The preeminence of ethnic diversity in scientific collaboration

Bedoor K. AlShebli1, Talal Rahwan1,2 & Wei Lee Woon1,3

Inspired by the social and economic benefits of diversity, we analyze over 9 million papers and 6 million scientists to study the relationship between research impact and five classes of diversity: ethnicity, discipline, gender, affiliation, and academic age. Using randomized baseline models, we establish the presence of homophily in ethnicity, gender and affiliation. We then study the effect of diversity on scientific impact, as reflected in citations. Remarkably, of the classes considered, ethnic diversity had the strongest correlation with scientific impact. To further isolate the effects of ethnic diversity, we used randomized baseline models and again found a clear link between diversity and impact. To further support these findings, we use coarsened exact matching to compare the scientific impact of ethnically diverse papers and scientists with closely-matched control groups. Here, we find that ethnic diversity resulted in an impact gain of 10.63% for papers, and 47.67% for scientists.
Diversity is highly valued in modern societies1–6. Social cohesion, tolerance, and integration are linked to tangible benefits—such as improved economic vibrancy5,8 and innovativeness5,9–11. Far from being an abstract ideal, this conviction has guided many governmental and hiring policies and can have broad and long-lasting effects on society12,13. However, diversity is a complex issue, as groups can be diverse in terms of various attributes, such as ethnicity, gender, age, and socio-economic background. It is also unclear if all forms of diversity are beneficial. For instance, ethnic density has been associated with positive outcomes in terms of health14–15, while ethnic polarization has a negative effect on economic development16. Furthermore, diversity can be a divisive topic that is clouded by emotion, partisan loyalties, and political correctness, all of which can hinder impartial discussions17. The factors above strongly motivate an objective study on the value of diversity, and on whether more diverse groups achieve greater success.

One domain in which this question can be effectively addressed is academia18,19. The structure of academic collaboration is observable via co-authorships, which frequently involve scientists from different locations, disciplines and backgrounds20,21. Furthermore, academic output has an objective, widely accepted measure—citation count22,23. This amenability to analysis has already attracted attempts at identifying the factors which underlie success in academia, an enterprise known as the Science of Science24. Although many such factors have been studied, including gender25, academic age26, team size27, inter-disciplinarity28, ethnicity29, and affiliation30,31, the study of these factors is extremely complex and many questions remain unanswered.

Our study seeks to address this shortcoming from a number of hitherto unexplored perspectives. Firstly, we compare homophily in scientific collaborations from the perspectives of age, gender, affiliation, and ethnicity. We find clear signs of homophily in the cases of ethnicity, gender, and affiliation. However, in only one case, ethnicity, was homophily found to be increasing steadily over time. Secondly, we examine the relationship between various classes of diversity and research impact at the level of scientific fields. Remarkably, we found that ethnic diversity is most strongly associated with scientific impact. Thirdly, we compare the benefits of diversity on groups vs. individuals, and find that the former outweighs the latter. Finally, we study the evolution and effect of diversity over time, team size, and number of collaborators, and verify that the above findings persist across all of these dimensions. The results of these multiple angles of analysis are combined to form a far richer picture of diversity than has been possible in the past.

Results

Exploring homophily. A natural starting point for our study of diversity is to establish the extent to which homophilii32 exists in academia—i.e., whether scientists tend to collaborate more frequently with similar others—which would lead to an overall lack of diversity in scientific collaborations. We use the Microsoft Academic Graph dataset (available at: https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/), and analyze 1,045,401 multi-authored papers (see Supplementary Figure 1 for the distribution of papers by year), written by 1,529,279 scientists, spanning eight main fields and 24 sub-fields of science. We analyzed diversity in terms of these five attributes: ethnicity (eth), discipline (dsp), gender (gen), affiliation (aff), and academic age (age); see Supplementary Note 1. Here, the abbreviations in parentheses are used in subsequent mathematical expressions to indicate the associated attribute. These attributes reflect many technical and social factors that influence teamwork and collaboration. Affiliation indicates the geographic location, and may even reflect the way collaborative work is carried out—from the style and culture of collaboration to its mundane details, such as the medium used to collaborate, e.g., face-to-face interactions vs. telecommunication or email. Academic age is not only indicative of the amount of experience that a scientist has, but is also typically associated with actual age. Discipline may reflect a scientist’s substantive knowledge and his/her acquired skills through training, as well as the culture in which collaborative work is carried out. Finally, ethnicity and gender may play a role in shaping scientists’ social identities, knowledge, and biases. To quantify diversity in terms of any of the aforementioned attributes, we use the Gini Impurity33, resulting in the following group diversity indices, $d^\text{eth}_{\text{gen}}$, $d^\text{age}_{\text{gen}}$, $d^\text{eth}_{\text{aff}}$, $d^\text{age}_{\text{aff}}$ and $d^\text{aff}_{\text{dp}}$ (an alternative diversity measure was also considered; see Supplementary Note 2 and Supplementary Figure 2).

To explore homophily, we generate different randomized baseline models whereby a particular attribute—be it ethnicity, gender, affiliation, or academic age—is shuffled. For example, in the case of ethnicity, this process is akin to creating a universe in which ethnicity is disregarded in the selection of co-authors, while retaining other criteria. To preserve the conditional distributions of the ethnicities, the shuffling process is constrained to only occur between authors of papers that have the same subfield, publication year, and number of authors; for full details, see Supplementary Note 3. This way, for every paper $p$ in the real dataset, there exists a matching paper $p'$ in the randomized dataset that may differ from $p$ in terms of ethnic diversity, but is identical to $p$ in terms of gender, affiliation, academic age, citations, publication year, and number of authors per paper. Importantly, while such a baseline model may produce homogeneous groups, the emergence of such groups is purely the result of random chance rather than homophily. As such, by comparing the real dataset with this baseline model, we can determine whether homophily exists, and if so, quantify the degree to which it is spread across academia. Figure 1a compares our real dataset with the randomized baseline model in terms of the cumulative distributions of $d^x$: $x \in \{\text{eth, age, gen, aff}\}$. As can be seen, for $x \in \{\text{eth, gen, aff}\}$, groups with low $d^x$ are more common in reality than would be expected by random chance, highlighting the fact that homophily does indeed exist in academia in terms of ethnicity, gender, and affiliation. However, for $x = \text{age}$, the opposite was observed (see Supplementary Figures 3–6 for subfield-specific distributions). These observations persist, regardless of the publication year (Fig. 1b), and the number of authors per paper (Fig. 1c). The temporal trends observed in Fig. 1b are particularly intriguing. For $d^\text{eth}$, while the population of scientists is becoming more ethnically diverse (see the steady increase in the red line), this trend is not reflected in the actual coauthor groupings, implying that ethnic homophily is steadily increasing. For $d^\text{gen}$, the actual level of diversity is greater than would be expected by random chance; this pattern is regularly observed in academia, e.g., consider the many publications resulting from advisor–advisee collaborations. For $d^\text{aff}$, although gender homophily continues to exist, it steadily decreases over time, suggesting that women are playing an ever greater role in scientific endeavors. Finally, for $d^\text{age}$, there is a marked decrease in affiliation homophily around the 1990s; this is consistent with the jump in multi-university collaborations in the 1990s due to the widespread of the Internet and other technologies that facilitate collaboration across geographically distant scientists30.

The link between diversity and scientific impact. Having explored homophily in academia, we now study the effects of
homophily (and diversity) on research impact, measured by the number of citations received within 5 years of publication, denoted by \( d^G_x \) (see Supplementary Note 4 and Supplementary Figure 7). Using the same dataset and notation described earlier, we study the relationship between a subfield’s diversity and its academic impact. Here, we distinguish between two notions of diversity, especially given that ethnicity is arguably unrelated to technical competence.

For each subfield, Fig. 3a depicts the mean group diversity indices, \( \langle d^G_x \rangle : x \in \{ \text{eth, age, gen, dsp, aff} \} \), against the mean 5-year citation count, \( \langle d^G_5 \rangle \), taken over papers in that subfield (notation summary and formal definitions are in Supplementary Table 1 and Supplementary Note 2, respectively). Remarkably, we find that a subfield’s ethnic diversity is the most strongly correlated with impact \( (r = 0.77) \); the positive correlation persists even when the subfields are studied in isolation (Supplementary Figures 8 and Supplementary Table 2), regardless of the number of authors per paper (Supplementary Figure 9). These findings are further supported by the regression analysis in Table 1. While these findings do not imply causation, it is still suggestive that one can largely predict scientific impact based solely on average ethnic diversity, especially given that ethnicity is arguably unrelated to technical competence.

Having studied group diversity, we now move our attention to individual diversity. Here, we analyze scientists with at least 10 collaborators over 9,472,439 papers (see Supplementary Table 3 for a set of experiments whereby \( d^G_x : x \in \{ \text{eth, age, gen, dsp, aff} \} \) differences in color represent differences in terms of \( x \). The group diversity index \( d^G_x \) of Paper A is higher than that of Paper B. The individual diversity index of Scientist C is higher than that of Scientist D.
Summary of all filters applied on the dataset. For each subfield, Fig. 3b depicts the mean individual diversity indices, \( \langle d_{\text{eth}} \rangle : x \in \{ \text{eth, age, gen, dsp, aff} \} \), against the mean 5-year citation count, \( \langle c_i \rangle \), taken over scientists in that subfield. As can be seen, a subfield’s ethnic diversity is again the most strongly correlated with impact \( (r = 0.55, p = 0.006) \), even when the subfields are studied in isolation (Supplementary Figure 10 and Supplementary Table 4).

The above results highlight a potential dysfunction. While homophily was observed for ethnicity, affiliation and gender, the only attribute for which it was found to be increasing over time was ethnicity, which seems strange given the apparent preeminence of ethnic diversity. Motivated by this observation, we further explore the relationship between ethnic diversity and scientific impact in the randomized universe used earlier in Fig. 1. Recall that, in such a universe, ethnicity is excluded as a criterion for selecting co-authors while the other factors are preserved. Hence, it stands to reason that any differences in impact between the randomized and real datasets can be attributed to ethnic diversity. To examine these differences, we partitioned the papers into two categories, labeled as diverse \( \langle d_{\text{eth}} \rangle \geq \tilde{d}_{\text{eth}} \rangle \) and non-diverse \( \langle d_{\text{eth}} \rangle \geq \tilde{d}_{\text{eth}} \rangle \), where the tilde denotes the median. The scientists were similarly partitioned into diverse \( \langle d_{\text{eth}} \rangle > \tilde{d}_{\text{eth}} \rangle \) and non-diverse \( \langle d_{\text{eth}} \rangle \leq \tilde{d}_{\text{eth}} \rangle \). We find that the diverse consistently outperforms the non-diverse, regardless of the year of publication (Fig. 4e), the number of authors per paper (Fig. 4g), and the number of collaborators per scientist (Fig. 4i). We replicated these plots using the randomized, instead of the real, dataset (Fig. 4f, h and j). As can be seen, the performance gap between the diverse and non-diverse is almost entirely disappears in the randomized dataset, suggesting that the observed impact gains in the real dataset could indeed be attributed to ethnic diversity. Note that, in the real dataset, a large proportion of papers have \( d_{\text{eth}} = 0 \) (see Fig. 4a), and a large proportion of scientists have \( d_{\text{eth}} = 0 \) (see Fig. 4c). As such, the observed performance gap between the diverse and the non-diverse could be predominantly due to these papers and scientists being less impactful than their counterparts.
### Table 1 Regression analyses of diversity classes on academic impact

|                        | Engineering & computer science | Health & medical sciences | Business, economics & management | Humanities, literature & arts | Physics & mathematics | Social sciences | Chemical & material sciences | Life sciences & earth sciences |
|------------------------|-------------------------------|---------------------------|---------------------------------|--------------------------------|-----------------------|----------------|-----------------------------|---------------------------------|
| **Author**             |                               |                           |                                 |                                |                       |                |                             |                                  |
| N                      | 139,705                       | 288,827                   | 38,938                          | 47,141                         | 146,574               | 158,479        | 88,300                      | 137,437                          |
| **Group discipline**   |                               |                           |                                 |                                |                       |                |                             |                                  |
| N                      |                               |                           |                                 |                                |                       |                |                             |                                  |
| **Group gender**       |                               |                           |                                 |                                |                       |                |                             |                                  |
| N                      |                               |                           |                                 |                                |                       |                |                             |                                  |
| **Group affiliation**  |                               |                           |                                 |                                |                       |                |                             |                                  |
| N                      |                               |                           |                                 |                                |                       |                |                             |                                  |
| **Group discipline**   |                               |                           |                                 |                                |                       |                |                             |                                  |
| N                      |                               |                           |                                 |                                |                       |                |                             |                                  |

The regression tables below present the effect of each of the five group diversity indices, $d_{i}$ ($i$ = eth, age, gen, off, dep), on the paper’s impact, $c_{i}^{2}$. Along with each class of diversity, the following predictor variables were used: university ranking, author’s prior impact, year of publication, and number of authors. Here, university rankings are based on the 2017 Academic Ranking of World Universities, also known as the Shanghai Ranking, whereas an author’s prior impact is measured as the annual number of citations that he/she accumulated prior to the year in which the paper was published. The columns correspond to papers from different fields. Of the five classes of diversity studied, ethnic diversity (A) was the only one for which all coefficients in the first row ($d_{i}$) are positive and significant. Standard errors in parentheses

$*$ $p < 0.05$; $** p < 0.01$; and $*** p < 0.001$
whose $d_{G_{eth}}^{G} > 0$ and $d_{G_{eth}}^{l} > 0$, respectively. To determine whether this is the case, we replicated the analysis of papers but after excluding those with $d_{G_{eth}}^{G} = 0$, and likewise replicated the analysis of scientists but after excluding those with $d_{G_{eth}}^{l} = 0$; see Supplementary Figure 11. As can be seen, even after this exclusion, the diverse mostly outperform the non-diverse, regardless of publication year, number of authors per paper, and number of collaborators per scientist.
Inferential causality. To provide further evidence of the link between ethnic diversity and scientific impact, we use coarsened exact matching\(^\text{34}\), a technique typically used to infer causality in observational studies\(^\text{35}\). Specifically, it matches the control and treatment populations with respect to the confounding factors identified, thereby eliminating the effect of these factors on the phenomena under investigation. In our case, when studying group ethnic diversity, the treatment set consists of papers for which \(d_{\text{eth}} = P_{100-\delta}(d_{\text{eth}}^{\text{G}})\), and the control set of papers for which \(d_{\text{eth}}^{\text{C}} \leq P_{\delta}(d_{\text{eth}}^{\text{G}})\), where \(P_{\delta}(d_{\text{eth}}^{\text{G}})\) denotes the \(\delta\)th percentile of \(d_{\text{eth}}^{\text{G}}\).

This process is repeated using \(i = 10, 20, 30, 40, 50\), corresponding to progressively larger gaps in ethnic diversity between the two populations. Thus, if ethnic diversity is indeed associated with increased scientific impact, we would expect to find a significant difference in impact between the two populations, and expect this difference to increase in tandem with the aforementioned gap in diversity. The confounding factors identified were the year of publication, number of authors, field of study, authors’ impact prior to publication, and university ranking. The same process was carried out for individual ethnic diversity, for which coarsened exact matching is particularly useful precisely because it allows us to establish causality despite such effects.

### Table 2: Coarsened exact matching of group ethnic diversity

| \(T\) | \(|T|\) | \(|I|\) | \(|C|\) | \(L_\text{c}\) | \(\delta\) | \(\text{CI}_{0.95}\) | \(p\) |
|---|---|---|---|---|---|---|---|
| \(d_{\text{eth}}^G > P_{90}(d_{\text{eth}}^{\text{G}})\) | 17,802 | 45,710 | 13,530 | 0.39 | 10.63 | [8.10, 12.38] | 0.003 |
| \(d_{\text{eth}}^C \leq P_{10}(d_{\text{eth}}^{\text{G}})\) | 24,827 | 45,710 | 18,965 | 0.38 | 10.22 | [8.12, 12.02] | 0.0009 |
| \(d_{\text{eth}}^G > P_{80}(d_{\text{eth}}^{\text{G}})\) | 56,662 | 58,889 | 51,782 | 0.27 | 4.93 | [3.74, 5.97] | 0.008 |
| \(d_{\text{eth}}^C \leq P_{20}(d_{\text{eth}}^{\text{G}})\) | 63,129 | 78,340 | 57,279 | 0.29 | 5.14 | [4.12, 6.17] | 0.003 |
| \(d_{\text{eth}}^G > P_{70}(d_{\text{eth}}^{\text{G}})\) | 63,129 | 127,629 | 58,292 | 0.27 | 3.37 | [2.45, 4.25] | 0.018 |

\(T\) and \(C\) are the treatment and control populations respectively; \(T\) and \(C\) are the populations of matched treatment and matched control papers, respectively; \(L_\text{c}\) is the multivariate imbalance statistic\(^\text{34}\); \(\beta\) is the relative impact gain of \(\beta\) over \(C\); i.e., \(\beta = 100\times((\langle\text{C}\rangle - \langle\text{D}\rangle) - \langle\text{C}\rangle)\). A test shows that \(\beta\) is statistically significant; see the resulting \(p\)-values. Since the academic impact \(\langle\text{C}\rangle\) is sensitive to extremal values, we bootstrap a 95% confidence interval \(\text{CI}_{0.95}\). Here, university ranking corresponds to the average rank of all universities in the paper, as opposed to the highest ranked university in the paper, which is the case in Supplementary Table 5. For more details, see Supplementary Note 6.

### Table 3: Coarsened exact matching of individual ethnic diversity

| \(T\) | \(|T|\) | \(|I|\) | \(|C|\) | \(L_\text{c}\) | \(\delta\) | \(\text{CI}_{0.95}\) | \(p\) |
|---|---|---|---|---|---|---|---|
| \(d_{\text{eth}}^G > P_{90}(d_{\text{eth}}^{\text{G}})\) | 113,883 | 68,563 | 16,512 | 0.47 | 47.67 | [44.49, 49.92] | 2.04e−39 |
| \(d_{\text{eth}}^C \leq P_{10}(d_{\text{eth}}^{\text{G}})\) | 139,015 | 136,837 | 65,412 | 0.35 | 43.54 | [42.61, 45.05] | 1.50e−156 |
| \(d_{\text{eth}}^G > P_{80}(d_{\text{eth}}^{\text{G}})\) | 223,747 | 205,686 | 128,001 | 0.32 | 28.75 | [28.10, 29.46] | 1.65e−211 |
| \(d_{\text{eth}}^C \leq P_{20}(d_{\text{eth}}^{\text{G}})\) | 280,514 | 274,209 | 184,749 | 0.29 | 23.86 | [22.86, 23.98] | 5.96e−218 |
| \(d_{\text{eth}}^G > P_{70}(d_{\text{eth}}^{\text{G}})\) | 356,564 | 329,066 | 242,123 | 0.28 | 15.77 | [15.21, 15.95] | 3.23e−158 |

The notation is as per Table 2.
Fig. 5 The interplay between group and individual ethnic diversity. The top part of the figure illustrates an example of 4 papers. The authors of paper A have different ethnicities, but each has ethnically homogeneous collaborators. Then, one could argue that paper A has high $d_{	ext{Geth}}^g$ but low $d_{	ext{Ieth}}^i$. Similarly, paper B has low $d_{	ext{Geth}}^g$ and low $d_{	ext{Ieth}}^i/paper$, paper C has low $d_{	ext{Geth}}^g$ and high $d_{	ext{Ieth}}^i/paper$, and paper D has high $d_{	ext{Geth}}^g$ and high $d_{	ext{Ieth}}^i/paper$. The matrix at the bottom-right corner represents the mean citation counts, $\langle c^G \rangle/paper$, of papers falling in different ranges of $d_{	ext{Geth}}^g$ and $d_{	ext{Ieth}}^i/paper$. 

**Paper A:**
- High $d_{	ext{Geth}}^g$
- Low $d_{	ext{Ieth}}^i/paper$

**Paper B:**
- Low $d_{	ext{Geth}}^g$
- Low $d_{	ext{Ieth}}^i/paper$

**Paper C:**
- Low $d_{	ext{Geth}}^g$
- High $d_{	ext{Ieth}}^i/paper$

**Paper D:**
- High $d_{	ext{Geth}}^g$
- High $d_{	ext{Ieth}}^i/paper$
that $\sum_{i=1}^{4} A_{1,i} < \sum_{i=1}^{4} A_{1,i}$ and $\sum_{i=1}^{4} A_{1,i} > \sum_{i=1}^{4} A_{1,i}$. Hence, while it appears that both group and individual diversities can be valuable, the former seems to have a greater effect on scientific impact. In other words, having co-authors who are inclined to collaborate across ethnic lines (i.e., co-authors whose individual ethnic diversity is high) appears to be not as important as the mere presence of co-authors of different ethnicities (i.e., co-authors whose group ethnic diversity is high).

**Discussion**

To summarize, this study is the first to cover five different classes of diversity, which allowed us to illuminate many interesting connections between diversity and scientific collaboration. It was also important to establish the occurrence of homophily, and this was achieved via a set of randomized baseline models. These were used to compare observed collaborations with simulated data where the attribute of interest was randomized while controlling for the relevant confounding variables. These comparisons revealed clear and consistent patterns of homophily in the cases of ethnicity, gender, and affiliation, and also revealed that ethnicity was the only attribute for which homophily is increasing over time. In the case of academic age, inverse homophily was found, i.e., scientists seem to prefer collaborating with individuals from different age groups, a possible reflection of the widely held practice of research students being mentored by, and collaborating with, more senior academics.

Armed with these results, we shifted our focus to the effect of homophily (and diversity) on scientific impact. This analysis was conducted using a number of different analytical tools, including regression analysis, randomized baseline models, and coarsened exact matching. Broadly, we found that diversity was positively correlated with impact, though the statistical significance of the observed effect varied significantly depending on the class of diversity and field of study. Overall, discipline and affiliation diversity were the least correlated with impact, a surprising finding given the apparent importance of these attributes. Conversely, ethnic diversity had the strongest correlation, which is especially surprising since ethnicity is not as related to technical competence as the other classes mentioned.

These findings have significant implications. For one, recruiters should always strive to encourage and promote ethnic diversity, be it by recruiting candidates who complement the ethnic composition of existing members, or by recruiting candidates with proven track records in collaborating with people of diverse ethnic backgrounds. Another implication is that, while collaborators with different skill sets are often required to perform complex tasks, multidisciplinarity should not be an end in itself; bringing together individuals of different ethnicities—with the attendant differences in culture and social perspectives—could ultimately produce a large payoff in terms of performance and impact. To put it differently, intangible factors, such as team cohesion and a sense of esprit de corps should be considered in addition to technical alignment.

The underlying message is an inclusive and uplifting one. In an era of increasing polarization and identity politics, our findings may positively contribute to the societal conversation and reinforce the conviction that good things happen when people of different backgrounds, cultures, and ethnicities come together to work towards shared goals and the common good.

**Data availability**

The details of all data and methods used are given in Supplementary Note 1.
Acknowledgements
We thank Steven Skiena and his team for providing access to their Name Ethnicity Classifier tool36,37. We also thank Kinga Makovi for suggesting the use of coarsened exact matching for causal inference.

Author contributions
B.K.A., T.R., and W.L.W. conceived and designed the experiments. B.K.A. and W.L.W. performed the coding of the experiments. B.K.A., T.R., and W.L.W. wrote the manuscript. B.K.A. and T.R. produced the figures and tables.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-07634-8.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018