Uncovering Niche Insights in Long-Tail Markets: A Unified LNRE Framework

Abstract

In long-tail markets, analysis predominantly is skewed towards the "head," while the "tail," rich with niche insights, often goes uninvestigated. The tail, however, stores valuable information that can expose emergent trends, sometimes contradictory to general findings derived from the head. This research highlights that while conventional machine learning algorithms struggle to model rare events due to token sparsity, our model successfully captures these elusive insights. By leveraging the mathematical properties of G-functions and Q-functions, this study integrates their behavior into the traditional fabric of machine learning frameworks, which improves rare event modelling. We show that uniform partitioning of data into subintervals facilitates granular analysis, enabling the detection of rare events. The findings highlight how long-tail contributions not only dominate but also grow with specific thresholds, uncovering emergent trends. Additionally, our approach analyzes temporal trends, dynamically tracking their evolution over time. This research highlights the need to integrate LNRE models with partitioning and convergence techniques, offering a robust tool for decoding the hidden value in long-tail markets and empowering businesses to access previously invisible opportunities.

Keywords: Long-tail markets, rare-event modelling, LNRE models, temporal trend analysis, niche insights, emerging trend analysis

1. Introduction

1.1 Background

Long-tail markets refer to a market condition where a significant portion of total contributions or revenue in a system comes not only from a few high-frequency events or popular items but from a large number of low-frequency or specific events. This concept, popularized by Chris Anderson (Anderson, 2006), highlights a steady observable transition from traditional markets to digital ecosystems where niche items rule. In the era of online platforms, digital consumerism has drastically reduced the cost of inventory, and distribution, allowing businesses to cater to a wide variety of niche demands over different geographical areas and time.

Nature of Online Long-Tail Markets

In the context of online businesses, long-tail markets thrive because they reduce physical and logistics constraints drastically. Traditional markets are often constrained by shelf space, production costs, consumer reach, and competition (Khan, 2020) forcing businesses to prioritize high-demand items, and dropping the lagging products which affect business in a dormant way (Morganti, 2023). However, digital platforms like Amazon, Netflix, and Google Play Store have disrupted this myth, offering limitless "virtual shelf space" to cater to diverse tastes. As a result, items that were previously ignored in mainstream markets now find sustainable demand leading to sustainable growth. For example, less popular books, independent or even art-house movies, or niche apps may each contribute small revenues individually but accumulate into significant value due to their sheer volume over time.

Long-tail markets can be bifurcated into two key categories:

1. **High-Frequency 'Head' vs. Low-Frequency 'Tail'**: While the head of the distribution captures the lion's share of attention, the tail contains countless low-frequency tokens (products, reviews, or events) with valuable yet unexplored insights (Sastry, 2012).
2. **Dynamic Growth of Niche Events**: Long-tail markets are not static; contributions from the tail evolve as new demands emerge and previously rare items gain attention. Temporal analyses reveal peaks at specific thresholds, indicating trends (Zhou, 2011). Thus, the methods described in this study are capable of capturing the trends, tastes, or preferences in budding stages. A correct identification of these patterns can help decision-makers to take proper actions at the right time, which might add significant value to the business.

Importance of Modelling Rare Events in Online Consumer Markets

Modelling rare events in long-tail markets is vital for business growth. Such activity can **reveal hidden opportunities, predict emerging trends, and optimize resource allocation**. Rare events, represented by low-frequency contributions in the tail, are often overlooked due to their minimal individual impact (Hinz et al., 2011). However, these events frequently contain:

- **Niche Insights:** Rare tokens highlight emerging pain points, unmet demands, or budding consumer preferences (Chen & Guo, 2014).

- **Growth Signals:** A sudden rise in rare events strongly indicates the potential for **niche markets** as they have started showing symbols of merging into mainstream opportunities (Peltier & Moreau, 2012)
- **Market Resilience:** By understanding long-tail behaviour, businesses can spread their offerings, reducing excessive reliance on a few high-performing items and neglecting those products which still have a market, so far invisible, unfortunately.

Quantitative modelling of rare events, particularly using mathematical frameworks like the **Q-function** (cumulative contributions) and **G-function** (weighted contributions), can empower businesses to analyze and prioritize long-tail contributions effectively.

1.2 Motivations

1.2.1 Problems in capturing niche insights using traditional techniques

1.2.1.1 Nature of LNRE

In the context of analysis of natural language, a user has the capacity to produce at any given time in their entire lifespan, an utterance they have not produced before. This is the result of the generative property of language which outlines that any user based on her/his finite linguistic experience can produce and comprehend an 'unbounded number of grammatically acceptable utterances' (Yang et al., 2017)

Though several attempts have been made to estimate the development of new utterances every year, it is computationally impossible to count them. Oxford Dictionary adds approximately 1000-1500 new utterances every year, which leads to an infinite combination of natural language. This productive process induces a distribution which assigns non-zero probability to all unseen events leading to a long tail of rare events. This long tail is a container of information-rich data which when mined properly can lead to better decisions, especially in consumer markets.

A sequence $\{v_n\}$ of random vectors $v_n = \{v_{1n}, v_{2n}, \dots\dots\dots, v_{Nn}\}$ is called a **Large Number of Rare Events** (LNRE) sequence if

$$\lim_{n\to\infty} \inf \frac{\mathrm{E}\,\mu_n(1)}{\mathrm{E}\mu_n} > 0 \text{ and } \lim_{n\to\infty} \mathrm{E}\ \mu_n = \infty$$

From the above definition, we can see that the expected value of the infimum is divided by n. Here, $\mathrm{E}\,\mu_n(1)$ is the expected number of events that occur exactly once in n trials. By dividing it with n, we deduce the exact proportion. This normalization allows us to compare it **across samples of different sizes**, making it more comprehensible to understand the 'rarity.'

From here, we will understand the concept of LNRE in a granular way.

$\lim_{n\to\infty} \inf \frac{\mathrm{E}\,\mu_n(1)}{\mathrm{E}\mu_n} > 0$ is the general expression which denotes that there is always a positive outcome of any event occurring exactly once. The second part of the definition $\lim_{n\to\infty} \mathrm{E}\ \mu_n = \infty$ denotes that the occurrence of a unique event grows infinitely with growth in n. On the basis of this, we can say that as we keep on increasing the size of the sample, the occurrence of unique events keeps on increasing infinitely, which is typical of consumer data.

If N is fixed or small, it implies that the total number of possible events gets constrained, as there is a finite pool of events from which occurrences are to be drawn. In such cases, each event i has a fixed probability p_{in} which does not increase even as n increases. In summary, we can say that N needs to be large which would help to maintain a pool of rare events. P_n should be structured in such a way that the events keep on occurring infrequently, yet they do not vanish.

1.2.1.2 Smoothed Language Models

Smoothed Language Models (SLM) is a method of handling problems of zero or rare probabilities, particularly when encountering rare or unseen words or phrases. If a word has occurred rarely during the training, the model finds it difficult to predict the importance of the next sentence.

Laplace Smoothing

Laplace Smoothing is a technique which adds a constant value to all events to prevent them from getting vanished. Thus, the words that occur very rarely get accommodated also in the training process.

$P(w_i|w_{i-1}) = \frac{count\ (w_{i-1},w)}{count(w_{i-1})}$, where, $count\ (w_{i-1},w)$ is the number of times a word pair occurs in a sequence and $count(w_{i-1})$ is the number of times w_{i-1} occurs as a context.

Thus, if a specific sequence $(w_{i-1}w_i,) =$ ("smells", "bad") does not appear in the training process in that way, the algorithm assumes $(w_{i-1}w_i,) = 0$.

Neural language models tend to underestimate the probability of rare events, though they are very powerful models. Rare words or events have less training data to support their patterns and hence they are quite difficult to learn. This is particularly problematic in low-entropy distribution where probability mass is centred on frequently occurring common events. This leads to a biased model, where events occurring rarely are grossly underrepresented. This reduces their likelihood and relative importance (Akutagawa et al., 2007).

In high-dimensional data, rare events often align with slow dynamics. This means that rare events occur with time. In contrast, frequent events align better with fast dynamics appearing regularly and dominating the pattern learnt by the model. Thus, frequent events eclipse the rare events making it more challenging for the model to learn patterns.

Another reason why Laplace smoothing leads to a problem lies in the fact that it assigns equal weightage to rare as well as unseen, improbable events. While it can avoid the problem of overfitting and assigning zero probabilities to unseen events, it tends to give undue weightage to completely implausible events.

1.2.1.3 Backoff and Interpolation

Backoff is a popular technique where the model backs off to a simpler model when it encounters a rare event. We can take an $n-$gram for example. If we have a situation where we define $P(w_3|w_1,w_2)$ and if the model does not have enough data to predict the probabilities, then it backs off to a 2-gram. If the model finds that further data is unavailable, then it backs off to a simple 1-gram. One of the most

popular techniques followed here is Katz Backoff. In Katz Backoff follows a simple backoff, but applies a discounting factor D to the observed n −grams. It subtracts some probabilities from the seen events and assigns it to the unseen n −grams.

In summary, it can be said that the Hessian matrix for Katz backoff is found to have negative diagonal entries, which suggests concave behaviour. Due to the discrete nature of the backoff, the log-likelihood is not globally concave or convex. This non-convexity arises from the piecewise definition of $P(w_i|w_{i-2}, w_{i-1})$ leading to a problem of local minima. Thus, the optimization process becomes increasingly difficult.

For rare events, the backoff mechanism overly depends on α and lower-bound probabilities. The Hessian terms associated with α is highly sensitive amplifying small changes in rare-event probabilities. This results in unstable gradients leading to suboptimal solutions.

LNRE models, on the other hand, assume a continuous probability distribution which mitigates the problem arising from the piecewise definition as we have seen in Katz backoff. LNRE models explicitly handle rare events by focusing on power-law distributions and other mathematical models that naturally model the frequency of rare events, for example, Zipf's law or Mandlebrot distributions. The advantage that LNRE enjoys is the fact that it smooths the likelihood function by assuming a continuous underlying distribution for all events including rare ones. Thus, the resulting smooth likelihood allows for a faster convergence and an enhanced interpretability of the model.

Katz backoff depends on hierarchical levels (unigram, bigram, or trigram) backoff weights. Thus, the model is more complex and difficult to interpret. Since LNRE models do not depend on hierarchical dynamics, they can handle large volumes of text. Empirical results show that LNRE-based models can outperform Katz backoff in data involving significant long tails.

1.2.1.4 LNRE vs LSTM vs BERT

LNRE vs LSTM

During training, LSTM operates with a fixed vocabulary V_{LSTM} extracted from the training corpus. Thus, the words outside V_{LSTM} are replaced with < UNK > token.

Let V_{LSTM} denote the size of the vocabulary after training and is constant irrespective of the size of N. On the contrary, Heap's Law suggests that $V(N)$ must grow as N increases. LSTM violates Heap's Law in this way.

However, there is more to it.

Heap's law suggests that rare words contribute significantly to vocabulary growth. On the other hand, if a rare word exists in the training data of LSTM, its embedding is executed poorly. It is weakly trained because of its sparse occurrences. We have already seen

$$V(N) = kN^{\beta}$$

$\frac{d}{dN}V(N) = k\beta N^{\beta-1}$, where $\frac{d}{dN}V(N) > 0$ because vocabulary size grows with N and because $\beta < 1$ the growth rate decreases as N increases.

In LSTM, once V_{LSTM} is fixed

$$\frac{d}{dN}\left(V_{LSTM}\right) = 0, \forall\, N > N_{train}$$

This shows that no unique words can be added once training has been completed. This violates Heap's law.

Proof

Case 1: $N \to \infty$

According to Heap's law as $N \to \infty$:

$$V(N) \to \infty$$

But in LSTM $V_{LSTM} = \text{constant}, \forall\, N > N_{train}$

Case 2: contribution of rare words

Let $n(W)$ denote the frequency of word w in the corpus. Heap's law accounts for rare words where $n(W)$ << 1 by allowing them to increase $V(N)$

$$V_{LSTM} = \text{constant}, \forall\, w, V_{LSTM}$$

$$\Delta V_{LSTM}(w) \approx 0$$

Whereas under Heap's law $\Delta V(N)(W) > 0$

Numerical Comparison

We have already seen that $N \to \infty, V(N) \to \infty$

$$\therefore\ V_{Heap}(N) = \alpha N^{\beta}$$

$$\therefore\ V_{LSTM}(N) = constant, \forall\, w$$

And for large N $\lim_{N\to\infty} \frac{V_{LSTM}}{V_{Heap}} = 0$ which indicates that LSTM underrepresents vocabulary growth.

LNRE vs BERT

The Zipfian rank-frequency distribution predicts a power-law decay

$$f(r) \propto \frac{1}{r^{s}}$$

For large r (rare words), $f(r)$ decreases rapidly. Because of vanishing gradients, LSTMs fail to capture $p(w) \to 0$ for rare words.

If $f_{LSTM}(r)$ is the frequency r predicted by LSTM

$$\frac{f_{LSTM(r)}}{f_{Zipf\,(r)}} = \frac{f_{LSTM}(r)}{\frac{1}{r^{s}}} \to 0 \text{ which violates Zipf's law}$$

LNRE assigns probability mass to rare events ensuring that $p(w)$ remains non-zero even for infrequent words which capture the long-tail behavior.

$$\lim_{r\to\infty} P_{LNRE}(r) \neq 0$$

1.2.2 The need for a robust LNRE framework

In summary, it can be said that Heap's law lays a strong foundation for understanding the vocabulary growth. This is critically pertinent in long-tail markets, where rare events dominate the distribution. LNRE models help to uncover niche insights, detecting emerging trends in long-tail markets.

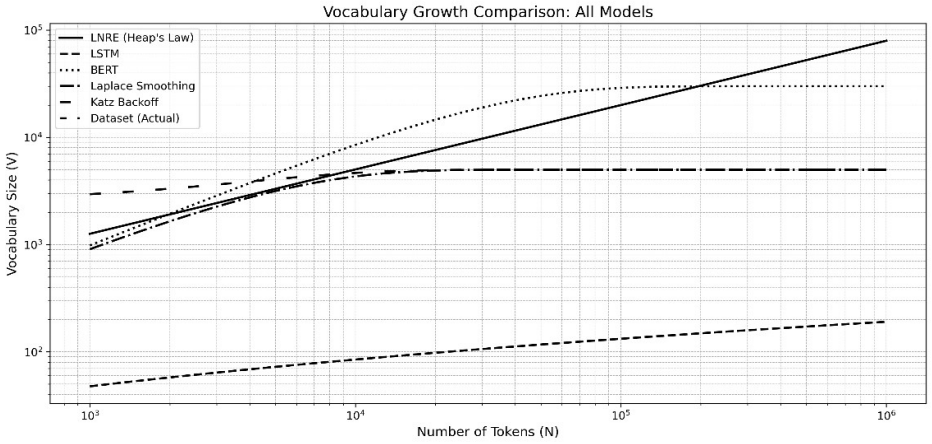


Fig I: vocabulary growth comparison across all models on the dataset

LSTM and BERT while being state-of-the-art models, fail to capture diverse long-tail phenomena. LSTMs rely heavily on sequential processing and finite memory capacity, leading to a plateau in vocabulary growth (Fig.I) as the contribution of the rare tokens is underestimated. Heap's law lays down a sublinear vocabulary growth, which the LSTM violates. BERT on the other hand with its subword tokenization technique limits vocabulary growth. Rare tokens in BERT are split into frequent subcomponents, masking the true distribution of rare events. A predefined vocabulary sets a hard upper limit, violating Heap's law which expects a continuous growth. Such models, while useful for general-purpose language modelling, suppress the natural dynamics for vocabulary in long-tail markets, where rare tokens play a crucial role.

The violation discussed in the earlier paragraphs highlights the limitations of the machine learning models in following foundational models like Heap's law. By emphasizing the capacity of LNRE to model rare event dynamics, this research validates the importance of implementing theoretical frameworks in modelling long-tail phenomena. LNRE models, combined with tools like uniform partitioning, temporal analysis, and variable substitution, provide innovative approaches for modelling and analysing long-tail data. These methods directly address the gaps left by LSTM and BERT, offering a more robust solution for niche market research.

Further, the data that was used for analysis has a long tail behaviour as indicated below

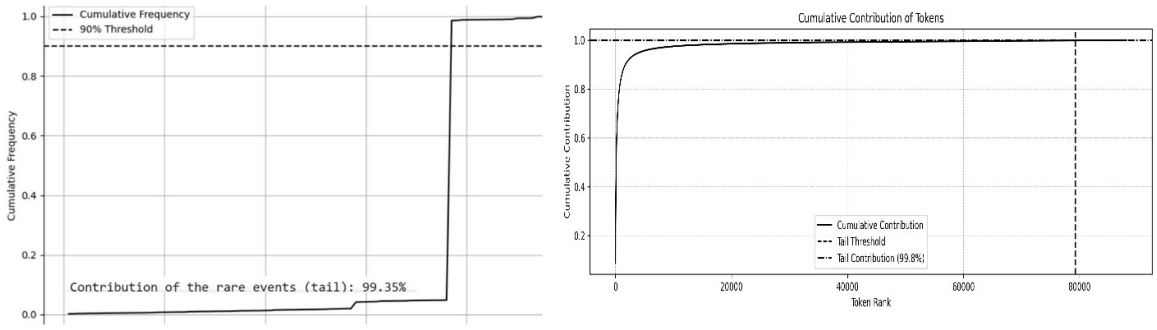


Fig 2: Contribution of rare events and contribution from the tail

1.3 Objectives

This research aims to unfurl how rare events within the long-tail distributions can help us to get ultra-specific or niche insights into consumer behaviour. Rare events hold valuable information regarding niche customer needs, pain points and unexplored avenues, even though most of the conventional machine learning algorithms consider rare events as noise. Thus, the importance of rare events is thoroughly overlooked. While serving the purpose as to how rare events can aid in unfurling niche insights, the study also aims to create an awareness to reassess the importance of rare events in future NLP research.

Another key objective is to identify the emerging trends in consumer preferences and behaviour in the long tail by analyzing the density distribution over time. Conventional models like BERT, or LSTM are unable to capture the temporal dynamics of low-frequency patterns. This research aims to analyze the contribution of the 'tail' market to 'head' markets. By implementing a two-way approach i.e., cumulative, and partition-specific approach, the study demonstrates the impact of low-frequency events in long-tail markets. Finally, the research aims to bridge the gap between theory and practice. By highlighting the misalignment between the theoretical foundation and general-purpose models, this research aims to reintegrate the theories into modern-day algorithms.

2. Related Works

2.1 Morphology of long-tail

The colossal growth of e-commerce websites has led to a deluge of data. The infinite inventory retailers such as Amazon, and Netflix have created a warehouse of data which contains textual data in them as significant parts. Such data shows two universal features:

a. In long-tail markets, most of the products are 'missed' by consumers. These are mostly niche products with low demand. These items do not cater to the masses but to specific, smaller groups of people (Lambersona, 2017).
b. While individual demand for these products or their impact on sales might be low, their cumulative effect on total sales, might be substantial. Goel et al., (2010) observes that 30% of Amazon's sales and 25% of Netflix's sales come from products which are not available in a brick-and-mortar store.

However, there is a strange pattern observed in the consumption pattern. Most of the consumers are satisfied with common products while also exhibiting a propensity for niche products. Thus, while maintaining a small inventory might satisfy most people nearly all the time, there are still some days on which some of the customers will be vexed. This calls for a need to draw a clear line between the two. In our study, we have studied the tail behaviour of data and found overwhelming evidence that in many cases the tail availability tempting the e-commerce retailers to maintain a large inventory will boost head sales (Dai & Taube, 2019).

Long-tail markets usually exhibit a power-law distribution, where the rank-frequency relationship exhibits a heavy-tailed curve. In these markets, while a few products, events, or trends dominate in terms of visibility or consumption, the long "tail" comprises a massive volume of low-frequency insights. For businesses, understanding this tail behaviour is critical, as it often encompasses untapped opportunities, emergent trends, and niche customer needs. Mathematically, long-tail markets align with the Large Number of Rare Events (LNRE) framework. This background sets the stage for analysing rare events, where contributions from the tail, although individually small, exhibit persistent and significant cumulative growth.

3. LNRE framework

The LNRE framework was introduced formally by Estate V. Khmaladze (Khmaladze, 1989) which mainly served to understand the behaviour of low-frequency, high-impact events that can have far-fetched implications. The importance of studying such events lies in the fact that while modelling such events or simulating their results is extremely difficult, the impact that they have on business can be extremely valuable. One of the major challenges in modelling rare events is the lack of historical data as we have seen in the earlier section. The two pillars on which machine learning algorithms stand are the availability of known events and understanding their behavior. In the case of LNRE, we have neither. Thus, it becomes increasingly important to study the LNRE framework

3.1 G-function and Q-function

3.1.1 Definition and derivations

The LNRE framework stands on two pillars – the G-function and the Q-function.

The G-function can be defined as $G_n(z) = \sum_{i=1}^{N} 1\{P_{in} > z\}$, i.e., the number of probabilities greater than a specified threshold.

The Q-function can be defined as $Q_n(z) = \sum_{i=1}^{N} P_{in} 1\{P_{in} \leq z\}$, i.e., the total number of probabilities that occur rarely.

Since a discrete variable might be sometimes less convenient, we divide it into two densities - P_n & f_n.

$\therefore P_n(t) = \sum_{i=1}^{N} P_i^n \, I\{i-1 \leq t < i\}$ and $f_n(t) = \sum_{i=1}^{N} nP_i^N I\{\frac{i-1}{n} \leq t < \frac{i}{n}\} = np_n(nt)$

We have already specified $G_n(z)$ and $Q_n(z)$. Using Lebesgue integrals, we have

$$G_j(z) = \int I\{f(t) > z\}dt$$

$$Q_j(z) = \int I\{f(t) \leq z\}dt$$

$G_f \downarrow G_f \leq \frac{1}{z}, \inf\{z: G_f(z) = 0\} = \operatorname{ess\,sup} f$ and $G_f(0^+)$ is equal to the length of the support of f.

In the Lebesgue, we have seen that $I\{f(t) > z\}$, which measures the size of the set, where $f(t)$ exceeds a certain value. For any given value of z, the integral cannot grow infinitely as it is bounded above. Thus, it can be regarded as a total measure of set where the density $f(t)$ exceeds a certain threshold. This clearly indicates that as z increases, $f(t) > z$ becomes smaller.

It is also observed that the infimum of z for which $G(z)$ becomes zero is the essential supremum of f which is the largest value achieved almost everywhere. The essential supremum reflects most of its support, even if it is not maximum.

Change of variable

$$\int \phi f(t)dt = -\int \phi(z)G_f(dz)$$

Change of variable refers to the mathematical function that permits us to study cumulative contribution changing the variable of integration. It connects the density function to its cumulative representation in terms of a new variable z. The negative sign appears in this case because $G_f(z)$is a non-increasing function of z. As z increases, $G_f(z)$ decreases.

In long-tail markets, $f(t)$ often represent the density of rare events. In a real-life market situation, this refers to low-demand products. Thus, analyzing the cumulative contribution might be infeasible or unintuitive. By transforming the variable, it is possible to express the cumulative contribution in terms of a new threshold z, where z might refer to thresholds like 'order volume' or 'minimum demand.'

Another advantage that we might enjoy from this variable change is the fact that we can compute the tail contributions into the total contributions. If $\phi f(t)$ is the revenue generated by the products with frequency $f(t)$, the right-hand side represents revenue in terms of cumulative contributions $G_f(z)$ given different values of z.

Convergence of Distribution

If I_a is a distribution concentrated at a and $\{F_n\}$ is a sequence of absolutely continuous distributions, then

$$F_n \rightarrow I_a \Rightarrow G_{F_n} \rightarrow 0 \; on \; (0, \infty)$$

$F_n \rightarrow I_a$ means that the sequence of distributions F_n converges weakly to Dirac measure I_a concentrated at a. This in turn indicates that if this convergence occurs, then $G_{F_n} \rightarrow 0$ for all $z > 0$.

In the above convergence the Dirac measure I_a is a distribution where the entire mass is concentrated at a.

$\therefore x < a: F(x) = 0$ which means no mass exists below a

$x = a : F(x) = 1$ which means the mass is concentrated at this point.

If F_n is a sequence of c.d.f. that converge to I_a, F_n can be interpreted as:

$n \to \infty$, the mass of F_n tends to concentrate at a and the limit $F_n(x)$ approaches $I_a(x)$ where:

$$I_a(x) = \begin{cases} 0 \; if \; x < a \\ 1 \; if \; x \geq a \end{cases}$$

Now,

$$F_n(x) = \; \mathrm{P}(X \geq x) = \int_{-\infty}^{\infty} f_n(t)dt$$

$\therefore$ for a given value z

$$G_{F_n}(z) = \int_0^z f_n(t)dt$$

Thus as $F_n \to I_a$, the density function $f_n(t)$ and the G-function $G_{F_n}(z)$ must reflect the concentration of mass.

3.2 Q-function

We define the Q −function as

$$Q_f(z) = \int_z^{\infty} xf(x)dx, \qquad z \geq 0$$

which quantifies the weighted 'tail' of the distribution in long-tail markets. It is calculated over the distribution $f(x)$, where $xf(x)$ represents the contribution of the variable x scaled by it p.d.f. This function is critically important to calculate how much of the total mass lies beyond z. In real-world practical scenarios, the Q-function measures the cumulative impact of niche events, allowing for a greater understanding of the tail's contribution to the overall behaviour.

The relation between $\boldsymbol{G(z), Q(z) and\; f(x)}$

Let us assume $u = x, so\; that\; du = dx$

$and\; dv = G(x)dx$ so that $v = \int G(x)dx = G(x)$

Applying this rule, we get

$$\int_0^z xG(x)dx = \; [xG(x)]_0^z - \int_0^z G(x)dx$$

$$\therefore \; Q_f(z) = \int_0^z xG(x)dx = \; -zG(z) + \int_0^z G(x)dx$$

In cases where $f(x)$ is the density associated

$$Q(z) = \int_0^z f(x)dx$$

$$\therefore f(x) = \frac{d}{dx}\, Q(x)$$

Now we can say

$$G(z) = \int_z^{\infty} \frac{1}{x} Q(x)dx \; or \; f = \; G^{-1}$$

Here the weight function $\frac{1}{x}$ is introduced because it stresses on smaller values of x which are typically important in long-tail distributions. It ensures that $G(z)$ discounts larger values of x which reduces the impact of 'frequent' events in LNRE.

$$\therefore Q(x) = \int_0^x f(t)dt$$

$$G(z) = \int_z^\infty \frac{1}{x}((\int_0^x f(t)dt)dx$$

3.3 Uniform partitioning

3.3.1 Definition and Derivations

In this section, we will see how we can further bolster the analysis by breaking the observations into small partitions. A detailed analysis of the subintervals narrows down the scope of analysis and opens space for a granular analysis.

Let $X_1, X_2, \ldots\ldots\ldots\ldots\ldots, X_n$ be independent variables distributed over [0,1] and f be the density of it.

$$\therefore P_i = F\left(\frac{i}{N}\right) - F(\frac{i-1}{N})$$

$$f_n(t) = N_{P_i}$$

$$\frac{i-1}{N} \leq t < \frac{i}{N}$$

We have seen that as $N \to \infty$, the empirical frequency $v_{i,n}$ converge to theoretical probability density $f(t)$ within their subintervals. In such cases, we break the observation into smaller subintervals or small subsets of population. By estimating the density f, rare or niche patterns – those concentrated in certain subintervals can be detected. In a real-life market, this might be a pain point which correspond to unexpected high density in the long-tail distribution.

So, we can sum up the condition as

$$v_{i,n} \approx \int_{\frac{i-1}{N}}^{\frac{i}{N}} f(t)dt$$

Significant increase or decrease in $v_{i,n,t}$ over time indicate shifts in density revealing how the focus of distribution shifts between various regions.

Now as $N \to \infty$, the interval shrinks approaching a single point in the limit. The relative frequency $v_{i,n,t}$ falling in the i-th interval converges to the integral $f_t(t)$. Any noticeable change in $v_{i,n,t}$ over subinterval i and time t denotes to a rise or fall in interest.

$$\therefore \Delta v_{i,n,t} = v_{i,n,t\_k} - v_{i,n,t} \text{ , where } \Delta v_{i,n,t} > 0$$

If $\Delta v_{i,n,t}$ is seen to be spiking in any such region where it was previously negligible, which could lead to the emergence of new trends. Analyzing local densities and density shifts while comparing them to $\Delta v_{i,n,t}$ across time can indicate the rate of change.

$$\therefore \frac{\Delta v_{i,n,t}}{\Delta t} = \frac{v_{i,n,t\_k} - v_{i,n,t}}{k}$$

3.4 Non-increasing & Inconsistent Density

Non-increasing densities refer to those analyses which aim to identify sub-interval regions where most rare events are concentrated. By analyzing such probabilities, we can pinpoint where high-density concentration is noticed shedding light on rare-event distributions.

3.4.1 definition and derivation

A non-increasing density refers to a p.d.f. $p(t)$ such that

$$p(t_1) \geq p(t_2) \text{ for } t_1 < t_2 \text{ where } t_1, t_2 \in (0, \infty)$$

the most frequent occurrence, i.e., the highest density happens at the lower end ($t \to 0^+$) and their frequency diminishes as $t \to \infty$. In such cases, low-frequency events dominate and high-frequency ones diminish.

$$\therefore P_i = \int_i^{i+1} p(t)dt$$

Further, it is noticed that tracking changes in P_i helps to detect the rate of shift in density contribution.

$$\therefore \Delta P_i = P_i(t_2) - P_i(t_1)$$

Where a positive value of ΔP_i indicates an emerging trend.

In long-tail markets, an inconsistent density refers to the conditions where the p.d.f. fail to capture the distribution's behaviour across different regions. These inconsistencies are common because the densities in the tail regions often exhibit a non-decreasing behaviour, which violates the assumptions of monotonic decay. Standard density estimators rely heavily on smoothness which underestimates the contribution from the tail leading to biased insight.

$$\therefore \left|\left|\hat{P}_n - P_n\right|\right| = || \hat{f}_n - f_n|| = \sum_{i=1}^{Nn} (\hat{v}_{i,n} - P_{i,n})$$

The norm $\left|\left|\hat{P}_n - P_n\right|\right|$ measures how far the estimated probabilities deviate from the true probabilities. A smaller $\left|\left|\hat{P}_n - P_n\right|\right|$ denotes that a model is performing well, which in turn aids in the identification of the patterns in the long tail.

3.5 Relevance of LNRE to Long-tail Markets

In summary, it can be said that LNRE models provide greater flexibility in terms of granular analysis of rare events. In real-life market conditions, an effective support defines the extent of the tail. The essential supremum in an LNRE model can define how far the tail is stretched. This is critically important for understanding the niche behaviour in long-tail markets. The monotonicity of the LNRE further bolsters its capacity to capture the diminishing contribution of rare items as z increases.

The Lebesgue integrals which the LNRE uses add to its analytical capacity. In long-tail markets, the transformations are capable of capturing the shifts in consumer preferences. For example, transforming $f(t)$ under λ^{-1} can help us in studying consumer preference in alternative representations without losing the overall structure. The variable substitution strengthens the spine of the system. Substitution allows the reframing of cumulative measures (for example sales from niche products) in terms of alternate thresholds. Another reason why we are proposing an LNRE model is the reason that the tail behaviour of non-increasing densities captures the sparsity of events as $t \to \infty$.

4. Propositions & Lemmas

4.1 Uniform partitioning and convergence

Proposition 1: *For any dataset which shows long-tail behavior in customer data, any uniform partitioning of frequencies ensures that the cumulative contribution of $Q(z)$ are dominated by low-frequency tokens when the threshold z is sufficiently large.*

Proof: by definition, a long-tail dataset always displays power-law properties or some other heavy-tail distributions, where most tokens are rare. Individually they contribute less, but cumulatively, they convey a lot of information. Thus, to get an idea of the contribution made by each token we isolate the tokens by using uniform partitioning. This helps us to get an idea of the contribution of tokens below a threshold z.

$$\therefore Q(z) = \int_0^z f(x)dx$$

With the growth of z, the value of $Q(z)$ grows proportionally to the contribution of rare tokens, since the frequency of high-contribution tokens decays rapidly as compared to the long-tail. The ratio of the cumulative contribution from the tail $Q_{tail}(z)$ to total contribution $Q(z)$ tend to move towards 1 with $z \to \infty$.

Lemma 1: *In the case of a long-tail dataset which shows a power-law decay, cumulative $Q_{tail}(z)$ for tokens below the threshold z tends to stabilize as $z \to \infty$.*

Proof: let $f(x) = kx^{-\alpha}, \alpha > 1$

$\therefore$ cumulative distribution up to z is

$$Q(z) = \int_0^z kx^{-\alpha}dx = \frac{k}{1-\alpha}[z^{1-\alpha} - 0^{1-\alpha}]$$

For large z, $Q(z) \sim z^{1-\alpha}$ and the contribution below z becomes

$$Q_{tail}(z) = \int_0^z f_{tail}(x)dx$$

$\because$ $f_{tail}(x) \sim x^{-\alpha}$, the growth behaviour is identical as $Q(z)$ which in turn stabilizes $Q_{tail}(z)$.

Proposition 2: *For a dataset that exhibits long-tail behaviour, uniform partition opens the scope for granular analysis. Assume $Q(z)$ is the cumulative contribution of items up to a threshold z and $G(z)$ is the cumulative contribution of items beyond z. Thus, if we can divide the dataset uniformly over n partitions, the convergence of G(z) and Q(z) in each partition gives a detailed scope for a granular analysis of rare events.*

Lemma 1:

Let z_i be the extreme endpoints of uniform partition $\{[z_{i-1}, z_i]\}$, $i = 1,2, \ldots \ldots, n$ such that $z_i = z_{max} \cdot \frac{1}{n}$. Now if $Q(z)$ is differentiable and $Q(z) \to 0$, we can sum up the partition as

$$\sum_{i=1}^{n} Q(z_i) \to Q(z_{max})$$

and the tail contribution becomes

$$\sum_{i=1}^{n} G(z_i) \to Q(z_{max})$$

which can guarantee the separability of granular rare events in long-tail distributions.

Proof: First, we partition the interval $[0, z_{max}]$ into n segments

$$\Delta z = \frac{z_{max}}{n}, z_i = i \cdot \Delta z$$

$\therefore$ cumulative contribution in i −th partition becomes

$$Q_i = Q(z_i) - Q(z_{i-1})$$

while the tail contribution is:

$$G_i = G(z_i) - G(z_{i-1})$$

So, for differentiable $Q(z)$, the Riemann sum converges to:

$$\sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} [Q(z_i - Qz_{i-1}] = Q(z_n) = Q(z_{max}) \; as \; n \to \infty$$

Similarly, $\sum_{i=1}^{n} G_i \to G(z_{max})$. Since uniform partition leads to an isolation of specific contributions in Δz, local variations in long-tail contributions become more visible.

Proof: In long-tail markets $Q(z)$captures the aggregated behaviour while $Q(z)$ focuses on tail contributions, partitioning it uniformly gives:

$$\{[z_{i-1}, z_i]\}, i = 1, \ldots \ldots, n$$

such that contributions of rare frequency tokens are localized. This convergence of $Q(z) \& G(z)$ as $n \to \infty$ makes sure that the contribution from the rarest token is considered with due importance. Such granular convergence helps in detecting changes in tail structure over time while revealing early warnings.

Proposition 3: *In an LNRE framework, the time-dependent relative frequencies $v_{i,n,t}$ over uniformly partitioned sub-intervals can help us identify emerging trends and reallocate the densities $f_t(t)$. By effectively analyzing $v_{i,n,t}$, we can identify areas of high activity.*

Lemma 1: Approximation of uniform partition

Let the segment of interest be partitioned into subintervals $[\frac{i-1}{N}, \frac{i}{N}]$. As $N \to \infty$, the relative frequency approximates the density $f_t(t)$.

$$\therefore\ v_{i,n,t} \approx \int_{\frac{i-1}{N}}^{\frac{i}{N}} f(t)dt$$

Proof: we have already seen that

$$v_{i,n,t} = \frac{number\ of\ obervations\ in\ the\ subinterval\ [\frac{i-1}{N}, \frac{i}{N}]}{total\ obervations}$$

Thus, $f_t(t)$ gives a fair idea of the continuous probability distribution over the domain and when N is sufficiently large $\frac{1}{N}$ attains infinitesimally small value

Now

$$\lim_{N \to \infty} v_{i,n,t} \approx \int_{\frac{i-1}{N}}^{\frac{i}{N}} f(t)dt$$

Thus, the approximation holds for finite N as:

$$v_{i,n,t} \approx \int_{\frac{i-1}{N}}^{\frac{i}{N}} f(t)dt$$

Lemma 2: **Temporal density shifts and emerging trends**

For subinterval $[\frac{i-1}{N}, \frac{i}{N}]$, any significant change over time i.e., $\Delta v_{i,n,t} > 0$ indicate a growing density $f_t(t)$.

Proof: the change in relative frequency can be defined as

$$\Delta v_{i,n,t} \approx \int_{\frac{i-1}{N}}^{\frac{i}{N}} f(t)dt$$

so, if $\Delta v_{i,n,t} > 0$, we can say

$$\int_{\frac{i-1}{N}}^{\frac{i}{N}} f_t(t+k)dt > \int_{\frac{i-1}{N}}^{\frac{i}{N}} f_t(t)dt$$

which suggests an upward shift in density within subinterval.

4.2 Tail Behavior through Cumulative Contribution

Proposition 5: *For any dataset that exhibits long-tail behaviour and modelled by LNRE, the cumulative contribution* $Q(z) = \int_0^z f(x)dx$ *and weighted contribution* $G(z) = \int_0^z xf(x)dx$ *can be transformed into logarithmic and normalized domains, highlighting the supremacy of tail behaviour. These will help in the identification of rare events and analysing them.*

Proof: Let $Q(z) = \int_0^z f(x)dx$ represent the cumulative contribution of events up to z and let $G(z) = \int_0^z xf(x)dx$ represent the weighted contribution of events.

Let $u = \log(x)$ and then $Q(z)$ after log transformation becomes

$$Q(z) = \int_{\log(1)}^{\log(z)} f(e^u)e^u du$$

and after normalized transformation, we get

$$v = \frac{x}{z}$$

$$G(z) = z^2 \int_0^1 vf(vz)du$$

These transformations simplify the analysis, emphasizing tail behaviour.

4.3 Non-increasing Density for latent interests

Proposition 6: *For long-tail datasets, a non-increasing density of rare event contributions leads to a strong presence of rare events in the cumulative contribution $Q(z)$, especially when z is significantly large.*

Proof: Let $f(x)$ be a non-increasing frequency density function

$$\therefore f(x_1) \geq f(x_2), \text{ where } x_1 < x_2$$

and the cumulative distribution of $Q(z)$ up to threshold z becomes

$$Q(z) = \int_0^z f(x)dx$$

Since, $f(x)$ is non-increasing, we can say that when z is sufficiently large, contribution from the long-tail dominates. It can also be said that when z is sufficiently large, the cumulative contribution of rare events from the tail converges to the total cumulative contribution $Q(z)$. When we deal with higher ranks, $f(x)$ approaches zero.

Lemma 1: The dominance of rare events from $Q_{tail}(z)$ in the long-tail datasets is proportional to the non-increasing nature of density $f(x)$.

Proof: for non-increasing property, we can say

$$f(x) = kx^{-\alpha}, \alpha > 1$$

$$Q(z) = \int_0^z kx^{-\alpha}dx = \frac{k}{1-\alpha}[z^{1-\alpha} - 0^{1-\alpha}]$$

For some baseline z_0, the tail contribution can be summed up as

$Q_{tail}(z) = Q(z) - Q(z_0)$ where $Q_{tail}(z)$ dominates as $z \to \infty$

4.4 Inconsistent Density Estimates

Proposition 7: *Inconsistent density estimates in long-tail markets lead to the creation of unstable contributions in the cumulative contribution $Q(z)$ for smaller z, while stabilizing for larger z.*

When density estimates $\hat{f}(x)$ fluctuates for smaller x, the contribution to $Q(z)$ become unstable.

$$Q(z) = \int_0^z \hat{f}(x)dx$$

Thus, the inconsistencies in $\hat{f}(x)$ lead to uneven growth in $Q(z)$lead to uneven growth in $Q(z)$ for smaller z.

Stabilization for large z

When z is sufficiently large $Q(z)$ inconsistencies are mitigated

$$\therefore \lim_{z\to\infty} Q(z) \approx \int_0^\infty f(x)dx$$

Lemma 1: Inconsistent density estimates $\hat{f}(x)$ stabilize die to cumulative averaging in $Q(z)$.

Proof: Stabilization mechanism:

The variance in $\hat{f}(x)$ mitigates as z tends to increase because

$$\text{Var}\left(Q(z)\right) \to 0 \; as \; z \to \infty$$

When z is sufficiently large cumulative average ensures $Q(z)$ reflects the underlying distribution

5. Experiment and Results

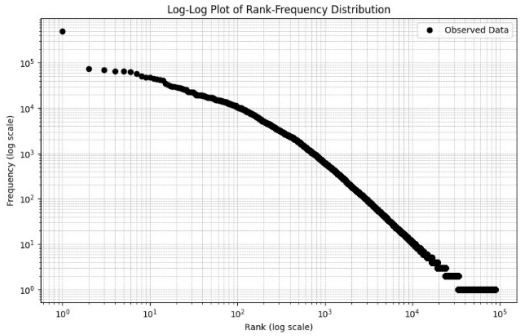
This section attempts to validate the theoretical foundations of the LNRE. In the previous section, we have discussed several propositions and established those propositions and their lemmas mathematically. In this section, we will attempt to establish those deductions in a practical way which will bolster the utility of LNRE-based algorithms in an efficient way.

5.1 Dataset

The data which was used for the research is a review of the dating platform Tinder which was collected using Google Play Scrapper, a Python library for scrapping publicly available app-related data. The dataset provides insight into user details, experience, and reviews. The dataset contains details of user name, images, rating, review content and review data version. It contains detailed and valuable insights for understanding niche customer insights and emerging trends. It was also tested for studying the temporal attributes which will help us in understanding the sentiment and sentiment shifts over time.

Long-tail Behavior of the data

The data was tested for long-tail behavior which is central to this research. We tested the data for its alignment with the power-law distribution.



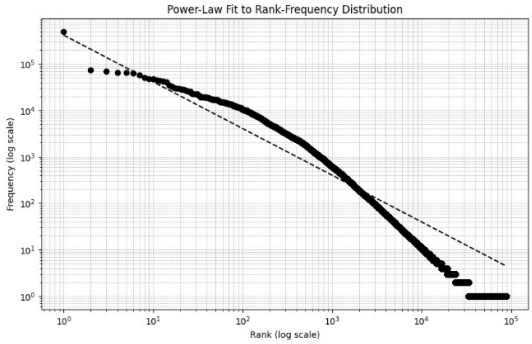


Fig. 2 & 3: test for long-tail behaviour of the data

The log-log plot reveals a linear trend over a broad range which indicates the presence of a heavy long tail. The curve is seen to flatten at the end which suggests that rare tokens are significant contributors in the dataset. The fit demonstrates that the majority of the tokens are situated in the tail region which is a signature of heavy long-tail behaviors. The power law generally aligns well with the observed data, except for the head and the tail which indicates noise or structural complexities. The gradual tapering off of the frequency supports the claim that rare tokens dominate the dataset.

5.2 Uniform Partitioning

In this section, we will discuss the cumulative contribution of the events while uniform partitioning is applied. As discussed earlier, we have seen that uniform partitioning is a crucial method of dividing the observations into equal-sized intervals which will facilitate the study of the contributions to detect patterns, especially in long-tail markets.

Our analysis suggests that most of the observations come from the tail as evident from the plot below.

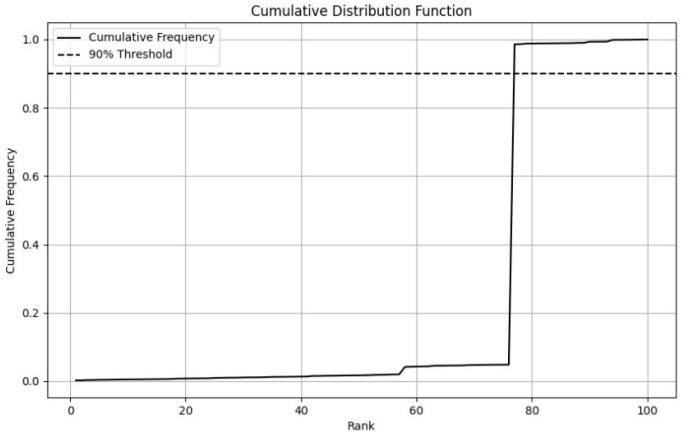


Fig 4: cumulative distribution of the dataset

The solid line represents the cumulative contribution of the events up to z, while the dashed line $Q_{tail}(z)$ denotes the contribution from the low-frequency tokens beyond the threshold z. The stepwise increment of $Q(z)$ and $Q_{tail}(z)$ indicates uniform partitioning, with new partitions added with increments in z. It is demonstrated that for smaller z, contributions remain fairly steady. As z increases, contributions increase more steadily which also demonstrates the contribution from the tail.

The partitioning proposition states that in any dataset that exhibits a long tail behavior the contributions are dominated by low-frequency tokens when z is sufficiently large. This uniform partitioning simplifies the process of isolating the contributions from different regions of the frequency spectrum. This in turn aids in studying the tail's contribution which directly supports the proposition.

We further analyzed the dataset and tried to visualize the cumulative distributions across uniform partitions which will enhance the interpretability of low-frequency versus high-frequency events. We will concentrate on the plot below and divide the discussion into several steps.

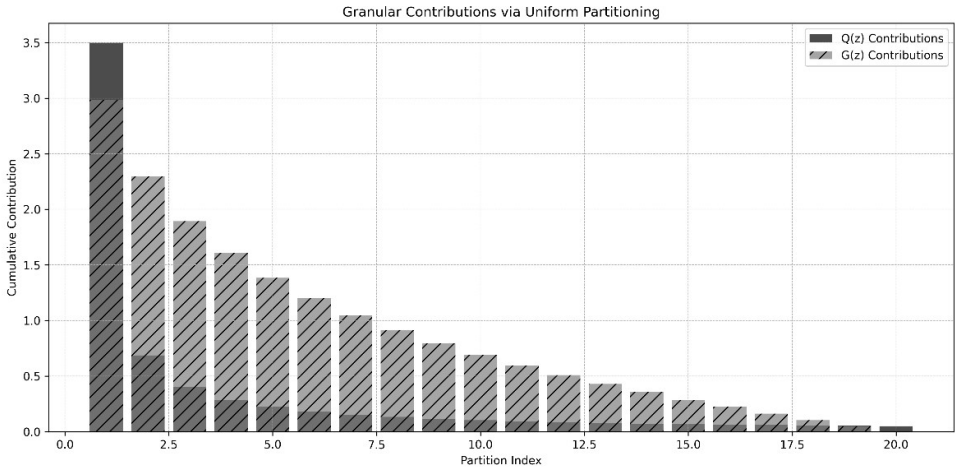


Fig. 5: visualization of the uniform partitioning

It is seen that the left-hand side bins exhibit a higher contribution to both G and Q-functions. The steep drop in contribution strongly supports the proposition. As we move to the higher partition indices the contributions to G and Q functions decrease steadily. The sharp fall in contributions with the partition indices increasing indicates that the uniform partitioning highlights the cumulative impact of these events.

The trends that we noticed from such partitioning indicate a falling popularity of Tinder among the users. It is strongly indicated that the users are complaining that the app has indulged in 'fake' activities and that it is probably charging the users exorbitant 'amount' of money from the user. From a period of 2014 to 2024, the popularity of Tinder seems to be decreasing as indicated by the plot.

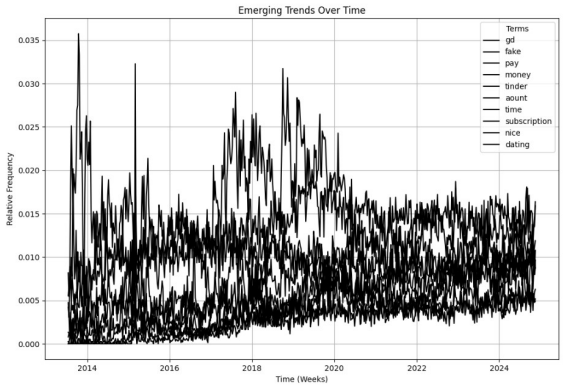


Fig. 6: Emerging trends after partitioning

5.3 Analysis of the Token Frequency Distribution

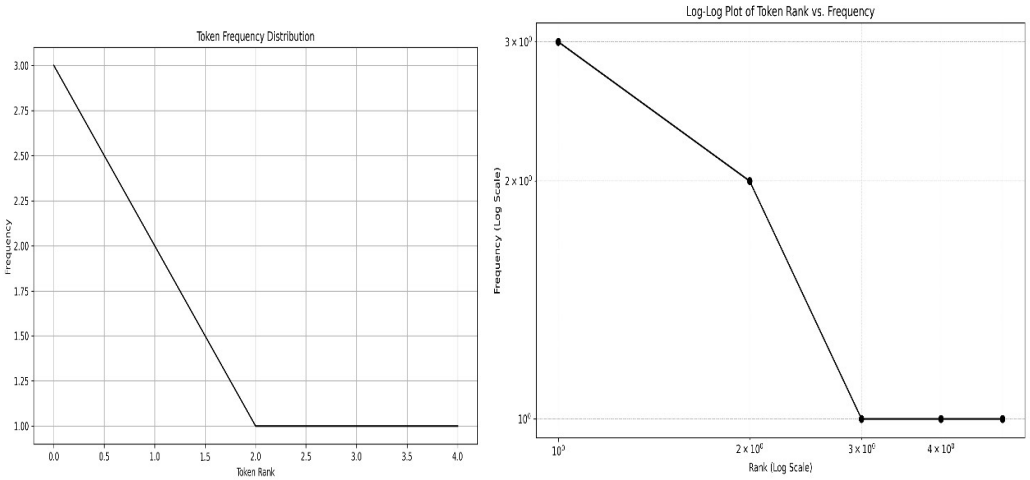


Fig 7 (a) & (b): Visualization of the token frequency

The plot on the left-hand side shows a sharp fall in frequencies with only top-ranked tokens contributing significantly, while the remaining tokens contribute minimally. This steep fall again aligns with the long-tail distribution where a small number of high-frequency tokens dominate the distribution and a large number of low-frequency tokens lead to the creation of the tail. The plot on the right-hand side reveals a linear trend for most of the distribution which aligns clearly with the power law property.

Uniform partitioning ensures that the cumulative contribution from the tail (denoted earlier in our lemmas as $Q(z)$) grows as the z threshold increases. This directly supports the proposition of uniform partitioning and tail dominance. the results strongly show that a strong focus on tail activities is not only mathematically just, but also capable of capturing customer insights. This visualization strongly shows that there is a good return on dissecting the data into granular partitions, each corresponding to a segment of the tail which will quantify the contributions of the rare events in a systematic way.

5.4 Convergence Properties

In this section, we will discuss, how the model aligns with theoretical foundations. We will show that the method we have built follows the theoretical basics and that there are certain advantages of using it which spans practical and analytical improvements in modelling and interpreting rare events.

Here we have studied convergence which refers to the capacity of the method to approach the theoretical $C(z)$ function. *G(z)* captures the observed cumulative contributions with an emphasis on low-frequency tokens, while *C(z)* provides a theoretical baseline to validate the behaviour of the system.

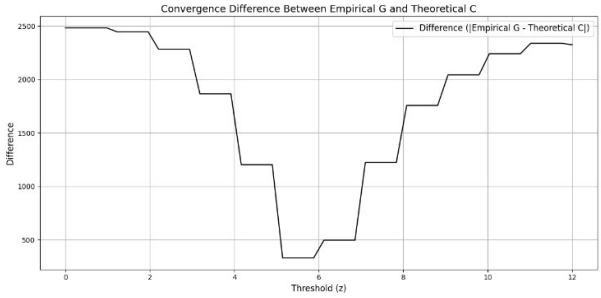


Fig 8: visualization of the convergence of *G-function* to *C-function*

The plot shows that initially a difference between the G(z) and C(z) is noticed. This large divergence at small z can be attributed to the dominance of high-frequency tokens (or events), which skew the empirical $G(z)$. These frequent events contribute disproportionately to G(z)leading to a deviation from the theoretical model. As z increases, the difference between the empirical and theoretical functions steadily decreases. This behaviour suggests that the empirical *G(z)* is gradually aligning with *C(z).* The decline in difference indicates that the influence of high-frequency events diminishes as z moves deeper into the long tail, where rare events dominate.

The importance of the plot lies in the fact that it directly supports the proposition that transformations aid us in analysis by quantifying the cumulative contributions of *G(z).* This in turn helps in the isolation of rare events. The decreasing trend in difference supports the notion that the weighting mechanism in *G(z)* (via the integration over *xC(x)*) effectively reduces the dominance of frequent events, enabling a clearer focus on the tail. This is directly supporting our proposition that transformations simply long-tail.

The plot highlights how the convergence difference changes across threshold z, with a clear minimum at $z \approx 6$. This region of alignment indicates a density shift toward rare events, supporting the lemma's focus on granular convergence as a tool for identifying emerging trends. The larger differences at lower and higher z emphasize regions where frequent or extremely sparse events dominate, helping to pinpoint areas of focus for trend analysis.

5.5 Q-function across different threshold and time

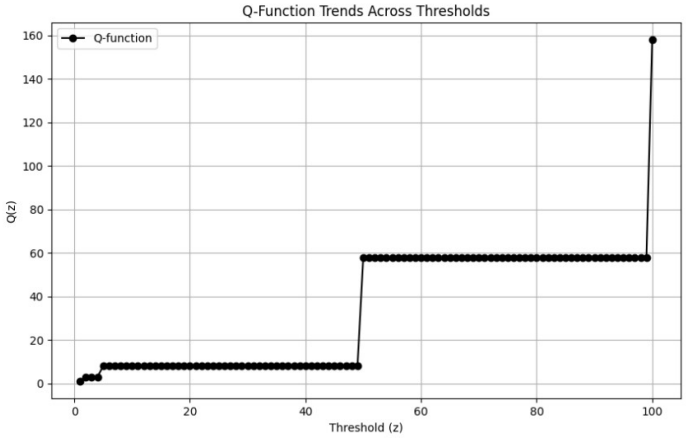


Fig 9: Q-function across the threshold

Since we are dealing with multiple thresholds, it was important to study the cumulative effect of $Q(z)$ on multiple thresholds. Fig.8 exhibits a piecewise growth pattern. The steep increases at certain thresholds *(z=40,100)* indicate that a certain number of events dominate the contributions in these regions. This reflects the **long-tail nature of the dataset**, where high-frequency events at specific thresholds heavily influence the cumulative contributions. The rapid increase at *z= 100* indicates that the cumulative contributions are dominated by a small number of high-frequency tokens (or events) at the tail-end of the distribution.

A comparatively flat growth of *Q(z)* for *z < 40* reflects the **limited cumulative impact of rare events** within this range. Conversely, the sharp increases at higher thresholds suggest shifts in density and the growing influence of specific events. The non-linear growth of *Q(z)* aligns with the long-tail market characteristics, where a few dominant events contribute disproportionately to the overall density. Analysing these contributions over time can expose budding customer preferences or market trends. If tracked over time, this plot can indicate **emerging trends** by identifying thresholds where *Q(z)* begins to increase sharply. For example, a new peak in *Q(z)* at an intermediate threshold could indicate an emergence of niche events. Decomposing *Q(z)* into contributions from smaller partitions can provide deeper insights into which segments dominate at specific thresholds.

To serve this objective, we further decomposed $Q(z)$ to explore the presence of any niche or emergent events. The curve indicates that the cumulative contribution expands as the threshold surges and two significant peaks have been detected:

$z = 11, Q(z) = 1073898$

$z = 21, Q(z) = 1413231$

Q(z) increases rapidly at lower thresholds and then plateaus as the contributions stabilize.

Detected Peaks: Peaks are marked at *z=11* and *z=21*, highlighting thresholds where cumulative contributions experience significant upward shifts. These peaks signify the presence of new niche events that contribute disproportionately to the cumulative total.

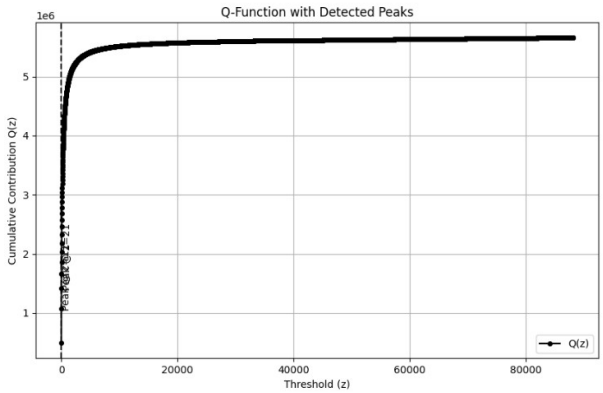


Fig 10: Detection of peaks in *Q*-function

Analysis of the new tokens from both the *z* threshold reveals that new niche ideas are more easily tracked when analyzed using LNRE-based methods. It is seen that tokens like **"even," "profile," "nice," "great," "money," and "work"** signify new patterns or increasing interest areas. *"profile"* might indicate rising concerns about profile management or user experience while *"money"* suggests discussions about monetary dissatisfaction issues, pricing issues, or poor financial aspects of the service.

"money": Evaluate whether users are talking about price sensitivity, exorbitant prices, or a need for a better value proposition. It might call for better research on how tailored plans can be introduced.

"profile": Calls for a need to dedicate resources to improve user profile features or fix associated bugs.

"work": If users highlight functionality or workability issues, focus on enhancing system performance or features.

While most of the text of the reviews might indicate a trend of overall satisfaction as indicated by the overall sentiment, the app developers should take serious note of the growing dissatisfaction as indicated by the heavy tail. It also sheds light on the capacity of the LNRE models to unfurl contradictory patterns in data. While analysis of the 'head' reflects a slightly positive valence, the 'tail' behaviour exposes the hidden reality often overlooked by conventional machine learning techniques.

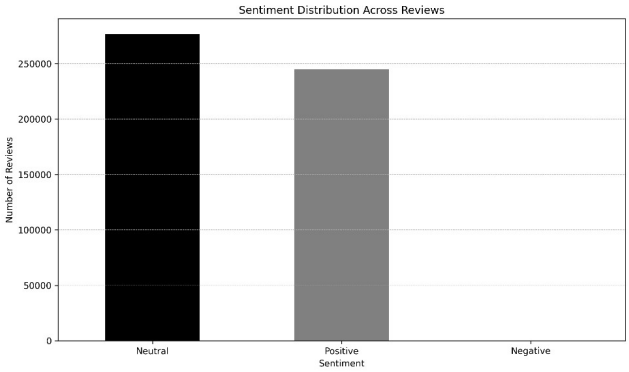


Fig 11: General sentiment of the reviews

5.6 Temporal Analysis of Rare Events

Temporal analysis of the rare events helps us to trace the changes in the contribution of the rare events over time and how they evolve with increasing thresholds.

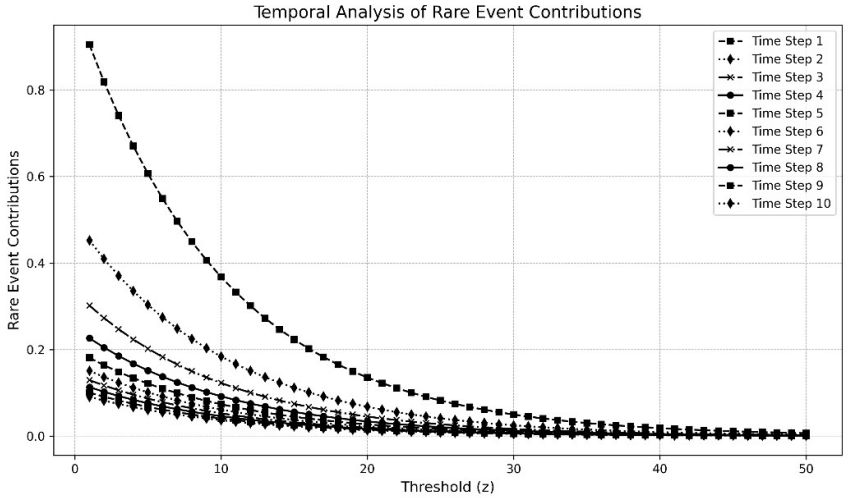


Fig.12: Temporal Analysis of the contribution of rare events

At **lower values of threshold z** (leftmost portion of the plot), rare event contributions dominate, especially during **early time steps** (e.g., Time Step 1 or 2).

As thresholds surge, the contributions show signs of steep decay, and the rate of decay becomes less steep with subsequent time steps.

Temporal Evolution:

- **Time Step 1** displays the highest contributions for rare events, suggesting that initially, rare events dominate the contributions in the dataset.
- As we proceed (Time Steps 2, 3, ..., 10), the contributions decrease significantly, indicating a **shift** in the focus of the data.

- Later time steps show that contributions stabilize at very low values as z increases, indicating to a more uniform temporal distribution.

Behaviour Across Thresholds *z*:

- Contributions diminish **exponentially** with increasing thresholds. This supports the claim of the behaviour expected in **long-tail distributions** where smaller z-values dominate.
- Contributions for each time step converge gradually to zero as z approaches 50, indicating the diminishing impact of rare events beyond a certain threshold.

Time-Specific Peaks:

- The curves suggest that contributions for rare events are most pronounced in the **initial time steps** which possibly shows the emergence of niche events that gradually lose prominence over time.
- This decay aligns with the **Q-function's property** of capturing cumulative contributions in long-tail markets.

The temporal analysis highlights the dynamic evolution of rare event contributions in long-tail markets. Initially, rare events dominate the cumulative distribution, but this dominance fades over time as new data emerges and stabilizes. This analysis provides valuable insights into **emerging trends** and shifts in data focus.

5.7 Analysis of the G-functions and Q-functions

In the next step, we analyzed various trends and metrics of G-functions and Q-functions which sheds light on the behavior of both the G −functions and Q −functions.

We started the discussion with the temporal evolution of the G-functions.

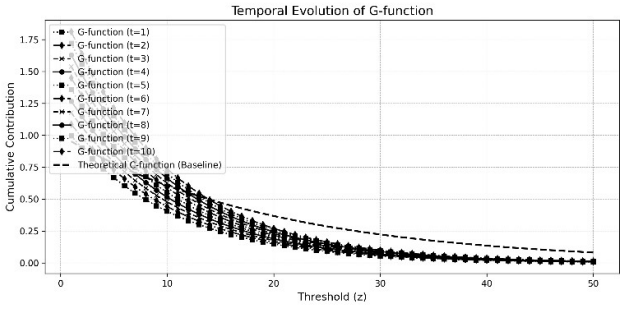


Fig. 13: Temporal evolution of G-functions

It is seen that for each time step t, the **G-function** initiates with higher cumulative contributions at lower thresholds and subsequently decreases as z increases, which again reveals the **dominance of frequent events** in the lower thresholds, while rare events in the long-tail contribute less as z grows. Empirical G-functions diverge significantly from theoretical *C-functions* at minor thresholds but converge at higher

thresholds. This strongly supports our claim that **rare events** influence the empirical *G*-function more prominently than the theoretical model.

At **earlier time steps**, the *G*-function shows higher cumulative contributions, especially at smaller thresholds. As we progress through time, the G-function reduces in magnitude, indicating that the contribution of high-frequency tokens or events diminishes. This pattern suggests **dynamic shifts** in event distributions, where the dominance of previously frequent events declines, and the focus shifts to rarer events.

In a 3-D rendering, we see that across the time steps, the *Q(z)* function alters its trajectory. Some intermediate thresholds show rapid growth in cumulative contributions. This trend highlights **emerging niche events** that gain importance over time, aligning with our primary objective of **monitoring trends in long-tail markets**.

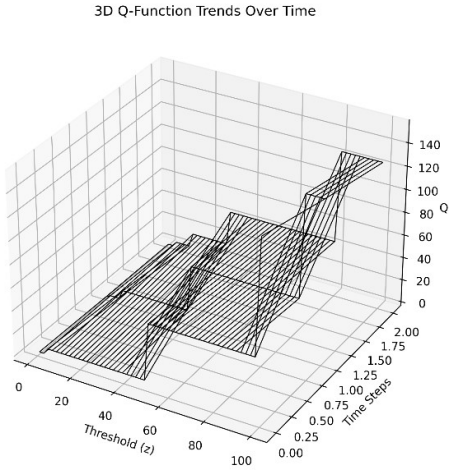


Fig. 14: 3-D plotting of temporal changes in *Q*-function

Thus, it was important for us to analyze the rate at which *Q*(z) changes over time. It was found that at the smaller threshold, the cumulative contributions are volatile, which possibly indicates a short-term fluctuation because more frequent tokens have started to lose their importance over time. At medium thresholds, cumulative contributions stabilize more over time, showing emerging niche events have started to gain importance. At higher thresholds, contributions from rare or niche tokens dominate, and their growth persists consistently across time steps. The thresholds *z* = 30 and *z* =50 are **critical** for identifying significant contributions from niche and rare tokens. Lower thresholds like *z* =10 contribute little insight into the long-tail behaviour and are less relevant for rare event analysis.

Rate of change table			
Q(z = 10) Q(z = 30) Q(z = 50)			
Time steps	***z* =10**	***z* =30**	***z* = 50**
1	2.0	5.0	10.0
2	3.0	10.0	20.0
3	-6.0	15.0	20.0
4	-3.0	10.0	20.0
5	-2.0	15.0	20.0

Table 1: Rate of change table

By analysing the trend of *Q(z)* across thresholds, researchers can dynamically identify which ranges capture meaningful contributions for long-tail analysis. The results show that there is a growing need to focus on tail contributions for **early detection** of emerging trends, anomalies, or underrepresented behaviours. Apart from studying the 'head' behaviour, there is a need to study the 'tail' behaviour.

The *Q*-function sums up the **cumulative contributions** from **all events** (both frequent and rare) up to a certain threshold. The steep rise shows that contributions from **rare events** dominate as the threshold increases. The *G*-function, being a **weighted measure** (e.g., weighted by x values), progresses more homogenously across thresholds. It emphasizes **frequent events** at lower thresholds and **discounts the impact of infrequent events**.

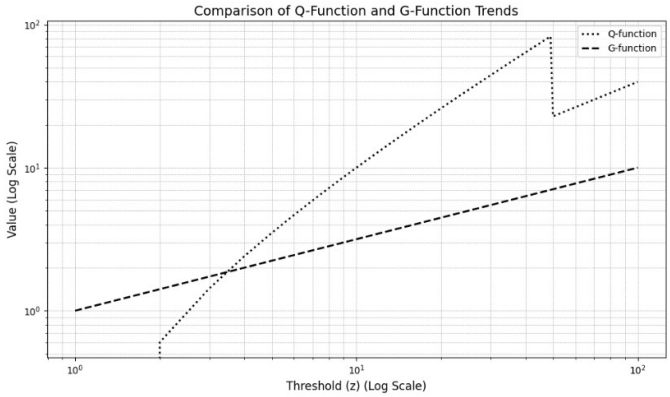


Fig 15: visualization of the performance of *G*-functions and *Q*-functions

6. Discussions

This research proposed to study the dynamics of long-tail markets by introspecting the less discussed ***Q*-function** (cumulative contributions) and ***G*-function** (weighted contributions). It considers both theoretical and empirical results, across datasets and thresholds. It was observed that cumulative contributions *Q(z)* and weighted contributions *G(z)* demonstrated strong alignment with the theoretical intuition, particularly at higher thresholds, where rare events dominate. The findings strongly support the idea that **long-tail contributions**, especially from rare events, dominate and grow considering specific thresholds, extracting emergent trends and niche insights critical to market research.

The results were further reinforced by temporal and comparative analyses, which emphasized how contributions evolve and vary across datasets. Temporal trends of *Q(z)* revealed the persistence of rare events, with sudden peaks signalling the rise of **niche phenomena**. Similarly, misalignments between empirical *G-functions* and the theoretical *C(z)* baseline emphasized dataset variability and the importance of optimizing thresholds to identify trends. Sentiment analysis and clustering enhanced these findings, showing clear patterns of frequent vs. rare token behaviour. It also highlighted the contradictory behaviour of conventional machine-learning techniques and LNRE-based models. While the traditional sentiment analysis showed that Tinder has a positive to neutral sentiment, the LNRE-based analysis showed budding discontent among the users. These insights not only validate the

mathematical propositions but also provide a robust framework for future studies in long-tail market analytics.

References

Akutagawa, K., Ishida, M., & LeBrun, C. (2007). Perelman's invariant, Ricci flow, and the Yamabe invariants of smooth manifolds. *Archiv Der Mathematik*, *88*(1), 71–76. https://doi.org/10.1007/S000013-006-2181-0

Anderson, C. (2006). *The Long Tail: Why the Future of Business is Selling Less of More*.

Chen, J., & Guo, Z. (2014). Online Advertising, Retailer Platform Openness, and Long Tail Sellers. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.2694073

Dai, S., & Taube, M. (2019). The long tail thesis. *Chinese Management Studies*, *14*(2), 433–454. https://doi.org/10.1108/CMS-03-2019-0109

Goel, S., Broder, A., Gabrilovich, E., & Pang, B. (2010). Anatomy of the long tail: Ordinary people with extraordinary tastes. *WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, 201–210. https://doi.org/10.1145/1718487.1718513

Hinz, O., Eckert, J., & Skiera, B. (2011). Drivers of the Long Tail Phenomenon: An Empirical Analysis. *Journal of Management Information Systems*, *27*(4), 43–70. https://doi.org/10.2753/MIS0742-1222270402

Khan, S. (2020). An overview of long tail marketing strategy and its challenges: A qualitative study. *South Asian Journal of Marketing and Management Research*, *10*(3), 31. https://doi.org/10.5958/2249-877X.2020.00012.0

Khmaladze, E. (1989). *The Statistical Analysis of Large Number of Rare Events*.

Lambersona, P. J. (2017). *Winner-take-all or long tail? A behavioral model of markets with increasing returns*.

Morganti, P. R. (2023). *Quality Choices Along the Long Tail*. https://doi.org/10.2139/SSRN.4331087

Peltier, S., & Moreau, F. (2012). Internet and the 'Long Tail versus superstar effect' debate: evidence from the French book market. *Applied Economics Letters*, *19*(8), 711–715. https://doi.org/10.1080/13504851.2011.597714

Sastry, N. (2012). How To Tell Head From Tail in User-Generated Content Corpora. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1), 567–570. https://doi.org/10.1609/ICWSM.V6I1.14333

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., & Bolhuis, J. J. (2017). The growth of language: Universal Grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews*, *81*, 103–119. https://doi.org/10.1016/j.neubiorev.2016.12.023

Zhou, W. (2011). *International Conference on Information Systems (ICIS) 1-1-2009 Product Variety , Online Word of Mouth and Long Tail : An Empirical Study on the Internet Software Market*.

Appendix

Log-likelihood of Katz Backoff

Case 1: seen events where $(count(w_{i-2}, w_{i-1}, w) > 0)$

$$\log P(w_i|w_{i-2}, w_{i-1}) = \log(D\ \frac{count(w_{i-2}, w_{i-1}, w)}{count(w_{i-2}, w_{i-1})})$$

Case 2: seen events where $(count(w_{i-2}, w_{i-1}, w_i) = 0)$

$$\log P(w_i|w_{i-2}, w_{i-1}) = \log(\alpha(w_{i-2}, w_{i-1})P(w_i|w_{i-1}))$$

Full log-likelihood:

$$\log \mathcal{L} = \sum_{i=3}^{n} \begin{Bmatrix} \log\left(D\ \frac{count(w_{i-2}, w_{i-1}, w)}{count(w_{i-2}, w_{i-1})}\right), if\ (count(w_{i-2}, w_{i-1}, w) > 0) \\ \log(\alpha(w_{i-2}, w_{i-1})P(w_i|w_{i-1})), if\ (count(w_{i-2}, w_{i-1}, w_i) = 0) \end{Bmatrix}$$

Deriving the Hessian

First derivative (Gradient)

Case 1: seen events

$$\frac{\partial}{\partial D}\log \mathcal{L}_{seen} = \sum_{i=3}^{n} \frac{1}{D}$$

Unseen events

$$\frac{\partial \log \mathcal{L}_{unseen}}{\partial \alpha(w_{i-2}, w_{i-1})} = \sum_{i=3}^{n} \frac{1}{\alpha(w_{i-2}, w_{i-1})}$$

Second derivative (Hessian)

Case 1: seen events

$$\frac{\partial^2 log\ \mathcal{L}_{seen}}{\partial D^2} = -\sum_{i=3}^{n} \frac{1}{D^2}$$

Case 2: unseen events

$$\frac{\partial^2 \log \mathcal{L}_{unseen}}{\partial \alpha^2} = -\sum_{i=3}^{n} \frac{1}{\alpha(w_{i-2}, w_{i-1})^2}$$

Change of variable

$$\int \phi f(t)dt = -\int \phi(z)G_f(dz)$$

Let $G_f(z)$ represent the c.d.f. of $f(t)$

$$G_f(z) = \int f(t)dt$$

Let $f(t) = z, G_f dz = -\frac{d}{dz}G_f(z)dz$

$$\therefore \int \phi f(t)dt = \int \phi(z)(-\frac{d}{dz}G_f(z))dz$$

$$= -\int \phi(z)G_f(dz)$$

Conflict of Interest Statement: The author declares that there is no conflict of interest regarding the publication of the paper.