Background

There has been substantial discussion about the level of geographical decentralisation, both in Lido research forums and in the more general Ethereum research forums. The reason behind such concerns is the possibility of some legislation to try to ban blockchain, as has occurred with cryptocurrency mining in the past. But beyond regional jurisdiction, other parallels, like cloud providers and software centralization, also have inherent risks that should be avoided. Simon Brown (Consensys) has been looking into these risks and wrote an interesting article with a first analysis of Ethereum's contemporary level of centralisation. This article caught Vitalik's attention, and he provided some feedback on novel statistical tools that could be used to extend the previous work. This project builds on top of the previous work as a collaboration between MigaLabs and Consensys, and with support and feedback from Vitalik, to use new datasets and new mathematical tools to get a better idea of the level of (de) centralisation of the Ethereum's staking node operators.

Problem Statement

We want to measure the level of (de) centralisation in the Ethereum validator set by calculating the market concentration of node operators, adjusted for how correlated they are to each other. The reason for this is that even if we have thousands of different node operators, each with a relatively equal market share (under 1%), there could still be a relatively high risk of centralisation if they are all based in the US, all using AWS, and all using the same client software. Measuring the market share alone doesn't reflect this nuance and can result in a "false positive" regarding the effective level of decentralization in the network.

Approach

Initially, we will start by just calculating the market concentration of node operators. To do this, we use the Herfindahl–Hirschman Index (HHI) to measure market concentration based on the market share of each node operator. We then re-calculate the HHI after adjusting the market share of each node operator by applying a correlation coefficient. The correlation coefficient for a node operator iscalculated from the sum of each node's correlation scores under the operator's control. This gives us a single index based on the market share for each node operator, adjusted for how correlated they are to each other in terms of geographical location (legislation), cloud provider, client software used, and MEV relays. This index number tells how truly concentrated or decentralized the network is.

Deliverable

MigaLabs will work on this project with@orbmis (Consensys), and we will deliver a full research paper explaining the methodology used, including the datasets used for the research, the mathematical tools used, the implementation of the data analysis, and our conclusions. The objective is to publish this article in arXiv for the community to access it easily and submit it to an internationally renowned conference. Additionally, we can present the results of this research in one of the Lido node operator's calls.

Resources

For this project, we are asking for a timeline of three months and DAI$16,000 from Lido paid to the address

0x492d683a51613aBcef3AD233149d69b7FE60FBd7. LEGO will

be part of the paper’s acknowledgements in the arXiv version and at the conference.