Exploring Energy Flow Classifier to Identify Fraudulent Cryptocurrency Transactions

Kevin S. Araujo[1], **Rodrigo Bonifacio de Almeida**[1], **Fabiano Cavalcanti Fernandes**[2]

[1]Departamento de Ciências da Computação – Universidade de Brasília (UnB) – Campus Universitário Darcy Ribeiro, Brasília-DF

[2]Instituto Federal de Brasília (IFB) – Taguating, DF – Brazil

kevin.araujo@aluno.unb.br, rbonifacio@unb.br, fabiano.fernandes@ifb.edu.br

Abstract. *This is a work in progress.*

1. Introduction

In 2008, a global financial crisis in the real estate sector occurred [Bordo 2008], which was caused by the State providing easy credit [Murphy 2008]. A cryptographic document began circulating on a mailing list for cryptographers, signed by the pseudonymous Satoshi Nakamoto, which compiled detailed data and discoveries made through cypherpunk innovations. Nakamoto utilized this knowledge to create an electronic transaction system that did not require the involvement of a third-party moderator. In essence, Nakamoto's work involved using math, programming, and cutting-edge cryptography to publish a map for removing governmental presence from financial transactions. The recent collapse of the economy had demonstrated that governments cannot be trusted, and Nakamoto's solution was to create a currency that was mathematically impossible to be corrupted — Bitcoin [Nakamoto 2008]. Although Bitcoin does present a solution to the corruptible nature of money, it does possess flaws and is target of more sophisticated frauds.

Detecting such sophisticated frauds requires advanced analytical techniques. Identifying anomalous patterns within complex data streams, such as network traffic or financial transactions, remains a critical challenge. Traditional methods often struggle with the high dimensionality, evolving nature, and sheer volume of modern data. In the domain of Network Intrusion Detection (NID), flow-based analysis offers a valuable abstraction by aggregating packet-level information into connection summaries, reducing data complexity while retaining essential behavioral characteristics. Within this context, a promising approach grounded in principles from statistical physics is the Energy-based Flow Classifier (EFC). Originally proposed using the Inverse Potts model, EFC characterizes the probability distribution of normal network flows through an energy function derived from observed data patterns [Pontes et al. 2019]. Configurations representing typical, legitimate flows are assigned low energy values by the model, whereas anomalous or potentially malicious flows, deviating from the learned normality, manifest as high-energy states. Subsequent research has further explored the capabilities of this energy-based framework, highlighting its potential for open-set recognition the challenging task of identifying novel anomalies not encountered during the model's training phase [Souza et al. 2022]. The fundamental principle of assigning an energy score as a measure of typicality provides a robust and theoretically grounded mechanism for distinguishing normal system behavior from potentially illicit activities.

While Bitcoin was designed to circumvent traditional financial system vulnerabilities [Nakamoto 2008], it is not immune to manipulation and anomalous activities, necessitating robust detection mechanisms [Zhang et al. 2020, Zainal et al. 2018]. Addressing this challenge within the Bitcoin ecosystem, this paper investigates the application of the Energy-based Flow Classifier (EFC), a technique adept at identifying deviations from established norms in complex data [Pontes et al. 2019, Souza et al. 2022]. Leveraging its foundation in statistical physics, EFC quantifies the typicality of data points through an energy score, whereby normal, expected transaction patterns correspond to low energy states, and significant deviations—potentially indicative of fraud or manipulation—manifest as high-energy anomalies. The central objective of this work is, therefore, to explore the efficacy of EFC in identifying such anomalous operations within a real-world Bitcoin transaction dataset, assessing its potential as a tool for enhancing the security and integrity of cryptocurrency exchanges.

2. Data and Methods

This section details the data and methods employed in our study, which builds upon the foundational research presented in "Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity" [Lorenz et al. 2021]. That seminal work explored the use of various machine learning classifiers (e.g., Random Forest, SVM, MLP) applied to engineered features from the Elliptic dataset to identify illicit Bitcoin transactions, specifically tackling the inherent challenge of label scarcity. While demonstrating the potential of standard ML techniques, their approach relied on supervised or semi-supervised frameworks requiring at least some labels. Our research, forming part of a larger dissertation effort, diverges by investigating the Energy Flow Classifier (EFC), introduced in Section 1. EFC provides an alternative approach rooted in statistical physics, designed to model the typical 'energy' profile of normal transactions. By focusing on identifying high-energy deviations representing anomalies, EFC inherently addresses the label scarcity problem from an unsupervised or anomaly detection perspective, offering a framework potentially better suited for identifying novel or diverse fraudulent behaviors without relying heavily on pre-existing fraud labels. The following subsections detail our specific dataset, preprocessing steps, EFC implementation, evaluation task, and computational setup.

2.1. Dataset Description

This study utilizes the Elliptic dataset, a publicly available graph dataset of Bitcoin transactions introduced by Weber et al. [Weber et al. 2019] and subsequently used in foundational studies on machine learning for Bitcoin money laundering detection, including the work by Lorenz et al. [Lorenz et al. 2021] which highlighted the challenges of label scarcity.

Source and Scope: The dataset represents a temporal subgraph of the public Bitcoin blockchain, focusing on transactions involving entities identified by Elliptic Ltd., a company specializing in blockchain analytics and financial crime prevention. It captures transaction patterns over 49 distinct time steps, where each step corresponds roughly to a two-week period. The full dataset comprises 203,769 transaction nodes and 234,355 directed edges representing the flow of Bitcoin between transactions.

Features: Each transaction (node) in the graph is described by a set of 166 anonymized features. One feature explicitly denotes the time step (1 to 49). The remaining 165 features are local transactional properties, including aggregated information about the transaction's inputs and outputs (e.g., number, amounts, fees) and potentially aggregated statistics from its immediate neighborhood in the transaction graph. These features are provided in a normalized or standardized form, obscuring raw values but preserving relational patterns crucial for machine learning analysis. The graph structure itself, defined by the edges connecting transactions where the output of one becomes the input of another, provides crucial contextual information about the flow of funds, although our EFC implementation primarily focuses on the node features.

Labels: A key characteristic of the Elliptic dataset, and a central challenge addressed by Lorenz et al. [Lorenz et al. 2021] and relevant to our EFC approach, is the presence of label scarcity. While the dataset contains over 200,000 transactions, only a subset is explicitly labeled. Based on the analysis by Weberet al. [Weber et al. 2019], approximately 46,564 transactions around 23% were initially labeled. These labels classify transactions into two main categories:

- **Licit:** Transactions associated with known legitimate entities such as exchanges, miners, wallet providers, and other regulated services (approx. 42,019 instances in the original labeled set).
- **Illicit:** Transactions linked to known illicit activities, including scams, ransomware, terrorist financing, Ponzi schemes, and dark market operations (approx. 4,545 instances in the original labeled set).

The vast majority of transactions (over 150,000) remain unlabeled. Within the labeled portion, illicit transactions represent a small minority approx. 9.8% of labeled data, or just over 2% of the total dataset. This significant class imbalance and the large volume of unlabeled data make the Elliptic dataset a realistic and challenging benchmark for evaluating fraud detection techniques, particularly unsupervised or anomaly-based methods like EFC that do not strictly rely on extensive labeling. Table 1 provides a summary of the dataset statistics based on the original publication [Weber et al. 2019].

Table 1. Summary Statistics of the Elliptic Dataset (based on [Weber et al. 2019]).

Characteristic	Value
Total Transactions (Nodes)	203,769
Total Edges	234,355
Time Steps	49
Features per Node	166
Labeled Transactions	46,564 (˜23%)
- Licit	42,019 (˜90.2% of labeled)
- Illicit	4,545 (˜9.8% of labeled)
Unlabeled Transactions	157,205 (˜77%)

2.2. Data Preprocessing

Preparing the Elliptic dataset for the Energy Flow Classifier (EFC) involved several key steps focused on handling labels, selecting relevant features, scaling the data appropriately, and partitioning it for training and evaluation in a temporally meaningful way.

First, addressing the labels described in Section 2.1, we filtered the transactions based on their classification. Given that EFC operates by learning the characteristics of 'normal' data and identifying deviations, transactions labeled as 'Licit' were designated as the normal class for model training. Transactions labeled as 'Illicit' were designated as the anomalous class, primarily used during the evaluation phase to assess detection performance. The large portion of transactions with 'Unknown' labels were excluded from both training and testing in this study to ensure a clear evaluation based on known ground truth. The binary nature of the task (Licit vs. Illicit) required mapping these labels to numerical values (e.g., 0 for Licit, 1 for Illicit) for evaluation purposes.

Second, feature selection and transformation were performed. From the 166 features available for each transaction, the feature explicitly indicating the time step (ranging from 1 to 49) was separated. This temporal information was crucial for partitioning the data but was not used as a direct input feature to the EFC model itself, as the model focuses on the intrinsic properties of the transaction rather than its absolute time position. The remaining 165 anonymized features, representing transactional and local graph properties, were retained as input for the EFC. Although the original dataset description mentions some form of normalization [Weber et al. 2019], to ensure consistency and potentially improve the stability of the energy calculations within the EFC framework, these 165 features were scaled to the range [0, 1] using Min-Max scaling. This scaling was applied separately to the training and test sets (fitting the scaler only on the training data) to prevent data leakage.

Third, a temporal data split was implemented, following common practice for this dataset [Weber et al. 2019, Lorenz et al. 2021] to simulate a realistic scenario where a model trained on past data is used to detect fraud in future transactions. Transactions belonging to time steps 1 through 34 were allocated to the training set. Transactions from the subsequent time steps, 35 through 49, constituted the test set.

Crucially, the EFC model was trained *only* on the 'Licit' (normal) transactions within the training time steps (1-34). The test set (time steps 35-49) contained both 'Licit' and 'Illicit' transactions, allowing for the evaluation of EFC's ability to assign higher energy scores to the unseen illicit transactions compared to the unseen licit ones.

These preprocessing steps are summarized in Table 2.

Table 2. Summary of Data Preprocessing Steps for the Elliptic Dataset.

Step	Description
Label Handling	- Selected transactions labeled 'Licit' (Normal) and 'Illicit' (Anomaly).
	- Excluded transactions labeled 'Unknown'.
	- Mapped Licit to 0, Illicit to 1 for evaluation.
Feature Selection	- Retained 165 anonymized transactional/local features.
	- Excluded the 'Time Step' feature from model input.
Feature Scaling	- Applied Min-Max scaling to the 165 features, scaling to range [0, 1].
	- Scaler fitted only on the training data.
Data Splitting	- Temporal split: Time steps 1-34 for Training, 35-49 for Testing.
	- Training Set Composition: 'Licit' transactions only from steps 1-34.
	- Test Set Composition: 'Licit' and 'Illicit' transactions from steps 35-49.

2.3. Energy Flow Classifier (EFC) Implementation

For detecting illicit transactions within the Elliptic dataset, we employed the Energy Flow Classifier (EFC), leveraging the Python package implementation [efc 2021] based on the principles described by Ponts et al. [Pontes et al. 2019, Souza et al. 2022]. EFC operates on the premise that normal system behavior corresponds to low 'energy' states, while anomalies or deviations manifest as high-energy states. Our implementation specifically utilizes EFC as a one-class anomaly detector, tailored to the label scarcity challenge inherent in the dataset.

Model Instantiation and Interface: We primarily utilized the class-based interface provided by the package, specifically the `EnergyBasedFlowClassifier` class. As indicated in our experimental setup code [**?**], an EFC instance was configured with specific hyperparameters:

- `n_bins`: This parameter controls the discretization of the input features. Each feature's range is divided into `n_bins` intervals, forming the basis for calculating the system's state probabilities and energy. Based on our experiments and configuration ('constants.py'), a value of `N_BINS = 30` was used.
- `cutoff_quantile`: This determines the energy threshold for classifying a sample as anomalous. After fitting the model on normal data, the energy distribution of this training data is computed. The threshold (`cutoff_`) is set at the energy value corresponding to the specified quantile (e.g., `CUTOFF_QUANTILE = 0.9` in 'constants.py', meaning energies above the 90th percentile of training energies are considered potentially anomalous).
- `pseudocounts`: To avoid issues with zero probabilities when calculating energies (especially for states not observed in the training data), a small pseudocount is added. We used a value of `PSEUDOCOUNTS = 0.1` ('constants.py').

While some experimental scripts might show usage of the lower-level functional interface (`one_class_fit`, `one_class_predict`) with potentially different parameter names (like `Q` possibly relating to bins, and `LAMBDA` for pseudocounts), the core experiments relied on the `EnergyBasedFlowClassifier` class with the parameters defined above.

Training Process: A key aspect of our implementation, driven by the goal of anomaly detection under label scarcity, was the training procedure. Following the data split described in Section 2.2, the EFC model's `fit` method was called *exclusively* using the 'Licit' (normal) transactions from the training time steps (1-34). This aligns with the one-class classification paradigm: the model learns the energy landscape characteristic of normal Bitcoin transactions based solely on examples known to be legitimate within the training period. The illicit transactions were entirely withheld during this phase.

Prediction Process: During the evaluation phase, the trained EFC model's `predict` method was applied to the test set (transactions from time steps 35-49), which contained both 'Licit' and 'Illicit' instances. For each test transaction, EFC calculates an energy score based on its features and the learned probability distributions from the training phase. If a transaction's energy score exceeded the pre-determined `cutoff_` threshold (derived from the `cutoff_quantile` applied to the training data energies), it was classified as anomalous (predicted 'Illicit', label 1); otherwise, it was classified as

normal (predicted 'Licit', label 0). The energy scores themselves (`predict_energies` method) were also used for evaluation metrics like AUC that rely on ranking rather than a hard classification threshold.

Comparison to Original EFC Package: Our usage adheres closely to the standard application of the EFC package for one-class classification/anomaly detection. We did not modify the core EFC algorithm (energy calculation, probability estimation). The primary "customization" lies in the specific application context:

1. **Strict One-Class Training:** Explicitly training only on the 'Licit' subset of the temporally defined training data.
2. **Dataset Specificity:** Applying EFC to the high-dimensional, anonymized features of the Elliptic dataset.
3. **Parameter Tuning:** While using standard EFC parameters, the specific values (`n_bins=30`, `cutoff_quantile=0.9`, `pseudocounts=0.1`) were chosen based on experimentation within this project's context.

Essentially, we used the EFC package "as intended" for anomaly detection but carefully configured its training data and parameters for the specific task of identifying illicit Bitcoin transactions in the Elliptic dataset scenario. The helper functions described in [**?**] primarily manage the data loading, preprocessing, splitting, evaluation, and visualization around the core EFC calls, rather than altering the EFC mechanism itself.

2.4. Task

The primary task addressed in this study is to evaluate the effectiveness of the Energy Flow Classifier (EFC), implemented as described in Section 2.3, for identifying illicit (anomalous) transactions within the Elliptic Bitcoin dataset. This aligns with the broader goal of exploring alternative methodologies, particularly those suited for label scarcity, compared to the supervised approaches examined by Lorenz et al. [Lorenz et al. 2021].

Specifically, the task is framed as a **one-class anomaly detection problem**. Having trained the EFC model exclusively on 'Licit' transactions from the initial time steps (1-34), the objective is to assess its ability to distinguish between 'Licit' and 'Illicit' transactions in the subsequent, unseen time steps (35-49) of the test set.

This evaluation involves two main perspectives:

1. **Classification Performance:** Using the energy threshold derived from the training data (based on the `cutoff_quantile`), we assess how well EFC classifies unseen transactions as either 'Licit' (below threshold) or 'Illicit' (above threshold). Performance is measured using standard classification metrics suitable for imbalanced datasets, such as Precision, Recall, F1-Score, and potentially Balanced Accuracy, calculated on the labeled test set.
2. **Ranking Performance:** Independent of a specific threshold, we evaluate EFC's ability to assign consistently higher energy scores to 'Illicit' transactions compared to 'Licit' transactions in the test set. This is primarily assessed using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which measures the model's ability to rank anomalies higher than normal instances across all possible thresholds.

Success in this task would demonstrate EFC's potential as a viable tool for flagging potentially fraudulent or anomalous activities in Bitcoin transactions, leveraging its unsupervised, energy-based approach to overcome challenges like label scarcity and potentially detect novel deviations from normal behavior. The results will provide insights into how EFC performs in this financial forensics context compared to implicitly known benchmarks from related studies using the same dataset.

3. Background and Related Work

WIP

4. Results

WIP

5. Conclusion

WIP

6. Future Work

WIP

7. Reproducibility

WIP

7.1. Computational Environment

WIP

References

(2021). EFC-package: Energy-based Flow Classifier. `https://github.com/EnergyBasedFlowClassifier/EFC-package`. Version 0.1.0, Accessed: [18/04/2024].

Bordo, M. D. (2008). An historical perspective on the crisis of 2007-2008. Technical report, National Bureau of Economic Research.

Lorenz, J., Silva, M. I., Aparício, D., Ascensão, J. T., and Bizarro, P. (2021). Machine learning methods to detect money laundering in the bitcoin blockchain in the presence of label scarcity.

Murphy, A. (2008). An analysis of the financial crisis of 2008: causes and solutions. *An Analysis of the Financial Crisis of*.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. `https://bitcoin.org/bitcoin.pdf`.

Pontes, C. F. T., Gondim, J. J. C., Bishop, M., and Marotta, M. A. (2019). A new method for flow-based network intrusion detection using inverse statistical physics. *CoRR*, abs/1910.07266.

Souza, M. M. C., Pontes, C., Gondim, J., Garcia, L. P. F., DaSilva, L., and Marotta, M. A. (2022). A novel open set energy-based flow classifier for network intrusion detection.

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., and Leiserson, C. E. (2019). Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics.

Zainal, A., Kamruzzaman, J., and Sarker, R. A. (2018). A review on machine learning techniques for the detection of financial statement fraud. *International Journal of Financial Studies*, 6(3):70.

Zhang, X., Wen, Y., Zhou, J., Zhang, W., and Liao, X. (2020). Financial fraud detection in cryptocurrency exchanges: A comprehensive survey. *IEEE Access*, 8:193150–193172.