Flipside User Scores: A Bayesian Framework for Assessing User Quality

Angela Minster, PhD and Eric Stone, MS
Flipside Crypto

July 2024

Flipside uses a Bayesian framework to assign a score to every user on every chain. We believe that a user is valuable to a chain when they generate significant and sustained economic value. Our scores are designed to help chains find the users that generate that value. Currently we calculate a daily score for over 150 million addresses across 12 chains, every day[1]. Our scoring model leverages a set of metrics that, while simple to interpret, are underpinned by a robust statistical framework designed to deliver insights that are both immediately actionable and deeply informative.

With this framework, our scores are meaningful at four distinct levels:

- **Personalized Insight:** Users gain a clear understanding of their own scores, enhancing personal engagement and contribution to the chain's ecosystem

- **Intra-chain Dynamics:** Enables effective comparison of users within a chain, at both point in time and longitudinally

- **Cross Chain Analysis:** Our approach also supports meaningful cross-chain comparisons, both at the user level and the chain level

- **Correlation with Success:** The resulting scores are correlated by chain with important metrics of success including market cap and user LTV

In practice these scores are used in a variety of ways including the following:

- Airdrop program development and success measurement
- User segmentation creation and analysis
- Longitudinal user-quality improvement through incentive programs
- Cross chain user promotion
- Intelligent incentive funding structure development
- In-depth evaluation of user engagement with the chain and protocol

Introduction to Scores

The scores are calculated using 15 metrics across 5 categories (activity, token accumulation, defi, nfts, gov). Addresses earn one point for each metric once they achieve a certain threshold in a rolling 90-day window. We remove any labeled addresses2 and contracts to get closer to a set of addresses that might be reasonably described as users.

Once the user achieves a minimum number of items to achieve a metric it then becomes less important how many they do and more important what they do, specifically the breadth of what they do. Simply giving a point to every NFT purchase has diminishing returns because the 7th to 10th NFT buys are less important than the 1st to 3rd liquid stakes.

Figure 1 shows the distribution of user scores on four selected chains and Figure 2 shows the average score for each chain that we currently score.

Figure 1: % of Users by Score

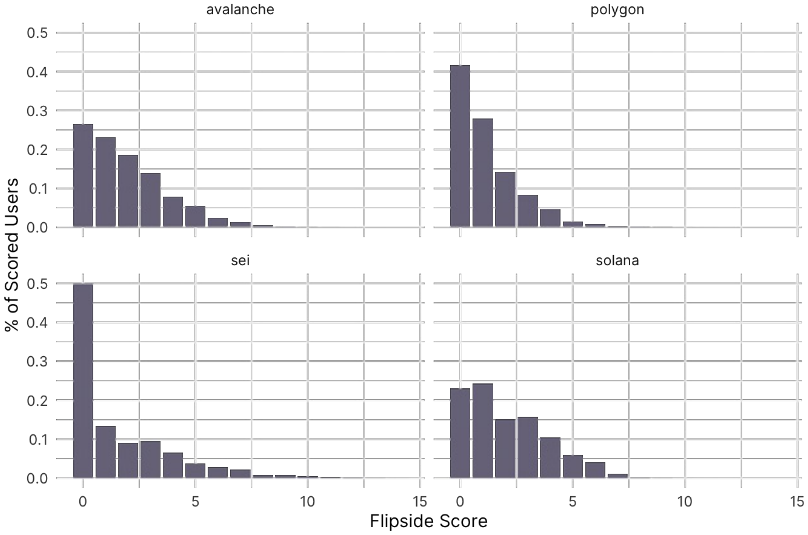
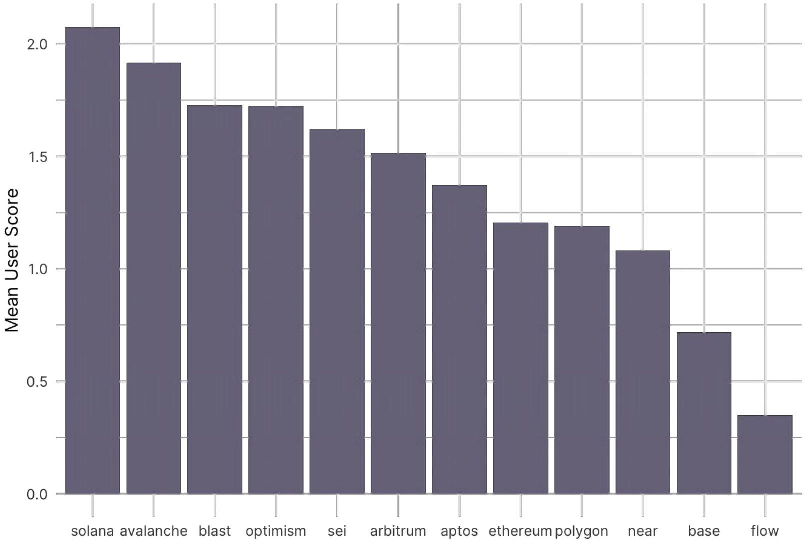


Figure 2: Mean User Score Across Scored Chains



Metric and Threshold Selection

Surprisingly, the soundness of the scoring framework remains relatively invariant to the specific metrics chosen, as long as the metrics are uncorrelated within each chain. This means that, assuming the mathematical framework in the next section is upheld, you can select a variety of metrics, and as long as they are uncorrelated at the user level, correlation to things like LTV and market cap is preserved. The interpretability at the user, protocol, chain, and cross-chain levels also remains consistent.

Subject matter expertise is necessary in constructing metrics; however, assuming the metrics are in fact uncorrelated, the math of threshold selection becomes more important than the specific metrics chosen. This flexibility in metric selection offers a significant advantage in constructing scores across chains, allowing for the incorporation of scores from different ecosystems while maintaining cross-ecosystem analysis.

In this paper, we will treat the specific metrics used by Flipside as proprietary. However, we will disclose that the Flipside Scores rely exclusively on transaction counts, rather than token or USD volumes. This approach offers two main advantages: 1) it is simple to calculate, explain, and understand; and 2) despite the low correlation between token volumes and transaction counts at the user level, our scoring methodology shows a correlation between scores and user lifetime value, for example, demonstrating that we are able to extract valuable information from previously obscured data.

This brings us to the most difficult and important piece of the process which is the selection of thresholds necessary to users to achieve points on each metric. As mentioned above, it is only through a strong mathematical framework that the scores become meaningful as a quantitative metric. The following section details this framework.

The Math

The key point to understand this section is that in order to calculate scores for a blockchain we need to estimate an unobservable feature: the distribution of address quality across the chain. While there are numerous qualitative methods for estimating user quality[3] with scores, we seek a data-driven approach so that we can facilitate the use of scores in further analysis and statistical modeling.

The math is actually quite simple. First we select a Poisson-Gamma distribution to model user scores as calculated from the selected set of metrics, with the assumption that scores on all chains are drawn from the same Poisson-Gamma distributed population. We select the Poisson-Gamma distribution because it is well-suited for modeling overdispersed count data, which we find on all blockchains. The goal then becomes estimating the parameters of this distribution: r and λ, for each chain. We use a Hierarchical Bayes model to estimate the parameters, i.e. to estimate the posterior distribution of scores on each chain. Finally we use an optimization procedure to select thresholds that create scores that fit as closely as possible to the estimated posterior distribution of each chain.

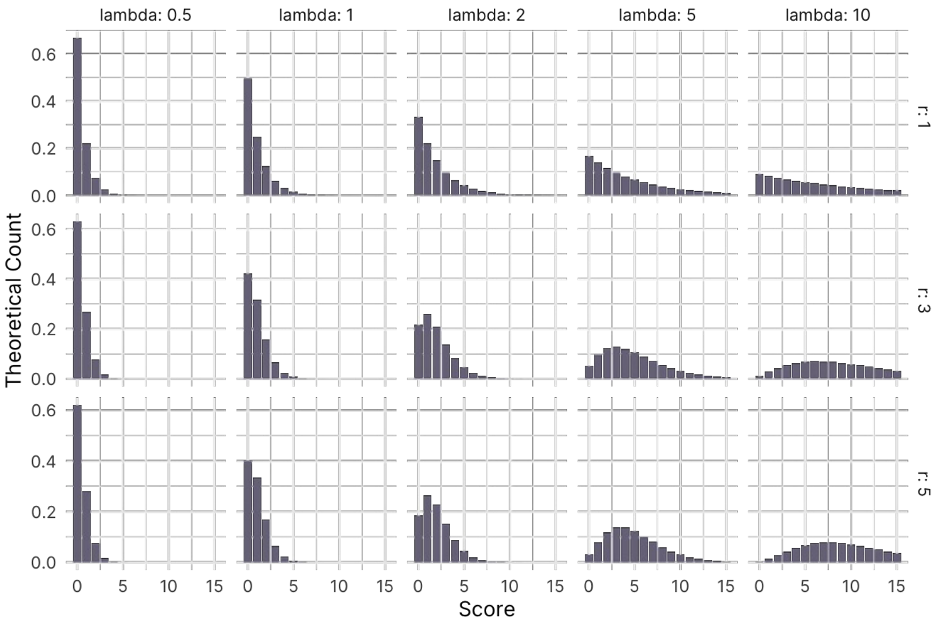
As with any Bayes solution, we estimate the posterior distribution from the empirical evidence (likelihood) and the prior distribution(s).

By utilizing a hierarchical Bayesian framework, we can account for variability across chains while also leveraging shared information to improve the accuracy of our estimates. The empirical evidence will be derived from the chains themselves, which presents a non-trivial task as the data we're modeling is latent and unobservable. The hierarchical Bayesian model allows us to estimate the posterior distributions of the parameters by incorporating both the data and the prior distributions. The initial assumptions for the priors are also informed by the data, ensuring that the model is grounded in empirical evidence. Although we know the form of the prior distributions, finding reasonable estimates for the hyperparameters is crucial for the model's accuracy and robustness.

Let's lay this all out more clearly:

To construct estimates for the shape and rate parameters of the Poisson-Gamma distribution, we need to gather two pieces of evidence. The shape parameter, r, helps us understand the concentration or dispersion of user activity quality on a chain. The rate parameter, λ, provides insights into the average rate of user activity. Figure 3 shows how different values of r and λ will affect the shape of the distribution of user scores.

Figure 3: Poisson-Gamma Distribution Across Varying λ and r



Evidence for Parameter Estimation

So what evidence do we have to estimate r and λ? The two metrics we will use come from the following observable distributions:

Fees Paid by All Score-able Users Over the Last 180 Days:

This distribution provides the mean (μ) and variance (σ2) of the fees:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \text{fees}_i$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (\text{fees}_i - \mu)^2$$

Distribution of "Naive" Scores: Naive scores are calculated with all thresholds set to 0. We look at the percentage of users scoring from 1 to 3, denoted as4.

The formula for p is:

$$p = \frac{\text{Number of users with naive scores } \leq 3}{\text{Total number of users}}$$

Estimating Empirical r and θ

From these distributions, we estimate the empirical parameters r and θ as follows. For the Poisson-Gamma (Negative Binomial) distribution, the mean (λ) is related to r and θ by:

$$\lambda = \frac{r}{\theta}$$

The probability p is related to λ and θ by:

$$p = \frac{\theta}{\lambda + \theta}$$

We can solve for r and θ using the mean (μ) and variance (σ2) of the fees:

$$r = \frac{\mu^2}{\sigma^2}$$

$$\theta = \frac{\mu}{\sigma^2}$$

Substituting $\lambda = \frac{r}{\theta}$ into $p = \frac{\theta}{\lambda+\theta}$:

$$p = \frac{\theta}{\frac{r}{\theta} + \theta} = \frac{\theta^2}{r + \theta^2}$$

Summary Formulas

To summarize, we estimate the parameters r and θ using:

$$r = \frac{\mu^2}{\sigma^2}$$

$$\theta = \frac{\mu}{\sigma^2}$$

And the relationships:

$$\lambda = \frac{r}{\theta}$$

$$p = \frac{\theta^2}{r + \theta^2}$$

Likelihood, Priors and Posterior Distribution (it's Bayes Time)

From here we could proceed directly to threshold optimization, but doing so would mean missing out on the valuable information that can be extracted through the hierarchical Bayesian model. This model leverages the assumption that all chains originate from the same population thus allowing us to extract more robust information. Without incorporating this step, the scores would lack the necessary comparability and reliability across different chains.

Likelihood
As described above, we model (estimate) two main aspects of our data: the mean scaled fees and the
p-estimate of low-performing addresses.

Mean Scaled Fees: We assume that the mean scaled fees (μ) for each chain follow a Gamma distribution with shape parameter r and rate parameter θ. The Gamma distribution is chosen due to its flexibility in modeling positively skewed data.

$$\mu_j \sim \text{Gamma}(r_j, \theta_j)$$

- The log link function is used to model the mean scaled fees on the log scale, allowing us to handle a wide range of values.

$$\log(\mu_j) = \alpha + \beta_{\text{chain}_j}$$

p-**Estimate:** The p-estimate represents the proportion of users scoring from 1 to 3 in the naive scores. We assume that the p-estimate for each chain follows a Beta distribution, which is appropriate for modeling proportions.

$$p_j \sim \text{Beta}(\alpha_p, \beta_p)$$

Similar to the fees, the model includes a logit link function for the p-estimate.

$$\text{logit}(p_j) = \alpha_p + \beta_{\text{chain}_j}$$

Priors

We specify the following priors for the parameters in our hierarchical model:

Weighted Mean Score: Intercept:

$$\alpha \sim \text{Normal}(\mu_{wms}, \sqrt{\sigma_{wms}^2})$$

where μwms is the average weighted mean scores and $\sigma^2{}_{wms}$ is the variance of the weighted mean scores.

Standard deviation of the chain-specific effects:

$$\sigma_{\text{chain}} \sim \text{Student-t}(3, 0, 2.5)$$

Shape parameter of the Gamma distribution:

$$r \sim \text{Gamma}\left(\frac{\mu_{wms}^2}{\sigma_{wms}^2}, \frac{\mu_{wms}}{\sigma_{wms}^2}\right)$$

p-**Estimate:** Intercept:

$$\alpha_p \sim \text{Normal}(\mu_{p\_est}, \sqrt{\sigma_{p\_est}^2})$$

Standard deviation of the chain-specific effects:

$$\sigma_{\text{chain}} \sim \text{Student-t}(3, 0, 2.5)$$

Phi parameter of the Beta distribution:

$$\phi \sim \text{Gamma}\left(\frac{\mu_{p\_est}^2}{\sigma_{p\_est}^2}, \frac{\mu_{p\_est}}{\sigma_{p\_est}^2}\right)$$

Posterior

The posterior distribution combines the likelihood and the priors to update our beliefs about the parameters given the observed data. The joint posterior for our model is expressed as:

$$p(\alpha, \sigma_{\text{chain}}, r, \alpha_p, \phi \mid \text{data}) \propto p(\text{data} \mid \alpha, \sigma_{\text{chain}}, r, \alpha_p, \phi) \cdot p(\alpha) \cdot p(\sigma_{\text{chain}}) \cdot p(r) \cdot p(\alpha_p) \cdot p(\phi)$$

### Model Fitting and Outputs

We fit the joint model using Markov Chain Monte Carlo (MCMC) sampling to obtain samples from the posterior distribution. This involves running multiple chains and iterating to ensure convergence and accurate parameter estimates.

The output of the Hierarchical Bayesian model provides posterior distributions for the key parameters of interest. These include the mean scaled fees and the pp-estimate for each chain, as well as the hyperparameters that describe the overall distribution across chains.

Posterior Distributions

The posterior distributions allow us to derive credible intervals and point estimates (e.g., mean, median) for each parameter:

Weighted Mean Score (μj): - For each chain j, we obtain the posterior distribution of the weighted mean score. - This distribution provides a range of plausible values for μj, along with a central tendency measure (e.g., posterior mean or median).

p-**Estimate (pj)**: - For each chain j, we obtain the posterior distribution of the p-estimate. - This distribution provides the range and central tendency measure for the proportion of low-performing users (users scoring from 0 to 2).

Hyperparameters: - We also obtain posterior distributions for the hyperparameters ($\alpha,\sigma_{\text{chain}},r,\alpha_p,\phi$) that describe the overall distribution across chains. - These hyperparameters help us understand the variability and general characteristics of the chains.

Using the Model Output in the Optimization Step

The posterior distributions obtained from the Hierarchical Bayesian model are used to inform the threshold optimization process for each chain. The goal is to find the optimal thresholds for user scores that align the empirical distribution with the theoretical Poisson-Gamma distribution derived from the model.

Optimization Process

1. **Initial Parameters:**

- Use the posterior means (or medians) of r_j and $\lambda_j = r_j / \theta_j$ for each chain j as the initial parameters for optimization.

2. **Objective Function**

- The objective function is the chi-squared statistic, which measures the difference between the observed (empirical) distribution of user scores and the expected (theoretical) distribution based on the Poisson-Gamma model.

$$\chi^2 = \sum_{k=0}^{K} \frac{(O_k - E_k)^2}{E_k}$$

where O_k is the observed frequency of users with score kk and $EkEk$ is the expected frequency calculated from the Poisson-Gamma distribution.

3. **Expected Frequencies:**

- Calculate the expected frequencies E_k using the Poisson-Gamma distribution with the posterior estimates of r_j and λ_j:

$$P(Y = k \mid r_j, \lambda_j) = \binom{k + r_j - 1}{k} \left(\frac{\lambda_j}{\lambda_j + r_j}\right)^k \left(\frac{r_j}{\lambda_j + r_j}\right)^{r_j}$$

- Multiply the probabilities by the total number of users N to get the expected counts:

$$E_k = N \cdot P(Y = k \mid r_j, \lambda_j)$$

4. **Optimization Algorithm:**

- Adjust the thresholds T_m for each metric m to minimize the chi-squared statistic. This can be done using numerical optimization methods such as gradient descent or other suitable algorithms.

5. **Iterative Process:**

- Repeat the optimization step for each chain, using the posterior distributions to update the thresholds and ensure that the empirical distributions align with the theoretical Poisson-Gamma distributions across all chains.

6. **Calculate Scores:**

- Once the thresholds are determined for each blockchain, calculate daily scores for users across chains using the chain-specific thresholds.

This process is repeated periodically to ensure that structural changes in the distributions of chains are reflected in the scores.

Results

The results of step 6 for the most recent Flipside optimization and score calculation are presented in Figure 4 below. Note that once optimization is completed and thresholds are set, scores can and will continue to change every day as different types of users become active and dormant within the ecosystem. The goal of a chain should be to growh a healthy use base over time, shifting the curve to the right. This makes Flipside's scores both roadmap for improving a chain (or protocol)'s users and a metric to use to measure success over time.

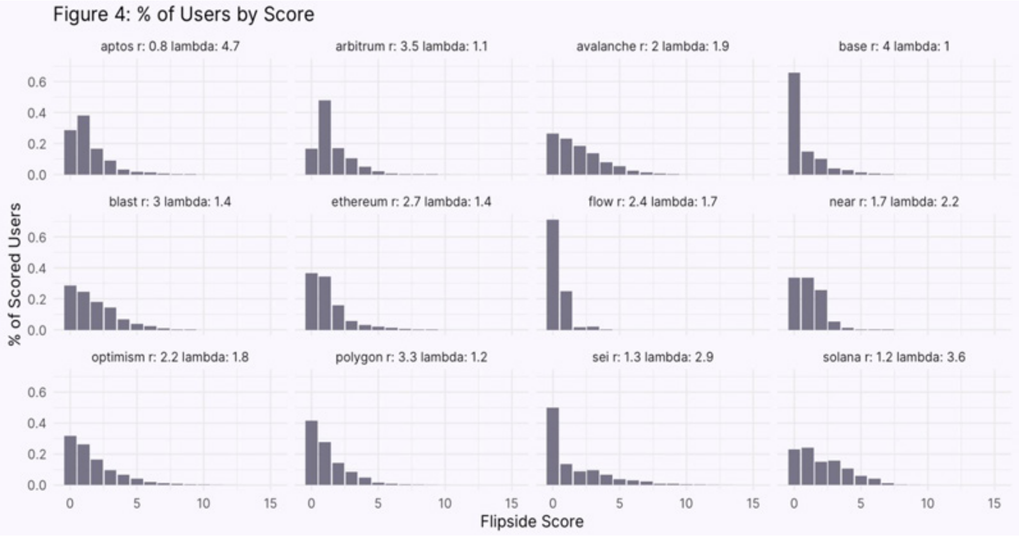


Figure 4: % of Users by Score

Final Thoughts

In this paper, we outlined a robust hierarchical Bayesian framework to estimate user activity scores across blockchains. Our primary objectives were to model the distribution of user scores and optimize the thresholds for these scores to align the empirical distributions with theoretical distributions derived from our model.

With this approach we have scored over 150 million users on 12 chains (and counting) and the resulting scores, due to their theoretical foundations, are useful for a variety of quantitative tasks not limited to: cross chain analysis, incentive program development, airdrop program development and other statistical modeling and machine learning based analysis.

Footnotes

[1] We are regularly adding new chains.

[2] Flipside maintains an extensive collection of blockchain addresses labels which are available at flipsidecrypto.xyz.

[3] We consider that most scoring metrics, while technically using numbers, are still qualitative because there is no theoretical framework behind them that would allow them to be soundly used as a quantitative metric, i.e., they lack the rigorous foundation needed for consistent and reliable numerical analysis.