Using Topological Data Analysis to Identify Distinct MEV Behaviors

0xEvan

https://twitter.com/EvanDeKim

November 8, 2022

Abstract

We use topological data analysis to identify statistically different MEV behaviors as well as transition states between these behaviors. To the best of the authors knowledge, this is the first application of the Mapper algorithm on MEV data or in a blockchain setting. We create a topological data analysis pipeline with Subgrounds [1] and Giotto-tda [16]. Then we use the Mapper algorithm [15] to analyze the underlying topological and geometric characteristics of a MEV trading dataset via a simplicial complex.

We identify six distinct MEV behaviors that all have statistically diverse profiles. We attribute positive MEV volume as positive behaviors to be encouraged and negative MEV volume as negative behaviors to be avoided. Intuitively, these distinct MEV behaviors can be considered as emotional behavior of MEV bots. For example the emotions happiness and excitement are both distinct, but positive emotions. In contrast, happiness and despair are opposite emotions - positive and negative. Emotions that are closer in similarity will also appear closer together as seen in Figure 2.

Keywords— Topological Data Analysis, Mapper, Reeb Space, Simplicial Complex, DeFi, MEV, Arbitrage, Liquidation

Introduction

MEV bot behavior is difficult to characterize. Previous work has shown that has elucidated negative relationships between MEV bots and human trading volume [20, 18]. Little is known about positive relationships between MEV arbitrageur and human trading volume within decentralized exchange (DEX) ecosystems. Initial work has been started to characterize the different types of OHM trading volume. This work extends the MEV research by applying topological data analysis to identify distinct MEV behaviors within the OHM trading and liquidation volume data. We use the Mapper algorithm to analyze the topological and geometric structure of the data to categorize MEV behavior by exploring the relationship between MEV arbitrage and liquidation volumes in a MEV dataset of OHM volume.

Maximal Extractable Value (MEV)

Maximal extractable value (MEV) refers to the process of bots extracting value from blockchain events [9]. Although MEV can be adversarial to the blockchain users, processes such as atomic arbitrage and liquidations are vital mechanics to the health of DeFi protocols. The existence of atomic arbitrage enforces the no arbitrage bounds and implies a valid market model [2]. In this article, we refer to arbitrage as a transaction that happens atomically within a single block between two on-chain entities such as DEX to DEX. Liquidations are required to restore a lending protocol's debt health when the collateral ratio becomes too high during adverse market conditions [13]. We are specifically interested in fixed-spread liquidation mechanics used by lending protocols such as Liquity, Vesta, Aave, and Compound which use the fixed-spread as an economic incentive to motivate MEV bots to liquidate unhealthy positions.

OlympusDao

OlympusDao [10] is a major DeFi protocol and has amassed hundreds of millions of assets on their treasury balance sheet. All OHM liquidity is protocol owned (POL) by OlympusDao with the majority of liquidity accessible only on-chain [5]. In contrast, other cryptocurrencies such as Ethereum and Bitcoin have significant levels of liquidity both on-chain and off-chain. Note that in the remainder of this article, OHM and gOHM will be used interchangeably. It is assumed that there exists an arbitrage opportunity between OHM and gOHM that keeps the price relatively 1:1 with each other.

Recently, OlympusDao treasury policy has been to stabilize the price of OHM around the market price of the treasury's liquid assets also known as the liquid backing. When OHM is below the liquid backing, inverse bonds are deployed which allows users to trade in OHM for stable treasury assets [12]. When OHM is trading at a premium compared to backing, LP (liquidity pool) and reserve bonds are deployed in which users trade in non-OHM assets in return for OHM.

OHM is also used as a collateral asset where users can permissionlessly borrow against their overcollateralized positions. More recently OlympusDao has introduced Flex Loans [11], which is a financial product. Flex loans allow DAOs to borrow against their OHM positions provide OHM-based liquidity for their native tokens. More recently OlympusDao has also started a permissionless undercollateralized borrowing market for OHM in collaboration with Sentiment [14]. Given that the majority of POL trading volume comes from MEV arbitrageur bots [5] and all of the liquidation volume comes from MEV liquidation bots, it is imperative to understand how MEV bots behave under different market conditions and how a combination of negative MEV trading volume and mass liquidations can quickly lead to a liquidation cascade.

Topological Data Analysis (TDA)

Topological data analysis (TDA) is a new, fast-growing field providing a set of tools that can extract topological shapes and geometric features from complex data using algebraic and combinatorial topological techniques. TDA started to receive widespread attention after Mapper was used to identify a highly treatable cluster of breast cancer patients [8].

Mapper is an algorithm developed in 2007 [15] that turns a dataset into a simplicial complex that embeds information about topological covers and nerves. Datasets of point clouds is usually a continuous Reeb space. In contrast, the simplicial complex is a discrete space. The Mapper output is a simplicial complex, a discrete combinatorial graph. By tracking topological covers and nerves, topological and geometric properties are preserved during the transformation from a continuous space to a discrete space.

Another popular TDA technique is called persistent homology, but is out of the scope of this work. Persistent homology has been used to predict asset bubbles in both the traditional stock markets and blockchain assets [19, 6].

Reeb Space

Let $f : \mathbb{X} \to \mathbb{M}$ be a continuous mapping from a topological space $\mathbb{X}$ to a metric space $\mathbb{M}$ such as $\mathbb{R}^d$. The Reeb space of $\mathbb{X}$ is the quotient space $R_f(\mathbb{X}) = \mathbb{X}/_{\sim_f}$ where equivalent points are quotiented out of the space. Two points $x, y \in \mathbb{X}$ are equivalent if $f(x) = f(y)$ and x and y belong to the same path connected component of the pre-image $f^{-1}(f(x)) = f^{-1}(f(y))$ [7].

The Reeb space reflects the evolution of level sets of f on the manifold. Equivalent points are zeroes of the function f. Since the zeroes are quotiented out of the underlying space, the Reeb space is differentiable and well-defined. In general, the ambient space of most point cloud datasets such as the MEV dataset come from Reeb spaces.

Simplicial Complex

Simplicial complexes are higher-dimensional generalizations of neighboring graphs that can be built into standard data analysis tools and algorithms as an additional feature extraction layer. Given a set $\mathbb{X} = \{x_0, \ldots, x_k\} \subset \mathbb{R}^d$ of $k + 1$ affinely independent points, the k-dimensional simplex $\sigma = [x_0, \ldots, x_k]$ spanned by $\mathbb{X}$ is the convex hull of $\mathbb{X}$ [3]. A geometric simplicial complex $K \in \mathbb{R}^d$ is a collection of simplices that have well defined properties:

1. any face of a simplex of K is a simplex of K ()
2. the intersection of any two simplices of K is either empty or a common face of both

Given that the blockchain is a countably discrete graph structure, a simplicial complex is a natural discrete graph structure for holding topological and geometric properties embedded in the blockchain data.

Mapper Algorithm

Given a finite open cover $\mathcal{U} = \{U_\alpha\}_{\alpha \in A} = \mathbb{X}$ over a topological space $\mathbb{X}$, the simplicial complex that corresponds to the nerve of the cover $\mathcal{U}$ is defined as $Nrv(\mathcal{U}) = \{\sigma \subset A | \bigcap_{\alpha \in \sigma} U_\alpha \neq \emptyset\}$. Given a continuous map $f : \mathbb{X} \to \mathbb{Y}$ where $\mathbb{Y}$ is a topological space equipped with a cover $\mathcal{U}$, we write $f*(\mathcal{U})$ as the cover of $\mathcal{X}$ obtained by considering the path connected components of $f^{-1}(U_\alpha)$ for each α. Mapper, defined as M, refers to the nerve of $f^*(\mathcal{U})$ where $M(\mathcal{U}, f) := Nrv(f^*(\mathcal{U}))$ [15].

The nerve tracks connected components of covers and records the patterns of intersections between them. Both the nerve and cover have theoretical basis in category theory as categorical

objects with universal functorial relationships between them. Mapper utilizes this functorial relationship between categorical objects to transform datasets based in continuous Reeb spaces to simplicial complexes, which are discrete spaces.

Data Availability and Collection Methodology

Historical liquidity pool data for OHM was retrieved from the Sushiswap subgraph. Prior analysis was completed on the data to create a MEV dataset. The MEV dataset categorizes MEV trading volume with the Flashbots arbitrageur address list [5] and aggregated to a daily time frame for TDA. OHM and GOHM liquidations were taken from lending protocols Vesta (Arbitrum) and Rari (Ethereum) subgraphs respectively.

Subgraph data was retrieved by Playground Analytics Subgrounds library [1]. Respective data and code can be found here [4]. Subgrounds provides a pythonic data access layer for applications querying subgraphs from The Graph Network [17].

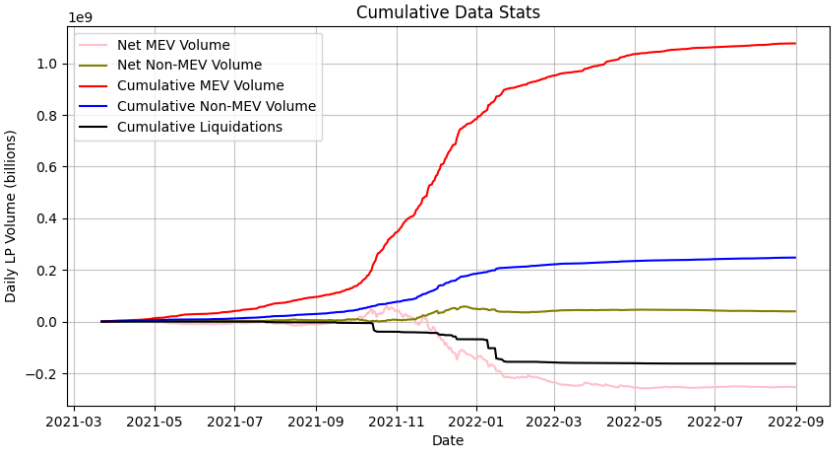


Figure 1: Historical Cumulative Stats

Historical Data Overview

Historical data was observed from March 22, 2021 to August 31, 2022. Cumulative MEV Volume and Non-MEV Volume reached 1.077b billion and 247m respectively. After accounting for positive and negative trading flows, net MEV Volume and Net non-MEV Volume were -253m and 39m respectively. Total liquidations across Ethereum and Arbitrum blockchains during this period were 162m.

Figure 1 shows a time series chart of the cumulative data. MEV volume clearly dominates the amount of non-MEV volume, but it is not immediately obvious what the quality of this volume is. Although the net volume lines are more comparable to the cumulative amount of liquidation volume, it is still not clear what relationships, if any, there are within the dataset.

It is interesting to see that negative MEV volume was a negative value whereas the non-MEV volume was positive. This immediately implies that the majority of negative selling pressure originated from MEV bots.

Analysis and Results

Giotto-tda

Giotto-tda is a Python TDA library for machine learning that is built on top of scikit-learn. As well as offering integration into machine learning pipelines, giotto-tda offers state of the art performance via C++ implementations and allows for parallelization.

The giotto-tda pipeline for Mapper can be described as follows:

1. Input a dataset with an (arbitrary) clustering method and filtering function. We use a vanilla projecting filtering and the default density-based spatial clustering of applications with noise (DBSCAN)

2. Build a simplicial complex on top of the dataset that highlights the underlying topology and geometry. We keep the cover intervals to the default of 10, but boost the overlap measure from 0.1 to 0.2. This was done ad-hoc after playing around with different values. In general, trying different values did not have a significant change of the results.

3. Explore nodes of interest from the simplicial complex to better understand the dataset.

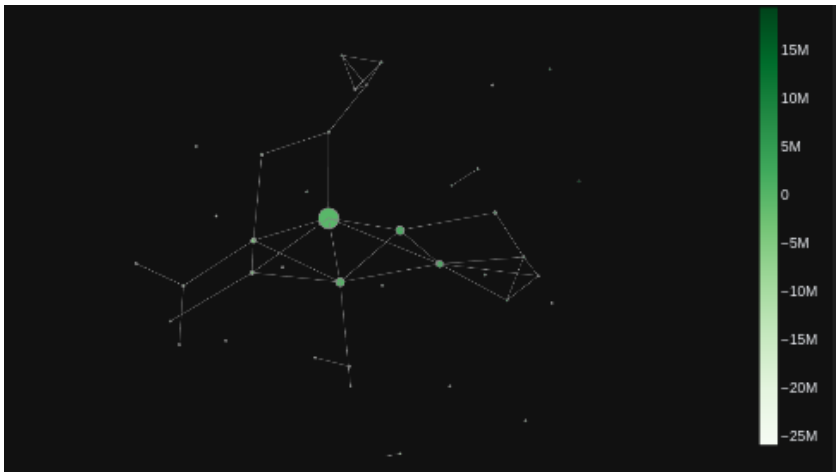


Figure 2: Simplicial Complex from Mapper Output colored by MEV volume

Mapper Simplicial Complex Output

Applying Mapper to the dataset outputs a simplicial complex as shown in Figure 2. The nodes in the simplicial complex are used to identify interesting connected clusters of data points. Although

advanced filter functions such as principal component analysis (PCA) can be implemented, we initially use the default parameters for filter functions and partial clustering inputs for simplicity.

The simplicial complex shows significant levels of connectivity around the larger nodes, sized with respect to MEV volume. Within each node there exists a set of data points. Nodes that contain intersecting data points are connected by an edge. The more data points shared between nodes, the more connected and close the nodes will appear in the graph.

There are a total of 40 nodes in the simplicial complex. Many of these nodes are not connected to the wider simplicial complex. 22 of these nodes have less than 3 data points, which isn't enough information to draw statistical examples from. Since we are interested in exploring the connectivity relationships between nodes, the nodes are sorted out for the top 8 largest nodes where 8 was chosen after inspecting the simplicial complex. Any nodes that intersect 100% with the base node are discarded for the initial analysis.

Base and Non-Base Nodes

The largest node is referred to as the base node or node 0 respective and contains 407 daily data points. At 77% of overall datapoints, the base node represents the majority of data. As a result, we consider the base node as the statistical control group. We expect the base node MEV behavior as a default behavior that occurs on most days.

The remaining node data is aggregated into the table below with the number of data points and amount of intersection with the base node:

Node	Total Data Points	Base Node Intersection Rate
Node 0 (base)	407	100%
Node 1	72	81.9%
Node 2	69	79.7%
Node 3	43	58.1%
Node 4	26	26.9%
Node 5	14	21.4%

Higher levels of intersections imply more connectedness and thus more similarity between the two nodes. Intuitively this similarity can be thought of as emotions. The emotions happiness and excitement are distinct but share much more similarity than happiness and despair. This is further validated geometrically as shown in Figure 4(a).

Identifying MEV Behavior

This section focuses on MEV behavior identified by the distinct statistical properties. Figure 3 shows the statistical properties after removing the Intersecting points. After applying Mapper, we characterize MEV behavior by the nodes connected to the base node. Some data points belong in multiple nodes. These intersecting data points suggest a transition state from one MEV behavior to another. We look at the nodes and examine beneficial or negative traits associated with them. Positive MEV volume is a desirable MEV behavior because it implies that MEV bots are contributing positive buying pressure. Conversely negative MEV would not be a desirable MEV behavior because that implies heavy selling pressure coming from bots.

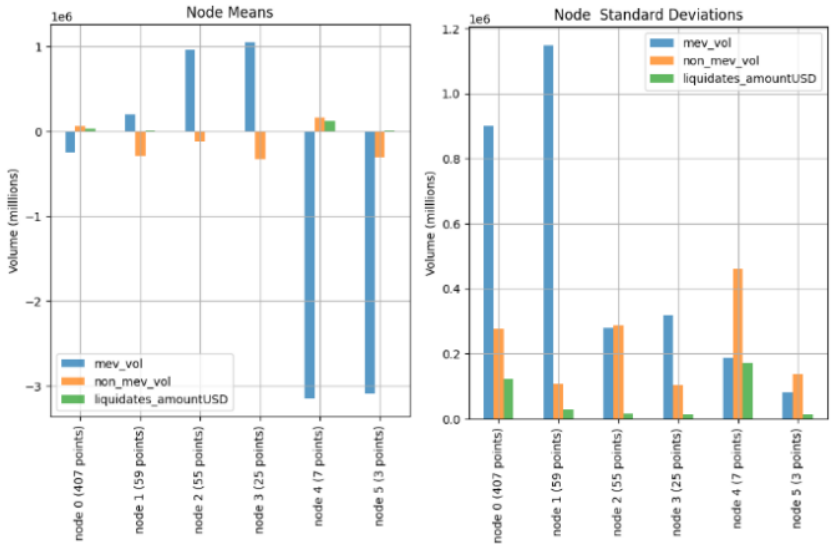


Figure 3: Node Behavior Statistics

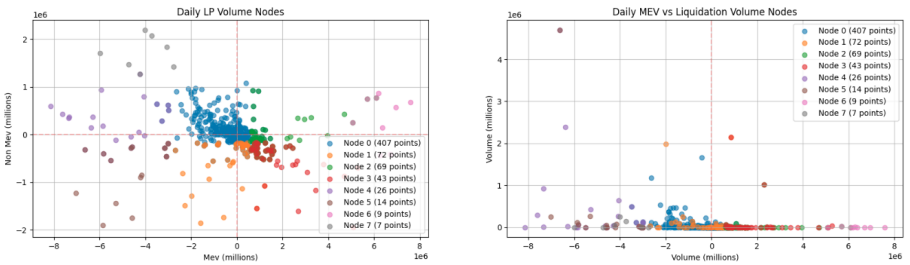
Node 1 appears the most similar to the base node. Indeed node 1 has a 81.9% intersection rate with the base node. In contrast, whereas the base node has a negative MEV volume mean, node 1 has a positive MEV volume and appears to be counteracting larger negative non-MEV mean volume.

Nodes 2 and 3 also intersect significantly with the base node at 79.7% and 58.1%, respectively. They are statistically different from the base node because they have the largest positive MEV volume means, whereas the base node's mean volume is negative. Nodes 2 and 3 have a much higher positive MEV volume mean than node 1 and share similar characteristics where non-MEV volume is negative, and liquidations are largely non-existent. Where Nodes 2 and 3 differ are from non-MEV volume. Node 3 has significantly more negative non-MEV volume and a lower standard deviation than Node 2.

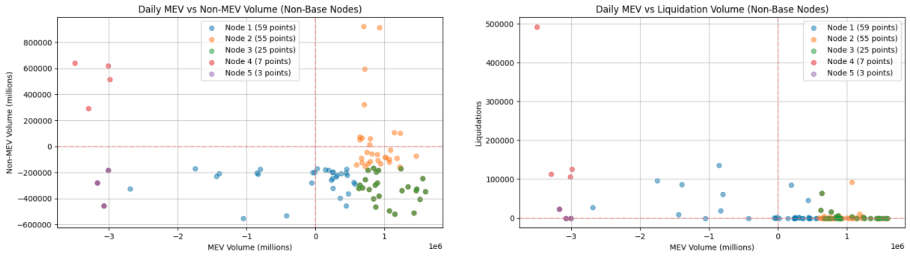
Nodes 1,2, and 3 are all beneficial MEV behaviors and counteract negative human selling pressure. However, based on the limited data in the MEV dataset, it is unclear what other external circumstances, such as the broader market, are also influencing this positive MEV volume behavior. It would be interesting to see if these behaviors hold when considering different sources of exogenous market data.

Nodes 4 and 5 are the negative outlier statistical distributions. Data points in these nodes have a significant amount of negative MEV volume with low standard deviation. Node 4 has the largest mean and standard deviations for liquidations, indicating negatively volatile days. In contrast, node 5 negative MEV volume is more correlated with negative non-MEV volume than liquidations. Nodes 4 and 5 are considered negative MEV behaviors because there appears to be a correlation between the negative MEV volume and liquidation amounts. Mass liquidations can lead to liquidation cascades and add a lot of sell pressure in the short term. Liquidation cascades imply a lot of market volatility, which runs orthogonally to OlympusDao's price stability policies.

Figure 4 shows the geometric separation of these statistical properties. Looking at 4(b) Daily MEV vs non-MEV Volume scatterplot, each cluster of node points exists within a distinct area.



(a) Nodes with Base Node and Intersection Points



(b) Nodes with Base Node and Intersection Points Removed

Figure 4: Distinct Node Behaviors

The 4(b) Daily MEV vs Liquidation Volume chart shows similar geometric separation, but appears more cluttered. These geometric distinctions create new ways to categorize heavily connected DeFi data and allows a distinct classification of MEV behavior based on daily data as well as offering a rich area of research to further understand how to incentivize more positive MEV behaviors such as Nodes 2 and 3 and avoid negative behaviors such as Nodes 4 and 5.

Conclusion and Discussion

In conclusion, Mapper builds a simplicial complex that contains the topological and geometric structure of the supplied MEV dataset. We explore and identify distinct statistical MEV behaviors contained within the simplicial complex. Although the dataset is focused on Olympus data, TDA is a general method and can be applied to analyze all aspects of blockchain data. Although this MEV dataset was relatively small and only had three columns (MEV volume, non-MEV volume, and liquidation volume), Mapper can handle much higher dimensional data and create a low dimensional representation with relevant topological and geometric structures contained within the simplicial complex.

Alternatively, persistent homology can be applied to the MEV landscape to capture the intensity of MEV behavior such as whether there was a sharp burst into a specific MEV behavior or if it is a more gradual change. Observing the level of "persistence" of topological features evolve over time corresponds directly to the predictability of the unobservable processes captured in the dataset such as asset bubbles.

References

[1] P. Analytics. Subgrounds - a pythonic data access layer for applications querying data from the graph network, 2022. URL: https://github.com/Protean-Labs/subgrounds.

[2] G. Angeris, H.-T. Kao, R. Chiang, C. Noyes, and T. Chitra. An analysis of uniswap markets, 2019. URL: https://arxiv.org/abs/1911.03380, doi:10.48550/ARXIV.1911.03380.

[3] F. Chazal and B. Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists, September 2021. URL: https://www.frontiersin.org/articles/10.3389/frai.2021.667963/full.

[4] E. Kim. Using mapper to analyze topological properties of olympus. URL: https://github.com/Evan-Kim2028/tda_ohm_analysis/blob/main/README.md.

[5] E. Kim. Mev arbitrage on olympus pol, October 2022. URL: https://mirror.xyz/evandekim.eth/Mc11J16dVP7Ervk1r2Sx_wkJ7dzb7Ce60Y2EpbRBlHY.

[6] Y. Li, U. Islambekov, C. Akcora, E. Smirnova, Y. R. Gel, and M. Kantarcioglu. Dissecting ethereum blockchain analytics: What we learn from topology and geometry of the ethereum graph?, 2020. URL: https://epubs.siam.org/doi/pdf/10.1137/1.9781611976236.59.

[7] E. Munch and B. Wang. Convergence between categorical representations of reeb space and mapper. *CoRR*, abs/1512.04108, 2015. URL: http://arxiv.org/abs/1512.04108, arXiv:1512.04108.

[8] M. Nicolau, A. J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, April 2011. URL: https://www.pnas.org/doi/10.1073/pnas.1102826108.

[9] A. Obadia, A. Salles, L. Sankar, T. Chitra, V. Chellani, and P. Daian. Unity is strength: A formalization of cross-domain maximal extractable value, December 2021. URL: https://arxiv.org/abs/2112.01472.

[10] OlympusDao. Olympus dao, 2021. URL: https://www.olympusdao.finance/.

[11] OlympusDao. Flex loans, 2022. URL: https://www.olympusdao.finance/flex-loans.

[12] O. Policy. Oip-94: Interim ranged stability policy levers, July 2022. URL: https://forum.olympusdao.finance/d/1250-oip-94a-amend-interim-ranged-stability-policy-levers.

[13] K. Qin, L. Zhou, P. Gamito, P. Jovanovic, and A. Gervais. An empirical study of DeFi liquidations. In *Proceedings of the 21st ACM Internet Measurement Conference*. ACM, nov 2021. URL: https://doi.org/10.1145%2F3487552.3487811, doi:10.1145/3487552.3487811.

[14] Sentiment. Sentiment, 2022. URL: sentiment.xyz.

[15] G. Singh, F. Mémoli, and G. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition, 2007. URL: https://research.math.osu.edu/tgda/mapperPBG.pdf.

[16] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. M. Medina-Mardones, A. Dassatti, and K. Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration, 2021. URL: http://jmlr.org/papers/v22/20-325.html.

[17] TheGraph. The graph network - apis for a vibrant decentralized future. URL: https://thegraph.com/en/.

[18] Y. Wang, P. Zuest, Y. Yao, Z. Lu, and R. Wattenhofer. Impact and user perception of sandwich attacks in the defi ecosystem, April 2022. URL: https://dl.acm.org/doi/abs/10.1145/3491102.3517585, doi:https://doi.org/10.1145/3491102.3517585.

[19] P. T.-W. Yen1 and S. A. Cheong. Using topological data analysis (tda) and persistent homology to analyze the stock markets in singapore and taiwan, March 2020. URL: https://www.frontiersin.org/articles/10.3389/fphy.2021.572216/full, doi:https://doi.org/10.3389/fphy.2021.572216.

[20] L. Zhou, K. Qin, C. F. Torres, D. V. Le, and A. Gervais. High-frequency trading on decentralized on-chain exchanges, 2020. URL: https://arxiv.org/abs/2009.14021, doi:10.48550/ARXIV.2009.14021.