Quantifying the Topic Disparity of Scientific Articles

Munjung Kim
Department of Physics
Pohang University of Science and Technology
Pohang, Republic of Korea

Jisung Yoon
Department of Industrial and Management Engineering
Pohang University of Science and Technology
Pohang, Republic of Korea

Woo-Sung Jung
Department of Physics
Department of Industrial and Management Engineering
Graduate School of Artificial Intelligence
Pohang University of Science and Technology
Pohang, Republic of Korea
wsjung@postech.ac.kr

Hyunuk Kim
Department of Administrative Sciences
Metropolitan College, Boston University
Boston, MA, USA
uk@bu.edu

ABSTRACT

Citation count is a popular index for assessing scientific papers. However, it depends on not only the quality of a paper but also various factors, such as conventionality, journal, team size, career age, and gender. Here, we examine the extent to which the conventionality of a paper is related to its citation count by using our measure, topic disparity. The topic disparity is the cosine distance between a paper and its discipline on a neural embedding space. Using this measure, we show that the topic disparity is negatively associated with citation count, even after controlling journal impact, team size, and the career age and gender of the first and last authors. This result indicates that less conventional research tends to receive fewer citations than conventional research. The topic disparity can be used to complement citation count and to recommend papers at the periphery of a discipline because of their less conventional topics.

CCS CONCEPTS

• **Information systems** → **Document representation**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

Neural embedding techniques, BERT, Microsoft Academic Graph

ACM Reference Format:
Munjung Kim, Jisung Yoon, Woo-Sung Jung, and Hyunuk Kim. 2022. Quantifying the Topic Disparity of Scientific Articles. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487553.3524655

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524655

1 INTRODUCTION

Citations have been used to evaluate articles, individual researchers, and organizations [1–3]. The assessments are based on the perspective that high-quality research receives more citations than low-quality research [4]. However, citations also vary by team size, gender, journal, career age, and conventionality. Large teams are likely to receive more citations compared to small teams [5–7], and female-authored papers tend to be less cited than male-authored papers [8–10]. Articles published in prestigious journals receive considerably more citations than second-tier journals [11–14]. Author's seniority also tends to increase citation counts in several disciplines [15, 16]. In addition, papers with highly conventional pairs tend to receive more citations than papers with unconventional pairs [17]. Papers with new pairs have greater chances of being highly cited but also have higher variances of citations [18], suggesting that novel research would be both risky and impactful.

Among the factors above, we focus on conventionality and suggest an alternative approach to quantifying conventionality by leveraging a neural embedding method that represents scientific texts as vectors [19]. Our measure is called *topic disparity* and based on actual texts rather than journal pairs. The topic disparity is defined as the cosine distance between a paper and its discipline on a vector space. Hence, the smaller the topic disparity is, the more conventional a paper is. We show that the topic disparity is negatively correlated with citation count, even team size, journal impact, and the career age and gender of the first and last authors are considered.

2 DATA AND METHODS

2.1 Microsoft Academic Graph

We retrieve journal articles published in 2019 from Microsoft Academic Graph (MAG; accessed on November 2, 2021). MAG is discontinued on December 31, 2021 and migrated to OpenAlex afterwards. MAG provides hierarchical discipline information of each article [21], namely "Field of Study (FoS)" ranging from Level 0 (Highest) to Level 5 (Lowest). Level 0 and Level 1 codes are named as fields and disciplines hereafter. A paper can belong to multiple fields and disciplines. There are 19 fields: Art, Biology, Business,

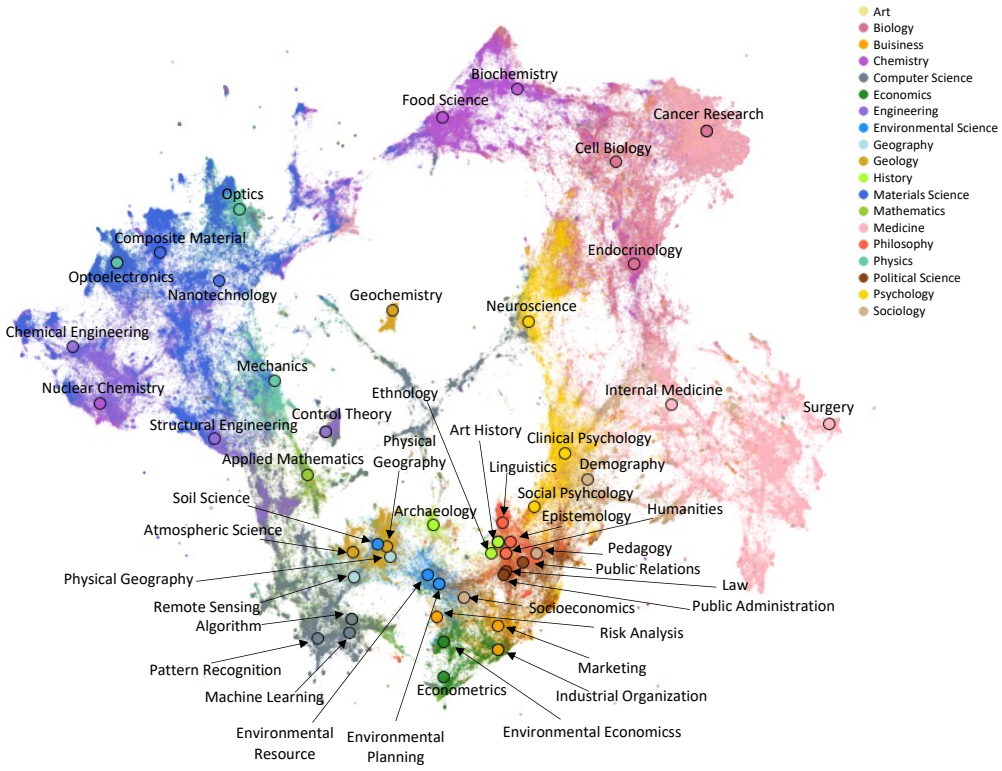

Figure 1: A UMAP [20] projection of 49 discipline vectors and the vectors of their 464,343 papers. Each small and large points correspond to a paper and a selected discipline, respectively.

Chemistry, Computer Science, Economics, Engineering, Environmental Science, Geography, Geology, History, Materials Science, Mathematics, Medicine, Philosophy, Physics, Political Science, Psychology, and Sociology.

2.2 Topic disparity

For each paper, we convert the concatenation of title and abstract into real-valued vectors by using SPECTER [19], a transformer-based machine learning model trained on scientific citations. We define the discipline vector V_j as the mean vector of papers that belong to discipline j as follows,

$$V_j = \frac{\sum_i v_{i,j}}{N_j}, \tag{1}$$

where N_j is the total number of papers in discipline j and $v_{i,j}$ is the embedding vector of paper i in discipline j. We assumed that V_j is the overall theme of discipline j.

The vector representations of papers and disciplines allow us to calculate D the topic disparity of a paper from its discipline. We define $D_{i,j}$ as the cosine distance between the vector of paper i and the vector of discipline j to which i belongs. We limit our analysis to the top three disciplines (Level 1 FoS) for each field (Level 0 FoS) with respect to the number of papers. As we have 19 fields, 57 unique disciplines are expected to be selected if a discipline belongs to only one field. However, eight disciplines – "Algorithm", "Art History", "Cancer Research", "Control Theory", "Environmental Planning", "Humanities", "Industrial Organization", and "Optoelectronics" – are one of the top three disciplines for two different fields. Hence, our analysis actually examines 49 unique disciplines.

2.3 Genders of the first and last authors

We use Genderize (https://genderize.io/) to infer the genders of the first and last authors of the papers in the selected 49 disciplines. We exclude papers of which authors are listed in alphabetical order (3% on average). For each given name, we assign either female or male if the probability returned from Genderize is higher than 0.7. If not, we use the Wiki-Gendersort algorithm [22] to fill missing genders as much as possible. Initials are removed from given names in order to reduce noises.

Papers are then categorized into FF, FM, MF, or MM, depending on the gender of the first and last authors. F and M stand for female and male, respectively. For instance, if a paper is written by a female first author and a female last author, this paper is classified as FF. Although the proportions of these categories vary by discipline, in total, FF and MF have smaller numbers of papers than FM and MM (14.2% FF, 25.0% FM, 15.2% MF, 45.6% MM).

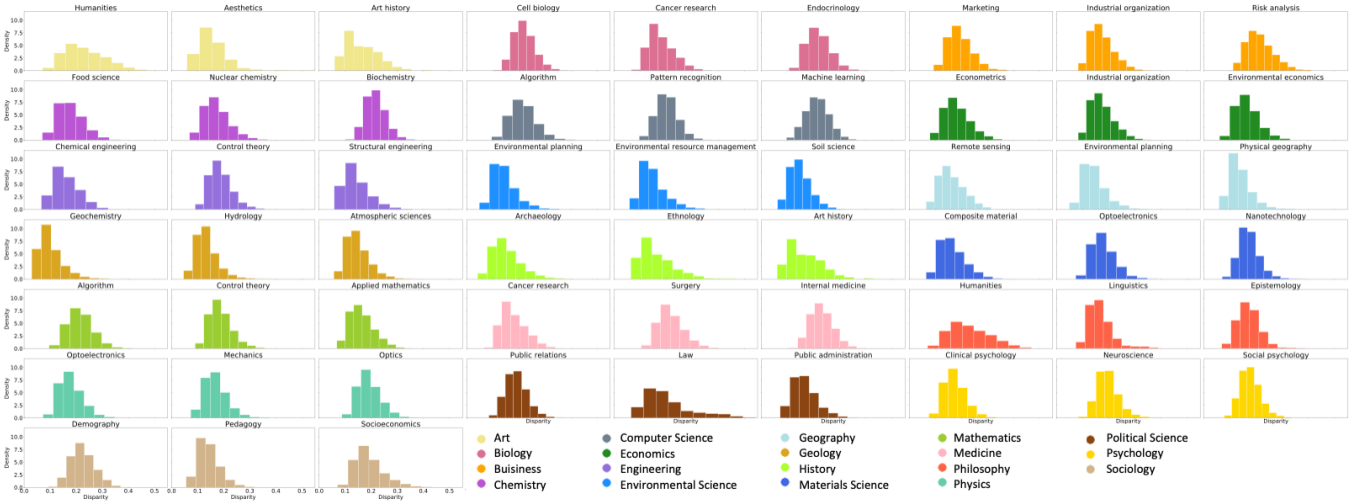

Figure 2: The topic disparity distributions of the top three disciplines for each field in terms of the number of papers. Note that field and discipline are the highest and the second highest "Field of Study" levels. Eight disciplines – "Algorithm", "Art History", "Cancer Research", "Control Theory", "Environmental Planning", "Humanities", "Industrial Organization", and "Optoelectronics" – which belong to two fields are shown twice in different colors.

2.4 Journal Impact

We use the average citation count as a proxy of journal impact. The average citation count is calculated from the papers published before 2019. For each discipline, we classify the journals into three groups: top, middle, and low impact journals.

2.5 Quantile regression

To examine the impact of topic disparity on citation count, we build quantile regression models for 0.50, 0.75, and 0.95 quantiles with the explanatory features in Table 1. The 0.25 quantile regression is not conducted as all citation counts are zero. We exclude papers of which the gender of the first or the last author is not inferred or the number of authors is larger than 10, in order to minimize the impact of missing values and outliers. The papers without journal information are also excluded. As a result, 464,343 papers from the 49 disciplines are used in the regressions.

Table 1: The explanatory features of the regression models.

Feature	Description
Discipline	The discipline which a paper belongs to.
Journal-High	1 if the journal impact is in the top 33.3% of the discipline. 0 otherwise.
Journal-Mid	1 if the journal impact is below the top 33.3% and over the bottom 33.3 % of the discipline. 0 otherwise.
Team Size	Normalized number of authors.
Career Age-First	Normalized number of articles written by the first author and published before 2019
Career Age-Last	Normalized number of articles written by the last author and published before 2019
Disparity	Normalized topic disparity calculated as Section 2.2
MF	1 if the gender of the first author is male and the gender of the last author is female. 0 otherwise.
FM	1 if the gender of the first author is female and the gender of the last author is male. 0 otherwise.
FF	1 if the genders of the first author and the last author are both female. 0 otherwise.

3 RESULTS

The paper and discipline vectors are projected onto a two dimensional space by the UMAP algorithm [20] (Figure 1). Large and small points correspond to the 49 disciplines and their 464,343 papers, respectively. Overall, the projection is consistent with existing maps of science [23–25].

For all disciplines, the topic disparity distribution has a peak in the middle of the value range and is right-skewed (Figure 2), implying most papers combine the overall theme of a discipline with some non-conventional components, while there is a small portion of papers pursuing genuinely novel topics.

From the regressions, we find that the topic disparity is negatively correlated with citation count ($p < 0.001$; Table 2). This result suggests that papers of high disparity values tend to receive less citations than papers of low disparity values. On the other hand, team size and the career age of both first and last author are positively associated with citation count ($p < 0.001$; Table 2), indicating large teams and senior researchers tend to receive more citations. The coefficient of the career age of the last author is higher than the coefficient of the first author, showing that the last author's seniority has a greater impact on citation count. In all regressions, the relationship between citation count and the journal impact is

Table 2: Quantile Regressions

(a) 0.50 quantile

Feature	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Journal-High	3.89*** (0.035)	3.67*** (0.043)	3.42***(0.039)	3.89***(0.036)	3.42***(0.039)	3.31***(0.043)
Journal-Mid	0.341*** (0.037)	0.421*** (0.046)	0.235*** (0.041)	0.371***(0.039)	0.237***(0.042)	0.293***(0.045)
Normalized Disparity	−0.225*** (0.007)	−0.295*** (0.008)	−0.259*** (0.007)	−0.240*** (0.007)	-0.259***(0.008)	-0.289***(0.008)
Normalized Team Size		0.555*** (0.008)				0.479***(0.008)
Career Age-First			0.360***(0.007)		0.357***(0.007)	0.336***(0.007)
Career Age-Last			0.742***(0.007)		0.745***(0.007)	0.690***(0.007)
FF				-0.161***(0.019)	-0.031(0.020)	-0.044*(0.022)
FM				-0.074***(0.015)	-0.044**(0.016)	-0.075***(0.018)
MF				-0.070***(0.018)	0.040*(0.019)	0.039(0.021)
Constant	−0.430** (0.138)	0.078(0.169)	0.152(0.154)	−0.393** (0.144)	0.155(0.154)	0.601***(0.169)
Discipline	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.082	0.085	0.087	0.082	0.087	0.089

(b) 0.75 quantile

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Journal-High	7.31*** (0.079)	6.73***(0.078)	6.20***(0.073)	7.31***(0.081)	6.20***(0.073)	5.84***(0.076)
Journal-Mid	1.15*** (0.086)	0.888*** (0.085)	0.749***(0.079)	1.14***(0.088)	0.750***(0.079)	0.595***(0.083)
Normalized Disparity	-0.442***(0.016)	−0.446*** (0.016)	-0.382***(0.014)	-0.442***(0.016)	-0.386***(0.014)	-0.404***(0.015)
Normalized Team Size		0.863*** (0.015)				0.675***(0.014)
Career Age-First			1.08***(0.010)		1.06***(0.010)	1.01***(0.011)
Career Age-Last			1.61***(0.012)		1.61***(0.012)	1.52***(0.013)
FF				-0.469***(0.043)	-0.110**(0.039)	-0.116**(0.040)
FM				-0.264***(0.034)	-0.134***(0.031)	-0.157***(0.032)
MF				-0.240***(0.040)	0.029(0.037)	-0.004(0.038)
Constant	−0.383(0.321)	0.693* (0.317)	1.01*** (0.295)	−0.112(0.328)	1.05***(0.296)	1.79***(0.308)
Discipline	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.097	0.100	0.107	0.098	0.107	0.109

(c) 0.95 quantile

Feature	Model 1	Model 2	Model 3	Model 4	Model 5	Model6
Journal-High	16.5*** (0.322)	15.4*** (0.298)	12.9***(0.253)	16.3***(0.310)	12.8***(0.256)	12.2***(0.246)
Journal-Mid	2.71*** (0.355)	2.10***(0.329)	1.67*** (0.279)	2.45***(0.342)	1.62***(0.283)	1.31***(0.272)
Normalized Disparity	−0.854*** (0.068)	-0.822***(0.063)	-0.620***(0.054)	-0.849***(0.066)	-0.611***(0.054)	-0.620***(0.052)
Normalized Team Size		1.75*** (0.059)				1.11***(0.048)
Career Age-First			4.92***(0.030)		4.88***(0.031)	4.81***(0.030)
Career Age-Last			5.14***(0.039)		5.15***(0.039)	5.02***(0.038)
FF				-1.86***(0.165)	-0.402**(0.137)	-0.362**(0.132)
FM				-1.07***(0.133)	-0.296**(0.110)	-0.361***(0.106)
MF				-1.08***(0.157)	-0.011(0.130)	0.005(0.125)
Constant	1.17(1.325)	3.55** (1.23)	5.36*** (1.05)	2.54* (1.28)	5.80***(1.06)	7.00***(1.02)
Discipline	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.127	0.130	0.152	0.128	0.152	0.153

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

significantly positive ($p < 0.001$; Table 2). Hence, papers published in high-ranked journals tend to receive more citations. The coefficient of FF is negative and significant, except Model 5 of the 0.50 quantile regression ($p < 0.05$; Table 2). The coefficient of FM is also negative and significant ($p < 0.01$; Table 2).

4 CONCLUSION

Here, we present a method for measuring the topic disparity of a paper by calculating the cosine distance between the paper and its discipline in a vector space. By applying this method, we examined the relationship between the topic disparity and citation count while considering journal impact, team size, career age, and gender.

Our result shows that there is a negative relationship between citation count and the topic disparity. Therefore, research papers focusing on topics far from the main research theme of a given discipline tend to receive fewer citations than papers dealing with conventional topics. A potential explanation for this observation is that areas studying less conventional topics are relatively small, so authors may receive fewer citations [26, 27].

Our approach can be extended to investigate the relationships between the topic disparity and other attributes in science such as the accessibility of research papers and the demographics of authors. Moreover, with domain knowledge, the topic disparity would provide detailed insights into a discipline and its characteristics. We expect that the topic disparity identifies marginalized papers and researchers developing novel perspectives.

REFERENCES

[1] Henk F Moed. *Citation analysis in research evaluation*, volume 9. Springer Science & Business Media, 2006.

[2] Jonathan R Cole. A short history of the use of citations as a measure of the impact of scientific and scholarly work. *The Web of Knowledge: A Festschri in Honor of Eugene Garfield*, pages 281–300, 2000.

[3] Alan Fersht. The most influential journals: Impact factor and eigenfactor, 2009.

[4] Lutz Bornmann and Hans-Dieter Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64:45–80, 2008.

[5] Vincent Larivière, Yves Gingras, Cassidy R Sugimoto, and Andrew Tsou. Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, 66(7):1323–1332, 2015.

[6] Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[7] Lingfei Wu, Dashun Wang, and James A Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, 2019.

[8] Jordan D Dworkin, Kristin A Linn, Erin G Teich, Perry Zurn, Russell T Shinohara, and Danielle S Bassett. The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, 23(8):918–926, 2020.

[9] Erin G Teich, Jason Z Kim, Christopher W Lynn, Samantha C Simon, Andrei A Klishin, Karol P Szymula, Pragya Srivastava, Lee C Bassett, Perry Zurn, Jordan D Dworkin, et al. Citation inequity and gendered citation practices in contemporary physics. *arXiv preprint arXiv:2112.09047*, 2021.

[10] Jacqueline M Fulvio, Ileri Akinnola, and Bradley R Postle. Gender (im) balance in citation practices in cognitive neuroscience. *Journal of Cognitive Neuroscience*, 33(1):3–7, 2021.

[11] Hendrik Van Dalen and Kène Henkens. What makes a scientific article influential? the case of demographers. *Scientometrics*, 50(3):455–482, 2001.

[12] Fereshteh Didegah and Mike Thelwall. Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, 64(5):1055–1064, 2013.

[13] Roosa Leimu and Julia Koricheva. What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, 20(1):28–32, 2005.

[14] Iman Tahamtan, Askar Safipour Afshar, and Khadijeh Ahamdzadeh. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, 107(3):1195–1225, 2016.

[15] Jacob B Slyder, Beth R Stein, Brent S Sams, David M Walker, B Jacob Beale, Jeffrey J Feldhaus, and Carolyn A Copenheaver. Citation pattern and lifespan: a comparison of discipline, institution, and individual. *Scientometrics*, 89(3):955–966, 2011.

[16] Ying Ding and Blaise Cronin. Popular and/or prestigious? measures of scholarly esteem. *Information Processing & Management*, 47(1):80–96, 2011.

[17] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[18] Jian Wang, Reinhilde Veugelers, and Paula Stephan. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436, 2017.

[19] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*, 2020.

[20] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[21] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246, 2015.

[22] Nicolas Bérubé, Gita Ghiasi, Maxime Sainte-Marie, and Vincent Larivière. Wikigendersort: Automatic gender detection using first names in Wikipedia, 2020.

[23] Ismael Rafols and Martin Meyer. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2):263–287, 2010.

[24] Katy Börner, Richard Klavans, Michael Patek, Angela M Zoss, Joseph R Biberstine, Robert P Light, Vincent Larivière, and Kevin W Boyack. Design and update of a classification system: The UCSD map of science. *PLoS One*, 7(7):e39464, 2012.

[25] Jaimie Murdock, Colin Allen, Katy Börner, Robert Light, Simon McAlister, Andrew Ravenscroft, Robert Rose, Doori Rose, Jun Otsuka, David Bourget, et al. Multilevel computational methods for interdisciplinary research in the HathiTrust Digital Library. *PLoS One*, 12(9):e0184188, 2017.

[26] Henk F Moed, WJM Burger, JG Frankfort, and Anthony FJ Van Raan. The use of bibliometric data for the measurement of university research performance. *Research Policy*, 14(3):131–149, 1985.

[27] Jean King. A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, 13(5):261–276, 1987.