1. Introduction

Spotify is a global music streaming platform where artists share their work with audiences worldwide, offering listeners an extensive collection of songs. To foster a strong community and support artistic creation, two-thirds of Spotify's revenue goes to a royalty pool, which is paid out to music rights holders (Spotify, n.d.). However, those with malicious intent take advantage of the system by running automated scripts to artificially inflate streams for financial gain. With the emergence of artificial intelligence, differentiating between genuine users and bots has become increasingly challenging (Sun, 2024).

2. Objective

Fraud detection models are not perfect and will misclassify fraudulent streams as legitimate. Rampant artificial streams will damage Spotify's branding, platform integrity and trust from their users and artists. As a result, it is imperative for Spotify's data science team to determine their current detection coverage to identify areas for improvement. The objective of this report is to propose a methodology to estimate the proportion of undetected fraudulent streams.

3. Randomized Controlled Testing

The proposed solution entails randomly assigning artists to either a test group or a control group. The initial two weeks is spent on data collection before informing the test group of a higher incentive for the next two weeks, encouraging fraudsters to increase their fraudulent activities. Afterwards, we compare the results between the two groups, as well as pre- and post-incentive data, to identify patterns of artificial streaming.

3.1 Sample Population

According to Spotify's December 2024 quarterly report (*Form 20-F* 2024), the platform has 675 million monthly active users, 12% growth from the previous fiscal year. Using the Sample Size Calculation for Proportions formula with a 99% confidence interval and 0.1% margin of error, the required sample size to estimate fraud, with a 5% fraud rate, is 316,179 listeners. To ensure more robust methodology**,** a more conservative approach of 50% fraud percentage requires 1.66 million listeners, rounded up to 2 million to enhance result accuracy.

Research by Gustar (2020) showed that artists' monthly average listeners follow a Power Law distribution, where a small group of top artists get most of the streams and exposure. We first generate a Power Law distribution of artists and exclude the bottom and top 25 percentile. Iteratively sample the remaining artist until their cumulative monthly listeners exceeds 2 million.

3.2 Metrics

The two key metrics used in this report are the total streams per artist and streams per song. These metrics will help to assess the impact of incentives on streaming behavior within both groups.

3.2 External Variables

To accurately measure the key metrics, we must first account for external variables that could influence the results, such as advertisement expenditure, virality, playlist placement and artist collaboration. Since different artists may respond differently to these factors, we apply Bayesian Regression to model each artist. The key variables as follows:

$$For\ artist\ i\ and\ metric\ j,$$
$$y_{i,j}(t) = Observed\ value$$
$$\alpha_{i,j}(t) = Baseline\ value$$
$$\beta_{1,i,j} = Effect\ of\ incentive$$
$$\beta_{2-5,i,j} = Effects\ of\ external\ confounders$$
$$x_{1,i}(t) = Post\ incentive\ indicator$$
$$x_{2,i}(t) = Ad\ spend$$
$$x_{3,i}(t) = Virality\ score$$
$$x_{4,i}(t) = Playlist\ placement\ count$$
$$x_{5,i}(t) = Collaborations\ with\ other\ artists$$
$$\epsilon_{i,j}(t) = Random\ noise$$

The experiment collects data on the external variables and the two key metrics throughout the entire study period, split into pre incentive and post incentive.

4.1 Bayesian AB testing

Bayesian approach is more suitable for AB testing in the context of fraudulent streaming detection than frequentist approach for multiple reasons. Instead of just rejecting a null hypothesis (frequentist), Bayesian methods allow us to estimate the probability that fraudulent streaming increased by a certain amount due to the incentive. Additionally, Bayesian approach can continuously update belief without affecting statistical validity (Fornacon-Wood et al., 2021) while frequentist suffers from the multiple testing problem; running tests repeatedly increases the probability of false positive (Growthbook).

The data collected is used to refine the models' parameters defined into three phases. The first step involves tuning the model using pre-incentive results to establish a baseline for fraudulent streaming activity before any incentives were introduced. Next, post-incentive data is used to integrate the incentive effect into the models. Lastly, post-incentive data from the test and control groups are compared to isolate the effect of the incentive from external factors.

Before estimating the baseline level for both the test and control group, it is necessary to specify appropriate priors for all model variables ($\alpha_{i,j}, \beta_{1,i,j}, \dots, \beta_{5,i,j}$) and define a likelihood function. These priors are the reasonable initial assumptions about the distribution of the variables, informed by historical data and prior studies. The likelihood function describes how the observed metrics are distributed given the model's variables. Since the number of streams is positive and exhibits high variance across artist, a gamma distribution may be a suitable likelihood function.

After defining the priors and likelihood function, Bayesian inference is performed during each of the four phases using Markov Chain Monte Carlo (MCMC) simulations to approximate the posterior distributions of the variables. The posterior distribution represents the updated beliefs about the variables after incorporating observed data. MCMC methods, such as the Metropolis-Hastings algorithm, are used to generate samples from the posterior distribution, allowing the models to iteratively adjust parameter estimates based on observed streaming data (Brownlee, 2019).

4.2 Baseline Effects (Pre-Incentive)

After establishing a preliminary model with defined priors and a likelihood function, pre-incentive data is continuously collected to dynamically refine the model's understanding of both the test and control group. This process involves updating $\alpha_{i,j}$ (baseline level) and $\beta_{2-5,i,j}$ (external confounders) parameters to reflect the pre-incentive state of fraudulent activity.

After the first week of data collection, perform Bayesian inference using MCMC methods to update the posterior distributions of the model's parameters. With the trained model, check the model's predictions are fitting the observed data using performance metrics such as accuracy and residual analysis. If the fit is poor, collect additional training data in the subsequent weeks, change the priors or adjusting the external variables. The estimated time duration for the first phase is from two to four weeks depending on the model's initial fit.

4.3 Post-Incentive

After implementing the incentive, post-incentive data is continuously collected to adjust the model's posterior distributions for the incentive effect ($\beta_{1,i,j}$) and external factors ($\beta_{2-5,i,j}$) on a regular weekly basis.

If the incentive effect ($\beta_{1,i,j}$) is small or test group's streaming activity did not increase substantially, the incentive strategy may require revision like increasing the royalty split or refining the incentive parameters. The post-incentive comparison would constitute approximately two to four weeks.

5. Test vs Control Comparison

In this report, Fraud Proportion is defined as the ratio of fraudulent streams to the total observed streams. After completing the models' adjustments, we obtain the posterior distributions' parameters for β_1 for each metric and artist.

5.1 Computing $\boldsymbol{\beta_1}$ in the Control Group

Within the control group, we first compute the mean and standard deviation for β_1:

$$\bar{\beta}_{1,j} = \frac{1}{N}\sum_{I=1}^{N} \beta_{1,i,j}$$

$$s_{\beta_{1,j}} = \sqrt{\frac{1}{N-1}\sum_{I=1}^{N} (\beta_{1,i,j} - \bar{\beta}_{1,j})^2}$$

If β_1 follows a normal distributed, the 95% confidence interval (CI) is given by:

$$\bar{\beta}_{1,j} \pm 1.96 \times \frac{s_{\beta_{1,j}}}{\sqrt{N}}$$

If the confidence interval contains 0, it suggests that external confounders have been well-accounted for. If the interval is centered around 0 with small variance, it shows strong evidence that incentive effects are isolated in the test group.

5.2 Handling Non-Normal Distribution Using HPD Interval

If β_1 does not follow a normal distribution, it is then necessary to compute the Highest Posterior Density (HPD) Interval using MCMC sampling. For each artist in the control, we obtain S posterior samples:

$$\beta_{1,i,j}{}^{(1)}, \beta_{1,i,j}{}^{(2)}, \ldots, \beta_{1,i,j}{}^{(S)}$$

These samples are pooled together across all artists in the control group and sorted in ascending order. The 95% HPD interval is determined by finding the smallest range that contains 95% of the samples (i.e. smallest difference between max and min β_1 values). The conclusions drawn from the HPD interval follows the same interpretation as the normal CI approach (Tolonen, 2019).

5.3 Assessing the impact of the incentive on Fraudulent Activity

If the incentive was substantial enough to trigger a significant increase in fraudulent streams, the pre-incentive undetected fraud can be considered negligible relative to the post-incentive surge in fraudulent activity. This assumption can be validated by analyzing the false negative

rate of the fraud detection model during training and comparing between detected fraudulent streams and industry benchmark for expected fraud levels. If the mean posterior estimate $\bar{\beta}_{1,j,test}$ in the test group is significantly greater (e.g. 5-10 times) than the expected number of undetected fraudulent stream ($false\ negative\ rate\ \times\ observed\ streams\ post\ incentive$), we can reasonably assume that the pre-incentive undetected streams are negligible in comparison to the rise in fraudulent stream post-incentive. Subsequently, we compute the estimate number of fraudulent streams within the test group and the entire artist population.

$$\bar{\beta}_{1,j,test} = \sum_{i=1}^{n} \bar{\beta}_{1,i,j,test}$$

$$\bar{\beta}_{1,j} = \sum_{i=1}^{N} \bar{\beta}_{1,i,j}$$

Where n is the sample size and N is the population size.

5.4 Estimating the Total Number of Fraudulent Streams Pre-Incentive

Detection rate for metric j is defined as:

$$Detection\ Rate_j = \frac{True\ Positive\ Post\ Incentive\ Test\ Group}{True\ Positive\ Post\ Incentive\ Test\ Group + \bar{\beta}_{1,j,test}}$$

Incorrect detections are easier to identify through user complaints and appeal tickets than false negatives, allowing us to estimate the number of true positives.

To estimate the total pre-incentive fraudulent streams, we divide the pre-incentive true positive by the detection rate.

$$\hat{F}_{pre,j} = \frac{True\ Positive\ Pre\ Incentive}{Detection\ Rate_j}$$

This estimation is repeated for both metrics, and the final estimate of undetected fraudulent streams pre-incentive is computed by a weighted combination of these estimates:

$$\hat{F}_{pre} = x\left(\hat{F}_{pre,streams\ per\ artist}\right) + (1-x)(\hat{F}_{pre,streams\ per\ song})$$

The weight factor x is determined based on empirical behavior study of fraudsters, where historical data and known fraud cases are analyzed to identify whether fraudulent streams are more concentrated at the artist level, such as bot-driven farms to artificially boost an artist's total

streams, or at the song level, where targeted fraud boosts a specific song's popularity. By incorporating insights from these studies, we can provide a refined estimate of undetected fraudulent activity.

6. Appendix

Spotify. (2024, December 31). Form 20-F. https://s29.q4cdn.com/175625835/files/doc_financials/2024/q4/8afe1e0f-192e-43ad-b8d1-aa947b389577.pdf

Sun, D. (2024, November 12). What happens when it becomes almost impossible to spot a scam, even for experts?. The Straits Times. https://www.straitstimes.com/singapore/what-happens-when-it-becomes-almost-impossible-to-spot-a-scam-even-for-experts

Royalties. Spotify. (n.d.). https://support.spotify.com/us/artists/article/royalties/

Gustar, A. (2020, July 6). *Fame, obscurity and power laws in music history: Empirical musicology review*. Fame, Obscurity and Power Laws in Music History | Empirical Musicology Review. https://emusicology.org/index.php/EMR/article/view/7003/5601

Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J., & Price, G. (2021, December 9). Understanding the differences between Bayesian and Frequentist Statistics - International Journal of Radiation Oncology, Biology, Physics. https://www.redjournal.org/article/S0360-3016(21)03256-9/fulltext

Growthbook. (n.d.). Multiple testing corrections. GrowthBook Docs. https://docs.growthbook.io/statistics/multiple-corrections#:~:text=What%20is%20the%20multiple%20testing,%2C%20or%20multiple%20comparisons%2C%20problem.

Tolonen, V. H. & T. (2019, March 13). Bayesian inference 2019. Chapter 3 Summarizing the posterior distribution. https://vioshyvo.github.io/Bayesian_inference/summarizing-the-posterior-distribution.html

Brownlee, J. (2019, November 5). A gentle introduction to Markov chain Monte Carlo for probability. MachineLearningMastery.com. https://machinelearningmastery.com/markov-chain-monte-carlo-for-probability/