the operating system. These systems use hooks into security access mechanisms to



write provenance to a buffer which is then typically translated into user space. In the *user space*, we can divide provenance into even more fine-grained layers. For example, one may want to isolate **application-layer provenance** to particular files the user is modifying. Details of what was modified will not be recorded at the kernel level, but we may want to semantically tie kernel-level security access with data modification at the application level. This can be further extended to browser-level provenance if we are using a browser application to view or modify the file, and may even go to **multi-host or network-level provenance** if we are sharing this file through a file transfer service like FTP. Part of the data engineering required for good data provenance is determining the granularity at which to capture metadata and how to semantically tie these layers together. On the one hand, capturing too little or too coarse an image of the system execution leads to over-summarization and can cause analysts to miss anomalies. However, too much detail or association leads to *dependency explosion*, or the phenomena of data becoming so entangled that the strongest semantic ties are lost.

In FLURRY, we are proposing to support a multi-plane approach for provenance graphs for representation learning, which to our knowledge is a novel addition to this subfield. At the kernel level, we use **CamFlow**, which is a Linux security module, to capture provenance through kernel hooks. Data provenance which is captured by this service include security access to system information like inodes, security contexts, timestamps and jiffies, which processes accessed kernel objects, and more. The full description of CamFlow provenance node and edge types is well-recorded in their documentation [21]. This provenance information is published to a ring buffer which is then captured by the CamFlow daemon. In the default configuration of FLURRY, we write this provenance to a log file on the virtual machine, but it may be additionally configured to use a unix pipe or a messaging service in order to convey this information dynamically. This allows the system to work in a non-stop manner without the need for static data files.

In addition to kernel-level provenance, FLURRY adds hooks to PHP-based source code in DVWA in order to capture application-level provenance. Such provenance allows us to trace the flow of data within the web application such as where they came from, how they came to be in their present state, who or what acted upon them. Here in our application, events such as who inputs the data, how it is transformed to an HTTP request, and subsequent server responses are recorded. This additional layer may be combined with kernel access information to provide semantic context to what is happening at the kernel level. In order to capture both layers of provenance information in our audit logs, we integrate the DVWA application source code with the Camflow Userspace API. For such integrations, we have also used a library called PHP-CPP, developed by Copernicus. This combination of application-level and kernel-level provenance gives a wider lens into system execution and a larger plane for anomaly detection.

Over time, storing provenance as log files is both space-consuming and unexpressive. Researchers may want to store provenance in a graph database as opposed to logs. This not only allows for provenance to be stored as nodes and edges, but also supports querying for specific graph structures from the large amount of data. For this reason, FLURRY can also be configured to push data to **Neo4j** [19], a

graph database which can be refined and queried. Storage can also be managed in the FLURRY environment through the **SPADE** [8] controller, an open-source provenance management software created by Gehani et al. SPADE offers different provenance capture methods including a CamFlow interface, analyzers, and transformers to perform execution partitioning and graph summarization. SPADE’s purpose is also to serve as a control interface to port CamFlow data into Neo4j or even relational databases like H2. As a research software, some of these capabilities remain unimplemented or run with error. In future work, we plan to expand on some of these additional capabilities and contribute to their improvement.

### 4.3 Provenance Graph Conversion

Related works in provenance graph representation learning have written their own **scripts** for conversion of provenance data into complex graphs [13, 14, 18]. While this serves as a stop-gap solution, ultimately these works have not focused on reusability or extensibility of conversion. As opposed to scripts for fixed use cases, we add to the field a software tool which supports conversion of data from many input forms created by mainstream provenance capture tools into software or serialized graphs **usable by** machine learning frameworks. In FLURRY v1.0, we call this **CF2G**, which stands for “CamFlow to graph”. CF2G accepts as input data serialized in the W3C-PROV [30] JSON format, or the SPADE [8] JSON format, both of which are published by the CamFlow daemon and configurable via the camflowd service. From the command line interface, CF2G parses the input graph and stores it as an internal graph. Using the CLI or the packaged library, a researcher may create JSON-exportable network representations of the stored provenance, call for the creation of a NetworkX [20] graph, or construct a Deep Graph Library [31] graph. CF2G also supports statistical output of the recorded provenance such as node types, edge types, and overall dimensionality. Currently, we support direct library calls or serialized and compressed export of the graphs to the next module.

### 4.4 Feature Engineering

Graph convolutional network models for learning node embeddings rely on iteratively aggregating a node’s information with the information of its neighbors. This process hinges on each node having an initial node feature vector which can be used in the aggregation process. For this work we use structural node information including node degree and custom spectral methods for heterogeneous graphs extended from [3], which performs an eigendecomposition using the LOBPCG [7, 27] estimator on a block matrix [4] built around the normalized Laplacian matrices of each relation type in the graph.

### 4.5 GNN Trainer

In addition to our system provenance collection and graph generation tools, we provide a unified graph neural network trainer based on DGL and PyTorch, inspired by the Trainer class provided by HuggingFace for training Transformer models on NLP tasks. Our training framework consists of a base Trainer class that defines the core training and evaluation functions for graph classification using GNN models, and a TrainingArguments class that defines training and evaluation hyperparameters and additional behavior.

We provide a set of default callback functions that define normal training operations, including logging, saving model checkpoints, and early stopping using validation metrics. Our framework also supports custom user-defined callback functions to further customize the Trainer’s behavior during model training and evaluation. Our Trainer framework utilizes DGL’s graph batching and message-passing functionality to provide a standardized yet flexible framework for training graph neural models. We support standard PyTorch loss functions, such as cross-entropy loss, and also custom-defined trainable loss functions that can be used for unsupervised training on graphs. Our framework currently supports graph classification tasks, and we plan to extend our Trainer framework to support node classification and edge prediction tasks in the future.

## 5 THREAT MODELS

The exact datasets we used for experiments will be publicly available. Table 1 shows the size of the graphs in terms of nodes and edges. Relation types are defined as unique source node, edge type, destination node three-tuples. The execution involves navigating to the tab on the DVWA website that corresponds to the attack and performing some attack or benign behavior. The attacks covered in this dataset and by the current version of FLURRY cover more than 50% of the most popular attack vectors cyber criminals choose for web-based applications [24].

### Cross-Site Scripting

Cross-site scripting allows an attacker to inject JavaScript in a web application in order to modify the web page viewed by the user or perform some other malicious action such as data exfiltration or modification. Like other injection attacks, the vulnerability usually exists because of a lack of data input sanitization from web forms or requests.

**XSS Reflected.** Reflected scripts are non-persistent, meaning that they are ephemerally injected into the HTTP client request processed by the web server. If the web server does not properly read the HTML escape sequences when reading the request, the JavaScript in the request will be executed on the page on the server side. In the FLURRY dataset, we create a benign execution by filling out and submitting the web form question (“What is your name?”). For the attack, we inject JavaScript in the web form to pop up a text box and corresponding message after submission.

**XSS Stored.** A stored scripted attack is not sent by the client, but rather stored in the server database and thus is persistent as it can be re-executed. For example, an attacker may post a message to a forum with an embedded script. When other clients go to read the message, the attacker’s script will be executed when the message is read/returned [25]. In FLURRY, we automate posting a non-malicious message to a forum on the XSS Stored tab for benign execution. For the attack, we inject a script in the message which creates a bot-style malicious post every time the tab is visited.

**XSS DOM.** A DOM-style attack can be modified on the client side and thus does not require access to the web server or poor sanitization on the server’s part. It relies on poor reading of HTML escape sequences in the client’s document object model. The XSS DOM tab in the Damn Vulnerable Web Application asks the user

to choose a language, and this setting is then part of the DOM and shown in the public-facing URL. In FLURRY, we modify the DOM by injecting the script into the web URL. Our script creates a pop-up similar to the XSS Reflected attack. For benign behavior, we automate selecting English as the default language.

### Injection

While cross-site scripting is a style of injection attack specific to JavaScript and the web page, other types of injections such as SQL for database access/modification and command line injection can be used to cause remote code execution, privilege escalation, data exfiltration, and other types of more consequential attacks.

**Command Line Injection.** When web submissions are not properly sanitized, system commands may be included and will be executed on the server machine. In DVWA, the Command Line Injection tab asks the user to provide an IP address to ping. For the benign execution, we provide localhost 127.0.0.1 for ping. For the attack, we inject `pwd` to print the current working directory where the server is executing the ping.

**SQL Injection.** In the DVWA SQL Injection tab, the user is prompted for an ID. In a benign execution, this ID is included as part of a SQL query which is looked up in the internal database. In the attack automation, we include a semicolon and additional SQL query for other user IDs.

### Password Cracking

In addition to injection-style attacks, we wanted to expand into password crackers, which make up a significant portion of web application-based cyber attacks [24]. In future versions of FLURRY, we may further look into more advanced rainbow table or dictionary style password attacks.

**Brute Force.** There are an array of automated tools for brute force; we use the Hydra [28] password cracking tool. We provide an attack script which executes the hydra command and provide a dictionary with the 100 most commonly used passwords to try. For benign execution, we automate logging in with the correct credentials.

## 6 EXPERIMENTS

In this section, we describe the anomaly detection experiments we perform on system provenance data collected by FLURRY. For each of the six attack types characterized in Section 5, we simulate 100 benign operations and 100 attack operations. We convert the system provenance data into heterogeneous graphs and partition the datasets into five cross validation folds to ensure consistent and reproducible results. For each attack type, we conduct graph classification experiments using our Trainer class with three different heterogeneous graph embedding frameworks from graph representational learning literature—R-GCN [26], HAN [32], and PROV-GE [15]—implemented in DGL. In each experimental setup, one cross validation fold is held out as a test set, and the graph embedding models are trained with the remaining provenance graphs using cross entropy loss. Our results on the graph anomaly detection task are presented in Table 2. We report the mean and standard deviation over five-fold cross validation of precision, recall, and F1

Attack Vector	Class Type	Avg. # of Nodes	Avg. # of Edges	Avg. # of Relation Types
XSS Stored	Benign	19,436	463,682	30
	Attack	19,547	480,179	31
XSS Reflected	Benign	33,481	824,276	31
	Attack	24,435	666,705	32
XSS DOM	Benign	31,243	753,519	30
	Attack	30,261	751,133	32
Command Line Injection	Benign	32,996	793,318	29
	Attack	25,720	631,633	30
SQL Injection	Benign	22,903	543,266	30
	Attack	29,576	733,375	30
Brute Force	Benign	21,518	517,101	30
	Attack	418	416	1

Table 1: Description of Attack Graph Datasets

	HAN			R-GCN			PROV-GEm		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
brute-force	99.51±0.97	100.0±0.0	99.35±1.29	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
cl-injection	63.35±8.07	81.44±15.97	65.57±12.98	70.60±8.26	69.16±23.59	79.84±12.95	58.08±9.56	63.72±15.37	60.00±12.76
sql-injection	71.32±17.85	71.43±16.39	83.30±17.35	81.66±9.19	73.07±15.60	89.26±3.39	71.30±10.61	84.59±5.55	69.41±10.42
xss-dom	91.23±4.33	84.83±8.43	97.50±5.00	80.70±9.97	81.66±16.51	82.92±11.95	91.23±4.33	84.83±8.43	97.50±5.00
xss-reflected	90.35±4.84	88.79±9.83	93.63±4.05	84.22±4.55	85.93±8.53	83.64±6.37	88.87±7.23	83.63±14.22	95.16±4.78
xss-stored	52.80±8.36	71.87±14.48	53.96±10.51	48.39±6.64	71.00±23.63	51.59±6.52	53.50±8.48	64.59±15.92	55.93±12.64

Table 2: Graph classification results using generated FLURRY data for six system attack types.

score. The code for our experiments along with our graph neural network training framework are provided at<sup>1</sup>

Due to the nature of brute force attacks, their system provenance generates graphs containing only a single relation type: (socket, WasDerivedFrom, socket). As such, these graphs are trivially easy to distinguish from benign system behavior, and all three of the graph neural network models were able to perform perfectly on the generated data. By contrast, for the two injection-style attacks—command line injection and SQL injection—anomalous behavior is much more difficult to identify from benign provenance graphs generated during normal application use. Both scenarios exhibited high variance in performance across cross validation folds, suggesting that both injection-style attacks can vary in their graph properties.

In our experiments, we found that the SQL injection and command line injection attacks were more difficult to detect than others from the datasets that were gathered. Considering the system execution, the kernel-level provenance and database-level provenance for both benign and attack scenarios are similar as both are ultimately SQL queries or file system reads/writes. It is evident from our experiments that single-layer provenance is not able to capture anomalies for all attack types; some attacks are more evident at different layers of the system execution. Currently, existing provenance capture tools have provided some anomaly detection capability, but there is a lack of tools for attack types which are not evident at the kernel level. Our results here are consistent with the postulation that multi-layer provenance would be more effective.

A similar phenomenon can be noted with our results in the stored cross-site scripting scenario. The attack behavior is a bot-style message. The attack itself is not much different from benign posting behavior, but the frequency of the message may be noted as anomalous in a graph learning system which models time-series information.

Our models performed best on the cross-site reflected and DOM scenarios. For the reflected threat model, the attack system execution behavior is distinctly different as it pulls up a pop-up window which does not exist in the benign version. Similarly, the DOM model also generates a pop-up window that does not exist in the benign scenario. Both of these generate unique provenance graph substructures which are detectable by the models as anomalous in the provenance layers we examined. A significant finding of our experiments is that whole-scope anomaly detection in provenance graphs will continue to require constant vigilance in multiple layers of the target system - no single layer solution will be sufficient to catch even a narrow variety of attacks.

## 7 DISCUSSION AND FUTURE WORK

The FLURRY system is a novel, initial contribution to automated provenance graph generation and machine learning and thus has room for testing and improvement. The system currently writes logs to files intermittently. This design is intentional as we do not provide graph summarization beyond the de-duplication that CamFlow [21] supports, and thus would overwhelm the downstream applications with log data if it were piped directly. Future versions

<sup>1</sup><https://github.com/NASCL/dgl-trainer.git>

will include a graph summarization element and the ability to work with streaming data as well as static graphs.

One aspect of browser-level provenance which we have not yet implemented in FLURRY is to capture and compare user entries into forms. We propose in future work to develop a browser provenance extension to FLURRY which will scrape these forms and use similarity metrics of routine entries as a potential learning feature. As a practical example, an injected SQL query will be quite dissimilar to an expected web form entry for a username. Locality sensitive hashing could provide both a distance measure that could improve model results for this attack type and obfuscation for user privacy. This experimental result and potential solution reveal the importance of multi-layer provenance, as different attacks will be more evident at different layers of the system.

In future versions of FLURRY, we intend to design a streaming graph approach to the pipeline in addition to segmented graphs in order to better capture attack rates. This approach has been explored in provenance graph learning systems [13, 18] and would be a useful addition to our automated framework. It gives an additional insight to data provenance that is temporal as opposed to solely spatial. This would also likely improve results in other frequency-based vectors like flooding in denial-of-service or brute-force attacks.

## 8 CONCLUSION

FLURRY advances the state of the art by extending both ways in the process to create a full data pipeline for provenance graph representation learning. First, we expand the user-friendliness and quality assurance of the traces by providing a GUI for ease of interaction with the attack scenarios. We also use a web driver to automate the attacks so that the researcher creating the data can watch the attack as it happens. The data is then captured by provenance tools pre-installed on our distributable VM image. While highly configurable, these tools are also ready at deployment to capture provenance at the click of a button. We provide additional tools to export the raw data or process it into provenance graphs. For maximum flexibility we provide a library tool for converting the raw logs into NetworkX or Deep Graph Library graphs, exportable JSON, or compressed and serialized formats which can be manually uploaded into other machine learning frameworks. Finally, we provide a collection of deep learning models for examination of these provenance graphs as they are generated. Our experiments and results reveal the capability of FLURRY as a provenance graph generation system from cyber attack to anomaly detection. We offer this system as a tool for the research community to continue to advance the state of the art in provenance graph representation learning.

## REFERENCES

- [1] Arunkumar Bagavathi and Siddharth Krishnan. 2019. Multi-Net: A Scalable Multiplex Network Embedding Framework. In *Complex Networks and Their Applications VII*, Luca Maria Aiello, Chantal Cherifi, Hocine Cherifi, Renaud Lambiotte, Pietro Lió, and Luis M. Rocha (Eds.). Springer International Publishing, Cham, 119–131.
- [2] Adam Bates, Dave Tian, Kevin R. B. Butler, and Thomas Moyer. 2015. Trustworthy Whole-System Provenance for the Linux Kernel. In *Proceedings of the 24th USENIX Conference on Security Symposium* (Washington, D.C.) (SEC'15). USENIX Association, USA, 319–334.
- [3] Mikhail Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* 15, 6 (2003), 1373–1396.
- [4] Ginestra Bianconi. 2018. *Multilayer Networks: Structure and Function*. Oxford university press.
- [5] digininja. 2020. Damn Vulnerable Web Application. <https://github.com/digininja/DVWA>.
- [6] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *SIGKDD*. 135–144.
- [7] Jed A Duersch, Meiyue Shao, Chao Yang, and Ming Gu. 2018. A robust and efficient implementation of LOBPCG. *SIAM Journal on Scientific Computing* 40, 5 (2018), C655–C676.
- [8] Ashish Gehani and Dawood Tariq. 2012. SPADE: Support for Provenance Auditing in Distributed Environments. In *Middleware 2012*, Priya Narasimhan and Peter Triantafillou (Eds.). Lecture Notes in Computer Science, Vol. 7662. Springer Berlin Heidelberg, 101–120. [https://doi.org/10.1007/978-3-642-35170-9\\_6](https://doi.org/10.1007/978-3-642-35170-9_6)
- [9] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.
- [10] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*. 1024–1034.
- [11] Xueyuan Han. 2018. StreamSpot Dataset. <https://doi.org/10.7910/DVN/83KYJY>
- [12] Xueyuan Han, James Mickens, Ashish Gehani, Margo I. Seltzer, and Thomas F. J.-M. Pasquier. 2020. Xanthus: Push-button Orchestration of Host Provenance Data Collection. *CoRR abs/2005.04717* (2020). arXiv:2005.04717 <https://arxiv.org/abs/2005.04717>
- [13] Xueyuan Han, Thomas F. J.-M. Pasquier, Adam Bates, James Mickens, and Margo I. Seltzer. 2020. UNICORN: Runtime Provenance-Based Detector for Advanced Persistent Threats. *CoRR abs/2001.01525* (2020). arXiv:2001.01525 <http://arxiv.org/abs/2001.01525>
- [14] Xueyuan Han, Xiao Yu, Thomas Pasquier, Ding Li, Junghwan Rhee, James Mickens, Margo Seltzer, and Haifeng Chen. 2021. SIGL: Securing Software Installations Through Deep Graph Learning. arXiv:2008.11533 [cs.CR]
- [15] Maya Kapoor, Joshua Melton, Michael Ridenhour, Siddharth Krishnan, and Thomas Moyer. 2021. PROV-GEM: Automated Provenance Analysis Framework using Graph Embeddings. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 1720–1727. <https://doi.org/10.1109/ICMLA52953.2021.00273>
- [16] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.
- [17] Zhijun Liu, Chao Huang, Yanwei Yu, Baode Fan, and Junyu Dong. 2020. Fast Attributed Multiplex Heterogeneous Network Embedding. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 995–1004.
- [18] Emaad A. Manzoor, Sadegh Momeni, Venkat N. Venkatakrishnan, and Leman Akoglu. 2016. Fast Memory-efficient Anomaly Detection in Streaming Heterogeneous Graphs. arXiv:1602.04844 [cs.SI]
- [19] Neo4j. 2012. Neo4j - The World's Leading Graph Database. <http://neo4j.org/>
- [20] NetworkX developer team. 2014. NetworkX. <https://networkx.github.io/>
- [21] Thomas F. J.-M. Pasquier, Xueyuan Han, Mark Goldstein, Thomas Moyer, David M. Eysers, Margo I. Seltzer, and Jean Bacon. 2017. Practical Whole-System Provenance Capture. *CoRR abs/1711.05296* (2017). arXiv:1711.05296 <http://arxiv.org/abs/1711.05296>
- [22] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.
- [23] Devin J. Pohly, Stephen McLaughlin, Patrick McDaniel, and Kevin Butler. 2012. Hi-Fi: Collecting High-Fidelity Whole-System Provenance. In *Proceedings of the 28th Annual Computer Security Applications Conference* (Orlando, Florida, USA) (ACSAC '12). Association for Computing Machinery, New York, NY, USA, 259–268. <https://doi.org/10.1145/2420950.2420989>
- [24] Positive Technologies. 2019. Attacks on web applications: 2018 in review. <https://www.ptsecurity.com/ww-en/analytics/web-application-attacks-2019/>
- [25] Positive Technologies. 2020. What is a cross-site scripting (XSS) attack? <https://www.ptsecurity.com/ww-en/analytics/knowledge-base/what-is-a-cross-site-scripting-xss-attack/#5>
- [26] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web*. Springer International Publishing, 593–607.
- [27] Andreas Stathopoulos and Kesheng Wu. 2002. A block orthogonalization procedure with constant synchronization requirements. *SIAM Journal on Scientific Computing* 23, 6 (2002), 2165–2182.
- [28] Kali Linux Tools. 2021. Hydra. <https://www.kali.org/tools/hydra/>
- [29] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lió, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.
- [30] W3C-PROV Working Group. 2013. PROV-Overview: An Overview of the PROV Family of Documents. <https://www.w3.org/TR/prov-overview/>
- [31] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, Ziyue Huang, Qipeng Guo, Hao Zhang, Haibin Lin, Junbo Zhao, Jinyang Li, Alexander J. Smola, and Zheng Zhang. 2019. Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs. *CoRR*



abs/1909.01315 (2019). arXiv:1909.01315 <http://arxiv.org/abs/1909.01315>

- [32] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *WWW. 2022–2032*.
- [33] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *International Conference on Learning Representations*.

## A APPENDIX

### A.1 Application Level Sample Provenance Graphs

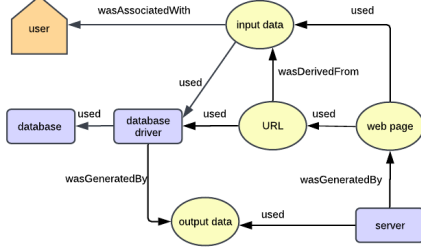


Figure 3: Provenance graph for SQL injection attack.

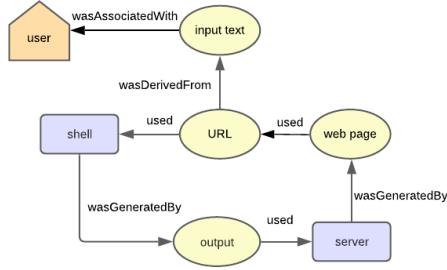


Figure 4: Provenance graph for command line injection attack.

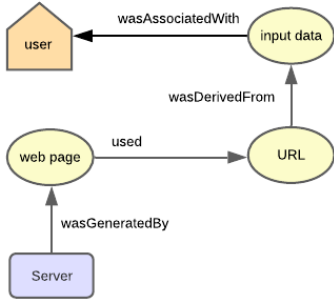


Figure 5: Provenance graph for cross site scripting attack.

### A.2 Implementation Details

The experiments were conducted on a 64core AMD EPYC 7742 CPU @ 2.25 GHz with an NVIDIA A100 GPU. Our models are implemented using PyTorch 1.10.2 and DGL 0.7.2 in Python 3.9. We use our Trainer and TrainingArguments classes to specify and define

training and evaluation behavior, and we use DGL to implement our message passing graph neural network models. Evaluation metrics are computed using scikitlearn *f1\_score*, *precision\_score*, and *recall\_score*. Experiments were conducted using five-fold cross validation with metrics reported as the mean and standard deviation averaged over the five folds.

*Parameter Configuration.* We set the hidden and embedding dimensions for all models to 256 dimensions for all experiments, and the attention dimension for HAN and PROV-GEm are set to 64 dimensions. A dropout out rate of 0.5 and L2 regularization weight decay of 0.005 were used as regularization during model training. Table 3 defines the individual training parameters used for each model for each experiment with the six attack-type datasets generated by FLURRY.

### A.3 Learning Parameters

Attack Type	Model	Epochs	Aggregation	Learn Rate
brute-force	HAN	20	sum	0.001
	R-GCN	10	sum	0.001
	PROV-GEm	20	sum	0.001
cl-injection	HAN	7	sum	0.0005
	R-GCN	7	sum	0.0005
	PROV-GEm	7	sum	0.0005
sql-injection	HAN	5	sum	0.0005
	R-GCN	20	mean	0.0005
	PROV-GEm	10	sum	0.0005
xss-dom	HAN	20	sum	0.001
	R-GCN	10	sum	0.0005
	PROV-GEm	20	sum	0.001
xss-reflected	HAN	20	sum	0.001
	R-GCN	20	mean	0.0005
	PROV-GEm	20	sum	0.001
xss-stored	HAN	7	sum	0.005
	R-GCN	7	sum	0.001
	PROV-GEm	7	sum	0.005

Table 3: Training Parameters for Graph Learning Models