Rank Distribution and RIF Index

Brian LLinas

2024-06-20

# Title, Abstract and Keywords

### Title

1. **Modeling Rank Distribution and the Relative Importance Factor Index in Power-Law Models: Application to Social Resilience Using Scopus Databases**
2. OTRO POSIBLE TITULO: Modeling Rank Distribution and Introducing the Relative Importance Factor Index in Heavy-Tailed Distributions: Application to Social Resilience Using Scopus Databases

### Abstract

Heavy-tailed distributions, such as the power-law distribution, are prevalent in both natural and human-made phenomena, characterized by their slow-decaying tails. Previous studies have primarily focused on modeling these distributions based on frequencies, often neglecting the estimation of rank distributions. This paper addresses this gap by introducing a novel approach to model the rank distribution of frequent factors and by defining the Relative Importance Factor (RIF) index. The first objective is to develop and validate a formula that accurately models the rank distribution of frequent factors. The second objective is to introduce the RIF index, comparing the probability of a factor occupying the first position with its proba-bility of being in any other position within the range. We apply our methodology to databases downloaded from Scopus, focusing on social resilience and analyzing the number of citations of articles. Using maxi-mum likelihood estimation (MLE) to determine power-law parameters and the Kolmogorov-Smirnov (KS) statistic to estimate the optimal threshold, we ensure robust modeling. Additionally, we employ bootstrap-ping techniques to assess the uncertainty of our estimates. By modeling the rank distribution and introduc-ing the RIF index, this study enhances the understanding of heavy-tailed distributions and provides a valua-ble tool for analyzing complex systems.

### Keywords

Power-law distribution; Rank distribution; Relative Importance Factor (RIF); Heavy-tailed distributions; Social resilience; Citation analysis

# 1. Introduction

Heavy-tailed distributions are common in both natural and human-made phenomena. These probability distributions are characterized by a slower decay in their tails compared to the normal or exponential distribution. Several distributions that fall into this category include the log-normal and stretched exponential distributions, among others. In addition to these, the most well-known heavy-tailed distribution is the power-law distribution (Jiang et al., 2023). Besides its ability to describe phenomena with heavy-tailed distributions, the power-law model has captured the interest of researchers in various disciplines because it can also be applied to analyze the behavior of many complex systems. This makes it ideal for describing phenomena in diverse areas (Corral and Gonzalez, 2019; Blasius, 2020; Chen, 2021; Jiang, and de Rijke, 2021; Sethna, 2022). However, limitations in data collection often result in empirical datasets that cover a narrow observation range, complicating the clear identification of power-law behavior (Navas-Portella et al., 2019).

Power-law distributions have several characteristics. One of them is their scale-invariance property, which means that the shape of the distribution remains constant regardless of the scale at which it is observed (Corral and Gonzalez, 2019; Banshal, 2020). Another is that when the entire range is plotted on a linear axis, the curve takes the shape of a perfect “L”. Moreover, when represented on a logarithmic scale, the curve always appears straight (Chen et al., 2020; Banshal, 2020).

In computer science, the power law has been applied to the study of networks, where it has been observed that a few nodes have an extremely high number of connections while most have few (Artico et al., 2020; Devlin, 2021; Somin et al., 2022). This pattern has also been observed in social networks, where a few users have a large number of followers, and most have few (Xu et al., 2019; Arthur and Williams, 2019; Rajput et al., 2020). Additionally, in risk management, the power law is used to model the occurrence of extreme events, such as natural disasters (Pisarenko and Rodkin, 2019; Yum, 2023; Sohn et al., 2023) or financial crises (Dufrénot and Paret, 2019; Taleghani, 2020; Ghosh et al., 2021; Ben Yaala and Henchiri, 2023).

In economics, it has been used to describe the distribution of wealth, where a small fraction of the population holds most of the financial resources while another suffers from extreme poverty (Masseran, 2019; Cardoso et al., 2020; Safari et al., 2020; Puttanapong et al., 2022; Kumer, 2024). Similarly, power laws have been used to study the distribution of crime, where a few criminals commit most of the crimes while the majority commit only a few (He et al., 2023; Ng et al., 2023). In ecology, it has been employed to model the distribution of forest types (Atkins et al., 2022), freshwater fishes (Baumgartner and Peláez, 2024), hot-spring microbiomes (Li and Ma, 2019), among others. In physics, power laws are fundamental for studying the behaviors of the decay rate as a link between dissociation energy and temperature (Fischer and Schweikhard, 2020), fully developed turbulent flows in a smooth pipe (Afzal et al., 2023), and nonlinear phonon hydrodynamics (Sciacca and Jou, 2024).

In the health sector, power laws have been used to model the distribution of epidemics, where a few outbreaks can infect a large number of people, while most outbreaks affect a much smaller number (Blasius, 2020; Neipel et al., 2020; Jha, 2020). This model has also been applied to analyze the distribution of health resources, such as the availability of hospital beds or the allocation of medications, where a few hospitals or health centers receive the majority of resources (Srivastav et al., 2021; Dong et al., 2021).

In the field of academic publications, the academic influence of articles, journals, authors, etc., has been studied for several decades. Currently, one of the most widely used metrics and long recognized as an important indicator for evaluating the impact of a journal or author is the number of citations (Zhao et al., 2019). Numerous studies have revealed that the citation distribution of scientific articles follows a power law (Thelwall and Nevill, 2018; Arroyo-Machado et al., 2020; Benatti et al., 2023). In particular, the number of articles with a specific number of citations is proportional to raised to a negative scale exponent (Popescu, 2003; Banshal, 2020).

It is emphasized that previous studies model the power-law distribution based solely on frequencies but do not estimate the distribution of the rank. Therefore, firstly, in this work, we present and demonstrate a formula that allows us to model the distribution of the rank of the most frequent factors. Secondly, we introduce a parameter, which we will call Relative Importance Factor index (RIF index). The RIF index compares the probability of a factor occupying a given position with the probability of it being in any higher position within the range. This innovation provides a new perspective for evaluating the relative prominence and importance of factors, an aspect that has not been considered in previous studies.

**[TWO OPTIONS ARE PRESENTED AS NAMES. For continuation, the first one (the preferred one) will always be mentioned**

**1. Relative Importance Factor index (RIF index).**

**2. Relative Prominence Factor index (RPF index). ]**

The proposed theory will be applied using databases downloaded from Scopus in the area of social resilience. In particular, we will focus on the number of citations of articles published in this area. This approach allows us to identify citation distribution patterns and evaluate how academic attention is concentrated on certain articles and specific topics but based on the RIF index. The choice of social resilience as a study area responds to the growing importance of this topic in the context of social sciences and its relevance for public policy formulation and the implementation of resilient practices in communities and organizations. According to Dagdeviren et al. (2020), social resilience depends on power relations, rules/institutions, and resource distribution, which are interconnected. Without favorable conditions in these three elements, individuals may be overwhelmed by crises or survive through harmful mechanisms.

The article is organized as follows. Section 1 contains the introduction. In Section 2, we explain the power-law model and the different methods for estimating the scale parameter and the optimal threshold of the model. In Section 3, we introduce the distribution of the rank and the so-called Relative Importance Factor index. In Section 4, we present an application. Finally, in Section 5, we present the conclusions.

# 2. Power-law for data frequencies

### 2.1 The power-law model

Power-law distributions can appear in two forms: continuous distributions, which govern continuous real numbers, and discrete distributions, where the quantity of interest can only take a discrete set of values, typically positive integers (Clauset et al., 2009). Variables of the problem are defined as follows: (1) denotes the variable of interest; (2) is composed of several factors , , , , ; and (3) is the frequency of occurrence of one factor within . Considering that our data capture the frequency with which each occurs, our study deals with a discrete problem. Based on that, the discrete power-law model to estimate the probability that the factor within appears with frequency can be defined as

where represents the random variable corresponding to the frequency of occurrences, is the exponent or scaling parameter for , indicating the steepness of the distribution, and is a normalization constant to ensure the probabilities sum up to 1 for . Both and depend on the distribution and can be found in Clauset et al. (2009). This distribution diverges to zero when , so there must be a lower bound for the behavior of the power-law. The normalizing constant is calculated as follows:

where is the generalized or Hurwitz zeta function (Zaghloul, 2019):

The cumulative distribution function of a power law is given by

### 2.2 Estimations of the parameters

To ensure accurate estimation of the power-law exponent and determine the range over which the power-law behavior is applicable, it is necessary to decide on the lower bound . This estimation helps identify the specific part of the media coverage distribution where the power-law model is valid. Moreover, obtaining an estimate is crucial for deriving an unbiased estimate of the power-law exponent . According to Clauset et al. (2009), if we assume that our data are sampled from a power-law distribution for values of greater than or equal to , the maximum likelihood estimator (MLE) for in the discrete case is defined as

where is the number of occurrences by the -th factor () and are the observed values of such that , where is the estimated value of . We estimate using the maximum likelihood (ML) method. The corresponding log-likelihood function for this estimation process is derived from the data of . Note in the above formula that, to calculate , we must first estimate . To estimate the lower bound on the power-law , we used a metric known as the Kolmogorov-Smirnov (KS) statistic (Ramos et al., 2024), which is defined as the maximum difference between , the cumulative distribution function (CDF) of the observations, and , the CDF of the power law that optimally fits the data . The KS statistic is defined as

The fitting process is sometimes performed by linear regression using logarithmically transformed variables. This approach is used because applying the logarithm to the power law function results in

Thus, a power law appears as a straight line with slope in a logarithmic plot. It is important to note that changes in the scale parameter can affect the slope of the curve in the log-log plot, resulting in changes in the shape and behavior of the distribution represented by the power law. The bootstrapping procedure is used to analyze the uncertainty associated with exponent estimation. It consists of randomly selecting data samples with replacement and then applying the MLE procedure with a KS cutoff on that sample. This process is repeated several times to evaluate the uncertainty. In this study, 1000 iterations were performed on all data sets. In addition, we perform a particular hypothesis test and provide the corresponding p-value to test : one power-law distribution fits adequately versus : another distribution might fit better. We will use the R 4.3.1 package poweRlaw (Gillespie, 2014) to conduct all the aforementioned analyses on datasets downloaded from Scopus, specifically in the field of social resilience.

# 3 Power-law for rank and RIF index

## 3.1 Relation between the Frequencies and the Rank

For the variable , which is composed of several factors , , we will suppose that the frequency of occurrences of each factor is a descending ordered series. Let denote the frequency of occurrence of the factor at rank , where is the rank of the factor within the variable . That is,

This means that the frequency of occurrence decreases as the rank increases. According to Popescu (2003), Zipf’s law of rank-frequency states that, in a generally ordered set of data, the frequency of occurrence of an element is inversely proportional to its rank :

where is the rank of the element (1 for the most frequent, 2 for the second most frequent, etc.); is the frequency of the element at rank for ; is the power law exponent, and is a constant of proportionality. This implies that

To estimate the parameters and , we only need to fit the regression model with the response variable and the explanatory variable . But the method of estimation for the parameters and can vary depending on the size of the data, i.e., the number of factors within . For datasets with a large number of factors, methods such as maximum likelihood estimation (MLE) are preferred due to their robustness and efficiency. MLE can provide precise estimates even with large sample sizes.nFor smaller datasets, ordinary least squares (OLS) regression can be used to estimate the parameters. OLS is simpler and computationally less intensive, making it suitable for smaller samples where computational resources may be limited. Additionally, bootstrapping techniques can be employed to assess the variability and confidence intervals of the estimated parameters, especially in cases where the sample size is small or the data exhibits significant variability. This approach involves repeatedly resampling the data with replacement and recalculating the estimates to build a distribution of the parameter estimates. In summary, depending on the number of factors in , different estimation techniques can be employed to accurately determine the parameters and , ensuring the reliability of the rank-frequency relationship described by Zipf’s law. Based on the above, once these estimates have been found, we can write

With these estimates, we can then estimate the rank distribution (see Section 3.2).

## 3.2 Estimation of the Distribution of the Rank

In this section, we will demonstrate one of the primary objectives mentioned in the introduction: estimating the distribution of the rank of the most frequent factors.

**Theorem 3.2.1.** Let be a variable composed of several factors , , with descending ordered frequencies of occurrence for each factor at rank , where is the rank of the factor within . That is, for all . Suppose that follows a power-law distribution with estimated scale parameter and threshold , and normalization constant , ensuring that the probabilities sum to 1 for . Then, for all ranks such that , the estimated probability that a factor occupies rank is given by:

1. .
2. , for all .

There represents the random variable corresponding to the frequency of occurrences, is a proportionality constant, is the normalising constant defined by:

and (real number) and are estimates such that , and is the power law exponent for the rank .

**Proof Theorem 3.2.1.** Given that for all , and assuming follows a power-law distribution, we have , where is a constant of proportionality and is the power law exponent. To estimate the distribution of the rank , we start with the relationship between the probabilities and . Since these probabilities are inversely proportional, we have:

where is a proportionality constant. This result demonstrates the part (a). Substituting into the probability , and using the normalization constant to ensure the probabilities sum to 1, we get:

where is the power law exponent for the occurrence frequency . Defining the estimation , we get , where is the normalizing constant defined as indicated by the statement of the theorem. Thus, we have shown that the probability of a factor being at rank follows the power-law distribution with exponent . This result demonstrates the part (b).

We now present a theorem that characterizes the properties of the rank distribution for a variable composed of several factors . This theorem provides conditions under which the mean, variance, moments, and cumulative distribution function (CDF) of the rank distribution exist and are well-defined.

**Theorem 3.2.2.** Let be a variable composed of several factors , where , each with descending ordered frequencies of occurrence at rank (the rank of the factor within ). That is, for all . Suppose that follows a power-law distribution with an estimated scale parameter , threshold , and normalization constant , ensuring that the probabilities sum to 1 for . Then, for all ranks such that , the following properties hold for the rank distribution:

1. For all , the mean of the rank distribution exists. It is given by , where is the Riemann zeta function, defined by
2. For all , the variance of the rank distribution exists. It is given by
3. For all , the -th moment of the rank distribution exists. It is given by .
4. For all , the cumulative distribution function (CDF) of the rank is given by

* where the Hurwitz zeta function, denoted by is defined for and by the infinite series

**Note.** The Hurwitz zeta function is a generalization of the Riemann zeta function, as when , it simplifies to .

**Proof Theorem 3.2.2.**

1. The mean of the rank distribution is given by

For the above series to converge, we require that . Thus, . Therefore, the mean is finite if and only if . Using the definition of the Riemann zeta function, we find the formula given in (a).

1. The variance of the rank distribution is given by

The above series converges if and only if . Therefore, the series converges if and only if . Using the definition of the Riemann zeta function, we have that . Using the result found in (a) and the property , we derive the formula given in (b).

1. The -th moment of the rank distribution is given by

For the above series to converge, we require that . Thus, . Therefore, the -th moment is finite if and only if . Using the definition of the Riemann zeta function, we find the formula given in (c).

1. The cumulative distribution function (CDF) of the rank is given by

The first sum converges for all and is calculated with Riemann zeta function . The second sum converges for and (since and in the rank notation), and is calculated using the Hurwitz zeta function . Therefore, converges for all . In summary, replacing both zeta functions, we obtain the formula in (d).

## 3.3. The Relative Importance Factor (RIF) Index

As mentioned in the introduction, previous studies have focused on modeling the power-law distribution based solely on frequencies, without estimating the distribution of the rank. In the previous section, we addressed this gap by presenting and demonstrating a formula to model the rank distribution of the most frequent factors. Now, we introduce a parameter known as the Relative Importance Factor (RIF) index, which will serve to compare the relative importance of factors across different ranks.

**Definition 3.3.1.** For all ranks such that , the estimated Relative Importance Factor (RIF) Index of a factor in rank with respect to another factor in rank , denoted as , is defined as:

where is the estimated probability of the factor occupying rank , and is the estimated probability of the factor occupying rank . When , we simply write and refer to it as the RIF index with respect to factor .

This estimate will help us explain how the probability of observing an element at rank compares to the probability of observing the element at rank . We know that . Then, in particular, . When , then , which is trivially 1, since we are comparing the probability of the first rank with itself. Suppose that . Then, the probability of observing the first element in the rank is times higher than the probability of observing the element in rank . In general,

For different values of , Figure 1(a) and Figure 1(b) show how the rank relates to both the RIF index and the probability . Given that is positive, the graph illustrates how the RIF index and the probability change as both and are increased.

**OJO. FALTAN LAS FIGURAS.**  Figure 1. Distribution of: (a) RIF Index and (b) Probability vs Rank, for different values of ; RIF Index vs (c) ; (d) .

Each line, color-coded according to the legend, represents the distribution for a particular value of . The dark line shows a steep increase (decrease) in RIF index (probability) with higher ranks, indicating rapid growth (decline), while other lines exhibit more gradual or stable trends in the odds. This suggests significant variability in how RIF index (probability) changes with rank across different values of , highlighting the unique impact of rank on RIF index (probability) for each type. As increases and becomes more positive, the RIF index increase (the probability decreases) exponentially. The higher the value of , the steeper the increase (decrease) in the RIF index (probability) with respect to . This indicates that for larger values of , the disparity between the probability of the highest-ranked event and lower-ranked events becomes much more pronounced. As increases, the RIF index diverge significantly and the probability converges to 0, especially for higher values, highlighting an exponential growth pattern. Figure 1(c) explores the relationship between the probability and the RIF index for different values of . As increases, the RIF index decrease sharply initially and then level off. The gradient color represents different values of , indicating that higher values of correspond to higher RIF index for a given . Figure 1(d) examines the relationship between the probability and the RIF index for different values of . As increases, the RIF index increase linearly. The gradient color represents different values of , showing that higher values of lead to higher RIF index for a given .

## 3.5 Domain and Ranges of the RIF index

The domain of is the set of all positive integer such that . The range of is . Since Since for any , the value of will always be 1 or greater. The range of values that the RIF index can take allows for several interpretations regarding the importance of factors:

* **Near 1:** When , it indicates that the probability of the factor being at rank is nearly the same as being at rank . This suggests factors with relatively stable importance.
* **High Values:** When , it indicates that the factor is much more likely to be at rank than at rank . This suggests a sharp decline in importance as the rank increases.

To provide clear and impactful terminology for discussing the importance levels of different factors within our analysis, we propose the following names for the ranges of , shown in the following definition.

**Definition 3.5.1.** Let be the estimated RIF index of a factor in rank with respect to another factor in rank . The receives the name (with respect to ) indicated below when the RIF meets the following conditions:

**ARE SUGGESTED RANKS AND NAMES. ONLY ONE MUST BE CHOSEN. AS A PREFERENCE, I WILL STICK WITH THE FIRST NAMES.**

1. (Factor *Balanced*, *in equilibrium*, or *Stable*): Indicates that the factor’s importance is nearly the same across different ranks. This suggests a stable distribution of importance and reflects minimal change in importance as the rank changes.
2. (Factor *Moderate*, *Intermediate*, or *Notable*): The factor has a slightly higher importance at rank compared to a rank . The difference in importance is noticeable but not extreme, representing a moderate decline in importance with increasing rank.
3. (Factor *Significant*, *Elevated*, or *Enhanced*): The factor is significantly more important at rank than at higher ranks . There is a clear elevation in importance at rank , indicating an enhanced level of importance at the rank .
4. (Factor *Critical*, *Major*, or *High*): The factor’s importance at rank is majorly higher than at other ranks . This reflects a critical difference in importance between the rank and others , with a steep decline as rank increases.
5. (Factor *Dominant*, *Supreme*, or *Peak*): The factor is overwhelmingly more important at rank than at any other rank . This represents a supreme level of importance at the top rank, with the importance peaking at rank and dropping sharply at higher ranks .

These categories can be used to effectively communicate the significance of our findings in terms of the relative importance of factor compared with various factors , for .

## 3.7 Values of for each range of

The value of will affect the values of , for all such that , as shown in the following theorem.

**Theorem 3.7.1.** Let be the estimated RIF index of the first factor in rank with respect to another factor in rank , where and . Let be defined as in previous sections. Then, for a fixed ratio and given estimations of , , , such that , the following holds:

where . Observe that .

**Proof Theorem 3.7.1.** This proof involves showing that the given bounds on translate directly to bounds on through the functional relationship defined between them. Since , taking the natural logarithm on both sides yields . Therefore, it is straightforward to see that the inequalities on correspond to inequalities on .

## 3.8. Theta values for PHi ranges

### 3.8.1 R functions for examples

**Calculate Theta for given Phi and one single fixed r**

# Define the function to calculate theta values  
calculate\_theta <- function(r, phi\_values) {  
 results <- data.frame()  
 # Calculate theta for each value in phi\_values  
 theta\_values <- log(phi\_values) / log(r)  
 temp\_df <- data.frame(r = r, Phi = phi\_values, Theta = theta\_values)  
 results <- rbind(results, temp\_df)  
 return(results)  
}

**Calculate Theta for given Phi and multiple fixed r**

calculate\_theta\_multiple\_r <- function(r\_values, phi\_values) {  
 results <- data.frame()  
   
 for (r in r\_values) {  
 theta\_values <- log(phi\_values) / log(r)  
 temp\_df <- data.frame(r = r, Phi = phi\_values, Theta = theta\_values)  
 results <- rbind(results, temp\_df)  
 }  
   
 return(results)  
}

**Calculate Min-Max Theta for given Phi and multiple fixed r**

calculate\_theta\_minmax\_multiple\_r <- function(r\_values, phi\_values) {  
 results <- data.frame()  
   
 for (r in r\_values) {  
 theta\_values <- log(phi\_values) / log(r)  
 min\_theta <- min(theta\_values)  
 max\_theta <- max(theta\_values)  
 temp\_df <- data.frame(r = r, Min\_Theta = min\_theta, Max\_Theta = max\_theta)  
 results <- rbind(results, temp\_df)  
 }  
   
 return(results)  
}

### 3.8.2 Numerical Examples: given Phi and one single fixed r

# Example usage  
r\_values <- 5 # Fixed value of r  
#phi\_values <- c(1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 10, 15.5, 20) # List of Phi values (including decimals)  
phi\_values <- seq(1, 2, by = 0.2)  
  
# Calculate theta values for the given Phi values and fixed r  
theta\_values <- calculate\_theta(r\_values, phi\_values)  
theta\_values

## r Phi Theta  
## 1 5 1.0 0.0000000  
## 2 5 1.2 0.1132828  
## 3 5 1.4 0.2090620  
## 4 5 1.6 0.2920297  
## 5 5 1.8 0.3652124  
## 6 5 2.0 0.4306766

# Display the results  
#data.frame(Phi = phi\_values, Theta = theta\_values)

### 3.8.2 Numerical Examples: given Phi and multiple fixed r

# Example usage  
r\_values <- 2:10 # Fixed values of r  
phi\_values <- seq(1.01, 2, by = 0.01) # List of Phi values within the range (1, 2]  
  
# Calculate theta values for the given Phi values and multiple r values  
results <- calculate\_theta\_multiple\_r(r\_values, phi\_values)  
  
# Display the results  
head(results)

## r Phi Theta  
## 1 2 1.01 0.01435529  
## 2 2 1.02 0.02856915  
## 3 2 1.03 0.04264434  
## 4 2 1.04 0.05658353  
## 5 2 1.05 0.07038933  
## 6 2 1.06 0.08406426

### 3.8.2 Min-Max Theta:

library(knitr)

## Warning: package 'knitr' was built under R version 4.3.3

# Example usage  
r\_values <- 2:10 # Fixed values of r  
phi\_values <- seq(2.01, 5, by = 0.01) # List of Phi values within the range (1, 2]  
  
# Calculate theta values for the given Phi values and multiple r values  
results <- calculate\_theta\_minmax\_multiple\_r(r\_values, phi\_values)  
  
# Display the results  
kable(results)

| r | Min\_Theta | Max\_Theta |
| --- | --- | --- |
| 2 | 1.0071955 | 2.3219281 |
| 3 | 0.6354696 | 1.4649735 |
| 4 | 0.5035978 | 1.1609640 |
| 5 | 0.4337755 | 1.0000000 |
| 6 | 0.3896364 | 0.8982444 |
| 7 | 0.3587703 | 0.8270875 |
| 8 | 0.3357318 | 0.7739760 |
| 9 | 0.3177348 | 0.7324868 |
| 10 | 0.3031961 | 0.6989700 |

### 3.8.2 Min-Max Theta:

library(knitr)  
# Example usage  
r\_values <- 2:10 # Fixed values of r  
phi\_values <- seq(5.01, 10, by = 0.01) # List of Phi values within the range (1, 2]  
  
# Calculate theta values for the given Phi values and multiple r values  
results <- calculate\_theta\_minmax\_multiple\_r(r\_values, phi\_values)  
  
# Display the results  
kable(results)

| r | Min\_Theta | Max\_Theta |
| --- | --- | --- |
| 2 | 2.3248106 | 3.321928 |
| 3 | 1.4667922 | 2.095903 |
| 4 | 1.1624053 | 1.660964 |
| 5 | 1.0012414 | 1.430677 |
| 6 | 0.8993595 | 1.285097 |
| 7 | 0.8281142 | 1.183295 |
| 8 | 0.7749369 | 1.107309 |
| 9 | 0.7333961 | 1.047952 |
| 10 | 0.6998377 | 1.000000 |

### 3.8.2 Min-Max Theta:

library(knitr)  
# Example usage  
r\_values <- 2:10 # Fixed values of r  
phi\_values <- seq(10.01, 20, by = 0.01) # List of Phi values within the range (1, 2]  
  
# Calculate theta values for the given Phi values and multiple r values  
results <- calculate\_theta\_minmax\_multiple\_r(r\_values, phi\_values)  
  
# Display the results  
kable(results)

| r | Min\_Theta | Max\_Theta |
| --- | --- | --- |
| 2 | 3.323370 | 4.321928 |
| 3 | 2.096813 | 2.726833 |
| 4 | 1.661685 | 2.160964 |
| 5 | 1.431298 | 1.861353 |
| 6 | 1.285655 | 1.671950 |
| 7 | 1.183808 | 1.539502 |
| 8 | 1.107790 | 1.440643 |
| 9 | 1.048407 | 1.363417 |
| 10 | 1.000434 | 1.301030 |

### 3.8.2 Min-Max Theta:

library(knitr)  
# Example usage  
r\_values <- 2:10 # Fixed values of r  
phi\_values <- seq(1.01, 2, by = 0.01) # List of Phi values within the range (1, 2]  
  
# Calculate theta values for the given Phi values and multiple r values  
results <- calculate\_theta\_minmax\_multiple\_r(r\_values, phi\_values)  
  
# Display the results  
kable(results)

| r | Min\_Theta | Max\_Theta |
| --- | --- | --- |
| 2 | 0.0143553 | 1.0000000 |
| 3 | 0.0090572 | 0.6309298 |
| 4 | 0.0071776 | 0.5000000 |
| 5 | 0.0061825 | 0.4306766 |
| 6 | 0.0055534 | 0.3868528 |
| 7 | 0.0051135 | 0.3562072 |
| 8 | 0.0047851 | 0.3333333 |
| 9 | 0.0045286 | 0.3154649 |
| 10 | 0.0043214 | 0.3010300 |

# Estimated Ranges of for Different Ranks Based on

Table 1. Estimated Ranges of for Different Ranks Based on

**OJO OJO FALTA LA TABLA**

The table below presents the estimated minimum and maximum values of for different fixed ranks across various ranges of . This estimation is based on the formula , which provides insights into how the parameter varies with different values of and . For the range , spans from 1.007 to 2.322 for , from 0.635 to 1.465 for , and continues to decrease as increases, indicating a lower bound near 0.303 for . In the range , values are higher, with estimates ranging from 2.325 to 3.322 for and from 0.700 to 1.000 for . For , further increases, with values ranging from 3.323 to 4.322 for and 1.000 to 1.301 for . Finally, for , is highest for the lower ranks and decreases significantly for higher ranks, starting from 0.014 to 1.000 for and 0.004 to 0.301 for . This analysis demonstrates the variability of and highlights the differences in the Relative Importance Factor index across different ranks and ranges of .

The table below presents the estimated minimum and maximum values of for different fixed ranks across various ranges of . This estimation is based on the formula , which provides insights into how the parameter varies with different values of and . For the range , spans from 1.007 to 2.322 for , from 0.635 to 1.465 for , and continues to decrease as increases, indicating a lower bound near 0.303 for . In the range , values are higher, with estimates ranging from 2.325 to 3.322 for and from 0.700 to 1.000 for . For , further increases, with values ranging from 3.323 to 4.322 for and 1.000 to 1.301 for . Finally, for , is highest for the lower ranks and decreases significantly for higher ranks, starting from 0.014 to 1.000 for and 0.004 to 0.301 for . This analysis demonstrates the variability of and highlights the differences in the Relative Importance Factor index across different ranks and ranges of . This behavior can be clearly visualized in Figure a (presented below).

# 4 Application to Scopus database

The proposed theory is applied using databases downloaded from Scopus in the area of social resilience. Specifically, we focus on the number of citations of articles published in this area. This approach enables us to identify citation distribution patterns and evaluate how academic attention is concentrated on certain articles and specific topics based on the Relative Importance Factor (RIF) index.

## 4.1 Data Collection and Preparation

The data for this study were collected from the Scopus database, encompassing a comprehensive set of articles published on social resilience. The dataset includes information on article titles, authors, publication years, journal names, and the number of citations each article received. The initial dataset underwent a thorough cleaning process to remove any missing, duplicated, or inconsistent data, ensuring the accuracy and reliability of the subsequent analysis.

## 4.2 Exploratory Data Analysis

We began with an exploratory data analysis (EDA) to understand the general patterns and characteristics of the citation data. Descriptive statistics such as mean, median, standard deviation, and interquartile range of the citation counts were computed. Additionally, we visualized the distribution of citations to identify any potential outliers and to get a sense of the overall citation distribution.

## 4.3 Topic Modeling with Python

To uncover the underlying topics within the articles, we applied topic modeling to the abstracts of the downloaded papers. This process was conducted using Python, specifically employing the Latent Dirichlet Allocation (LDA) algorithm. The steps involved in this process are as follows:

1. Preprocessing the Text: The abstracts were preprocessed to remove stop words, punctuation, and to perform stemming and lemmatization, ensuring that the text data was clean and suitable for analysis.
2. Vectorization: The cleaned text data was then vectorized using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) to convert the text into a numerical format that could be used for topic modeling.
3. Applying LDA: The LDA algorithm was applied to the vectorized text data to identify the underlying topics. This involved specifying the number of topics to be extracted and iterating through the algorithm to refine the topic distribution.
4. Interpreting the Topics: The resulting topics were interpreted by examining the most prominent words in each topic. This allowed us to label the topics and understand the main themes present in the abstracts of the papers.

## 4.4 Application of the Power-Law Distribution

Assuming that the citation counts follow a power-law distribution, we estimated the parameters of this distribution, including the scale parameter and the threshold . The normalization constant was also determined to ensure that the probabilities sum to one. Using these parameters, we calculated the estimated probability that a factor occupies rank within the dataset.

4.5 Relative Importance Factor (RIF) Index Calculation

Next, we computed the RIF index for each article to evaluate their relative importance within the citation distribution. The RIF index compares the probability of an article occupying a given rank with the probability of it being in any higher rank, providing a nuanced understanding of the prominence of each article. Articles with higher RIF index values were identified as significantly more influential within the field of social resilience.

# Conclusions and Discussions

**CONCLUSIONS AND DISCUSSIONS ON THE APPLIED CASE WITH SCOPUS PAPERS ARE LACKING. BELOW THERE IS AN IDEA**

This study has introduced a comprehensive methodology for estimating the rank distribution of the most frequent factors within a variable composed of several factors , ordered by descending frequencies . By assuming that follows a power-law distribution, we derived a formula for the estimated probability of a factor occupying rank . Our findings reveal that this probability can be expressed using a proportionality constant , a normalization constant , and the power-law exponent . This approach effectively addresses a significant gap in the literature where prior studies have primarily focused on frequency distributions without considering rank estimation.

The introduction of the Relative Importance Factor (RIF) index allows for a deeper evaluation of the prominence and significance of factors. The RIF index compares the probability of a factor occupying a specific rank to the probability of it being in any higher rank. This innovative parameter offers a fresh perspective for analyzing complex systems and provides valuable insights into the dynamics of factor importance across different ranks. Future research could build on these findings by applying the RIF index to various fields, conducting longitudinal analyses, and integrating it with other statistical models. Additionally, the potential applications of the RIF index in different geographical regions and its policy implications present promising avenues for further investigation.

Theorem 3.2.2 outlines the conditions under which the mean, variance, moments, and cumulative distribution function (CDF) of the rank distribution exist and are well-defined. Specifically, the mean exists if is greater than 2, while the variance requires to exceed 3. Additionally, the -th moment exists if is greater than , and the CDF is well-defined for . These results provide a robust framework for understanding the statistical properties of rank distributions in power-law models.

The results illustrated in Figure 1 offer a detailed understanding of how the RIF index and probabilities change with rank and the power-law exponent . Figure 1(a) shows the distribution of the RIF index across different ranks for various values. It is evident that as both the rank and increase, the RIF index rises significantly, indicating a higher relative importance of factors at higher ranks. This underscores the sensitivity of the RIF index to changes in rank and the power-law exponent, emphasizing the necessity of accurately estimating for reliable assessments of factor prominence. Figure 1(b) demonstrates the inverse relationship between rank and probability, with higher values leading to a steeper decline in probability as rank increases. Figures 1(c) and 1(d) illustrate how the RIF index varies with the estimated probabilities and , respectively. The observed patterns align with theoretical expectations, confirming that the RIF index effectively captures the relative importance of factors within a ranked distribution. This analysis offers valuable insights into the dynamics of factor prominence, providing a robust framework for future research across various fields, including social resilience, economics, and network analysis.

Theorem 3.7.1 provides a crucial theoretical foundation for understanding how the RIF index can be used to compare the relative importance of factors across different ranks. By establishing a relationship between the power-law exponent and the RIF index, this theorem allows us to delineate conditions under which the importance of one factor relative to another can be quantified. This advancement is significant because it moves beyond merely modeling frequencies and delves into the dynamics of rank distributions, offering a more nuanced perspective on factor prominence. The theorem’s result, which relates the RIF index to the power-law exponent , ensures that this comparison is both meaningful and mathematically sound. The introduction of fixed ratios and the bounds on further enhance the robustness of this comparison. These theoretical insights are not only valuable for academic research but also have practical implications in fields such as social resilience, economics, and network analysis. Future work can build on these findings by applying the RIF index to empirical data, exploring its applicability in various contexts, and refining the theoretical framework to accommodate more complex distributions and interactions.

The analysis presented in this paper extends the traditional understanding of power-law distributions by focusing not only on frequencies but also on the distribution of ranks. The provided table and accompanying figure illustrate the relationship between the RIF index and the power-law exponent for various ranks. Specifically, the table delineates the range of values for different categories of , offering a clear visualization of how the importance of factors changes with rank. The figure further elucidates this relationship, showing that as increases, the RIF index rises more steeply, indicating a sharper decline in importance as rank increases. This behavior is particularly pronounced in the “Dominant” category, where the RIF index exceeds 10.

The implications of these findings are multifaceted. Firstly, they provide a robust mathematical framework for evaluating the relative importance of factors across different ranks. Secondly, the ability to categorize factors based on their values offers practical insights for fields such as social resilience, economics, and network analysis. For instance, in social resilience studies, understanding the prominence of specific factors can help policymakers prioritize interventions. Future research could explore the application of the RIF index in other domains, refine the theoretical model to accommodate more complex interactions, and validate the findings with empirical data. The clear categorization of factors into “Moderate,” “Significant,” “Critical,” and “Dominant” based on their values provides a nuanced tool for analyzing factor importance, enhancing both theoretical and practical understanding.

The analysis, carried out using the Scopus database, revealed several key findings:

1. Citation Distribution Patterns: The citation distribution exhibited a heavy-tailed pattern, consistent with the power-law behavior, indicating that a small number of articles received a disproportionately high number of citations, while the majority received relatively few.
2. RIF Index Insights: The RIF index highlighted the most prominent articles, which tended to be those that addressed critical aspects of social resilience, such as power relations, rules/institutions, and resource distribution. These findings align with the observations by Dagdeviren et al. (2020) on the interconnected nature of these elements in determining social resilience.
3. Topic Modeling Results: The topic modeling process uncovered several key themes within the abstracts, including community resilience, disaster response, and policy interventions. These topics reflect the diverse aspects of social resilience and provide a deeper understanding of the focus areas within this field.
4. Policy Implications: The results suggest that academic attention is highly concentrated on a few pivotal topics within social resilience, which can inform public policy formulation and the implementation of resilient practices in communities and organizations. Policymakers can leverage these insights to prioritize areas that are critical for enhancing social resilience.

In summary, applying the proposed theory to the Scopus dataset on social resilience has provided valuable insights into the distribution of academic attention in this field. The use of the RIF index has enabled a detailed evaluation of the relative importance of articles, contributing to a deeper understanding of the dynamics of academic influence. The topic modeling process further enriched our analysis by identifying key themes within the abstracts, offering a comprehensive view of the research landscape. This approach not only fills a significant gap in the literature but also offers a robust framework for future research and practical applications in policy and decision-making contexts.

This work has addressed a notable gap in the literature by introducing both a formula to model the rank distribution of frequently occurring factors (an area previously overlooked in power-law distribution studies) and the Relative Importance Factor (RIF) index. The practical implications of this framework are significant. For instance, in social resilience studies, understanding which factors are consistently important (Balanced) versus those that dominate only at top ranks (Dominant) can inform policy decisions and resource allocation. Additionally, this methodology can be applied to various fields, such as economics, network analysis, and risk management, where understanding the dynamics of factor importance is crucial. Future research can expand on this work by applying the RIF index to different datasets and exploring its potential in predictive modeling and decision-making processes. The versatility and robustness of the RIF index make it a valuable tool for analyzing complex systems and their underlying factors.

# FUTURE WORKS

Based on the advancements and insights presented in this work, several avenues for future research can be pursued:

1. Application of the RIF Index to Other Domains: While this study focuses on social resilience using citation databases, the RIF index can be applied to other fields such as economics, biology, and network science to evaluate the relative importance of different factors within those contexts.
2. Longitudinal Studies: Conduct longitudinal studies to observe how the rank distribution and the RIF index evolve over time. This can provide insights into the dynamic changes in the prominence of factors across different periods.
3. Comparison Across Different Geographical Regions: Extend the application of the RIF index to compare the relative importance of factors across different countries or regions. This can help identify regional differences and similarities in the distribution of factors.
4. Integration with Other Statistical Models: Integrate the RIF index with other statistical and machine learning models to enhance predictive analysis and decision-making processes. For example, combining the RIF index with regression models or clustering algorithms.
5. Exploration of Other Heavy-Tailed Distributions: Investigate the applicability of the RIF index in the context of other heavy-tailed distributions such as the log-normal and stretched exponential distributions. This can broaden the scope and utility of the RIF index in various theoretical and practical applications.
6. Sensitivity Analysis: Perform sensitivity analysis to understand the robustness of the RIF index under different conditions and assumptions. This includes examining the impact of varying the threshold and scale parameters on the rank distribution and the RIF index.
7. Policy Implications: Explore the policy implications of the findings by using the RIF index to inform strategies for resource allocation, risk management, and strategic planning in various sectors such as healthcare, education, and public administration.

By pursuing these future research directions, the theoretical framework and practical applications of the RIF index can be further developed and refined, contributing to a deeper understanding of the distribution and importance of factors in complex systems.