More Details of the Datasets

In the Banking dataset, the data of April, May, and June 2017 are selected. The Banking dataset contains B2C transactional records for consecutive months, which are either labeled positive (fraudulent) or negative (normal). In the original dataset, each transaction is characterized by 64 features, where most of them have sparsely valid values (about 10% to 30% on average). We filter out the features with sparse data and then screen out some fields with few correlations to frauds by assessing the feature importance. Finally, we choose 8 features as the properties in the dataset to build our model, which have the highest correlation with fraud detection.

In the Kyoto dataset, the data of 27, 28, 29, 30, and 31 August 2009 are selected. The Kyoto dataset was collected from both honeypots and regular servers that were deployed at Kyoto University. Each network traffic data in this dataset has 24 different features. We also adopt 8 features as the properties for our method.

In the Darknet dataset, the data were captured regular, VPN and Tor traffic for seven diverse categories under respective applications. All the data is labelled as benign to present regular traffic or malicious to represent anonymous (Tor or VPN) traffic related to hidden services provided by darknet. The data provider extracted 83 features through CICFlowMeter, and gave the importance of those features. For this work, we only select the top 9 features of importance, so each event consists of 9 properties in our property graph.

In the Gowalla dataset, the data of June 2010 are selected. The Gowalla dataset contains the data of global users who share their locations by checking in social networking websites. For this work, we only select check-in records located in the United States. Each record consists of 5 properties.

In the CICDDoS dataset, all attacks were carried out by using TCP/UDP based protocols at the application layer. The data provider profile the abstract behaviour of human interactions and generate naturalistic benign background traffic, so it resembles the true real-world data. More than 80 traffic features are extracted by using the CICFlowMeter. In this work, we focus on detecting the NetBIOS and LDAP attacks, so we choose the first five features highly related to identifying anomalies for the two attacks.

In the CERT dataset, the data generator released multiple version (e.g., r3.1 and r3.2). Generally, later releases include a superset of the data generation functionality of earlier releases. We use r4.2 version of the synthetic dataset in this work, which contains behavior data for 1000 users over a period of 1.5 years.

The details of properties are shown in TABLE I.

TABLE I: The selected properties in five datasets.

Dataset	Property	Type	Continuous	Description
Banking	account id	Int	No	Each account id represents a user's account.
	merchant id	Int	No	Each merchant id represents a merchant in a B2C transaction.
	place id	Int	No	Each place id represents an issuing area of banking cards used for transactions.
	time	String	Yes	The exact time when the transaction occurred.
	amount	Float	Yes	The amount of money transferred to the merchant in a B2C transaction.
	ip	String	No	Whether a commonly used IP address or not in a transaction.
	last result	Boolean	No	Judgment of the last transaction in the relevant account id.
	type	String	No	Each type represents a transaction of different type.
Kyoto	service	String	No	The connection's service type, e.g., http, telnet, etc.
	destination bytes	Int	Yes	The number of data bytes sent by the destination IP.
	count	Int	Yes	The number of connections whose source IP address and destination IP address are the same to those of the current connection in the past two seconds.
	same srv rate	Float	Yes	% of connections to the same service in count property address.
	serror rate	Float	Yes	% of connections that have 'SYN' errors in count property.
	srv serror rate	Float	Yes	% of connections that have 'SYN' errors in srv count (the number of connections whose service type is the same to that of the current connection in the past two seconds) property.
	dst host srv count	Int	Yes	Among the past 100 connections whose destination IP address is the same to that of the current connection, the number of connections whose service type is also the same to that of the current connection.
	dst host serror rate	Float	Yes	% of connections that have 'SYN' errors in dst host count property.
Darknet	idle max	Int	Yes	The maximum amount of time time a flow was idle before becoming active.
	fwd seg size min	Int	Yes	The minimum size of segment in forward direction.
	bwd packet length min	Int	Yes	The minimum length of packet in backward direction.
	protocol	String	No	The captured protocols of network flow.
	idle mean	Int	Yes	The average amount of time time a flow was idle before becoming active.
	fwd init win	Int	Yes	The number of bytes in initial window in forward direction.
	fin flag count	Int	Yes	The number of flag FIN.
	subflow bwd bytes	Int	Yes	The number of bytes in a sub flow in backward direction.
	bwd packet length max	Int	Yes	The maximum length of packet in backward direction.
Gowalla	user	Int	No	The identification of a user check-in.
	check-in time	String	Yes	The occurrence time of a user check-in.
	latitude	Float	Yes	The latitude of a user check-in geolocation.
	longitude	Float	Yes	The longitude of a user check-in geolocation.
	location id	Int	No	The location identifier for a user check-in location.
CICDDoS	max packet length	Int	Yes	The maximum length of packet.
	fwd packet length max	Int	Yes	The maximum length of packet in forward direction.
	fwd packet length min	Int	Yes	The minimum length of packet in forward direction.
	average packet size	Int	Yes	The average size of packet.
	min packet length	Int	Yes	The minimum length of packet.
	fwd packets/s	Float	Yes	The average rate of packets in forward direction.
	min seq size forward	Int	Yes	The minimum size of segment in forward direction.
	protocol	String	No	The captured protocols of network flow.
	fwd header length	Int	Yes	The length of header in forward direction.
	fwd header length.1	Int	Yes	The length.1 of header in forward direction.
CERT	user	String	No	The identification of a user.
	weekday	Int	Yes	The day of the week when the user performed the operation.
	hour	Int	Yes	The hour of the day when the user performed the operation.
	pc	String	No	The device used by the user to perform the operation.
	type	String	No	The type of operation performed by the user.
	state	String	No	The active state of the operation performed by the user.