Demystifying Fraudulent Transactions and Illicit Nodes in the Bitcoin Network for Financial Forensics

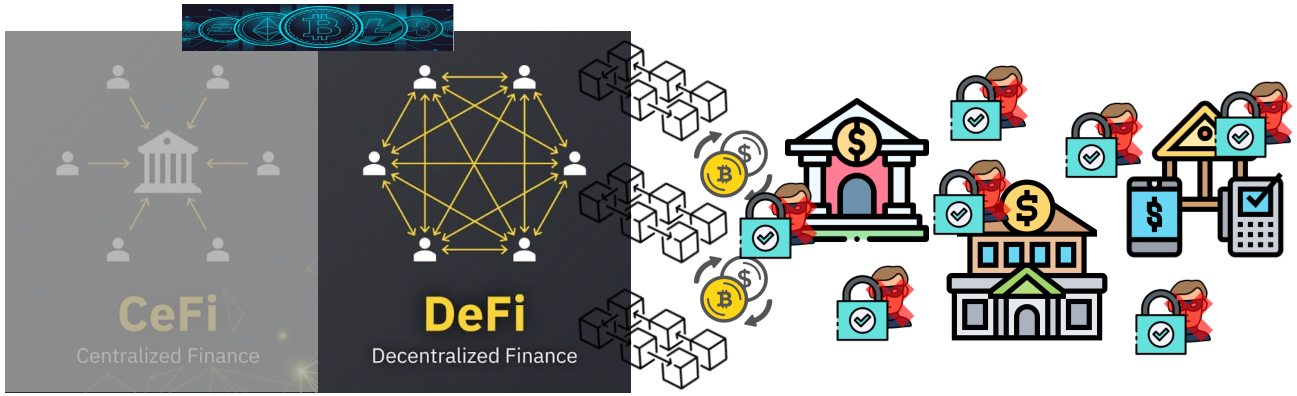
Youssef Elmougy* and Ling Liu

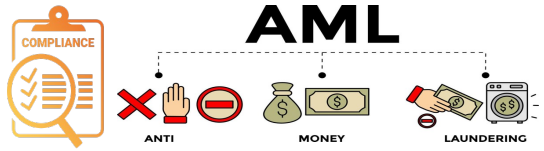
Georgia Institute of Technology, Atlanta GA USA

* Corresponding Author and Presenting Author

Georgia Tech

The Catalyst for Cryptocurrency Integration into Traditional Financial Institutions

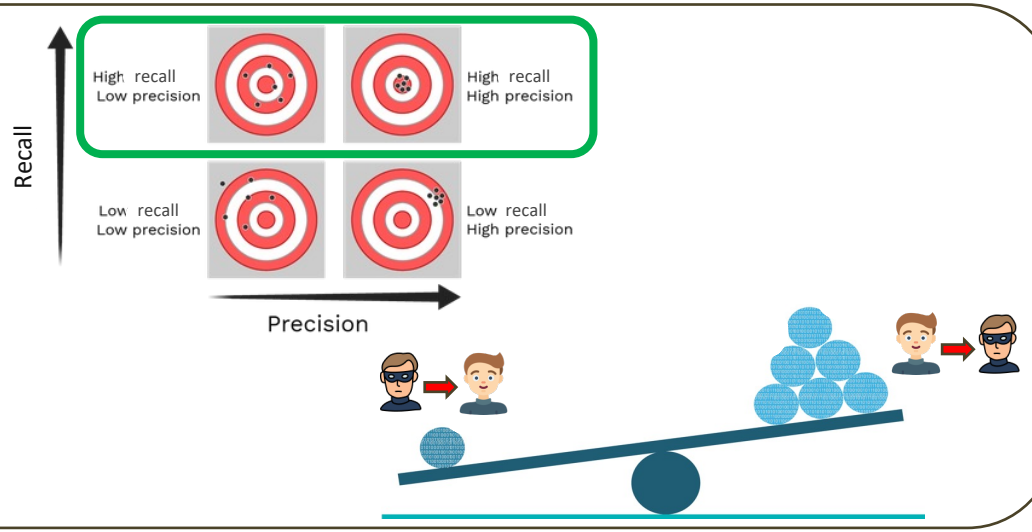




Fraud detection models trained for financial forensics should:

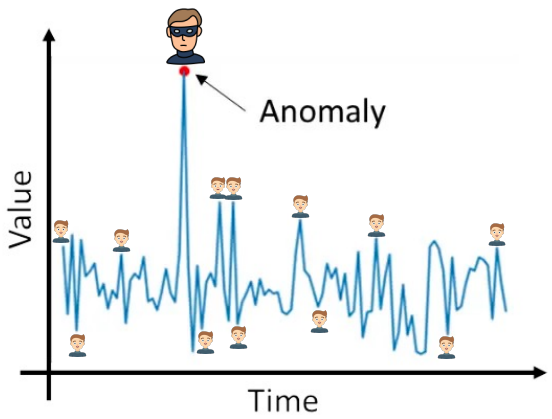
1

PRIORITIZING HIGH RECALL



2

EXPLAINABILITY AND PROOF



Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Downfalls of Available Financial Forensics Datasets

Largest labelled Bitcoin transaction dataset currently publicly available:

Elliptic Data Set

Bitcoin Transaction Graph

Consists of over $203k$ transactions labelled illicit, licit, and unknown

Available at:
https://www.kaggle.com/datasets/ellipticco/elliptic-data-set

✖ Elliptic Data Set consists of only Bitcoin transactions, without features of the addresses involved and the different interactions between pairs of addresses

Prominent downfall: when a model predicts an illicit transaction, the addresses responsible cannot be clearly identified since a transaction may be associated with several input and output addresses

GIVEN THIS MOTIVATION, WE MAKE TWO CONTRIBUTIONS

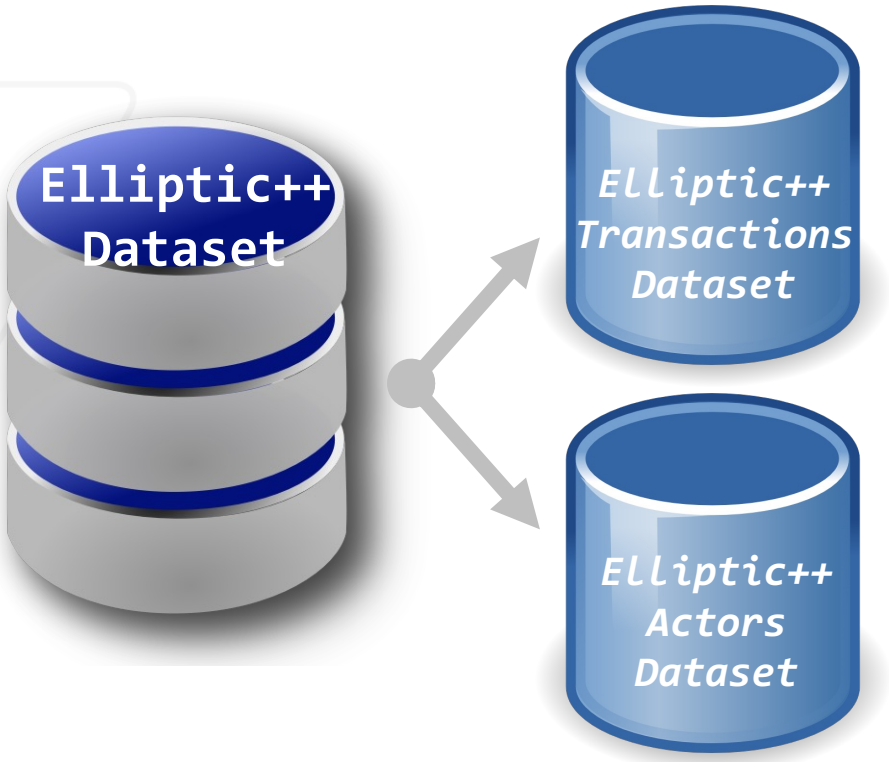
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

1st Contribution: The Elliptic++ Dataset

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses



- $203k$ transactions
- $234k$ payment flows
- 183 transaction features
- Labels: **illicit**, **licit**, **unknown**

- $822k$ wallet addresses
- $1.27M$ temporal occurrences
- 56 address features
- Labels: **illicit**, **licit**, **unknown**

Forensics performed using The Elliptic++ Dataset:

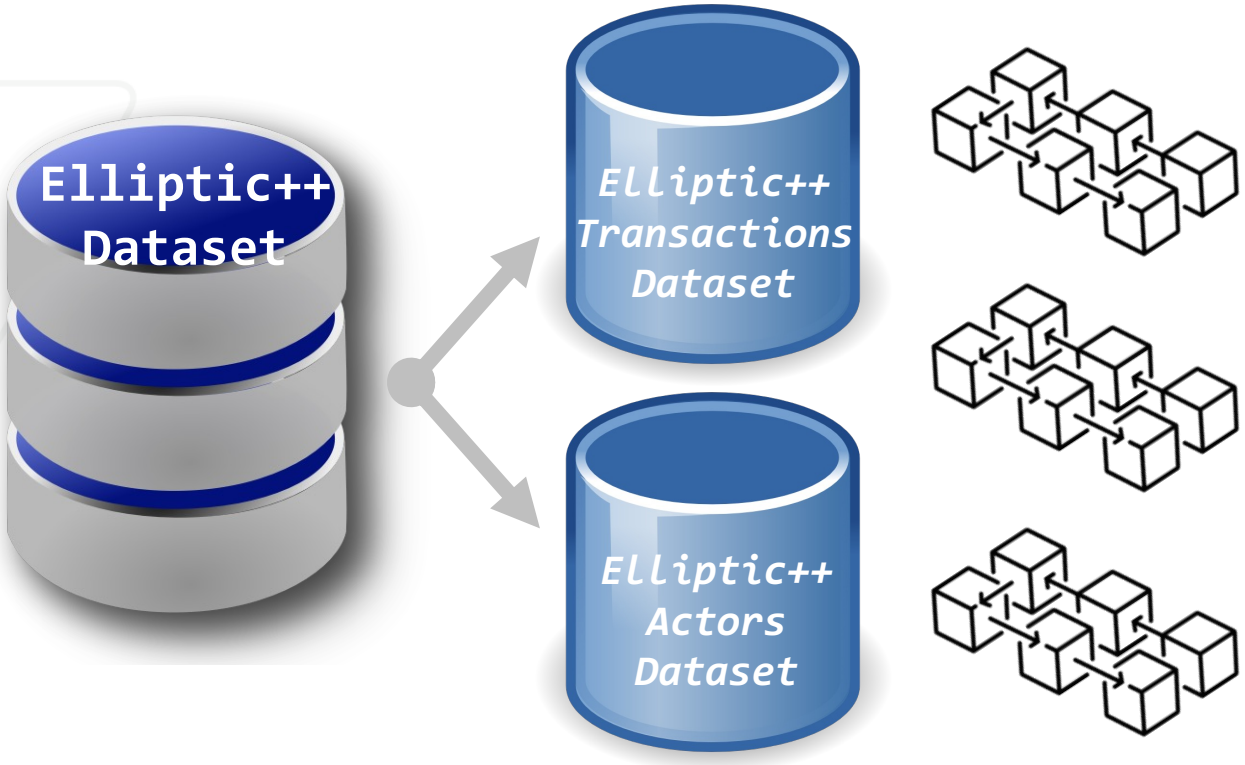
- Identifies *fraudulent transactions* and *dishonest actors* in the Bitcoin network
- Allows *explainability* as to why a wallet address is illusive
- Characterizes *addresses as the center* of the risk detection model

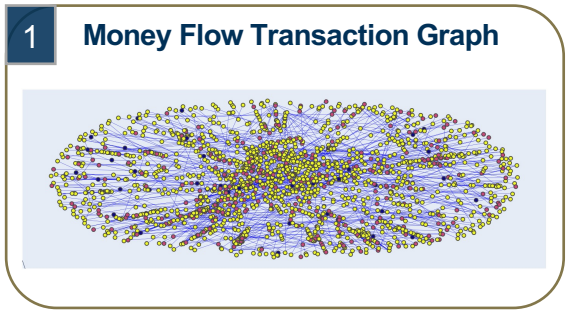
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

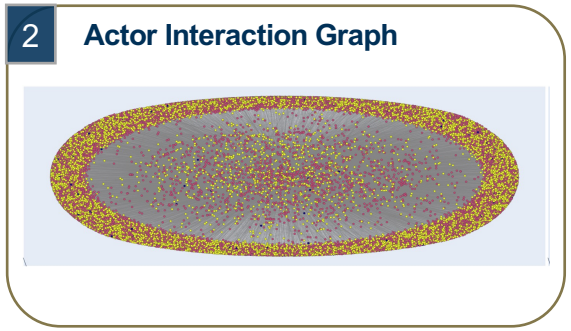
Georgia Tech

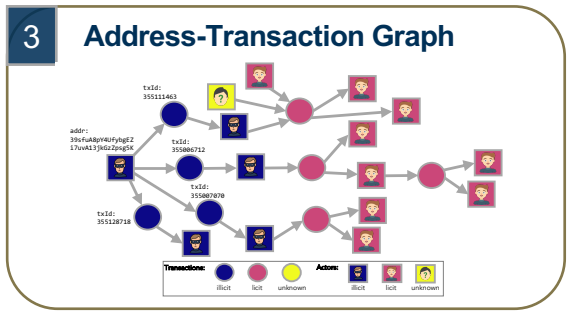
1st Contribution: The Elliptic++ Dataset

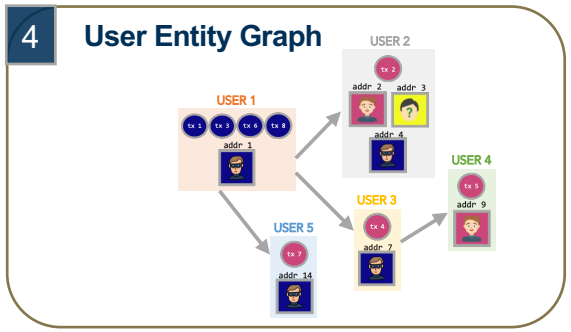
Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses











Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Georgia Tech

2nd Contribution: Fraud Detection using The Elliptic++ Dataset

Fraud detection by combining diverse ML algorithms and feature optimizations

Model	Precision	Recall	F1 Score	Micro-F1
RF^{TX}	0.975	0.719	0.828	0.980
$RF^{TX\psi}$	**0.986**	**0.727**	**0.836**	**0.981**
$RF+XGB^{TX}$	0.977	0.706	0.820	0.979
$RF+XGB^{TX\psi}$	**0.987**	**0.717**	**0.826**	**0.980**
$RF+MLP+XGB^{TX}$	0.962	0.723	0.826	0.980
$RF+MLP+XGB^{TX\psi}$	**0.968**	**0.729**	**0.834**	**0.980**

Model	Precision	Recall	F1 Score	Micro-F1
RF^{AR}	0.911	0.789	0.845	0.990
$RF^{AR\psi}$	**0.921**	**0.802**	**0.858**	**0.990**
$RF+XGB^{AR}$	0.959	0.530	0.682	0.982
$RF+XGB^{AR\psi}$	**0.967**	**0.543**	**0.686**	**0.982**
$RF+MLP+XGB^{AR}$	0.933	0.572	0.709	0.983
$RF+MLP+XGB^{AR\psi}$	**0.945**	**0.601**	**0.718**	**0.984**

Allows for in-depth understanding of the *root cause of fraudulent activities* in cryptocurrency transactions through *semantic and statistical explainability*, shining light on the strategies for fraud detection and prevention

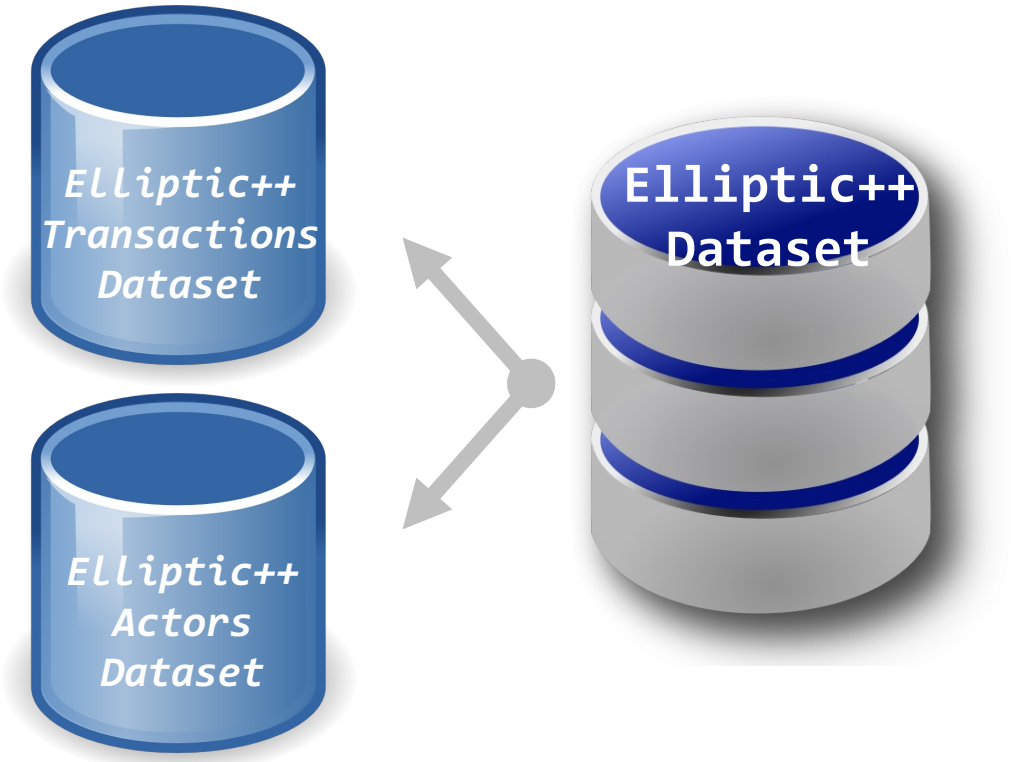
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

The Elliptic++ Dataset

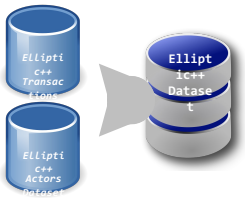
Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses



Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

The Elliptic++ Dataset: Collection Pipeline

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses



txs_edgelist

txId1	txId2
...	...
272145560	296926618
272145560	272145556
299475624	272145560
...	...

txs_features

txId	Time step	LF_1	...	AF_72
...	...	...	...	...
272145560	24	-0.155493	...	1.521399
...	...	...	...	...

PUBLIC DATASET — Elliptic Dataset

search edgelist with txId

PUBLIC DATASET — Txs De-anonymized Dataset

txId	txHash
...	...
272145560	56cba8aa85f90e1d8a703cc868dc79c11 3b289050da2958a64214e1efa5dd898
...	...

scrape blockchain with txHash

OUR BLOCKCHAIN TRANSACTION PARSER AND EXTRACTOR

/btc/address/, /btc_in/, /btc_out/,
/address_in/, /address_out/,
/btc_total/, /tx_fees/, /tx_size/

calculate TXS_in and TXS_out using edgelist

txId	Time step	LF_1	...	AF_72	TXS_in	TXS_out	...	BTC_out_total
...	...	...	...	...	...	...	...	...
272145560	24	-0.155493	...	1.521399	1	...	...	2.77279994
...	...	...	...	...	...	...	...	...

Elliptic++ Transactions Dataset

grab all addresses

txId	address
...	...
272145560	39sfuA8pY4UfybgEZi7uvA13jkGzZpsg5K
272145560	1MLvzsdbo2N12ANUcGTXkMCQ1GZbLTvkTL
...	...

scrape blockchain with address

OUR BLOCKCHAIN ADDRESS PARSER AND EXTRACTOR

/btc_sent/, /btc_received/,
/tx_fees/, /tx_fees_share/, /txs_total/,
/blocks_txs/, / blocks_input/, /blocks_output/,
/blocks_first/, /lifetime/, /blocks_last/,
/blocks_first_sent/, /blocks_first_receive/

address	class	TXS_input	...	Addr_interactions_median
...	...	...	...	...
39sfuA8pY4UfybgEZi7uvA13jkGzZpsg5K	1	420	...	1
1MLvzsdbo2N12ANUcGTXkMCQ1GZbLTvkTL	1	1	...	1
...	...	...	...	...

Elliptic++ Actors Dataset

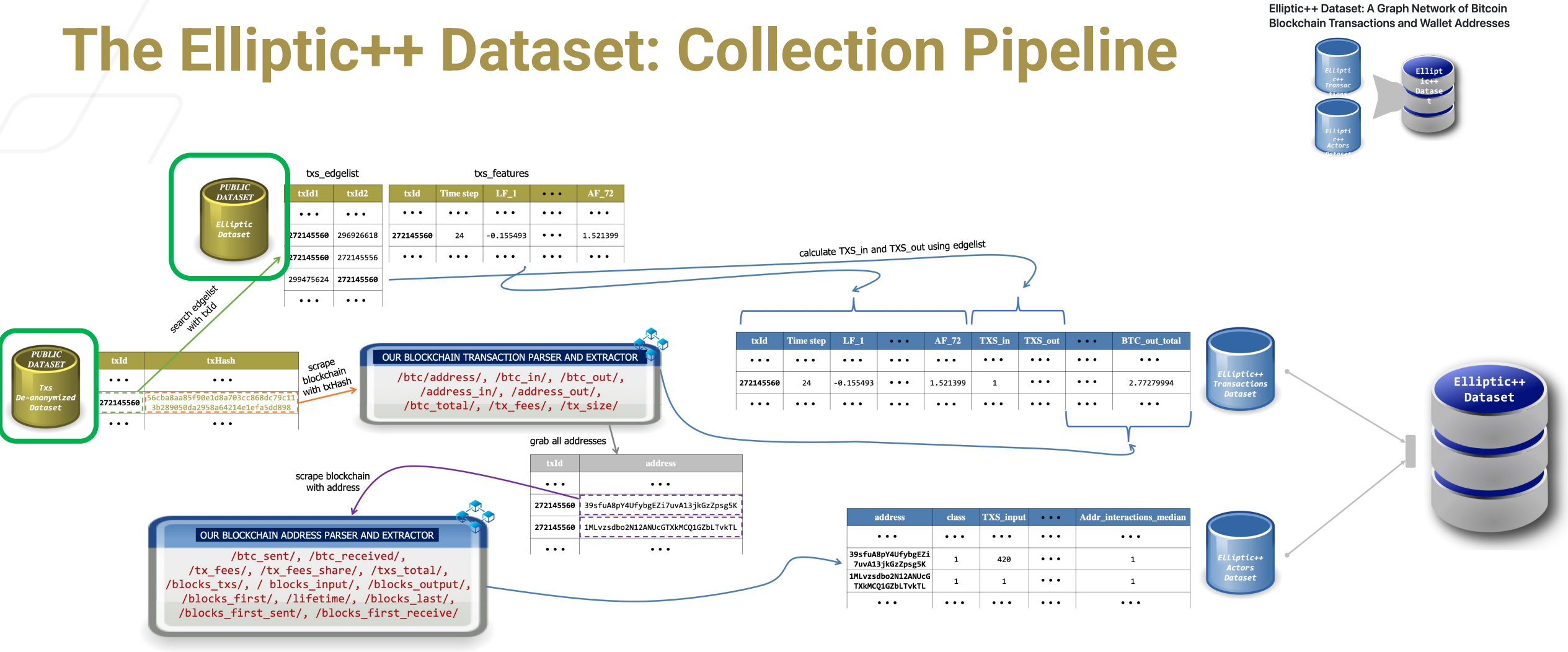
Elliptic++ Dataset

Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

The Elliptic++ Dataset: Collection Pipeline

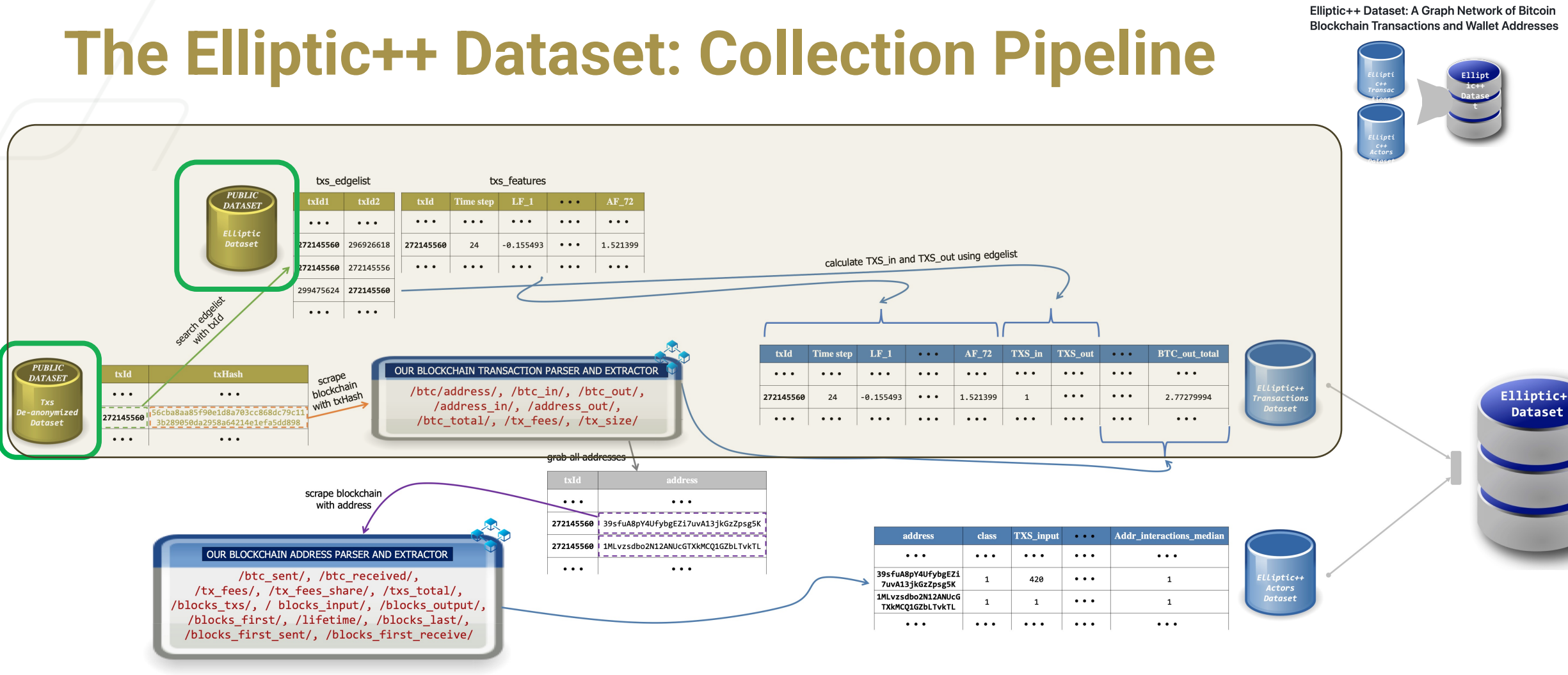


Elliptic Dataset: www.kaggle.com/datasets/ellipticco/elliptic-data-set
Txs De-anonymized Dataset: www.kaggle.com/datasets/alexbenzik/deanonymized-995-pct-of-elliptic-transactions
Bitcoin Blockchain: www.blockchain.com

Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Georgia Tech

The Elliptic++ Dataset: Collection Pipeline



Elliptic Dataset: www.kaggle.com/datasets/ellipticco/elliptic-data-set
Txs De-anonymized Dataset: www.kaggle.com/datasets/alexbenzik/deanonymized-995-pct-of-elliptic-transactions
Bitcoin Blockchain: www.blockchain.com

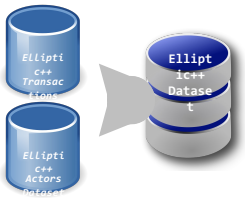
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

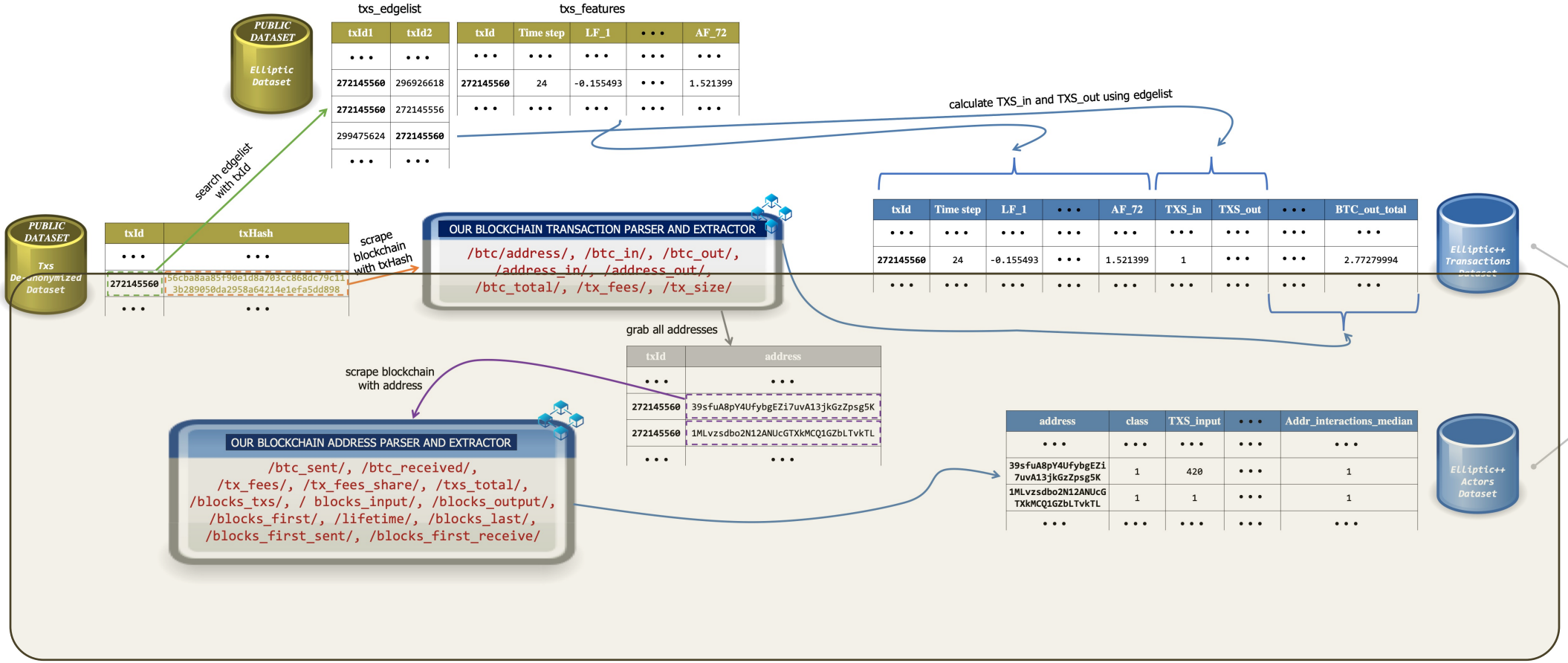
Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

The Elliptic++ Dataset: Collection Pipeline

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses





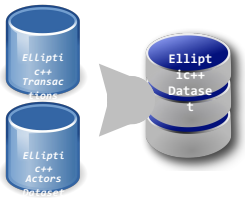
Elliptic Dataset: www.kaggle.com/datasets/ellipticco/elliptic-data-set
Txs De-anonymized Dataset: www.kaggle.com/datasets/alexbenzik/deanonymized-995-pct-of-elliptic-transactions
Bitcoin Blockchain: www.blockchain.com

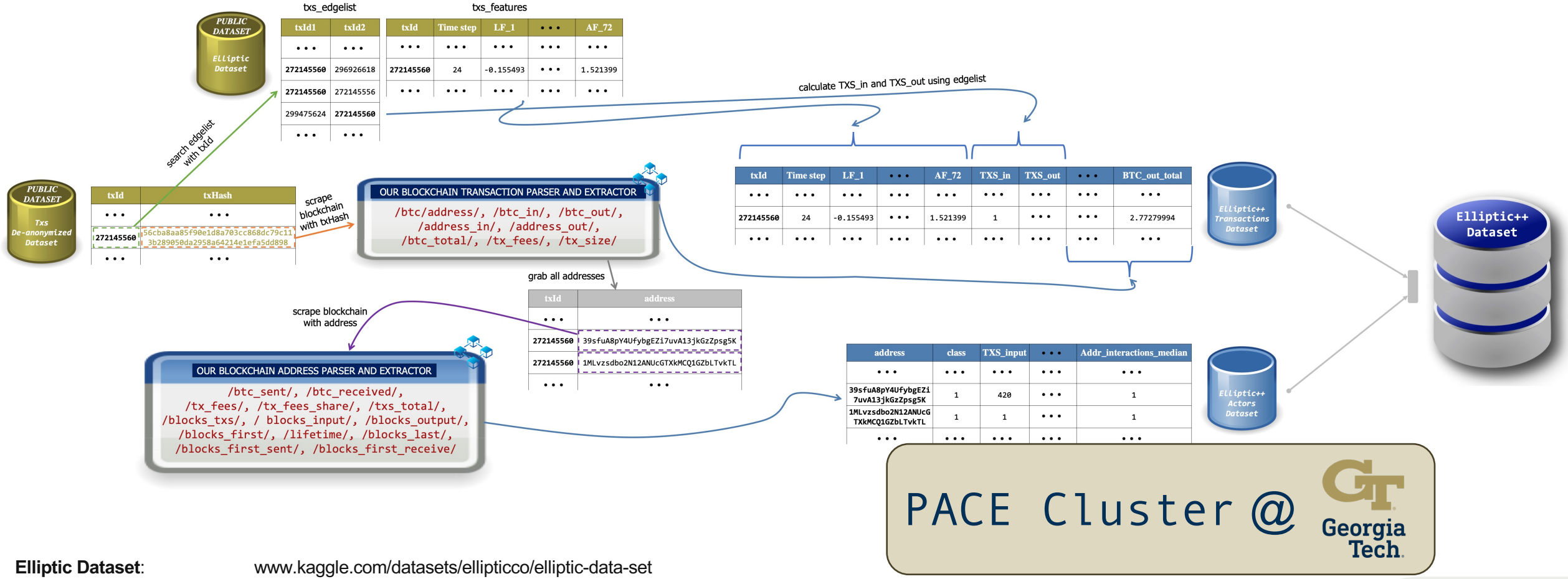
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

The Elliptic++ Dataset: Collection Pipeline

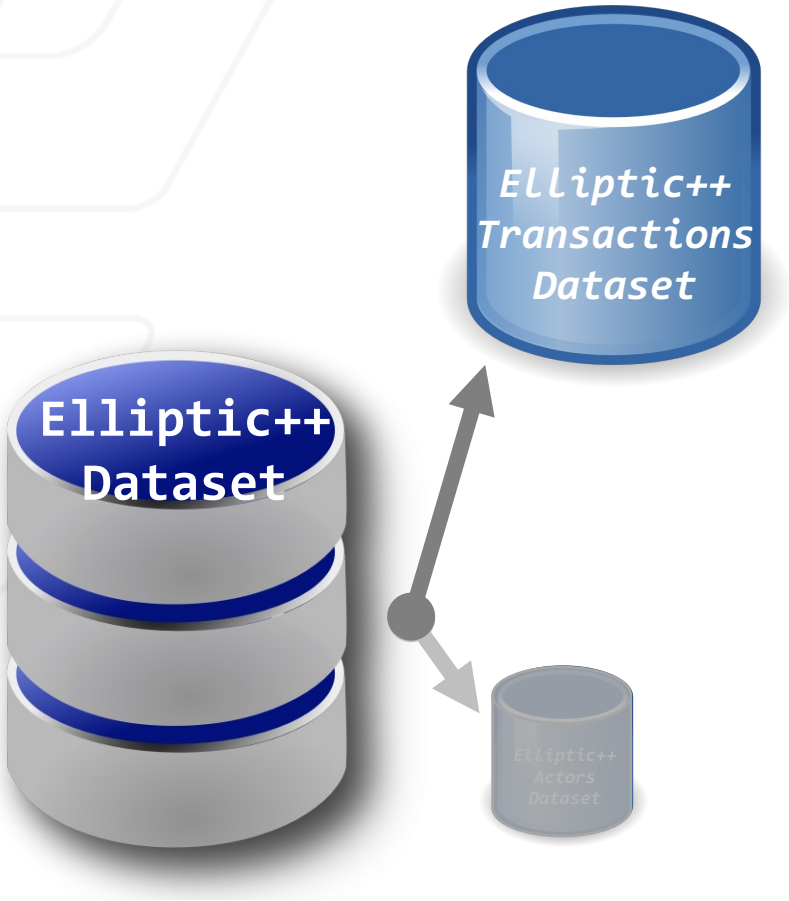




Elliptic Dataset: www.kaggle.com/datasets/ellipticco/elliptic-data-set
Txs De-anonymized Dataset: www.kaggle.com/datasets/alexbenzik/deanonymized-995-pct-of-elliptic-transactions
Bitcoin Blockchain: www.blockchain.com

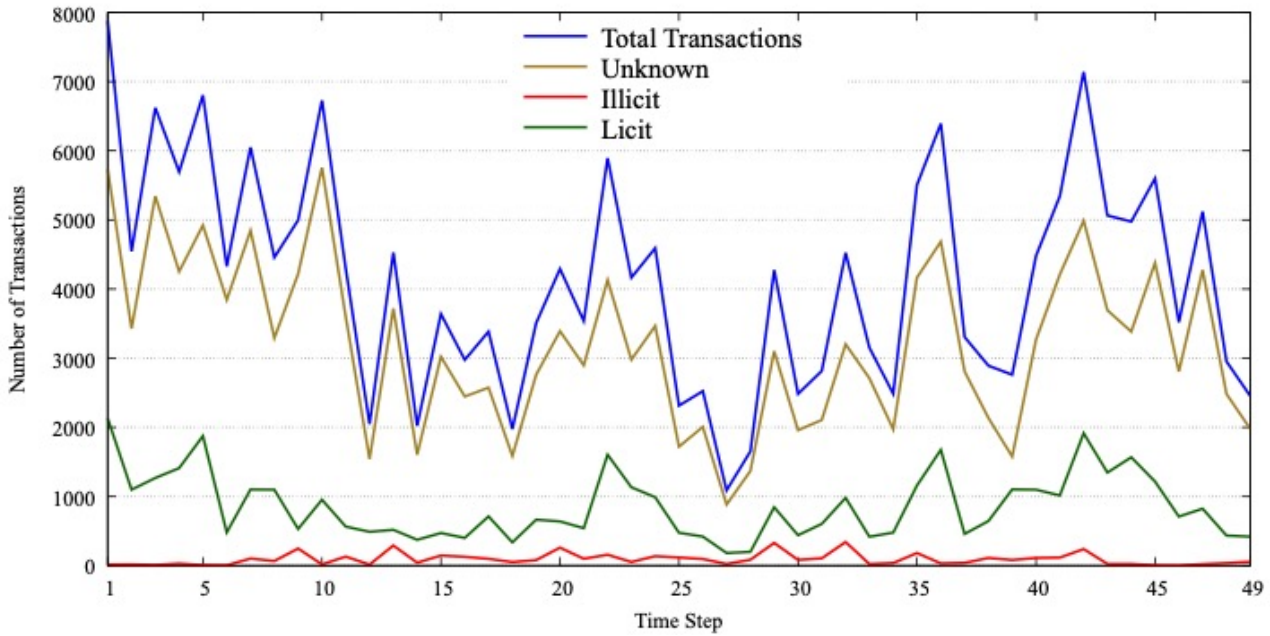
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

The Elliptic++ Dataset: Transactions Dataset



- **49** time steps
- **203,769** transactions
- **234,355** payment flows
- **183** transaction features

- Labels:
  - **2% illicit**
  - **21% licit**
  - **77% unknown**

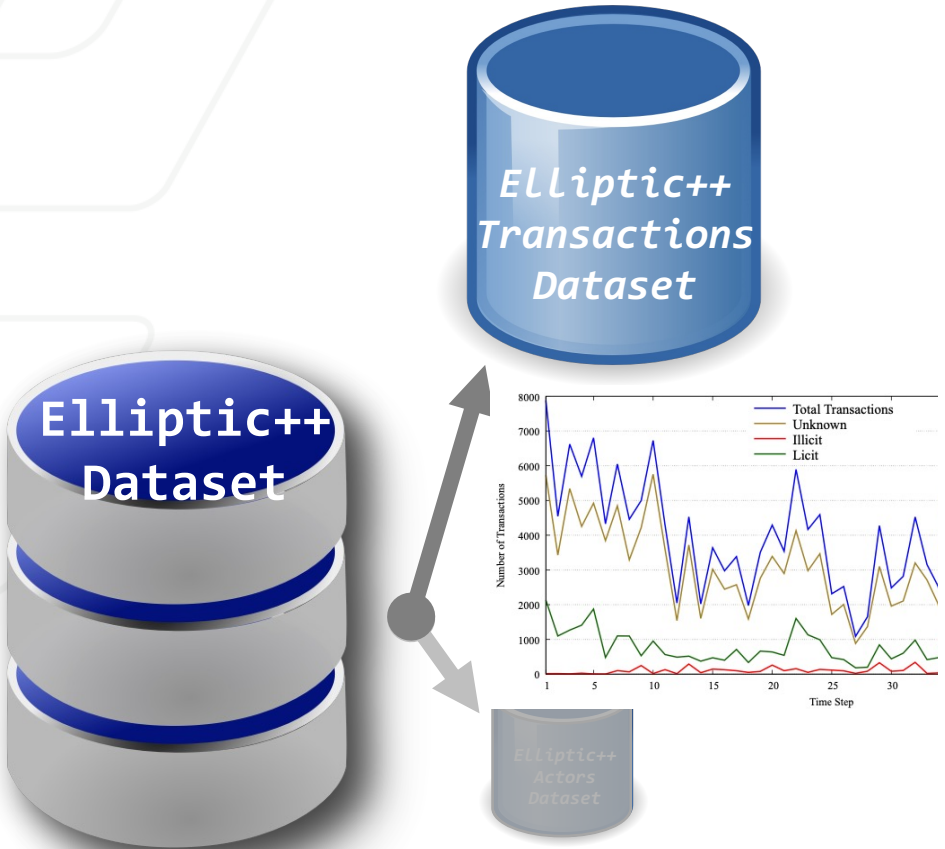


Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

The Elliptic++ Dataset: Transactions Dataset



- **49** time steps
- **203,769** transactions
- **234,355** payment flows
- **183** transaction features

- Labels:
  - **2% illicit**
  - **21% licit**
  - **77% unknown**

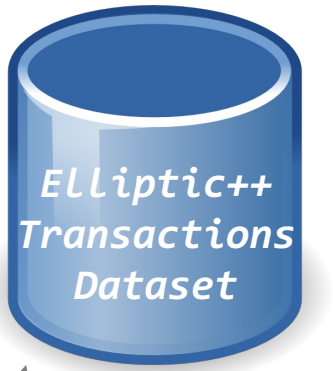
Feature	Description
BTC_{in}	Total BTC incoming
BTC_{out}	Total BTC outgoing
each has 5 values:	**total, min, max, mean, median**
Txs_{in}	Number of incoming transactions
Txs_{out}	Number of outgoing transactions
$Addr_{in}$	Number of input addresses
$Addr_{out}$	Number of output addresses
BTC_{total}	Total BTC transacted
$Fees$	Total fees in BTC
$Size$	Total transaction size
	single value

Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

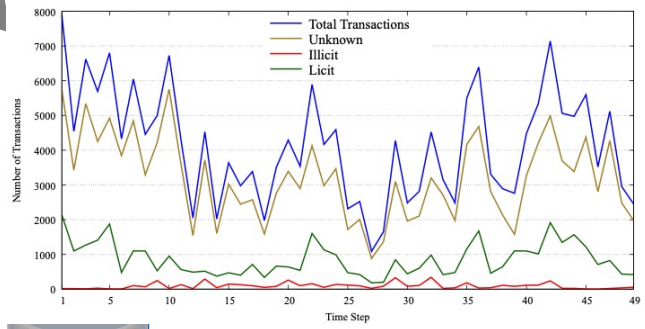
Georgia Tech

The Elliptic++ Dataset: Transactions Dataset



- **49** time steps
- **203,769** transactions
- **234,355** payment flows
- **183** transaction features

- Labels:
  - **2% illicit**
  - **21% licit**
  - **77% unknown**



Feature	Description
BTC_{in}	Total BTC incoming
BTC_{out}	Total BTC outgoing
each has 5 values:	**total, min, max, mean, median**
Txs_{in}	Number of incoming transactions
Txs_{out}	Number of outgoing transactions
$Addr_{in}$	Number of input addresses
$Addr_{out}$	Number of output addresses
BTC_{total}	Total BTC transacted
Fees	Total fees in BTC
Size	Total transaction size
	single value

txs_features.csv

txId	Time step	LF_1	···	LF_93	AF_1	···	AF_72	TXS_in	···	BTC_out_total
272145560	24	-0.155493	···	1.135279	-0.159681	···	1.521399	1	···	2.77279994

txs_edgelist.csv

txId1	txId2
272145560	296926618
272145560	272145556
299475624	**272145560**

txs_classes.csv

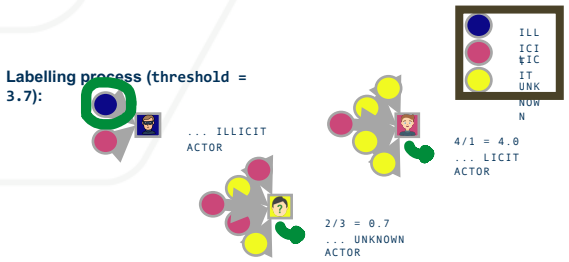
txId	class
272145560	1

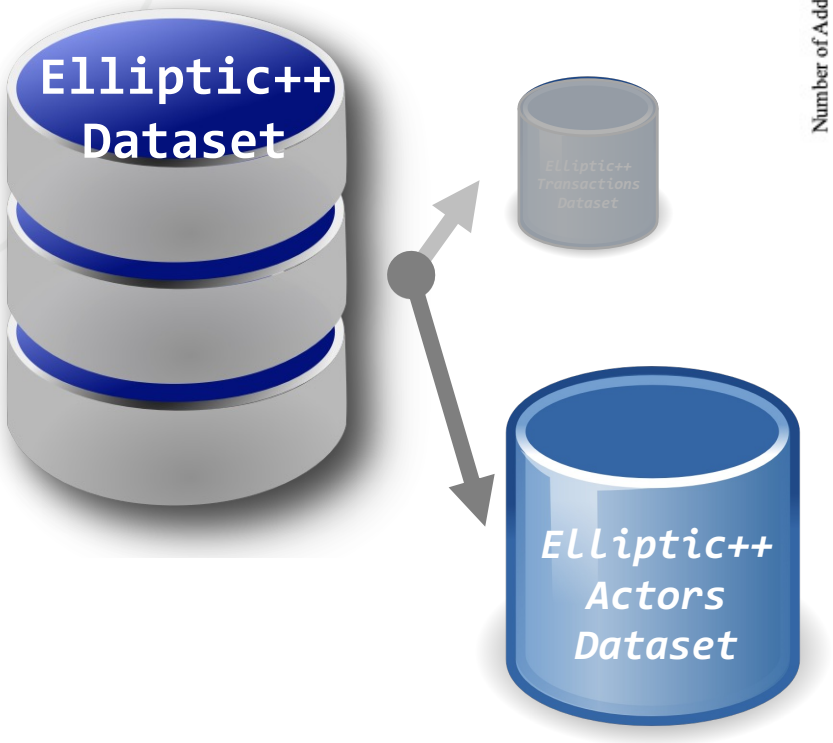
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

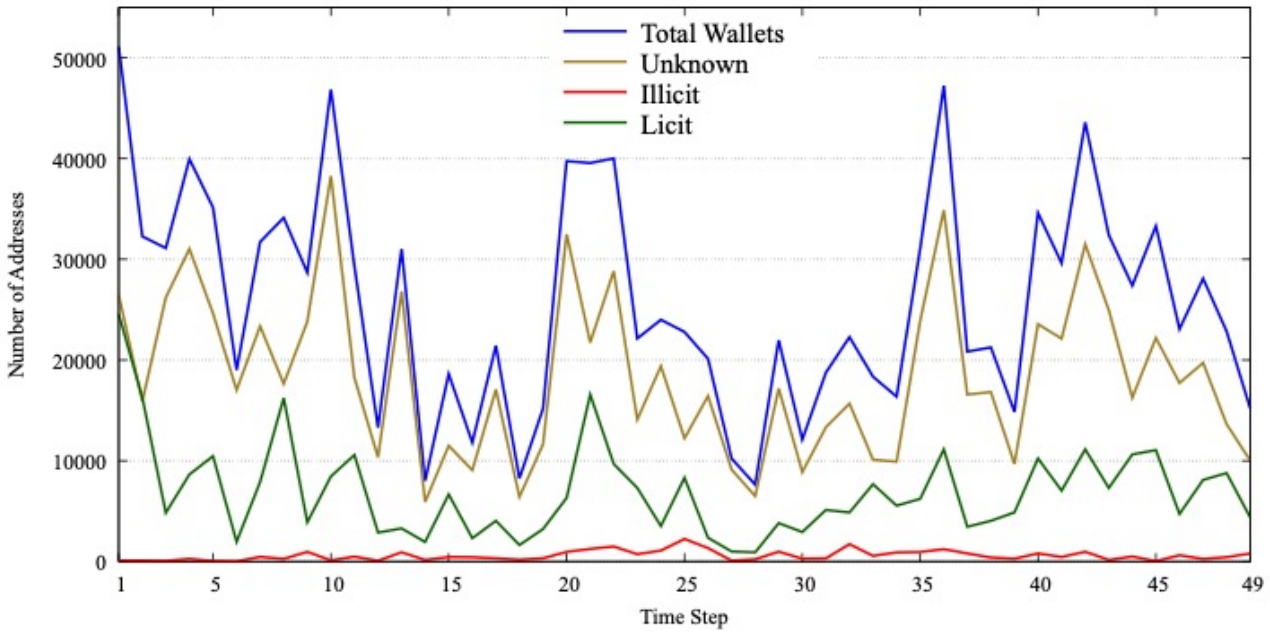
Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

The Elliptic++ Dataset: Actors Dataset







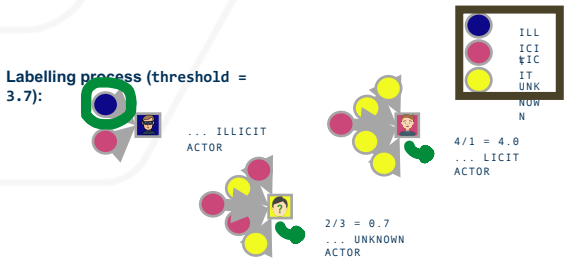
- **822,942** wallet addresses
- **1,268,260** temporal occurrences
- **56** address features
- Labels: **2%** **illicit**, **31%** **licit**, **67%** **unknown**

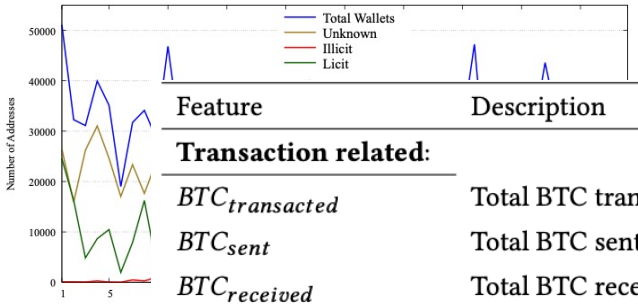
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

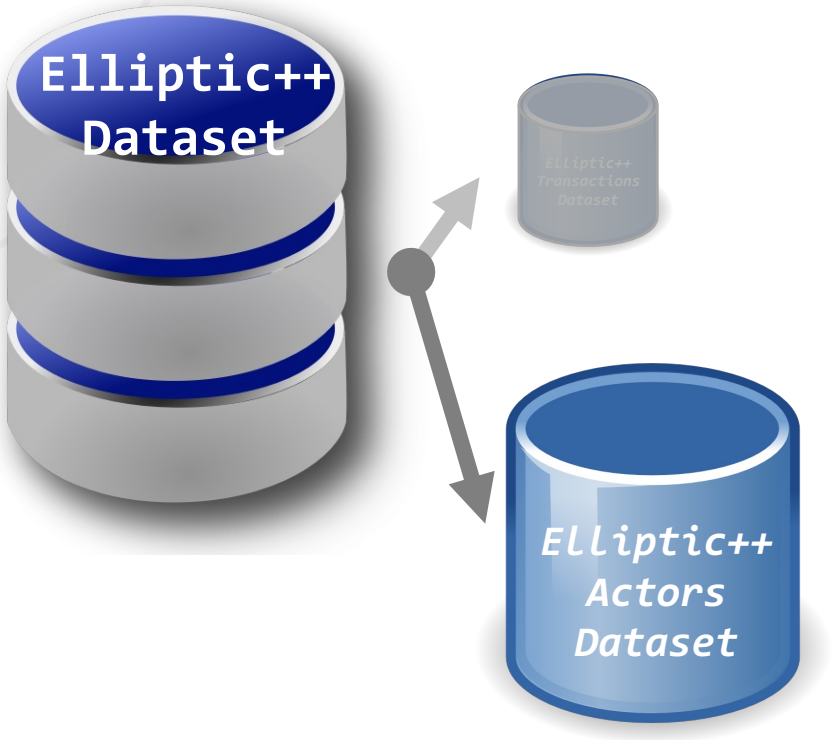
The Elliptic++ Dataset: Actors Dataset





Feature	Description
Transaction related:	
$BTC_{transacted}$	Total BTC transacted (sent+received)
BTC_{sent}	Total BTC sent
$BTC_{received}$	Total BTC received
$Fees$	Total fees in BTC
$Fees_{share}$	Total fees as share of BTC transacted
Time related:	
$Blocks_{txs}$	Number of blocks between transactions
$Blocks_{input}$	Number of blocks between being an input address
$Blocks_{output}$	Number of blocks between being an output address
Addr interactions	Number of interactions among addresses
each has 5 values:	**total, min, max, mean, median**

Class	class label: {illicit, licit, unknown}
Transaction related:	
Txs_{total}	Total number of blockchain transactions
Txs_{input}	Total number of dataset transactions as input address
Txs_{output}	Total number of dataset transactions as output address
Time related:	
Timesteps	Number of time steps transacting in
Lifetime	Lifetime in blocks
$Block_{first}$	Block height first transacted in
$Block_{last}$	Block height last transacted in
$Block_{first\ sent}$	Block height first sent in
$Block_{first\ receive}$	Block height first received in
Repeat interactions	Number of addresses transacted with multiple times
	single value



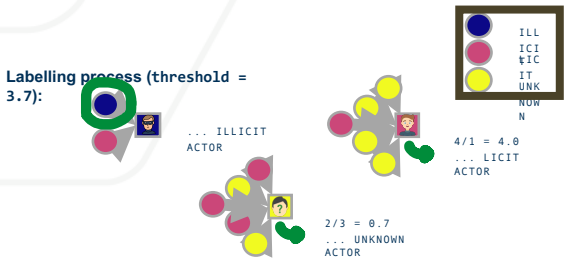
- **822, 942** wallet addresses
- **1, 268, 260** temporal occurrences
- **56** address features
- Labels: 2% **illicit**, 31% **licit**, 67% **unknown**

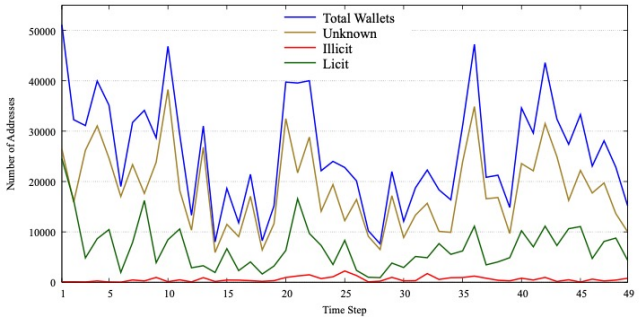
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

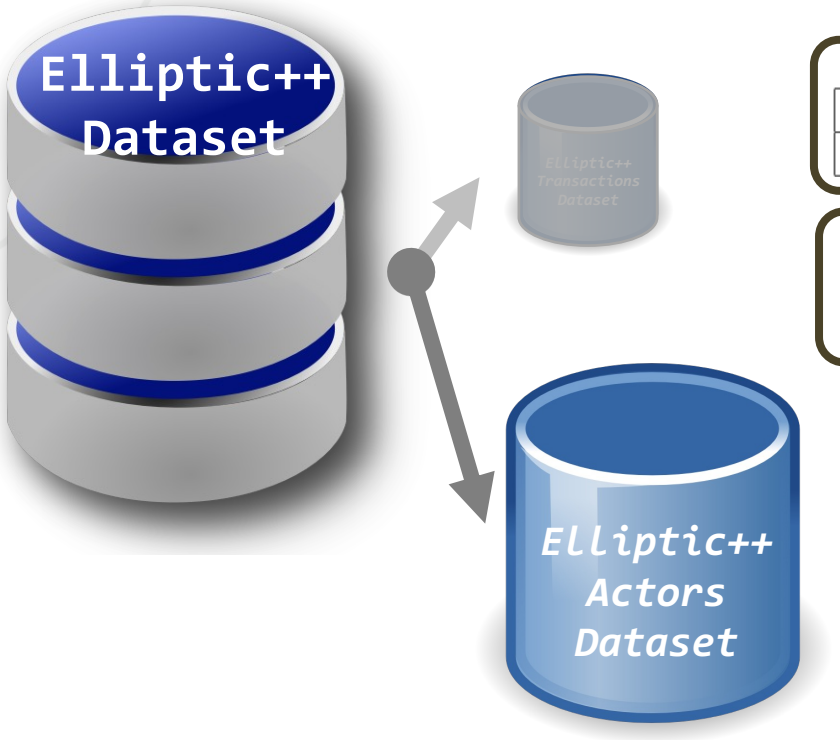
The Elliptic++ Dataset: Actors Dataset





Feature	Description
Transaction related:	
$BTC_{transacted}$	Total BTC transacted (sent+received)
BTC_{sent}	Total BTC sent
$BTC_{received}$	Total BTC received
$Fees$	Total fees in BTC
$Fees_{share}$	Total fees as share of BTC transacted
Time related:	
$Blocks_{txs}$	Number of blocks between transactions
$Blocks_{input}$	Number of blocks between being an input address
$Blocks_{output}$	Number of blocks between being an output address
Addr interactions	Number of interactions among addresses
each has 5 values:	**total, min, max, mean, median**

Class	class label: {illicit, licit, unknown}
Transaction related:	
Txs_{total}	Total number of blockchain transactions
Txs_{input}	Total number of dataset transactions as input address
Txs_{output}	Total number of dataset transactions as output address
Time related:	
Timesteps	Number of time steps transacting in
Lifetime	Lifetime in blocks
$Block_{first}$	Block height first transacted in
$Block_{last}$	Block height last transacted in
$Block_{first\ sent}$	Block height first sent in
$Block_{first\ receive}$	Block height first received in
Repeat interactions	Number of addresses transacted with multiple times
	single value



wallets_features.csv

address	time step	txs_input	···	lifetime_blocks	···	Addr_interactions_median
39sfuA8pY4UfybgEZi7uvA13jkGzZpsg5K	23	420	···	18145	···	1

AddrAddr_edgelist.csv

input_address	output_address
39sfuA8pY4UfybgEZi7uvA13jkGzZpsg5K	1ML...kTL

AddrTx_edgelist.csv

input_address	txId
39sfuA8pY4UfybgEZi7uvA13jkGzZpsg5K	272145560

TxAddr_edgelist.csv

txId	output_address
322554634	**39sfuA8pY4UfybgEZi7uvA13jkGzZpsg5K**

wallets_classes.csv

address	class
39sf...sg5K	1

- **822, 942** wallet addresses
- **1, 268, 260** temporal occurrences
- **56** address features
- Labels: 2% **illicit**, 31% **licit**, 67% **unknown**

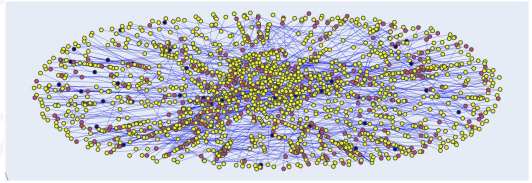
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Graph Visualization of The Elliptic++ Dataset

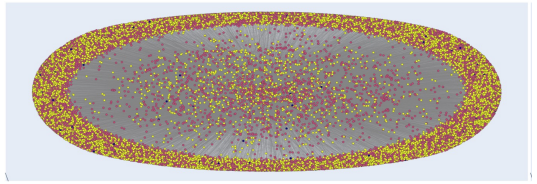
1 **Money Flow Transaction Graph**



BTC flow from one transaction to the next

Utility: Exploration of spatial and temporal patterns surrounding a transaction

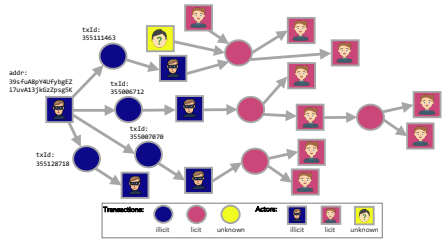
2 **Actor Interaction Graph**



Pairwise interactions among input and output addresses of txs

Utility: Shows density of k-hop neighborhoods of addresses

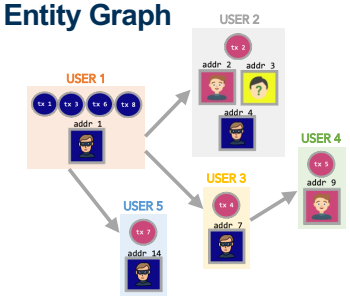
3 **Address-Transaction Graph**



BTC flow across txs and addresses

Utility: Evaluation of relationships among addresses of same transactions

4 **User Entity Graph**



Address cluster graphs

Utility: Linking of addresses controlled by a specific user (de-anonymization)

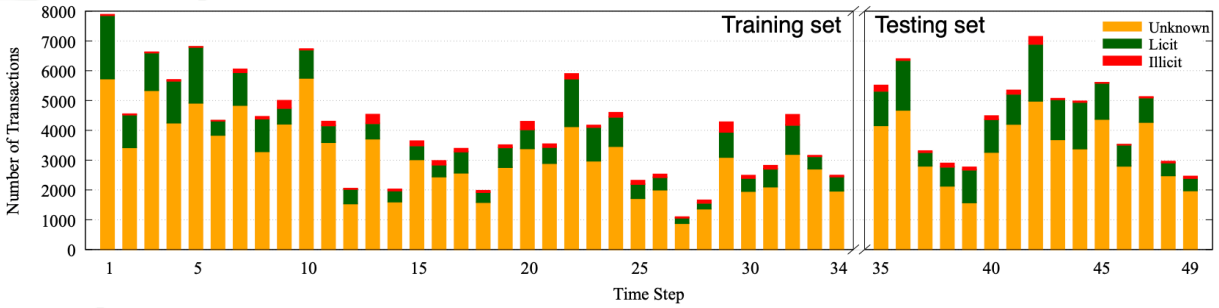
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

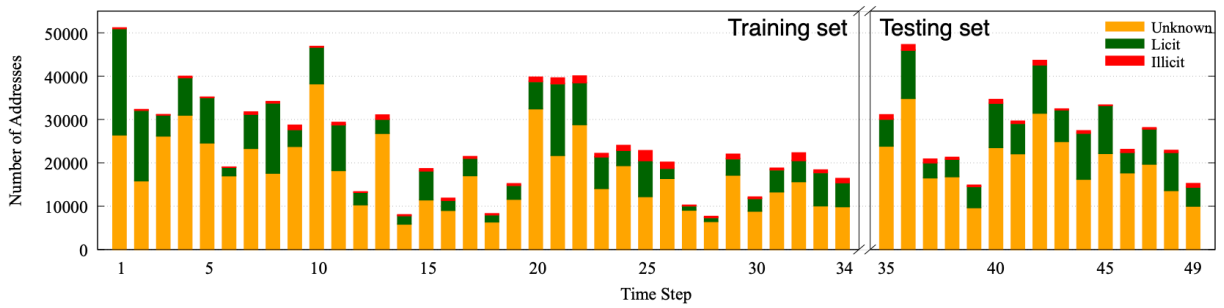
Georgia Tech

Fraud Detection Methodology

Dataset Preprocessing



TRANSACTIONS DATASET



ACTORS DATASET

Machine Learning Models

Random Forest (**RF**) Scikit-learn: default parameters, 50 estimators

Multilayer Perceptrons (**MLP**) 1 hidden layer w/ 50 neurons, 500 epochs, Adam opt., 0.001 LR

Long Short-Term Memory (**LSTM**) TensorFlow: sigmoid act., Adam opt., 30 epochs, binary cross-entropy loss, 15 emb. output dims

Extreme Gradient Boosting (**XGB**) XGBoost: default parameters, multi:softmax obj., 2 classes

Logistic Regression (**LR**) Scikit-learn: default parameters

Evaluation Metrics

Precision, Recall, F1 Score, Micro-Avg F1 Score, Matthews Correlation Coefficient (MCC)

Case studies:

(i) EASY cases: all models classify an illicit tx correctly

(ii) HARD cases: all models classify an illicit tx incorrectly

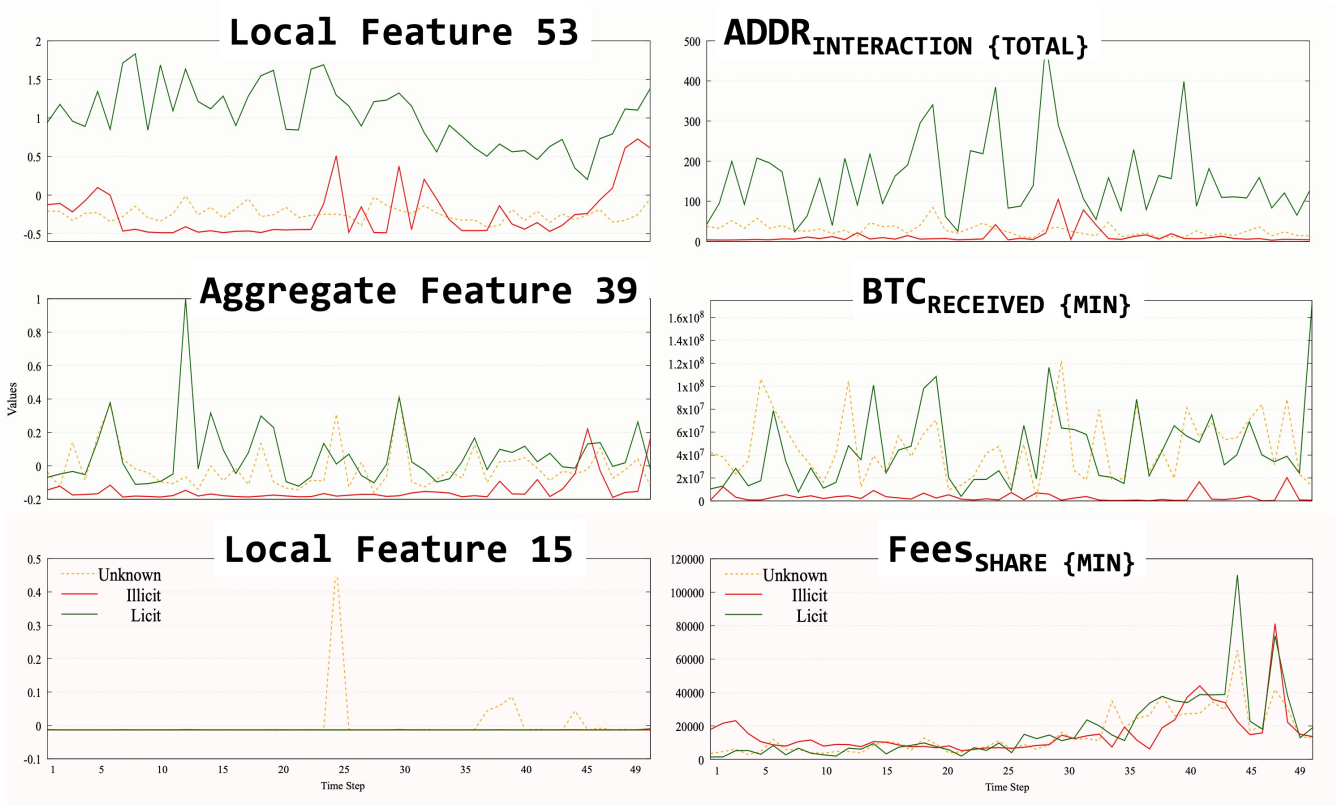
(iii) AVERAGE cases: some models failed to classify an illicit tx but ≥ 1 models classified correctly

Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Results: Statistical Analysis of the Dataset

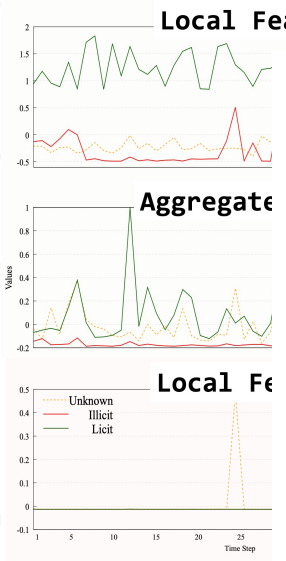


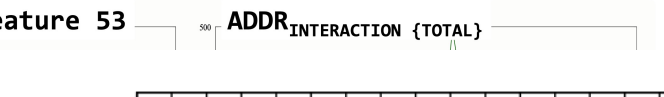
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

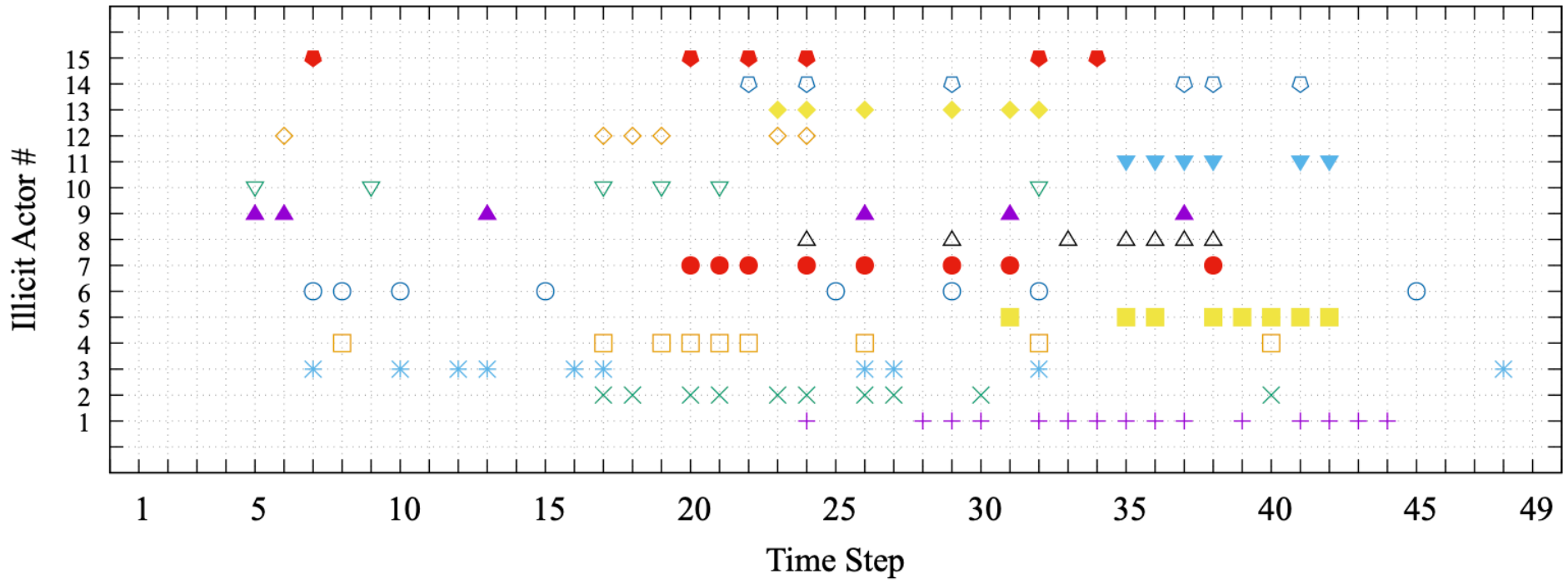
Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Results: Statistical Analysis of the Dataset





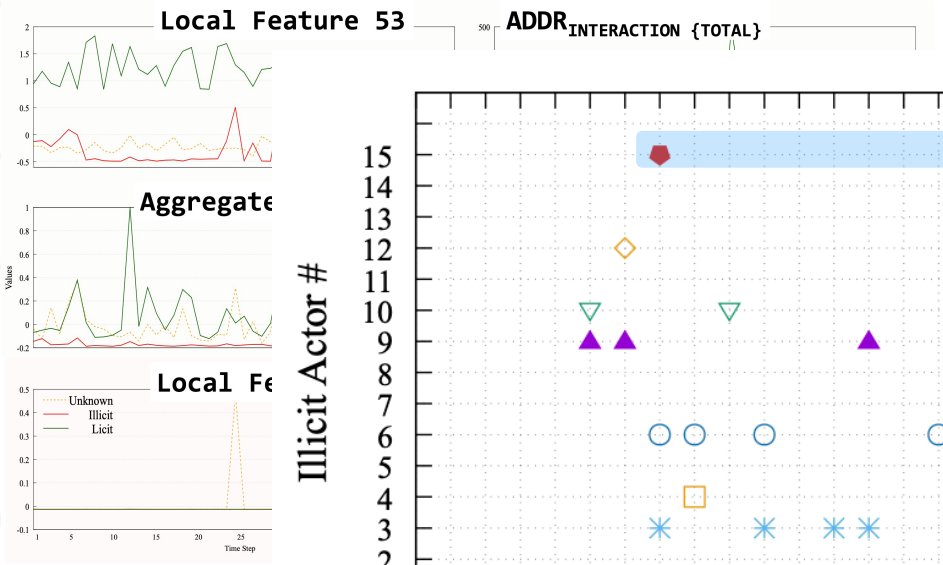


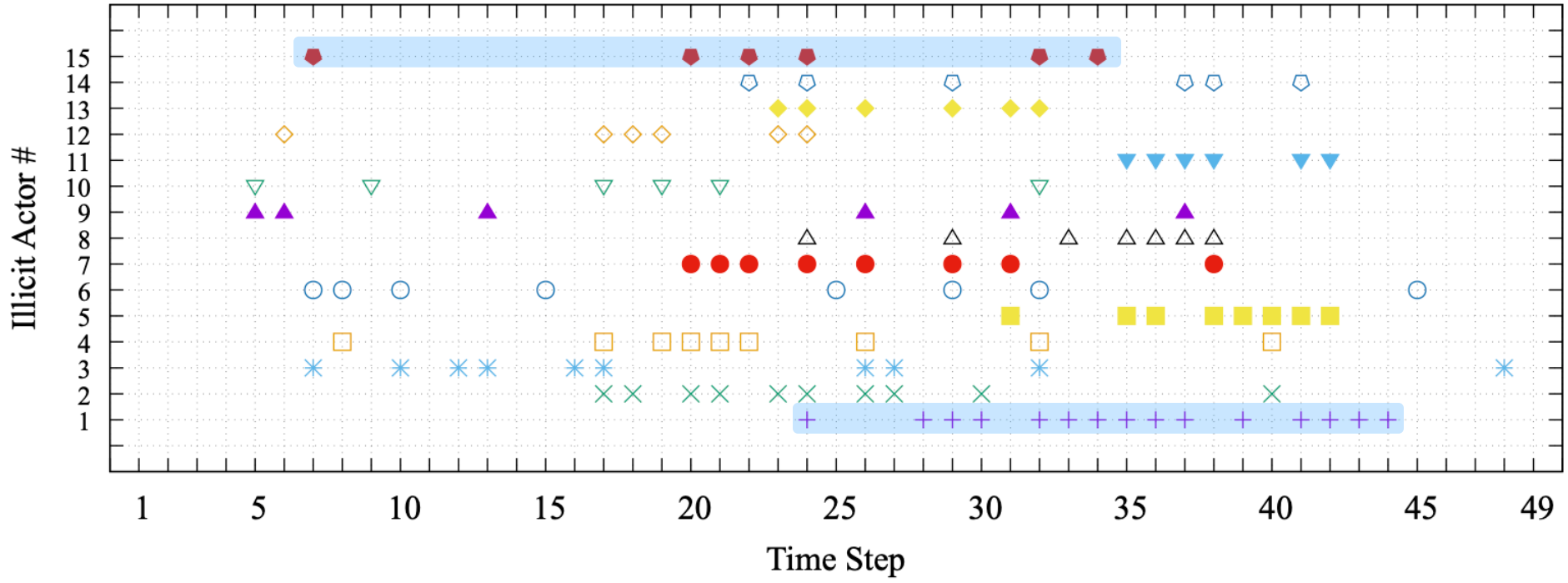
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Results: Statistical Analysis of the Dataset



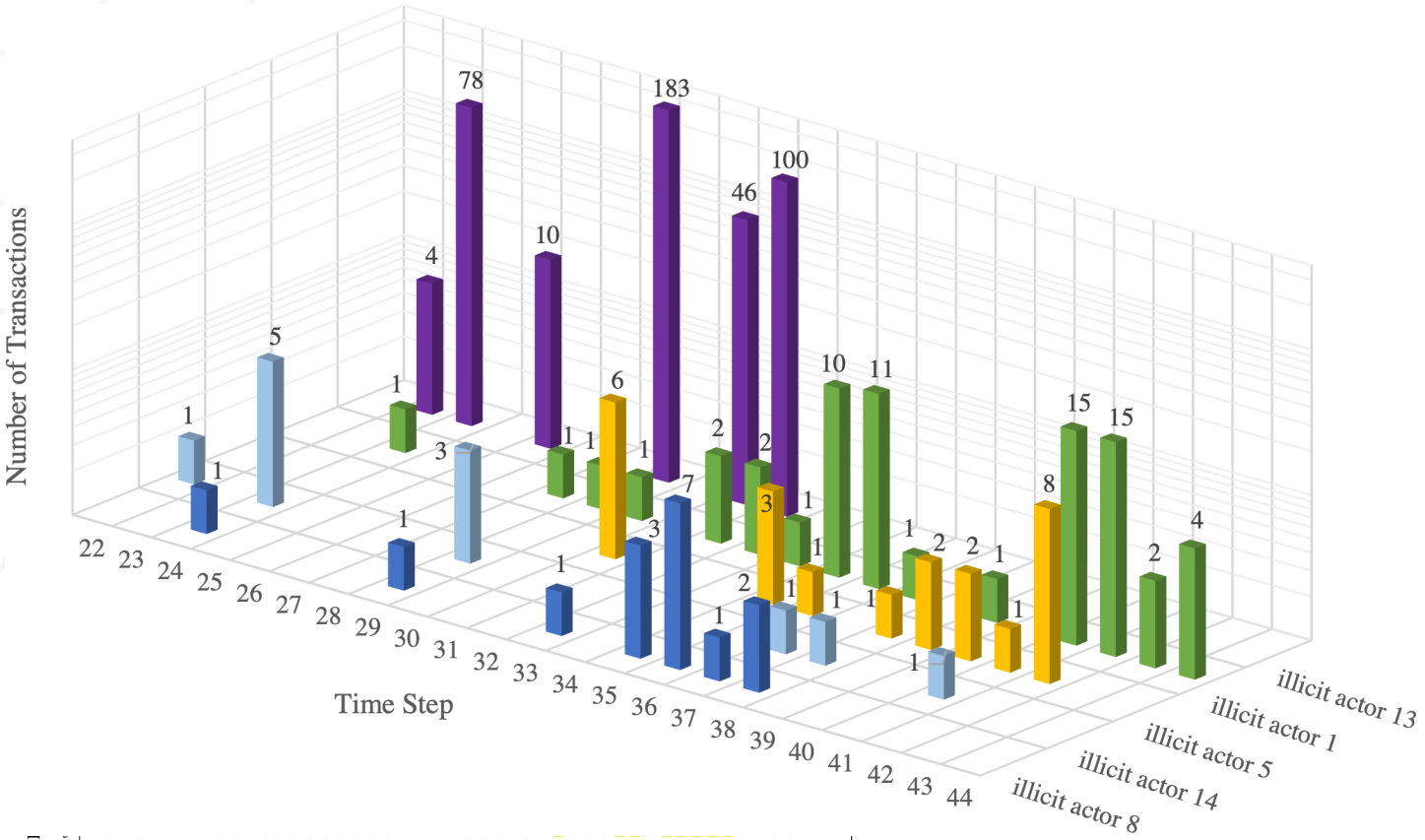


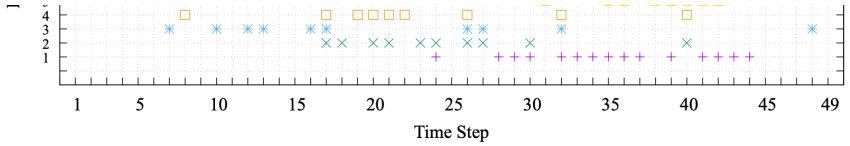
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Results: Statistical Analysis of the Dataset



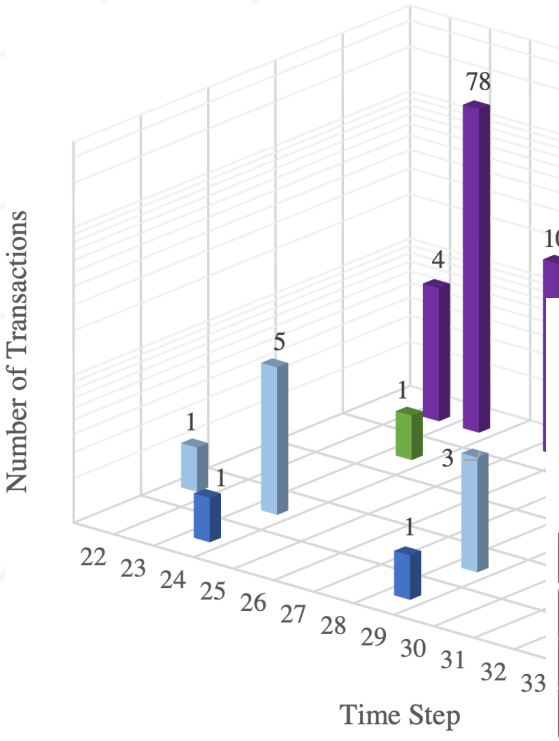


Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

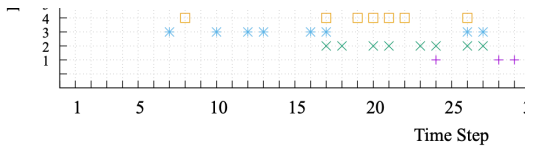
Georgia Tech

Results: Statistical Analysis of the Dataset



Time step	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
illicit actor appears in 1 time step	68	69	42	186	36	8	233	116	384	26	185	47	342	74	193	174	118
illicit actor appears in 2 – 4 time steps	5	1	3	15	2	1	28	7	21	11	7	2	4	0	6	8	8
illicit actor appears in ≥ 5 time steps	0	0	0	0	2	2	5	4	3	3	2	1	2	0	2	1	6

Time step	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
illicit actor appears in 1 time step	77	122	405	612	692	376	532	1096	704	43	97	372	130	131	759	270	447
illicit actor appears in 2 – 4 time steps	4	5	20	16	18	17	40	9	21	1	7	52	11	24	23	7	11
illicit actor appears in ≥ 5 time steps	2	4	5	4	5	3	12	1	9	2	1	9	4	6	8	5	2



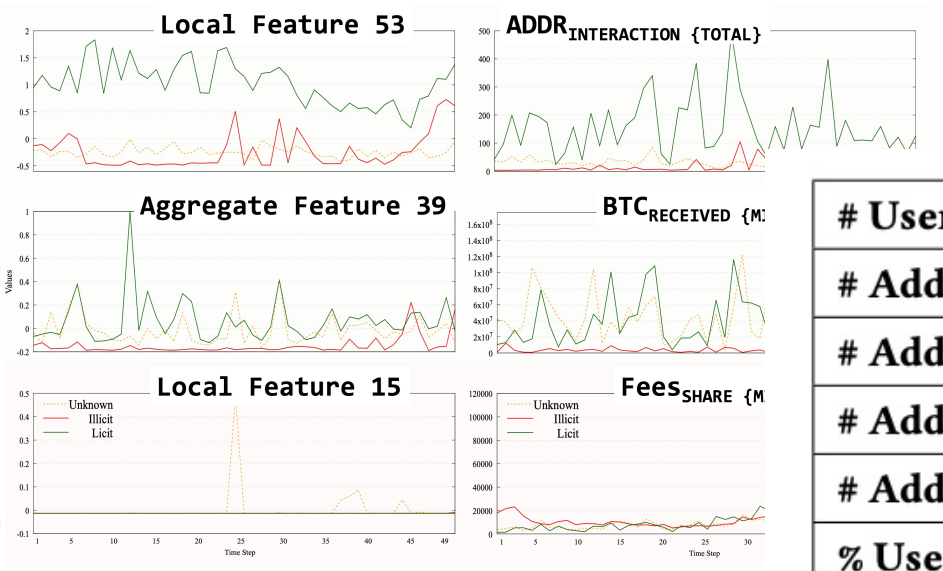
Time step	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
illicit actor appears in 1 time step	427	588	428	180	127	436	168	395	93	277	23	505	193	370	631
illicit actor appears in 2 – 4 time steps	9	5	5	9	0	14	8	12	11	6	9	4	11	17	20
illicit actor appears in ≥ 5 time steps	6	4	7	5	2	3	6	3	1	1	1	0	0	1	1

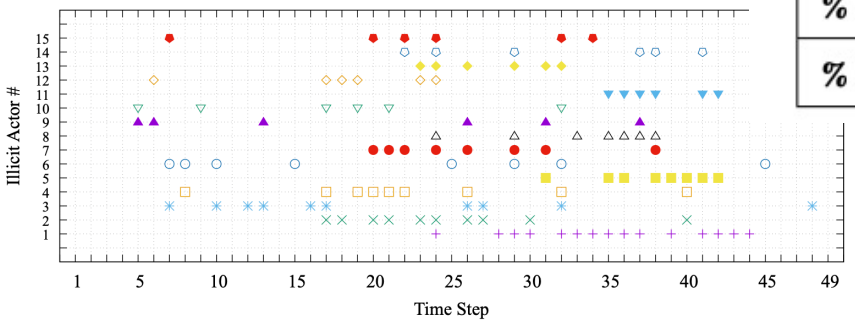
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

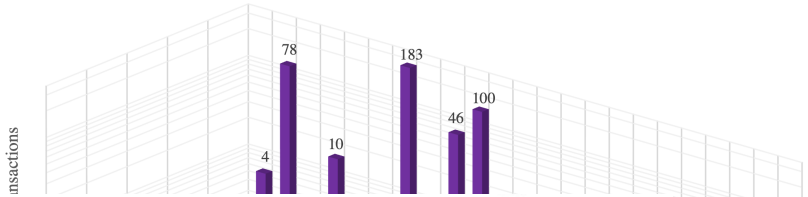
Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Results: Statistical Analysis of the Dataset







# Users	146, 783
# Addresses per User: Min	1
# Addresses per User: Median	1
# Addresses per User: Mean	2.73
# Addresses per User: Max	14, 885
% Users w/ 1 – 10 Addresses	98.72%
% Users w/ 11 – 1K Addresses	1.26%
% Users w/ 1K – max Addresses	0.02%

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	68	69	42	186	36	8	233	116	384	26	185	47	342	74	193	174	118
	5	1	3	15	2	1	28	7	21	11	7	2	4	0	6	8	8
	0	0	0	0	2	2	5	4	3	3	2	1	2	0	2	1	6

8	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
7	122	405	612	692	376	532	1096	704	43	97	372	130	131	759	270	447
4	5	20	16	18	17	40	9	21	1	7	52	11	24	23	7	11
2	4	5	4	5	3	12	1	9	2	1	9	4	6	8	5	2

	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
	427	588	428	180	127	436	168	395	93	277	23	505	193	370	631
os	9	5	5	9	0	14	8	12	11	6	9	4	11	17	20
s	6	4	7	5	2	3	6	3	1	1	1	0	0	1	1

Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

TX – Elliptic++ Transactions Dataset
EC – Elliptic Dataset
AR - Elliptic++ Actors Dataset

Results: Model Evaluation and Analysis

TRANSACTIONS DATASET

Model	Precision	Recall	F1 Score	Micro-F1
LR^{EC}	0.326	0.707	0.446	0.886
LR^{TX}	0.328	0.707	0.448	0.884
RF^{EC}	0.940	0.724	0.818	0.979
RF^{TX}	**0.975**	**0.719**	**0.828**	**0.980**
MLP^{EC}	0.476	0.673	0.558	0.931
MLP^{TX}	0.611	0.613	0.612	0.949
$LSTM^{EC}$	0.665	0.350	0.459	0.946
$LSTM^{TX}$	0.709	0.223	0.339	0.942
XGB^{EC}	0.812	0.717	0.761	0.971
XGB^{TX}	0.793	0.718	0.754	0.969
2 ***classifiers ensemble***, *selecting top 3 classifiers*				
$RF+MLP^{EC}$	0.987	0.624	0.765	0.975
$RF+MLP^{TX}$	0.989	0.635	0.773	0.975
$RF+XGB^{EC}$	0.960	0.704	0.812	0.979
$RF+XGB^{TX}$	**0.977**	**0.706**	**0.820**	**0.979**
$MLP+XGB^{EC}$	0.457	0.737	0.564	0.926
$MLP+XGB^{TX}$	0.974	0.596	0.739	0.972
3 ***classifiers ensemble***, *selecting top 3 classifiers*				
$RF+MLP+XGB^{EC}$	0.947	0.719	0.817	0.979
$RF+MLP+XGB^{TX}$	**0.962**	**0.723**	**0.826**	**0.980**

ACTORS DATASET

Model	Precision	Recall	F1 Score	Micro-F1
LR^{AR}	0.477	0.046	0.083	0.964
RF^{AR}	**0.911**	**0.789**	**0.845**	**0.990**
MLP^{AR}	0.708	0.502	0.587	0.974
$LSTM^{AR}$	0.922	0.033	0.064	0.965
XGB^{AR}	0.869	0.534	0.662	0.980
2 ***classifiers ensemble***, *selecting top* 3 *classifiers*				
$RF+MLP^{AR}$	0.967	0.403	0.568	0.978
$RF+XGB^{AR}$	**0.959**	**0.530**	**0.682**	**0.982**
$MLP+XGB^{AR}$	0.929	0.324	0.481	0.975
3 ***classifiers ensemble***, *selecting top* 3 *classifiers*				
$RF+MLP+XGB^{AR}$	**0.933**	**0.572**	**0.709**	**0.983**

Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Results: EASY, HARD, AVERAGE Cases Analysis

Time Step		35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	TOTAL
EASY		32	0	2	5	5	1	1	3	0	0	0	0	0	0	0	49
HARD		4	0	10	7	4	28	6	36	22	20	4	1	21	27	53	243
AVERAGE	LR	0	0	3	0	2	3	0	6	2	3	1	0	1	9	2	
	RF	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	
	MLP	0	0	1	1	0	2	0	2	0	0	0	0	0	0	0	
	LSTM	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	
	XGB	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	
	RF,XGB	4	0	0	1	2	1	17	2	0	0	0	0	0	0	0	791
	LR,MLP	1	0	0	1	0	2	0	2	0	0	0	0	0	0	0	
	RF,MLP,XGB	5	6	0	8	3	4	1	0	0	0	0	0	0	0	0	
	LR,RF,XGB	6	1	10	27	18	10	5	21	0	0	0	0	0	0	0	
	RF,MLP,XGB,LR	124	24	12	57	45	55	81	159	0	1	0	1	0	0	0	

<1%

71%

This motivates us to focus on optimization of the RF model with feature refinement

Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Model Optimization by Feature Refinement

TRANSACTIONS DATASET

ACTORS DATASET

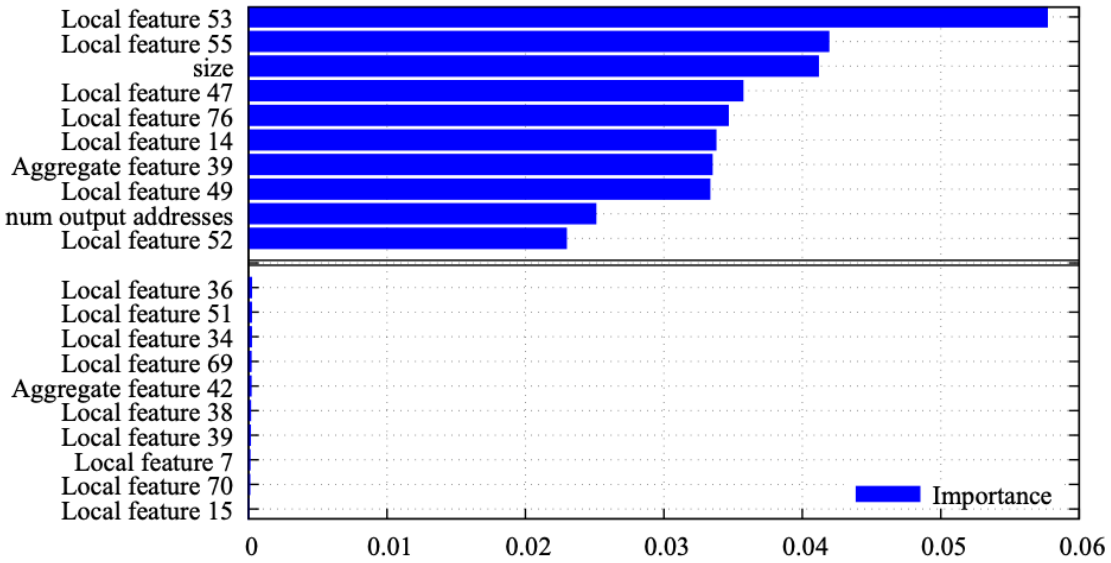
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

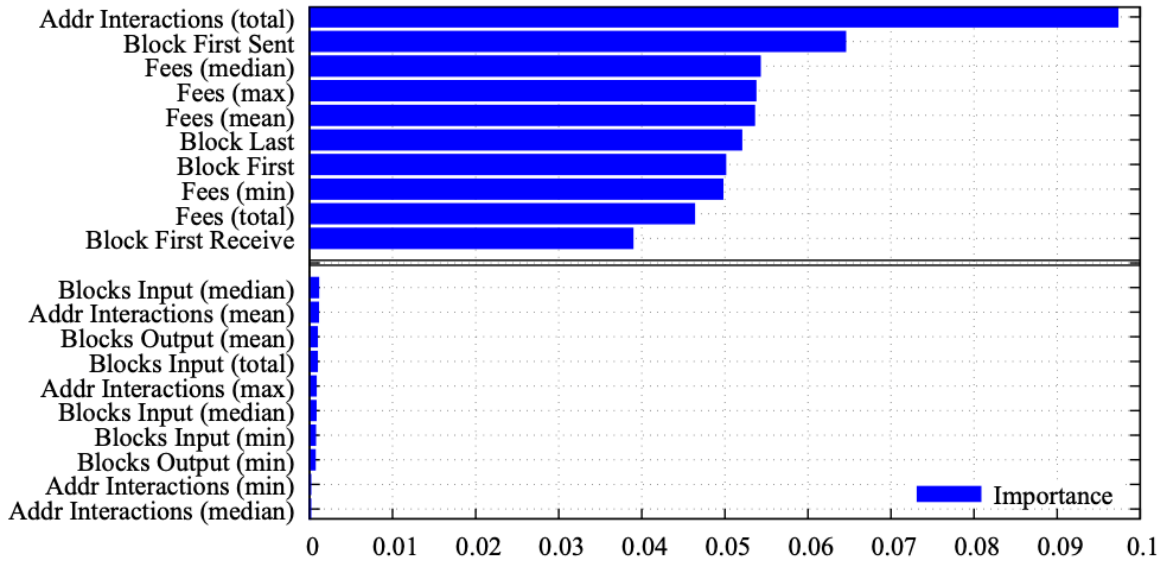
Georgia Tech

Model Optimization by Feature Refinement

TRANSACTIONS DATASET



ACTORS DATASET



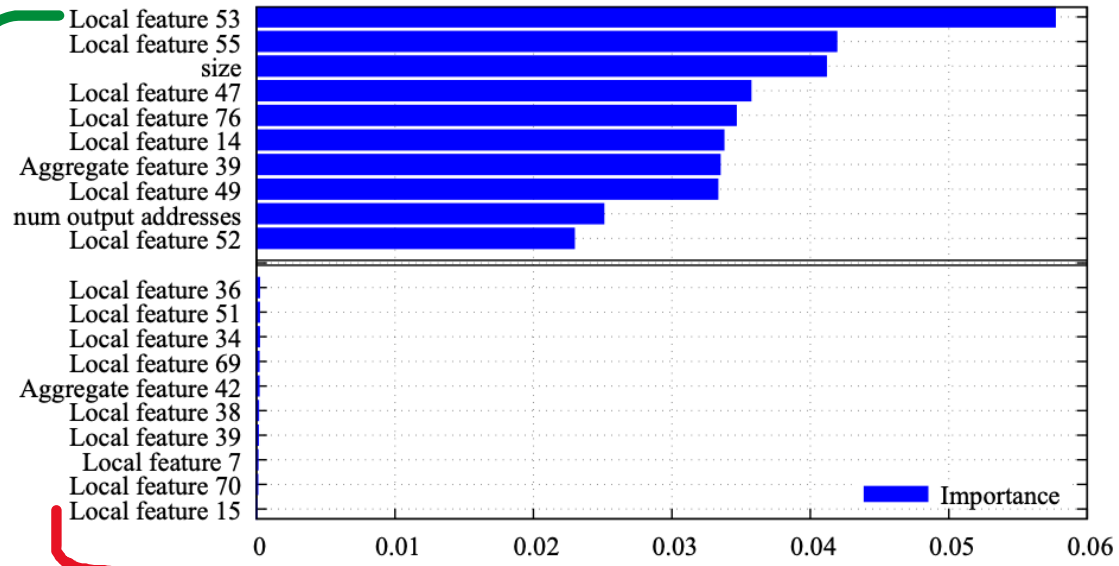
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

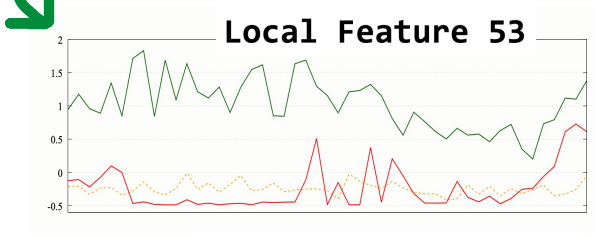
Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

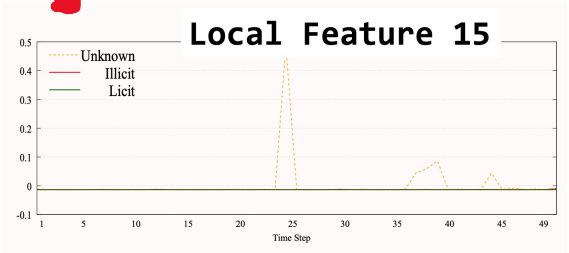
Georgia Tech

Model Optimization by Feature Refinement

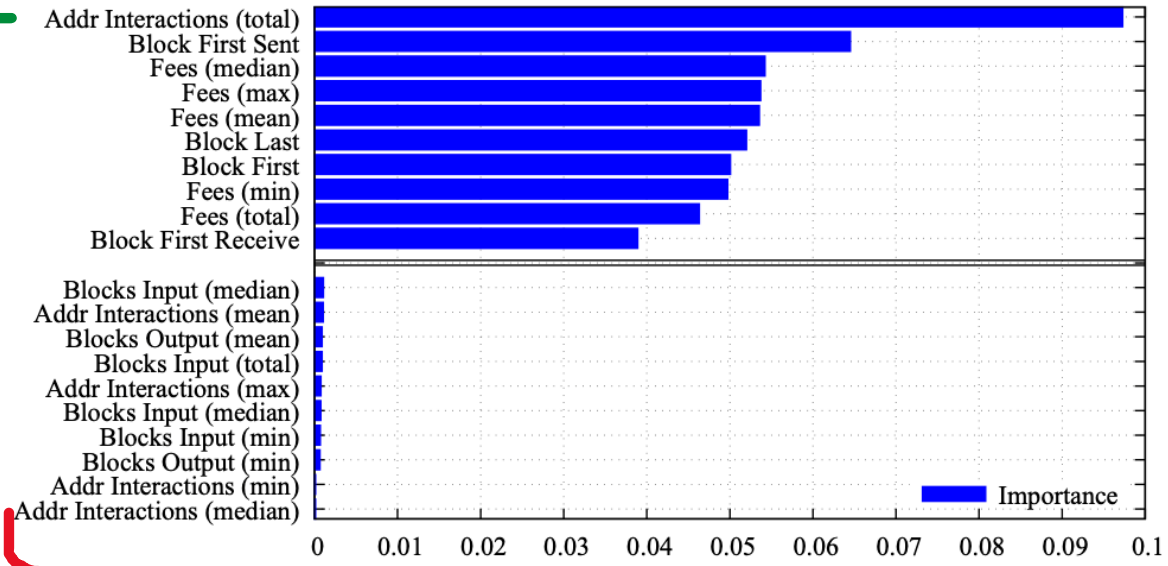
TRANSACTIONS DATASET

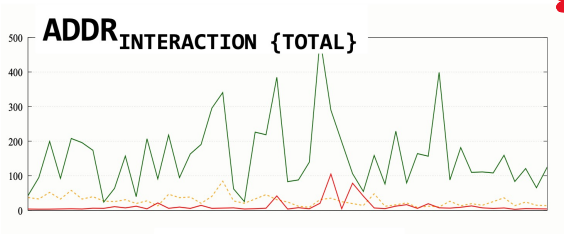


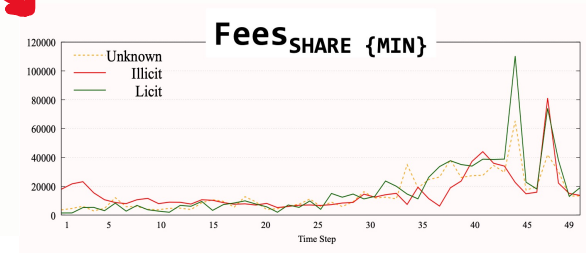




ACTORS DATASET







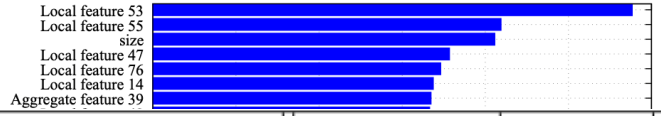
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

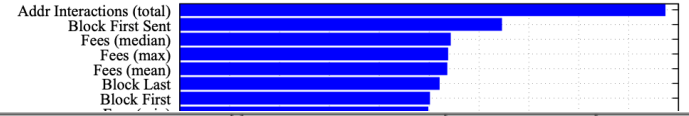
Model Optimization by Feature Refinement

TRANSACTIONS DATASET



Model	Precision	Recall	F1 Score	Micro-F1
RF^{TX}	0.975	0.719	0.828	0.980
$RF^{TX\psi}$	**0.986**	**0.727**	**0.836**	**0.981**
$RF+XGB^{TX}$	0.977	0.706	0.820	0.979
$RF+XGB^{TX\psi}$	**0.987**	**0.717**	**0.826**	**0.980**
$RF+MLP+XGB^{TX}$	0.962	0.723	0.826	0.980
$RF+MLP+XGB^{TX\psi}$	**0.968**	**0.729**	**0.834**	**0.980**

ACTORS DATASET



Model	Precision	Recall	F1 Score	Micro-F1
RF^{AR}	0.911	0.789	0.845	0.990
$RF^{AR\psi}$	**0.921**	**0.802**	**0.858**	**0.990**
$RF+XGB^{AR}$	0.959	0.530	0.682	0.982
$RF+XGB^{AR\psi}$	**0.967**	**0.543**	**0.686**	**0.982**
$RF+MLP+XGB^{AR}$	0.933	0.572	0.709	0.983
$RF+MLP+XGB^{AR\psi}$	**0.945**	**0.601**	**0.718**	**0.984**

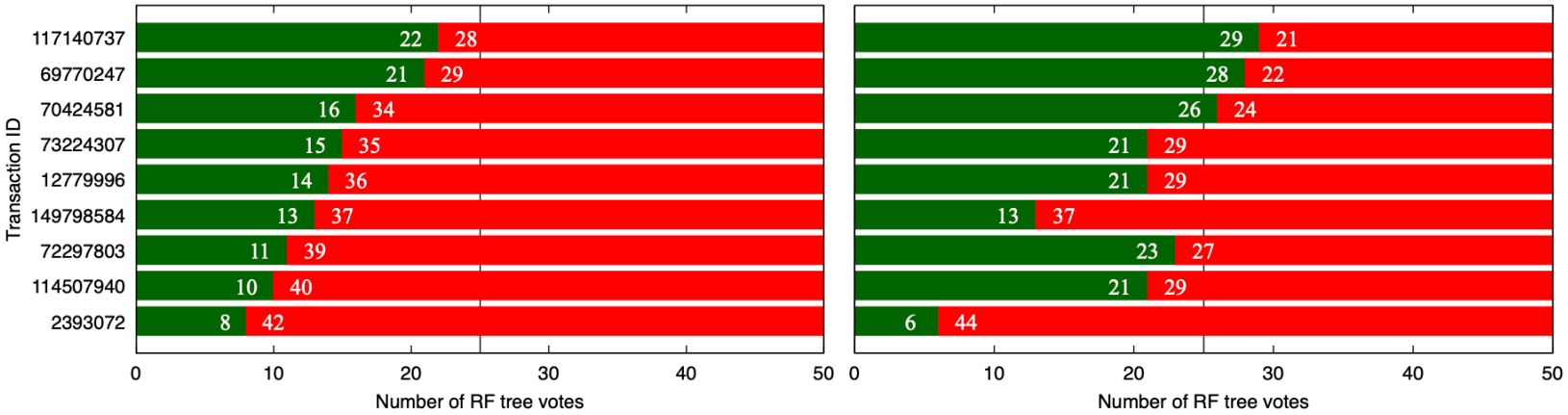
Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

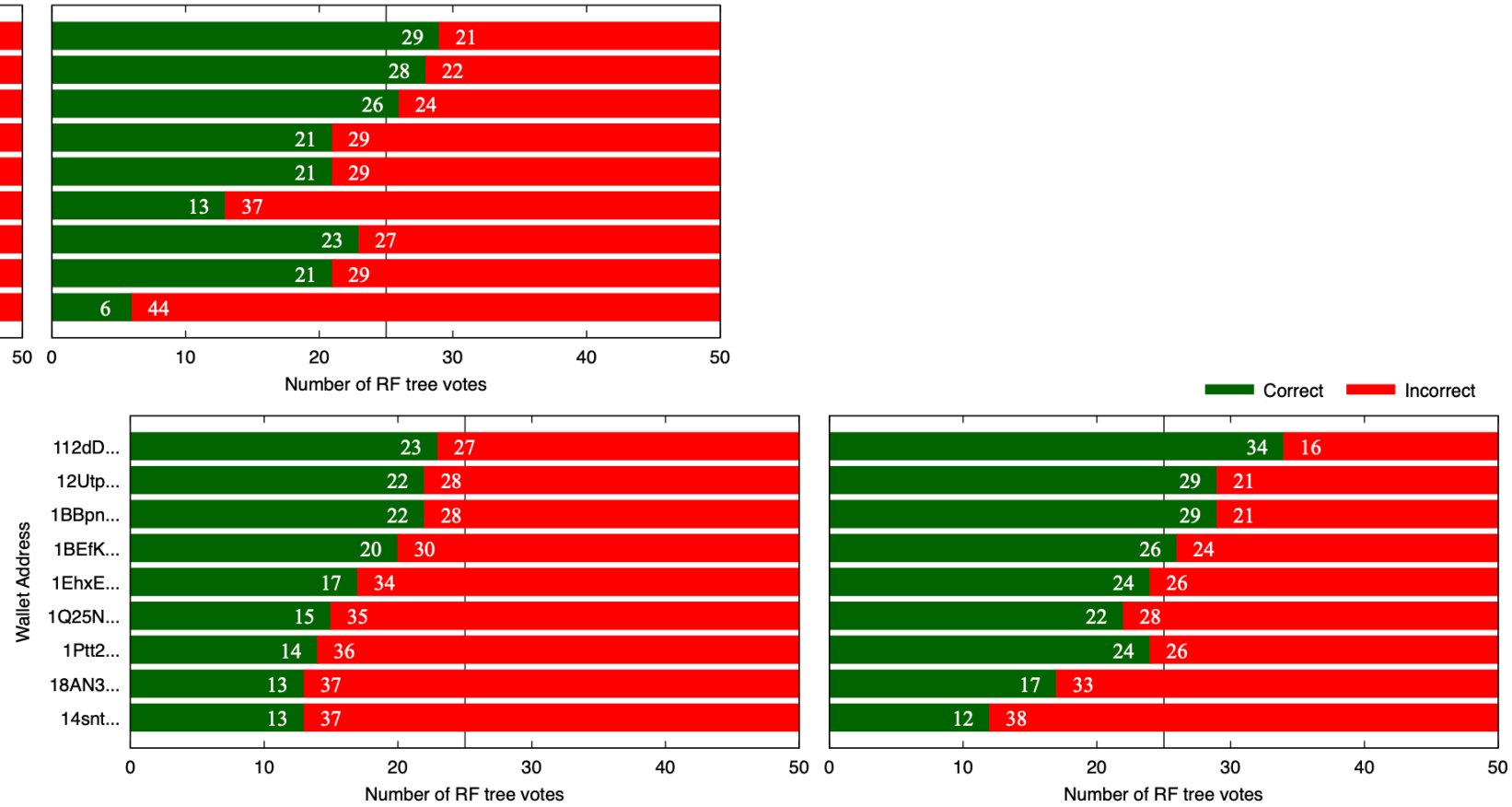
Georgia Tech

Model Optimization by Feature Refinement

TRANSACTIONS DATASET



ACTORS DATASET



Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

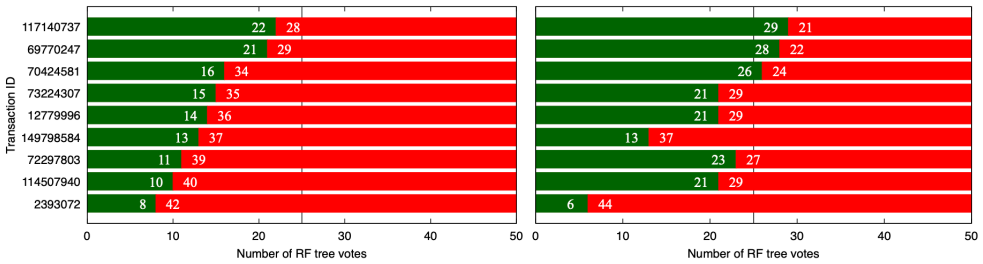
Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

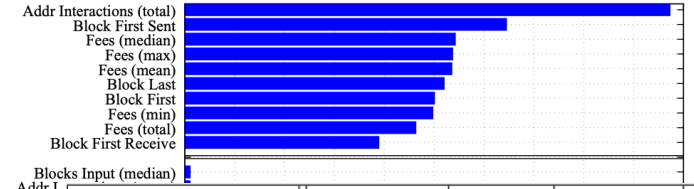
Model Optimization by Feature Refinement

TRANSACTIONS DATASET

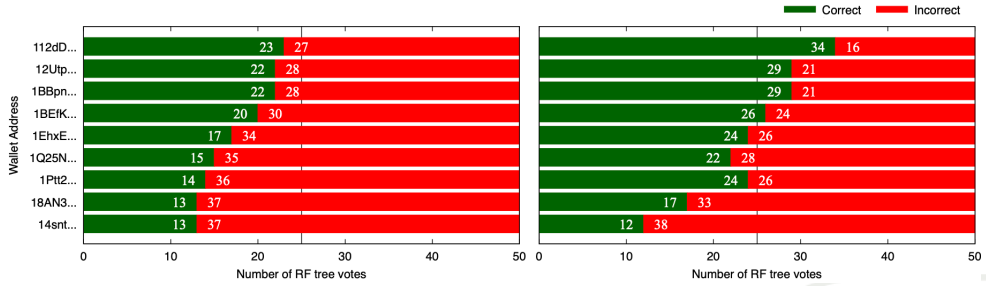
Model	Precision	Recall	F1 Score	Micro-F1
RF^{TX}	0.975	0.719	0.828	0.980
$RF^{TX\psi}$	**0.986**	**0.727**	**0.836**	**0.981**
$RF+XGB^{TX}$	0.977	0.706	0.820	0.979
$RF+XGB^{TX\psi}$	**0.987**	**0.717**	**0.826**	**0.980**
$RF+MLP+XGB^{TX}$	0.962	0.723	0.826	0.980
$RF+MLP+XGB^{TX\psi}$	**0.968**	**0.729**	**0.834**	**0.980**



ACTORS DATASET



Model	Precision	Recall	F1 Score	Micro-F1
RF^{AR}	0.911	0.789	0.845	0.990
$RF^{AR\psi}$	**0.921**	**0.802**	**0.858**	**0.990**
$RF+XGB^{AR}$	0.959	0.530	0.682	0.982
$RF+XGB^{AR\psi}$	**0.967**	**0.543**	**0.686**	**0.982**
$RF+MLP+XGB^{AR}$	0.933	0.572	0.709	0.983
$RF+MLP+XGB^{AR\psi}$	**0.945**	**0.601**	**0.718**	**0.984**



Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

Elliptic++ Dataset: A Graph Network of Bitcoin Blockchain Transactions and Wallet Addresses

Georgia Tech

Concluding Remarks

Our Contributions to financial forensics on blockchain networks:

1. **We release The Elliptic++ Dataset, combining over 203k transactions and 822k addresses, and provide four unique graph representations.**
2. **We demonstrate the utility of the dataset for detecting fraudulent transactions and illicit accounts using representative ML approaches.**
3. **We provide detailed analysis on improving the generalization performance of individual and ensemble ML algorithms on the dataset.**

Dataset: **https://www.github.com/git-disl/EllipticPlusPlus**

The Elliptic++ Dataset and Tutorials

Publicly available at

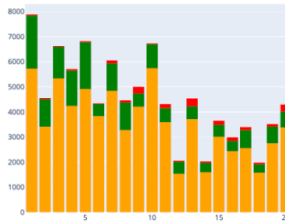
github.com/**git-disl/EllipticPlusPlus**

Along with several tutorials!!!

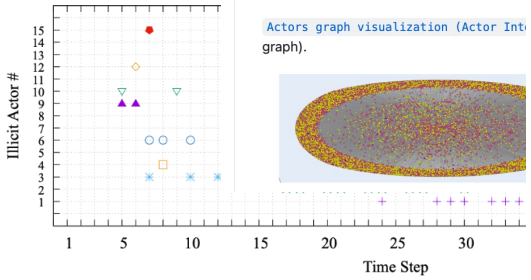
Dataset Tutorials

We are sharing tutorial notebooks for users and researchers to explore, study, and learn fr
notebooks are available for both datasets and cover dataset statistics, graph visualization
classification, case analysis, and feature refi

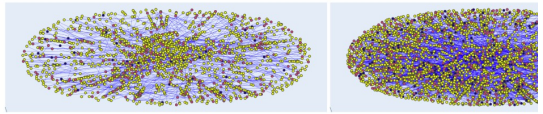
`Transactions dataset statistics` : overa



`Actors dataset statistics` : overall



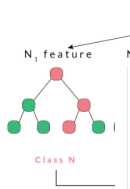
`Transactions graph visualization` : visualizations of the Money Flow Transaction grap



`Actors graph visualization (Actor Int` graph).

`Actors graph visualization (Addres` tx-addr graph).

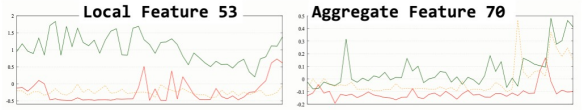
`Transactions classification` : model training and classification on the transactions data.



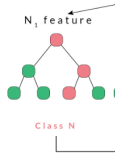
`Transactions case analysis` : unique case (EASY, HARD, AVERAGE) analysis using the transactions data.

Time Step	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	TOTAL
EASY	32	0	2	5	5	1	1	3	0	0	0	0	0	0	0	49
HARD	4	0	10	7	4	28	6	36	22	20	4	1	21	27	53	243
LR	0	0	3	0	2	3	0	6	2	3	1	0	1	9	2	

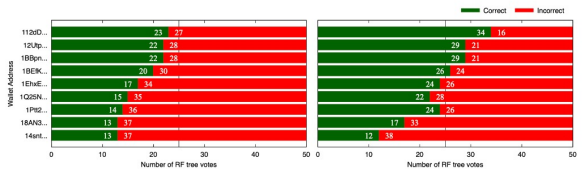
`Transactions feature analysis` : feature importance analysis of the transactions data.

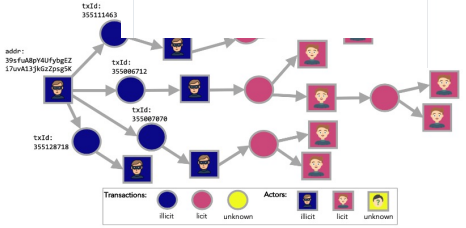


`Actors classification` : model training



`Actors feature analysis` : feature importance analysis of the actors data.





Georgia Tech

Demystifying Fraudulent Transactions and Illicit Nodes in the Bitcoin Network for Financial Forensics

Youssef Elmougy and Ling Liu

Georgia Institute of Technology, Atlanta GA USA

Dataset: github.com/**git-disl/EllipticPlusPlus**

Thank you for your attention!

Georgia Tech