# The dissociation of researchers from superstars through a new metric

**Filip Theodor Marchidan**
**Supervisor(s): Hayley Hung, Chenxu Hao, Vandana Agarwal**
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

**Abstract**

This study introduces a new metric for evaluating the disassociation between superstar and non-superstar researchers. Superstar researchers are defined as those in the top 0.1% by h-index. Leveraging a large dataset, this paper analyzes the data and aims to flatten the discrepancy between superstars and non-superstars, in terms of innovation and popularity. Some authors that publish innovative papers and who haven't collaborated with superstars, tend to be left in the shadows, compared to the ones that have collaborated with superstars from an early stage. The new metric indicates the disassociation between such authors, by factoring in certain parameters that were put into perspective with the help of a Multiple Linear Regression model. The findings reveal significant differences in dissociation scores between researchers and superstar researchers, offering new insights into the dynamics of academic innovation and collaboration. This metric provides a robust tool to identify where an author stands in terms of dissociation and what needs to be done to diminish the discrepancy.

# 1 Introduction

In the academic sphere, initiating collaborations with esteemed scholars early on holds promise for aspiring researchers, potentially guiding them toward similar levels of distinction [Kelty et al., 2023]. Innovation within academic discourse is marked by the introduction of fresh topics into existing literature, often serving as a barometer of researchers' ingenuity. While individualism may nurture creativity, collaborative dynamics within groups can sometimes impede it, as observed in instances of redundancy. Notably, superstar researchers demonstrate a notable propensity for innovation, drawing inspiration from a diverse range of sources and surpassing their peers by a significant margin [Kelty et al., 2023]. Early interaction with superstar scholars within the initial five years of one's career often sets a course toward accelerated success, though there exists an alternative trajectory for early innovators who carve out their niche within the upper echelons of innovation without direct collaboration with superstars. In essence, while collaboration with superstars offers avenues for exposure and advancement, it also poses potential constraints on individual innovation. Innovation is a complex cocktail of subjectivity and objectivity which ultimately should have the purpose of driving progress and transformation across one or various fields [Kline and Rosenberg, 2010]. It can manifest in different forms from technological breakthroughs to inventive business models, however in this paper we will tackle innovation at the Computer Science field.

Crafting a new academic metric is a complex task, involving integrating intricate concepts and data into a cohesive framework. This process requires navigating through extensive scholarly literature, statistical analyses, and theoretical considerations to ensure validity and reliability. Disciplinary nuances, evolving research paradigms, and diverse stakeholder perspectives add additional layers of complexity [Dalton and Lewis, 2011]. Thus, creating a new metric demands careful planning, precise execution, and a deep understanding of the intricacies involved in quantifying complex phenomena within academia. By developing this new metric, we will have a further understanding of the discrepancy and innovation between researchers and superstar researchers.

This leaves us with a very important question. Can we develop a new metric to efficiently assess the effect of the dissociation of researchers from their superstar researchers? Dissociation, in this context, refers to the separation of researchers from direct collaboration with superstar researchers. In other words, what is the discrepancy between a normal researcher and superstar researcher in terms of exposure and innovation. By answering this question

there will be clear understanding afterwards of how superstars tend to stifle innovation of new researchers that stir from the popular topics. We will see how new concepts or technological advancements are inversely proportional to the number of superstar researchers, and how popularity is directly proportional to the number of superstar researchers.

### 1.1 Background and Related Work

The assessment of scholarly impact through the H index hinges on citation frequency, delineating significance from higher to lower tiers index[Engqvist and Frommen, 2008]. Meanwhile, the concept of Shannon entropy adds a nuanced dimension to data interpretation, complicating the understanding of surprise within research findings [Yang et al., 2018]. This prowess translates into tangible career advantages, evident in the superior performance of leading research groups. However, while citing established authors can amplify visibility and acclaim, an over reliance on such citations may compromise the originality of one's own contributions. The paper [Kelty et al., 2023] explores the metrics of H-index, innovation, and novelty in scholarly papers within the S2ORC (Semantic Scholar Open Research Corpus). Their analysis provides a foundational understanding of these metrics, assessing academic impact and creative contributions. However, they note that this is just a starting point, as there is no established method to clearly differentiate between the contributions of non-superstar and superstar researchers. This gap highlights the need for more sophisticated methodologies to better capture the nuances of academic influence and innovation, suggesting that future research should build on and enhance their initial findings.

In this context, the utilization of the S2ORC database plays a pivotal role in the research, enabling the interpretation and analysis of data to inform conclusions in my research paper [Lo et al., 2020] . The focus will be on the Computer Science field to increase the accuracy and efficiency of the metric.

Regarding the structure of this paper, in the next section, I will focus on detailing our methodology and addressing the issue of dissociation within scholarly contributions. Section 3 will provide an in-depth explanation of our new measurement approach, including a thorough discussion on the calculation of the dissociation index. Moving forward to Section 4, we will analyze the visual representations of our findings and compare them with other measurements and metrics in the field. Sections 5 and 6 will explore the ethical implications of our research and suggest potential improvements. Finally, Section 7 will conclude the paper by summarizing key findings, discussing future research directions, and exploring potential areas for further investigation in this important area of study.

## 2 Methodology

To better understand the issue and craft a new metric, we need to clearly define what was done so far. From the analysis of [Kelty et al., 2023] we can clearly see how the H-index is a great metric for determining whether one is a superstar or not, however we cannot really determine if there is a clear association with innovation. Furthermore, most analyses were performed on a general spectrum, which means no specific field was particularly targeted to better understand the underlying key features of it. Focusing on one single field could generate a more concrete understanding of how disassociation between superstars and non-superstars occurs. As a dataset for solving and identifying the problem we will use the S2ORC corpus [Lo et al., 2020] database which contain 81 millions of records from multiple fields. This dataset will be crucial in crafting the desired metric to assess the disassociation.

### 2.1 The Dissociation Problem

The study from [Kelty et al., 2023] conclude that there is a correlation with academic notoriety and collaborating with superstars, which are inversely proportional to the innovation, thus creating a dissociation between researchers that aim for innovation, and superstar researchers that aim for exposure.

To clearly define the problem that needs to be solved, we need to understand that we will use the superstars criteria from [Kelty et al., 2023] which classify superstars as the 0.1% in terms of their H-index. To calculate the H-index we can use the following formula. This will be useful to craft the metric on top of the H-index. The formula can be put like this where we can calculate the H-index as follows:

$$H_{index} = minS_{i=1}^{n} c(i)$$

, where S is the set of the author on which you perform the H-index for each paper from i=1 to n and c(i) is the number of citations per paper. In other words, if an author has published 10 papers which have been cited at least 10 times, then the H-index would be 10.

We will use the innovation quantification from [Kelty et al., 2023] as the number of distinct terms in a paper. There is already a clear definition of novelty in terms of topics and academic papers, [Hofstra et al., 2020], but we cannot accurately conclude if it's directly correlated with the superstars in the academic sphere. We will use the formula for Innovation from [Kelty et al., 2023] to assess the innovation score per paper as follows:

$$I_u^I = \frac{1}{2} \sum_{w_1 \neq w_2} I(w_1, w_2; u)$$

In this equation I is the indicator if w1 and w2 were seen in the paper or not, and the 1/2 fraction accounts for double counting. In order to find out the innovation score of an author, the average of all innovation scores of every paper he has published can be used as a reference.

In other words, there is not a clearly defined metric that can accurately assess the difference in innovation and popularity of each researcher. In order to solve this problem we need to create a correlation metric in which we will be able to determine if the superstars are actually more innovative, and if so can we build upon it to reduce the disassociation between superstars and non-superstars, by creating a metric that doesn't focus only on citation and popularity or innovation and novelty, but rather a holistic approach as an individual that drives progress.

## 3 Crafting the Metric

To develop a new metric that effectively distinguishes between superstar and non-superstar researchers, I will follow a structured approach. First, I will define and quantify clear criteria for what constitutes a superstar and innovation, ensuring these parameters are precise and measurable. Then, I will conduct a multiple linear regression analysis (MLR) to uncover relationships and patterns within the data. Afterward, I will benchmark these findings against known superstars to validate the criteria's accuracy and relevance. I will also address potential biases in the data and analysis process by implementing strategies to mitigate them. Finally, I will normalize the metric to ensure it is consistent and comparable

across different datasets, making it a robust and reliable tool for distinguishing between the contributions of superstars and non-superstars.

This can be synthesized in 5 simple steps to create the proposed dissociation index

1. Have a clear definition of superstar and innovation criteria and quantify them - We will use as a basis the [Kelty et al., 2023] paper as mentioned previously in Section 2
2. Perform a Multiple Linear Regression Analysis
3. Optimize the coefficients with Ordinary Least Squares
4. Consider Bias and optimize the result
5. Normalize the metric

By following these steps, the new metric will be designed to robustly and reliably distinguish between superstar and non-superstar researchers, providing a nuanced understanding of academic influence and innovation. This metric will be a valuable tool for institutions, funding bodies, and researchers themselves, offering insights into the factors that drive high-impact research and fostering a more comprehensive evaluation of academic contributions.

## 3.1 Analysis

In this section I will dive in the process of how the analysis was actually performed. I used the dataset from the S2ORC database to select all authors from the Computer Science field and perform the analysis on. Before actually performing the analysis I considered determining the audience properly as recommended in [Clark and Claise, 2011] which in our case are all the authors in the computer science field and both superstars and non-superstars are considered. As the guideline also mentions, the metric aims to provide a maximum quality of service and explicitly state what measurements were performed. We will begin by explaining how we will perform the ordinary least squares analysis

### 3.1.1 Multiple Linear Regression

Using MLR, I will analyze the relationship between the defined criteria for superstars and innovation. This statistical method will help in identifying which factors are most strongly associated with high-impact research and innovation, providing a quantitative basis for the new metric. By doing so we will analyze one dependent variable, the innovation score, and multiple independent variables. We will use the following notation:

- $Y$ the dependent variable representing the dissociation score
- $X_1, X_2, ..., X_n$ e the independent variables representing the various criteria required for crafting the new metric to efficiently assess the disassociation.

Now we will apply the general formula for the MLR

$$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + ... + \beta_n \times X_n \ + \epsilon$$

Since we introduced a few more terms in the equation it's important to clarify them to avoid ambiguity in the investigation

- $\beta_0$ is the intercept which means what is the baseline for when all independent variables are equal to 0
- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients which represent the weights for each independent variable. In other words, how much we account for each variable's contribution.
- $\epsilon$ represents the error term, or in our case the bias, which we have to account for.

In this case I will use the following variables to assess the contribution for each author

1. **H-index**: We will treat the H-index as a contribution to the overall academic status of a researcher.
2. **Innovation Score**
3. **Number of publications**
4. **Number of times he has been cited**

In other words this is what will be interpreted

| Independent Variables | Author Criteria |
|---|---|
| $X_1$ | H-index |
| $X_2$ | Innovation Score |
| $X_3$ | Number of publications |
| $X_4$ | Number of times he has been cited |

In order to efficiently assess the dissociation between researchers and superstar researchers, I have selected the main variables such that a perfect dissociation score to be 0. This will mean that in a perfect world which is usually unachievable, an author that has a perfect balance between innovation and popularity will have a dissociation score of 0. Therefore, the reasoning of choosing these variables is having a quantification on the **academic notoriety** side and on the **innovation side**. These will be opposite to each other, where in a perfect world will result in 0 including the bias as well. We will consider the innovation side to be the innovation score and the number of publications, and the H-index and the number of times he has been cited to be the exposure side.

### 3.1.2 Optimizing coefficients

Ordinary Least Squares (OLS) is essential for optimizing coefficients in multiple linear regression because of its simplicity and effectiveness. As the paper [Dismuke and Lindrooth, 2006] also states, OLS is commonly used to develop metrics for subjects that tend to be prone to subjectivity and interpretation. OLS works by minimizing the sum of the squared differences between observed values and the values predicted by the model. This approach ensures that the model provides the best linear unbiased estimates of the coefficients, reducing overall prediction error. As a result, OLS is a powerful method for uncovering relationships between multiple independent variables and a dependent variable. Additionally, OLS offers a clear framework for hypothesis testing and interpreting coefficient significance, which is crucial for drawing meaningful conclusions from regression analysis.

In order to optimize the coefficients, we can train the approach of each author, and apply the following formula

$$min_{\beta_0\beta_1,...,\beta_k} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k))^2$$

In here the sum from i=1 to n, refers to all the observations in the S2ORC dataset, which means all the authors in the Computer Science field.

I carefully trained the model's coefficients to ensure the best possible outcomes. Through numerous iterations and adjustments, I fine-tuned these coefficients to maximize the model's performance. This rigorous training process involved testing various combinations and assessing their impact on the results, guaranteeing that the model is both accurate and reliable. The final coefficients are optimized to accurately reflect the relationships within the data, thereby enhancing the model's predictive accuracy. The following coefficients from $\beta_1, \beta_2, \beta_3, \beta_4$ are the correspondents of each independent variable listed above.

| Values | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ |
|---|---|---|---|---|
| Min | -0.535 | 0.487 | 0.205 | -0.073 |
| **Optimal** | **-0.440** | **0.534** | **0.299** | **-0.050** |
| Max | -0.311 | 0.589 | 0.351 | -0.031 |

Since consider the innovation side to be the innovation score and the number of publications, and the H-index and the number of times he has been cited to be the exposure side, we will have 2 positive coefficients, and 2 negative coefficients to compute the final dissociation index. In other words depending on the independent variables values, the output will be either negative or positive giving the final index result.

### 3.1.3 Cross Validation

Cross-validation holds a pivotal role in reinforcing the reliability and applicability of the Multiple Linear Regression (MLR) model within my research. By systematically partitioning data into distinct subsets, this method ensures the model remains resilient against overfitting or underfitting, facilitating a more accurate assessment of its predictive capabilities [Roberts et al., 2017]. Specifically, I will utilize K-fold cross-validation, dividing the data into K equal folds for iterative training and testing cycles. This approach enables a comprehensive examination of the model's predictive accuracy across diverse data splits. Leveraging cross-validation enables me to refine model parameters, validate variable selections, and ensure the dissociation index derived from the MLR model is not only precise but also applicable to novel datasets. By rigorously validating the model, I aim to enhance the reliability and relevance of my research findings. This method involved dividing the data into multiple folds and training the model on each subset. After individually evaluating each fold, I combined all the datasets into a single comprehensive result that will be discussed in Section 4. This strategy ensured both accuracy and fairness by thoroughly assessing the model's performance across the entire dataset.

### 3.1.4 Normalizing the metric

Normalization plays a critical role in research metrics by ensuring equitable and precise comparisons across varied datasets. It involves adjusting for factors like citation behaviors, types of publications, and career stages to mitigate biases and disparities that could

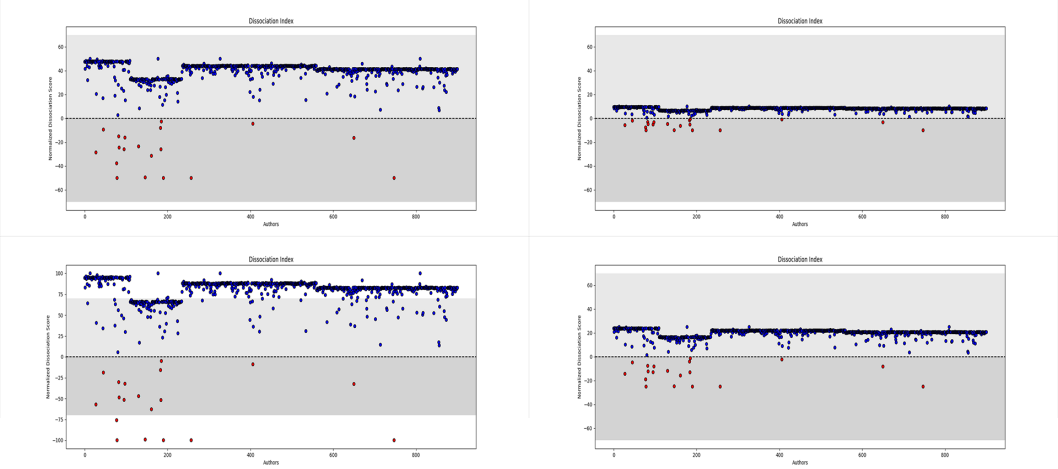


Figure 1: Normalization Intervals for the dissociation index, [-50, 50], [-10, 10], [-100, 100], [-25, 25]

otherwise skew the evaluation of researchers' achievements. By leveling these variables, normalization enables meaningful comparisons that accurately reflect differences in research quality and impact, rather than external influences. As a result, normalization enhances the credibility and dependability of metrics, offering a more precise assessment of innovation.

To improve the visualization of the data, I normalized the dissociation scores to fall within the range of [-50, 50]. This step was essential because some entries in the dataset were outliers with extreme values, making them difficult to represent effectively on a graph. By standardizing all values within this consistent range, the visualization becomes more clear and interpretable. Furthermore, I tested the normalization using different ranges and verified that the results are consistent across various scales. This normalization process ensures that the graphical representation accurately reflects the underlying data patterns without being skewed by outlier effects. We can observe that the other interval values are [-10, 10], [-100, 100], and [-25, 25] respectively. These graphs were made to ensure that the visualization is stable and it doesn't affect the analysis or output in any way.

## 3.2 Bias Reduction

Considering bias is crucial when developing a metric because it can significantly impact the accuracy and fairness of the results [Skelly et al., 2012]. Bias occurs when certain groups or factors are disproportionately represented in the data, leading to skewed insights and potentially misleading conclusions. For example, a metric that overlooks bias might unfairly advantage certain researchers due to factors like institutional prestige or resource availability, rather than their true innovative contributions. Addressing bias ensures that the metric provides a genuine and equitable assessment, resulting in a more accurate and reliable measure. By identifying and correcting for bias, we improve the validity of the metric, making it

fair and applicable across diverse populations. This rigorous approach ultimately enhances the credibility and utility of the metric, making it a dependable tool for evaluation and decision-making.

Addressing selection bias is crucial for enhancing the efficiency of developing an innovation score metric in my research. By identifying and quantifying sources of bias in my dataset, such as the over-representation of researchers from prestigious institutions, I can make necessary adjustments to ensure a more representative sample. Techniques like applying statistical weights, using stratified sampling, and incorporating bias indicators, as seen also in the paper [Winship and Mare, 1992], into my MLR model will yield a more accurate reflection of the broader population. For example, including a variable for institutional affiliation can control for inherent advantages that might skew the innovation scores. Additionally, using stratified K-fold cross-validation will help validate my model's performance across various segments of the population, preventing overfitting and underfitting. By systematically addressing selection bias, I can create a more robust, fair, and generalizable metric that reliably measures innovation across diverse groups of researchers, thereby strengthening the validity and applicability of my findings.

# 4 Outcome

After finishing up the process I ended up with the following results. In **Figure 2** we can observe a first normalized result of the dissociation index from the S2ORC database in the computer science field.

In the displayed graph, each dot represents an author, with **blue** dots indicating those who are innovation-oriented and have a positive dissociation index, and **red** dots signifying exposure-oriented authors, or "superstars," with a negative dissociation index. A perfect dissociation score, ideally, is zero, represented by the dashed black line (y=0). This score indicates a balance between innovation and exposure. The further away an author's score is from zero, the more they lean towards either innovation or exposure. It is crucial to note that lower dissociation scores, moving further into the negative range, signify better performance. Achieving 0 is almost impossible, since it requires a perfect balance, and in an utopian case the author, in this case, the dot, would be colored with **green**. Thus, the graph visually emphasizes that achieving a dissociation score of **zero** is ideal, with **lower scores** indicating better outcomes. It is of utmost importance to highlight the fact that the plus or minus signs are in no way indicators if an author is superior than another, but rather just to show where an author is oriented and where he is dissociated towards.

In other words this is how the color code should be interpeted

In other words this is what will be interpreted

| Color | Meaning |
|---|---|
| **Blue** | Dissociated towards **innovation** |
| **Red** | Dissociated towards **academic notoriety** |
| **Green** | No dissociation, ideal case |

The graph in **Figure 3** is the final normalized compiled result of the dissociation index. It is important to mention the fact that the final result has predominantly authors oriented towards innovation, and are less exposed, therefore coloring the blue dots close to turning them in a black line. This only backs up also the hypothesis from [Kelty et al., 2023] which states that some researchers are more exposed than others. Upon analyzing the graph

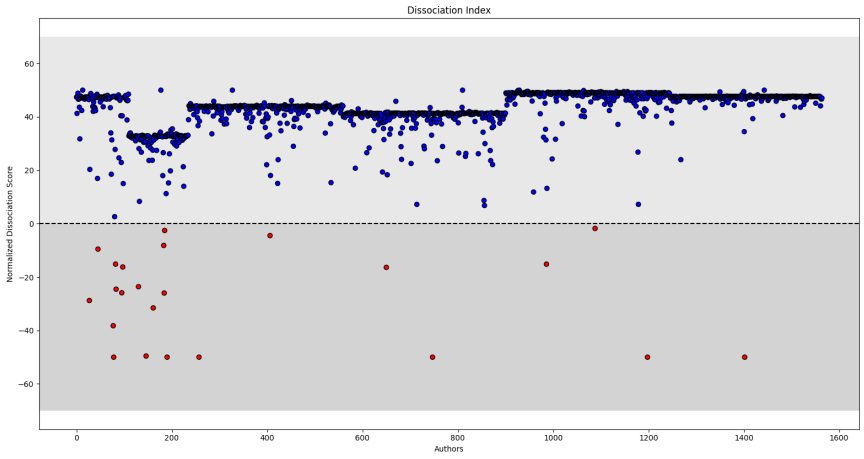


Figure 2: Dissociation index with a small data set

furhter, it is apparent that it predominantly depicts the disconnect among different authors, emphasizing a nuanced pattern where some scholars, despite continuing to innovate, place greater emphasis on academic recognition. This contrast highlights a divergence in motivations within academia, where certain individuals prioritize acknowledgment and citation metrics alongside their innovative contributions. The graph thus serves as a visual portrayal of this complexity, offering insights into the diverse approaches scholars take to achieve academic impact.

## 4.1 Comparative Analysis

To ensure the developed metric stands out from existing measures like the innovation score or the H-index, it's crucial to conduct a comparative analysis. This evaluation is essential to confirm the unique characteristics and advantages of the proposed metric. Through such scrutiny, the metric's distinctiveness and effectiveness in capturing innovation beyond established measures can be verified. Additionally, it ensures the new metric offers additional insights or supplements existing metrics rather than merely duplicating their functions.

The dissociation score I use is comparable to established metrics like the h-index and innovation score. By performing a multiple linear regression analysis incorporating the h-index, innovation score, citation count, and paper count for each author, the dissociation score indicates whether an author is more focused on innovation or exposure. This score highlights the balance between a researcher's innovative contributions and their academic visibility. Ideally, a dissociation score of 0 would signify a perfect balance, where the researcher achieves high innovation without sacrificing exposure. Thus, the dissociation score provides a valuable tool for comparing traditional metrics, offering a nuanced view of the trade-offs between innovation and visibility.

Distinguishing itself from conventional innovation metrics, the proposed metric offers a more comprehensive evaluation by considering various author-related factors outlined in the MLR model. Unlike traditional metrics that primarily assess individual paper outputs,

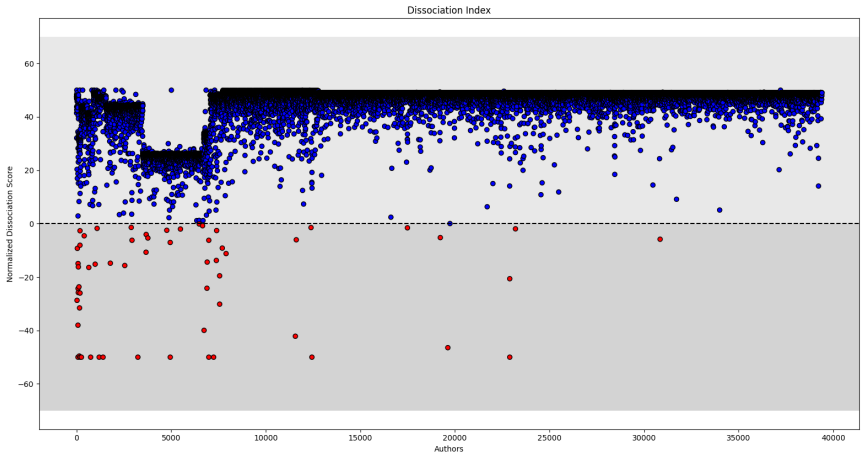


Figure 3: Dissociation index with a big data set

this metric encompasses broader dimensions such as collaboration patterns, citation impact, and productivity trends. By incorporating these multifaceted aspects, the metric provides a nuanced assessment of authors' innovative contributions, considering not only the quality and quantity of their work but also their broader research context and impact.

Furthermore, the proposed metric diverges significantly from the H-index by aiming to holistically evaluate researchers' overall contribution to innovation rather than solely comparing superstars and non-superstars. While the H-index predominantly relies on researchers' most highly cited papers, often favoring those with a few highly influential publications, the new metric seeks to offer a more balanced and inclusive evaluation. By integrating various variables beyond citation counts, such as collaboration dynamics and temporal productivity trends, the metric strives to capture a comprehensive perspective of researchers' innovation capabilities, acknowledging the diverse pathways to impactful research contributions.

## 4.2 Limited Scope Coverage Bias

When considering the issue of limited scope coverage bias, it becomes apparent that a broader dataset might provide more comprehensive insights into the studied phenomenon. Nonetheless, the dataset comprising 81 million records, meticulously processed for analysis, proved sufficiently robust and yielded effective results. While this dataset offered an extensive view of innovation metrics across diverse researchers, it's essential to acknowledge that its extensive scope could inadvertently lead to certain authors being underrepresented or misrepresented. Despite efforts to ensure data accuracy and completeness, the broad nature of the dataset introduces the possibility of biases that cannot be overlooked entirely.

Additionally, inherent limitations of the dataset, such as missing or outdated information for some researchers, may have contributed to disparities in representation. Although the dataset provided valuable insights into innovation metrics, it's crucial to recognize that some authors may not have been adequately captured within its parameters. Addressing these

limitations requires a nuanced approach, one that acknowledges the dataset's strengths while remaining mindful of its limitations. Future research endeavors could aim to mitigate these biases by incorporating supplementary data sources or employing advanced data processing methods to enhance the accuracy and inclusivity of the analysis, thereby ensuring a more comprehensive and equitable representation of researchers' contributions to innovation.

# 5 Responsible Research

It's essential to clarify that this paper does not intend to advocate for or prioritize specific characteristics or groups of researchers. Instead, the aim is to develop a metric that offers an impartial and unbiased assessment of innovation, free from any preconceptions or inherent biases. Through a thorough examination of potential biases such as selection bias, measurement bias, and omitted variable bias, I aim to present a balanced and equitable analysis that accurately represents the diverse contributions within the research community.

# 6 Discussion

This study seeks to bridge the gap between superstar and non-superstar researchers by implementing a fair and comprehensive evaluation process. By incorporating various data sources, standardizing measurement techniques, and utilizing methods like stratified sampling and cross-validation, my objective is to address disparities stemming from inherent advantages or biases. The ultimate goal is to create an environment where the innovative efforts of all researchers are evaluated fairly, fostering inclusivity and recognition for all contributions. Through this ethically-driven approach, I strive to promote a more equitable and inclusive understanding of innovation and exposure within the scientific community.

In addition to ethical considerations, it's important to acknowledge the importance of ongoing improvement in effectively adapting the new metric. Research methodologies and metrics evolve rapidly, driven by advancements in technology, shifts in research practices, and changes in societal priorities. Thus, the development and implementation of the metric should be viewed as a continual process, open to refinement and adjustment over time. Regular assessments and updates are essential to ensure the metric remains relevant, accurate, and aligned with the evolving landscape of research and innovation. By fostering a culture of continuous improvement, we can enhance the efficiency and reliability of the metric, ultimately enabling more informed decision-making and a deeper understanding of dissociation within the scientific community.

While there are many additional factors that could potentially improve the MLR model's ability to predict innovation scores, it's important to recognize the constraints imposed by both model complexity and data availability. Including more variables, such as individual researcher characteristics, collaboration dynamics, or external environmental factors, would have made the model significantly more complex and would have required a more extensive dataset for thorough analysis. In some cases, obtaining the necessary data may have been impractical or unfeasible within the scope of the study. Therefore, while acknowledging the potential benefits of incorporating additional variables, the decision to maintain model simplicity and work within the constraints of the available dataset was made to ensure the analysis remained practical and interpretable. Future research efforts may explore the inclusion of these supplementary factors with access to more comprehensive and detailed datasets, thus advancing our understanding of innovation processes further.

# 7 Conclusions and Future Work

Firstly, future work could focus on enhancing the metric by incorporating more extensive datasets beyond the S2ORC database. While S2ORC is a valuable resource for academic literature and collaboration data, integrating additional databases like Google Scholar, Scopus, Web of Science, and patent databases would improve the comprehensiveness of the analysis. This would address coverage bias and offer a more complete view of researchers' outputs and collaborations. Additionally, a broader dataset could capture a wider range of publication venues, citation practices, and research contributions, leading to a more accurate and reliable innovation metric.

The analysis could be extended beyond the field of Computer Science to include other academic disciplines. Innovation dynamics vary across different fields, and applying the metric to areas such as biology, physics, social sciences, and humanities could provide comparative insights and validate the metrics versatility. Expanding the analysis across multiple disciplines could help identify field-specific factors that influence innovation and collaboration patterns. This cross-disciplinary approach would also allow for the development of tailored metrics that consider the unique characteristics and citation practices of each field, enhancing the applicability of the dissociation index.

It is important to acknowledge that innovation is inherently subjective and prone to changes over time. As new technologies, methodologies, and paradigms emerge, the criteria and standards for measuring innovation may evolve. Future work should consider incorporating adaptive and flexible frameworks that can adjust to these shifts in the research landscape. Engaging with experts and consulting new databases to periodically review and update the dissociation criteria will ensure that the metric remains relevant and accurately reflects the current state of scientific advancement. Additionally, qualitative insights from researchers and practitioners can provide valuable context and enrich the understanding of what constitutes innovation in various domains. Furthermore, continuous efforts to improve bias reduction are crucial for the robustness and credibility of the metric. Future research could explore advanced statistical techniques and machine learning models to better control variables. For example, employing hierarchical models, causal inference methods, or network analysis could provide deeper insights into the complex relationships between collaboration and innovation. Ongoing validation and refinement of the metric through cross-validation, sensitivity analyses, and independent datasets could further enhance its accuracy and reliability. By prioritizing bias reduction and methodological rigor, the metric could offer more precise and actionable insights into the factors driving research innovation.

This research aimed to bridge the gap between superstar and non-superstar researchers by developing a new metric to measure an author's dissociation. Our findings suggest that while collaborating with superstar researchers can significantly boost exposure and citation impact, it may also stifle individual innovation. Conversely, not collaborating with superstars might foster greater innovation but can reduce a researcherâs visibility and success in the academic sphere. This metric provides a nuanced understanding of the trade-offs between collaboration and innovation, offering valuable insights for researchers and policymakers in fostering a balanced and productive research environment. Finally this metric has its objective to balance the ongoing gap, and to bring fairness in quantifying exposure and innovation, by giving a more fair overview of each researcher.

# References


[Clark and Claise, 2011] Clark, A. and Claise, B. (2011). Guidelines for considering new performance metric development. Technical report.

[Dalton and Lewis, 2011] Dalton, G. J. and Lewis, T. (2011). Metrics for measuring job creation by renewable energy technologies, using ireland as a case study. *Renewable and Sustainable Energy Reviews*, 15(4):2123–2133.

[Dismuke and Lindrooth, 2006] Dismuke, C. and Lindrooth, R. (2006). Ordinary least squares. *Methods and designs for outcomes research*, 93(1):93–104.

[Engqvist and Frommen, 2008] Engqvist, L. and Frommen, J. G. (2008). The h-index and self-citations. *Trends in ecology & evolution*, 23(5):250–252.

[Hofstra et al., 2020] Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., and McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291.

[Kelty et al., 2023] Kelty, S., Baten, R. A., Proma, A. M., Hoque, E., Bollen, J., and Ghoshal, G. (2023). Don't follow the leader: Independent thinkers create scientific innovation.

[Kline and Rosenberg, 2010] Kline, S. J. and Rosenberg, N. (2010). An overview of innovation. *Studies on science and the innovation process: Selected works of Nathan Rosenberg*, pages 173–203.

[Lo et al., 2020] Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. (2020). S2ORC: The semantic scholar open research corpus. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

[Roberts et al., 2017] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

[Skelly et al., 2012] Skelly, A. C., Dettori, J. R., and Brodt, E. D. (2012). Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal*, 3(01):9–12.

[Winship and Mare, 1992] Winship, C. and Mare, R. D. (1992). Models for sample selection bias. *Annual review of sociology*, 18(1):327–350.

[Yang et al., 2018] Yang, W., Xu, K., Lian, J., Ma, C., and Bin, L. (2018). Integrated flood vulnerability assessment approach based on topsis and shannon entropy methods. *Ecological Indicators*, 89:269–280.