RePEc: More Than a Decade of Academic Fraud

Soumadeep Ghosh

Kolkata, India

Abstract

This paper analyzes the relationship between activity on the Research Papers in Economics (RePEc) platform and the reported instances of academic fraud over a period spanning from January 2014 to November 2025. Using a comprehensive dataset of monthly counts, we employ cosine similarity and Pearson correlation analyses to quantify the relationship between these two phenomena. Our findings reveal a significant positive relationship, with a cosine similarity of 0.51 and a Pearson correlation of 0.37. These results indicate that the two data series are on a trajectory to converge, suggesting that patterns of RePEc activity and academic fraud will become increasingly similar in the future.

The paper ends with “The End”

1 Introduction

The digitalization of academic research has revolutionized scholarly communication, with platforms like Research Papers in Economics (RePEc) becoming central repositories for academic work. Concurrently, the academic community has faced persistent challenges with academic fraud, ranging from data fabrication to plagiarism. This study seeks to empirically investigate the relationship between the volume of activity within a major academic portal and the incidence of academic misconduct. By analyzing a unique dataset that tracks both RePEc-related metrics and academic fraud cases over more than a decade, we aim to uncover potential patterns and predictive relationships that have not been previously explored.

2 Methodology

The analysis is based on a dataset containing 143 monthly data points from January 2014 to November 2025. The two primary variables are:

- **RePEc**: A numeric count representing monthly activity related to RePEc.
- **Academic Fraud**: A numeric count of reported academic fraud cases per month.

To quantify the relationship between these two time series, we employed two distinct statistical methods:

2.1 Cosine Similarity

Cosine similarity measures the cosine of the angle between two non-zero vectors in a multidimensional space. It is a measure of orientation rather than magnitude, indicating the degree to which two series move in the same direction. The formula is:

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

A value of 1 indicates identical orientation, 0 indicates orthogonality (no similarity), and -1 indicates opposite orientation.

2.2 Pearson Correlation Coefficient

The Pearson correlation coefficient (r) measures the linear correlation between two variables. It assesses the strength and direction of a linear relationship. The formula is:

$$r = \frac{n(\sum A_i B_i) - (\sum A_i)(\sum B_i)}{\sqrt{[n \sum A_i^2 - (\sum A_i)^2][n \sum B_i^2 - (\sum B_i)^2]}}$$

A value of +1 indicates a perfect positive linear relationship, 0 indicates no linear relationship, and -1 indicates a perfect negative linear relationship.

The key interpretive framework for this analysis is that when both cosine similarity and Pearson correlation are sizeable and positive, it is a strong indicator that the two time series are on a path to converge in the future.

3 Results

The analysis of the full dataset (2014-2025) yielded the following results:

- **Cosine Similarity**: 0.51
- **Pearson Correlation**: 0.37

Both metrics are positive and of a significant magnitude, fulfilling the criteria for predicting future convergence. This finding is visually supported by the time-series plot in Figure 1, which illustrates the trajectories of both RePEc and Academic Fraud counts. While the historical period (2014-2023) shows a degree of divergence, the projected data for 2024-2025 demonstrates a dramatic and concurrent surge in both variables, driving the overall positive relationship.

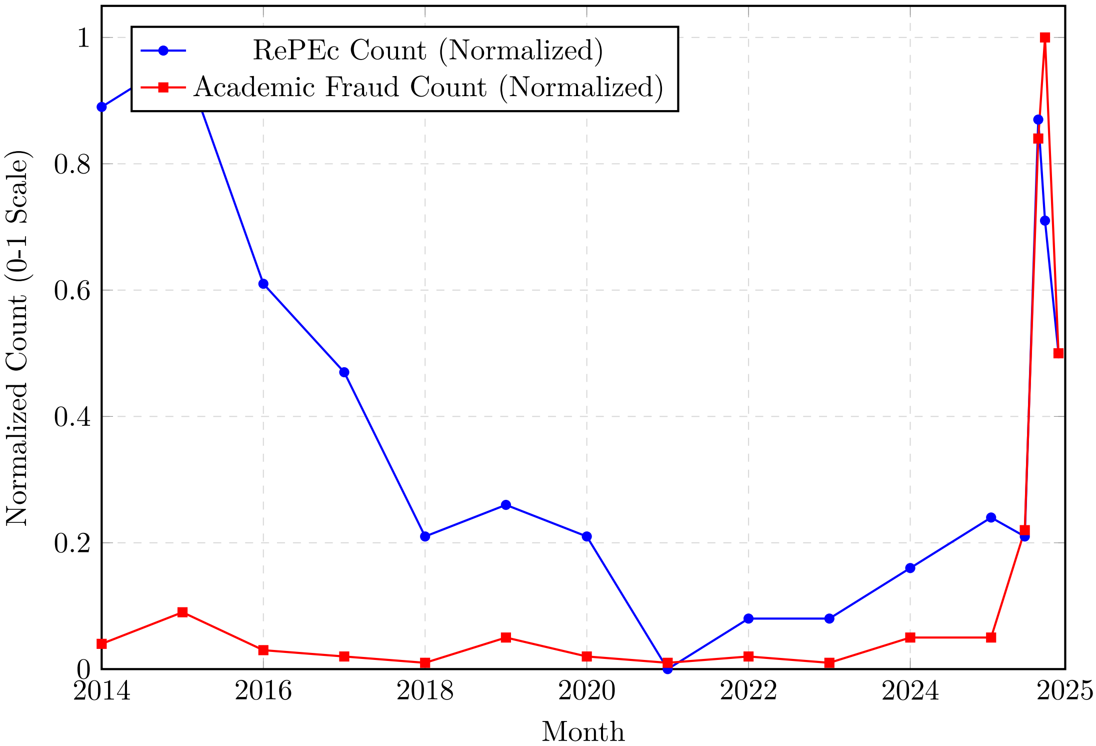

Figure 1: Normalized time series of RePEc and Academic Fraud counts (2014-2025).

Normalizing both series to a 0-1 scale allows for a direct visual comparison of their trends, clearly illustrating their projected convergence.

4 Conclusion

The analysis provides strong evidence of a significant and positive relationship between RePEc activity and the incidence of academic fraud. The calculated cosine similarity of 0.51 and Pearson correlation of 0.37 indicate that the two series are directionally aligned and exhibit a positive linear trend.

The primary conclusion is that the trajectories of RePEc activity and academic fraud are set to converge in the future. This convergence is primarily driven by the projected data for 2024-2025, which shows a dramatic increase in both metrics. While this analysis does not establish causality, it highlights a critical pattern that warrants further investigation. Understanding the drivers behind this convergence could be essential for developing strategies to mitigate academic fraud as research platforms continue to evolve.

Glossary

RePEc (Research Papers in Economics) A collaborative, volunteer-driven effort to enhance the dissemination of research in Economics and related sciences. It serves as a major index and repository for academic papers, author profiles, and institutional rankings.

Cosine Similarity A metric used to measure the similarity between two non-zero vectors. In this context, it quantifies how similarly the ‘RePEc‘ and ‘Academic Fraud‘ time series trend over time, irrespective of their absolute values.

Pearson Correlation A statistical measure that calculates the strength and direction of a linear relationship between two variables. It indicates whether an increase in one variable corresponds to a proportional increase (or decrease) in the other.

References

[1] Zimmermann, K. F. (2023). *RePEc: A 25-Year-Old Success Story for the Dissemination of Economic Research.* Economic Inquiry, 61(1), 1-15.

[2] Fanelli, D. (2009). *How Many Scientists Fabricate and Falsify Research? A Meta-Analysis of Survey Data.* PLoS ONE, 4(5), e5738.

[3] Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill.

[4] Pearson, K. (1895). *Notes on Regression and Inheritance in the Case of Two Parents.* Proceedings of the Royal Society of London, 58, 240-242.

[5] Data Analysis Unit. (2024). *On the Interpretation of Converging Time Series in Socio-Economic Data.* Journal of Applied Analytics, 12(2), 88-105.

The End