

The emergence of heterogeneous scaling in research institutions

Keith A. Burghardt ¹✉, Zihao He ¹, Allon G. Percus^{1,2} & Kristina Lerman¹

Research institutions provide the infrastructure for scientific discovery, yet their role in the production of knowledge is not well characterized. To address this gap, we analyze interactions of researchers within and between institutions from millions of scientific papers. Our analysis reveals that collaborations densify as each institution grows, but at different rates (heterogeneous densification). We also find that the number of institutions scales with the number of researchers as a power law (Heaps' law) and institution sizes approximate Zipf's law. These patterns can be reproduced by a simple model in which researchers are preferentially hired by large institutions, while new institutions complementarily generate more new institutions. Finally, new researchers form triadic closures with collaborators. This model reveals an economy of scale in research: larger institutions grow faster and amplify collaborations. Our work deepens the understanding of emergent behavior in research institutions and their role in facilitating collaborations.

¹Information Sciences Institute, University of Southern California, Marina del Rey, USA. ²Institute of Mathematical Sciences, Claremont Graduate University, Claremont, USA. ✉email: keithab@isi.edu

Scientific innovation and training require efficient and robust infrastructure. This infrastructure is provided by research institutions, a category that includes universities, government labs, industrial labs, and national academies^{1–5}. Despite the long tradition of bibliometric and science of science research⁶, the focus has only recently shifted from individual scientists^{7,8} and teams^{9–11} to how institutions affect researcher productivity and impact^{12,13}. Many gaps remain in our understanding of the role of institutions in the production of scientific knowledge, and specifically, how they form, grow, and facilitate scientific collaborations. These questions are important, because collaborations are increasingly prevalent in scientific research^{1,9,10} and produce more impactful and transformative work^{10,14}. Collaboration allows scientists to cope with the increasing complexity of knowledge¹⁵ by leveraging the diversity of expertise¹⁶ and perspectives offered by collaborators from different institutions¹⁷ and disciplines¹⁸.

To understand the evolution of research institutions and collaborations, we analyze a large bibliographic database spanning many decades and multiple scientific disciplines. The database contains millions of publications from which the names of authors (collaborators) and their affiliations (research institutions) have been extracted for each paper. Analysis of these data reveals strong statistical regularities. We find that collaborations scale superlinearly with institution size, i.e., faster than institutions grow, consistent with densification of growing networks^{19–21}. However, the scaling law is different for each institution, and as a result, different parts of the collaboration network densify at different rates. We also find that institutions vary in size by many orders of magnitude with an approximately power-law distribution, also known as Zipf's law²². The number of institutions, in contrast, scales sublinearly with the number of researchers, thus following Heaps' law^{23,24}. The sublinear scaling implies that, even as more institutions appear, each institution gets larger on average, but this average belies an enormous variance.

Finally, we create a stochastic model that helps explain how institutions and research collaborations form and grow. In this model, a researcher appears at each time step and is preferentially hired by larger institutions (e.g., due to their prestige or funding), which leads to the rich-get-richer effect creating Zipf's law. With a small probability, however, a researcher joins a newly appearing institution. The arrival of this new institution then triggers yet more new institutions to form in the future, which explains Heaps' law²⁵. Finally, once hired, researchers make connections to other researchers and their collaborators with an independent probability to explain collaborations scaling superlinearly with institution size. Despite its simplicity, the model reproduces a range of empirical observations, including the number and size of research institutions, and how pockets of increasingly dense structures form in collaboration networks.

These empirical results demonstrate universal emergent patterns in the formation and growth of research institutions and collaborations. Our model demonstrates that new institutions are critical to absorbing extra capacity by collecting researchers who do not join large institutions. At the same time, large institutions offer an economy of scale: they grow faster and provide more collaboration opportunities compared to smaller institutions.

Results and discussion

As the first step towards characterizing the complexity of institution scaling, we collect data from Microsoft Academic Graph²⁶ to capture how millions of collaborations evolve over time. Figure 1 shows the collaboration network at the institution level in the field of sociology. Figure 1a demonstrates a remarkable diversity of

institution size and growth, both in terms of the number of researchers (node growth) and collaborations between institutions (edge growth). Collaborations are clustered, with clear groups of interacting institutions. Research collaborations within an institution are equally complex. Figure 1b highlights the largest connected component of the collaboration network within Harvard. Individual researchers vary widely in the number of collaborators, with new collaborations appearing in clusters.

This dataset helps us capture how the number of collaborations scale with an institution's size, n . Figure 2a, b shows the number of internal and external collaborations versus n across four different disciplines: computer science, physics, math, and sociology. While each institution follows a scaling law $c \sim n^\alpha$ (R^2 is close to 1.0, see Supplementary Note 6), the exponents α differ substantially between institutions. This is shown in the insets of Fig. 2a, b where we collect scaling exponents across thousands of institutions and notice that their distribution stretches between zero (in which institutions do not gain any collaborations) to two (collaborations are extremely dense). In the thermodynamic limit, exponents cannot be larger than two, therefore values above two are due to finite-size effects.

To show that the scaling exponents of all institutions are different, we create a null model (see Supplementary Note 3) in which all institutions follow the same scaling law. In this null model, residuals of each institution's fitted scaling relation are reshuffled and added as noise onto a single scaling relation. Differences between fitted exponents in this model are due to statistical noise rather than different scaling laws. We find that the variance of the scaling laws across all institutions is much higher than this null model. We therefore reject the hypothesis that all the exponents within a field are the same within statistical error. We explore the dependence of scaling on final institution size in Supplementary Note 6, and find the scaling exponents are superlinear (approximately 1.2 on average) and do not depend strongly on the final size of the institution. Different parts of the collaboration network therefore densify at different rates, which extends on previous work that uncovered densification for many networks at the aggregate level¹⁹.

We find weak evidence that higher scaling exponents correspond to institutions with greater impact. In physics, the Spearman rank correlation, s , between mean paper impact after five years and internal collaboration scaling exponents is 0.09 (borderline significant, p -value = 0.06) and for external collaboration is 0.27 (p -value < 10^{-5}). Similarly, in sociology, the correlation is 0.19 (p -value = 0.03) between impact and internal collaboration exponents, and the same correlation value is found for external collaboration exponents. For all other fields, however, the correlations are not statistically significant (p -value \geq 0.20). Impact, a proxy of institution research quality, cannot fully explain why collaborations grow faster in some institutions and not others, but can give some insight into reasons for this diversity. These results suggest that highly impactful institutions seem to form collaborations more easily as they grow. Nonetheless, almost all institutions benefit from being larger, as the number of collaborations per person typically grows with size (Fig. 2a, b inset).

The superlinear scaling of collaborations cannot be explained by researcher productivity. The scaling exponents of output, i.e., the cumulative number of papers published by researchers affiliated with that institution at a given year, are centered around 1.0 (see Supplementary Note 4). Paper output per researcher is therefore approximately independent of institution size. The average team size per institution, however, increases with institution size (see Supplementary Note 5), which may help explain the scaling of collaborations. Namely, as institutions grow, they form larger teams for each paper. This, in turn, creates more collaborations (which are proportional to the team size squared).

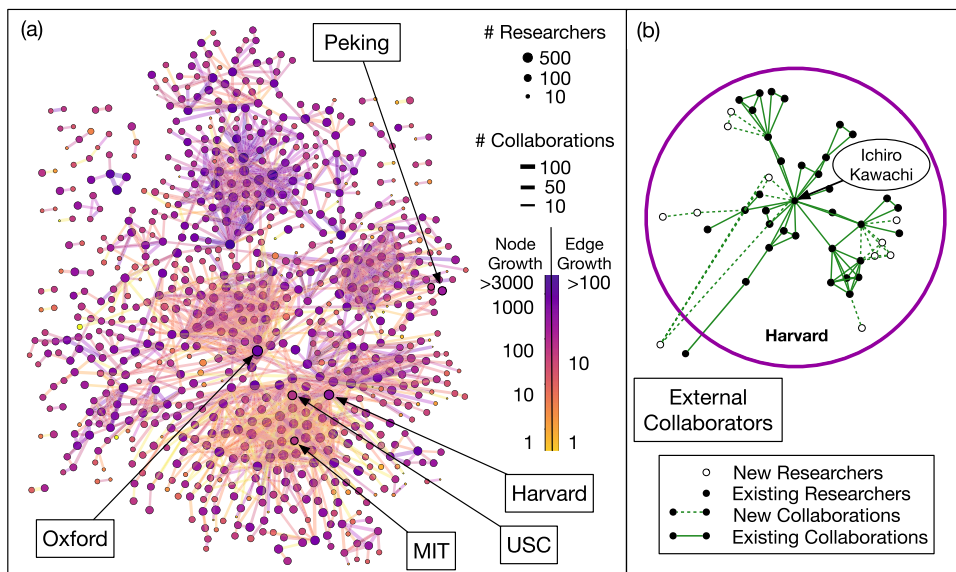


Fig. 1 Network visualization of the collaborations in the field of sociology in 2017. **a** Collaborations between institutions. Each node represents a research institution, and institutions with more researchers are represented by larger nodes. Each link represents collaborations between researchers at different institutions and more collaborations are represented by thicker lines. Darker nodes represent faster-growing institutions (defined as the number of new researchers added between 2012 and 2017), and darker links represent faster-growing collaborations (defined as the number of new inter-institution collaborations between 2012 and 2017). Links with fewer than 10 collaborations are removed, as are isolated nodes. A few major universities are labeled: Peking University, Oxford University, Harvard University, Massachusetts Institute of Technology (MIT), and the University of Southern California (USC). **b** The largest connected component of collaborations within Harvard University. Each node represents a researcher. Dashed lines represent new collaborations added between 2012 and 2017, while open circles represent new researchers added between 2012 and 2017. The highest degree node is Ichiro Kawachi, a highly cited sociologist.

We also find that the distribution of institution sizes (as of 2017) follows Zipf’s law (Fig. 2c), similar to the observed heavy-tailed distribution of city sizes^{22,27}. In Supplementary Note 1 and Supplementary Data 1, we show that while the largest institutions are intuitive, such as Harvard, the smaller institutions tend to be for-profit colleges, community colleges, and institutions without a formal department in the field of interest (e.g., an engineering school with papers in sociology). In addition, the number of institutions grows sublinearly with the number of researchers in each field (Fig. 2d). This feature, known as Heaps’ law, implies that quadrupling the number of researchers in a field roughly doubles the total number of institutions associated with that field. Exact scaling law values for each field can be found in Table 1, where Heaps’ laws are calculated for the total number of researchers in each field, N , greater than twenty and Zipf’s law is calculated for institution size, n , greater than ten.

A Model of Institution Growth. We now describe a stochastic growth model of institution formation that elucidates how institutions and collaborations jointly grow. We model institution formation and growth with a Pólya’s urn-like set of mechanisms described in ref. ²⁵, and we model the growth of collaborations with a network densification mechanism^{20,21}. Unlike existing models of network densification^{19–21}, however, our model reproduces the heterogeneous densification of internal and external collaborations, and the non-trivial growth structure on institutions. This is complimentary to a very recent model on heterogenous exploration²⁸, in which Pólya’s urn models vary as a function of a node’s position on a (static) network.

We imagine an urn containing balls of different colors. The balls can be thought of as the resources given to each institution, where each color represents a different assigned institution, as shown in Fig. 3a. Balls are picked uniformly at random with replacement, with each pick representing a newly-hired

researcher, and the ball color is recorded in a sequence to represent what institution hires the researcher. Afterwards, ρ balls of the same color are added to the urn to represent the additional resources and prestige given to a larger institution, known as reinforcement (left panel of Fig. 3a)²⁵. If a previously unseen color is chosen, then $\nu + 1$ uniquely colored balls are placed into the urn, a step known as triggering (right panel of Fig. 3a)²⁵. The new colors represent institutions that are able to form because of the existence of a new institution. This triggering, also known as adjacent possible²⁵, does not imply causality per se, e.g., the cause of the University of California Merced’s creation was not strictly because of previously established institutions. Instead, these institution-specific causes are represented as stochastic noise, a remarkable simplification that does not remove the observed statistical regularities. Triggering, however, agrees with anecdotal evidence, making it an intuitive factor behind the creation of institutions. For example, UC Davis was spun out of UC Berkeley, and USC Institute for Creative Technology was spun out of USC Information Sciences Institute, which itself was founded by researchers from the Rand Corporation. The model we describe is known as Pólya’s urn with triggering²⁵, and predicts Heaps’ law with a scaling relation $\sim N^{\nu/\rho}$ and Zipf’s law with scaling relation $\sim n^{-(1+\nu/\rho)}$. In our simulations, we arbitrarily chose ρ to be 4 and ν to be 2, which agrees well with the data shown in Fig. 2.

Next, we explain the heterogeneous and superlinear scaling of collaborations through a model of network densification. Building on the work of^{20,21}, we have each new researcher, represented as a node, connect to a random researcher within the same institution, as well as an external researcher picked uniformly at random (left panel of Fig. 3b). New collaborators are then chosen independently from neighbors of neighbors with probability p_i , where p_i is unique to each researcher’s institution (right panel of Fig. 3b). We let p_i be a Gaussian distributed random variable with mean,

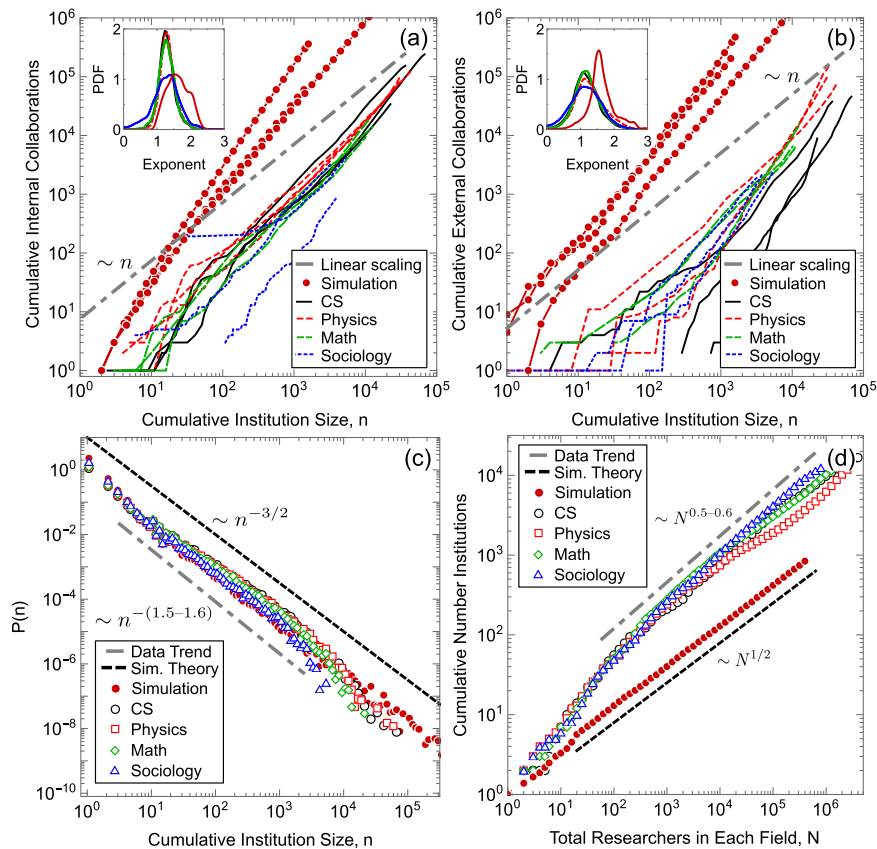


Fig. 2 Institutions densify at different rates but their size and frequency follow universal patterns. **a** Internal and **(b)** external collaborations versus institution size for three arbitrarily chosen institutions with more than 10^3 cumulative researchers in each field or simulation. Circle markers correspond to simulation data; solid lines, medium dashed lines, long dashed lines, and short dashed lines correspond to data from the fields of computer science (CS), physics, math, sociology, respectively. Dash-dotted lines report linear scaling, showing that institutions' scaling laws are super-linear. Insets: distribution of exponents across thousands of institutions (cf. Supplementary Note 1). **c** The distribution of researchers in each institution as of 2017 (Zipf's law), and **(d)** the number of unique institutions versus the total number of researchers in each field (Heaps' law). Closed circle markers correspond to simulation data; open circles, squares, diamonds, and triangles correspond to computer science (CS), physics, math, sociology, respectively. In addition, light dash-dotted lines indicate empirical trends while darker dashed lines indicate theoretical scaling law exponents $-1 - \nu/\rho$ and ν/ρ for **(c)** and **(d)** respectively²⁵. Simulation data in **(a)** and **(b)** are collected from four realizations and in **(c)** and **(d)**, from fifteen realizations (individual realizations show similar trends). Simulation parameters are ρ equals 4, ν equals 2, μ_p equals 0.6, and σ_p equals 0.25.

Table 1 Zipf's law and Heaps' Law exponents for research fields and simulation.

Discipline	Heaps' Law Exponent	Zipf's Law Exponent
Comp. Sci.	0.554 ± 0.004	-1.470 ± 0.005
Physics	0.501 ± 0.007	-1.474 ± 0.006
Math	0.549 ± 0.008	-1.516 ± 0.006
Sociology	0.622 ± 0.005	-1.603 ± 0.009
Simulation	1/2	-3/2

Each fit is a linear regression on log-scaled x and y axes for the number of researchers in each field above 100. Errors are standard errors of linear regression coefficients. Simulation scaling laws are theoretical exponents calculated for Polya's urn model with triggering with coefficients $\rho = 4$ and $\nu = 2$ ²⁵. See Results and Discussion for details of the mechanism coefficients.

$\mu = 0.6$, and standard deviation, $\sigma_\mu = 0.25$ and truncated between 0 and 1. Lambiotte et al.²¹ show that their equivalent to μ , when greater than 0.5, produces densification. We therefore choose $\mu = 0.6$ to ensure the network densifies. We show separately that p_i directly controls the heterogeneity we observe in internal collaboration scaling, but the heterogeneity in external collaboration scaling is an emergent outcome of this model²⁹.

To summarize, our model has four parameters: ρ (reinforcement), ν (triggering), and two parameters to explain collaboration densification heterogeneity, μ_p and σ_p . In the main text, we let ρ equal 4, ν equal 2, μ_p equal 0.6, and σ_p equal 0.25. These are arbitrarily chosen parameters meant to create statistical patterns that are qualitatively similar to empirical data. Namely, $\mu_p > 0.5$ ensures collaboration densification²¹, and $\sigma_p > 0$ ensures that densification scaling exponents vary between institutions. Interestingly, this model's Zipf's and Heaps' laws can be exactly calculated, as discussed by Tria et al.²⁵, with Zipf's law exponent equal to $-1 - \nu/\rho$ and Heaps' law equal to ν/ρ . This model qualitatively reproduces Zipf's and Heaps' laws (Fig. 2c, d and Table 1) and the heterogeneous scaling of internal and external collaborations shown in Fig. 2a, b. While other plausible mechanisms for Zipf's law^{30–32}, Heaps' law²⁴, or densification¹⁹ exist, the current model describes these patterns in a cohesive framework and explains the heterogeneous scaling we discover in the data. While this heterogeneity is built into our internal scaling laws, the external scaling heterogeneity is an emergent property within the model²⁹.

The model also reproduces qualitative trends of cross-sectional analysis. Specifically, the scaling exponents of internal collaborations produced by the model when measured at a specific point in

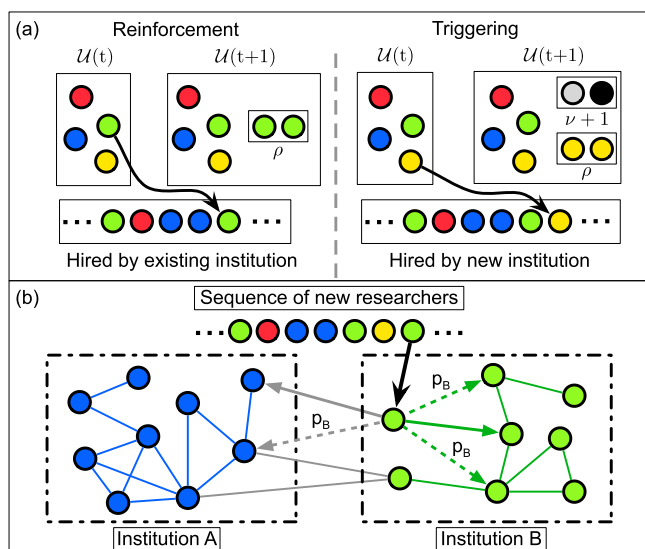


Fig. 3 Schematic representation of the institution growth model. **a** At time t a new researcher is hired, modeled as extracting a ball with uniform probability with replacement from an urn, U (black arrow). The ball color represents an institution. Hiring a researcher will always add ρ new balls of the same color to the urn in the next timestep (reinforcement). Hiring the first researcher at an institution (picking a ball color that has never been picked before), triggers $\nu + 1$ new colors to enter the urn, increasing the likelihood of more institutions to hire their first researcher (triggering). **b** Researchers within each institution (dash-dotted boxes) have both internal collaborators (darker solid lines) and external collaborators (gray lines). Once a researcher is hired, they choose one random internal and one random external collaborator (solid arrows). New collaborations (dashed arrows) are formed independently with probability p_A , if hired by institution A, and p_B if hired by institution B. These new connections form triangles.

time, i.e., in a cross-sectional setting, vary in time and are larger than scaling exponents of external collaborations and decrease over time (Supplementary Note 6), unlike what we see in data (Supplementary Fig. 3). These results are robust to stochastic variations of the densification mechanism (Supplementary Note 7). As a final comparison with data, we compared the growth of institutions and the ways links form to the model mechanisms and found broad agreement²⁹.

Conclusion

We identify strong statistical regularities in the growth of research institutions. The number of collaborations increases superlinearly with institution size, i.e., faster than institutions grow in size, though the scaling is heterogeneous, with a different exponent for each institution. Therefore, each institution has its own universal scaling, i.e., regardless of its size, it will always have the same percentage of new collaborations for each percentage increase in size. The super scaling is not explained by the increased productivity of researchers at larger institutions the number of papers per researcher is roughly independent of institution size. Instead, the growing collaborations are associated with bigger teams at larger institutions. The diversity in collaboration scaling exponents is partly explained by variations in institution impact. Institutions with higher impact papers also tend to have a larger scaling exponent. This provides evidence that a higher collaboration scaling exponent allows for collaborations to form more easily, and that in turn creates higher-impact papers. Further analysis is needed to test this hypothesis in the future.

When these observations are incorporated into a minimal stochastic model of institution growth, we are able to reproduce the

surprising regularity of research institution formation, growth and the heterogeneous densification of collaboration networks. That said, there is still room for improvements to this model, given quantitative differences between the model and data, such as the constant shift difference between the Heaps' laws (Fig. 2c), or the difference in the collaboration scaling law exponents (insets of Fig. 2a, b).

These findings support the idea that academic environments differ in their ability to bolster researcher productivity and prominence¹², and also demonstrate that institution size and ability to facilitate collaborations as a potential factor explaining differences in academic environments. Additional research is needed to identify other factors that contribute to an institution's success.

Methods

Data. We use bibliographic data from Microsoft Academic Graph (MAG), from which researcher names (authors), their institutional affiliation, and references made to other papers have been extracted^{26,33}. MAG data has disambiguated institutions and authors for each paper, allowing us to consider all authors with the same unique identifier to be the same researcher, and similarly for each institution. In these data, authors typically have only one affiliation at any time (see Supplementary Note 1). We focus on four fields of study: computer science, physics, math and sociology. After data cleaning, we have almost ten million papers published between 1800 and 2018 (see Supplementary Note 1). Our computer science data includes early research in topics relating to computers, including electrical engineering, and therefore stretches back to before 1900.

We define *institution size* in a given year as the number of authors who have been ever been affiliated with that institution up until that year. *Collaborations* are defined as two researchers who have co-authored a paper up until that year. We distinguish between internal collaborations (co-authors at the same institution) and external collaborations (co-authors affiliated with different institutions). Finally, to understand the relation between collaborations and institution size, we define output as the cumulative number of papers from researchers affiliated with an institution in a particular year.

Analysis. We use cumulative statistics to reduce statistical variations and to better compare to a stochastic growth model of institution formation. To check the robustness of results, we compare to an alternate yearly definition of institution size and collaborations (see Supplementary Note 2). We find all qualitative results are the same, in part because both definitions are highly correlated.

We present scaling results for longitudinal analysis, which tracks how collaborations evolve as individual institutions grow^{34–36}. This contrasts to cross-sectional analysis applied in previous work on city scaling^{37,38} and institution scaling^{2–4,39}, which measures collaborations as a function of the size of all institutions at a given point in time. We find that cross-sectional analysis identifies scaling laws that are not representative of the growth of most institutions (see Supplementary Note 7), and while simulations and empirical data give scaling exponents that are fairly constant in time for each institution, cross-sectional scaling exponents vary in time for both data and simulation. For these reasons, we focus on longitudinal scaling analysis in this paper, although scaling laws derived by either analysis method strongly relate to each other^{36,40}.

Data availability

Microsoft Academic Graph data can be accessed via the following link: <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/>²⁶. Replication data collected from Microsoft Academic Graph are available in the following repository: <https://github.com/KeithBurghardt/HeterogeneousScalingCode/tree/master/output>. Sample raw data for small institutions are available in Supplementary Data 1.

Code availability

Code for this study is available in the following repository: <https://github.com/KeithBurghardt/HeterogeneousScalingCode>.

Received: 2 February 2021; Accepted: 29 July 2021;

Published online: 02 September 2021

References

- Hicks, D. & Katz, J. S. Science policy for a highly collaborative science system. *Science and public policy* **23**, 39–44 (1996).
- Taylor, R. C. et al. The scalability, efficiency and complexity of universities and colleges: a new lens for assessing the higher educational system. Preprint at <https://arxiv.org/abs/1910.05470> (2019).

3. van Raan, A. F. J. Universities scale like cities. *PLoS ONE* **8**, e59384 (2013).
4. Jamtveit, B., Jettestuen, E. & Mathiesen, J. Scaling properties of European research units. *Proc. Natl. Acad. Sci.* **106**, 13160–13163 (2009).
5. Murray, D. et al. Unsupervised embedding of trajectories captures the latent structure of mobility. Preprint at <https://arxiv.org/abs/2012.02785> (2020).
6. Fortunato, S. et al. Science of science. *Science* **359**, ea0185 (2018).
7. Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).
8. Wang, D., Song, C. & Barabási, A.-L. Quantifying long-term scientific impact. *Science* **342**, 127–132 (2013).
9. Guimera, R., Uzzi, B., Spiro, J. & Amaral, L. A. N. Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697–701 (2005).
10. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
11. Milojević, S. Principles of scientific research team formation and evolution. *Proc. Natl. Acad. Sci.* **111**, 3984–3989 (2014).
12. Way, S. F., Morgan, A. C., Larremore, D. B. & Clauset, A. Productivity, prominence, and the effects of academic environment. *Proc. Natl. Acad. Sci.* **116**, 10729–10733 (2019).
13. Deville, P. et al. Career on the move: Geography, stratification, and scientific impact. *Sci. Rep.* **4**, 4770 EP – (2014).
14. Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382 (2019).
15. Jones, B. F. The burden of knowledge and the “death of the renaissance man”: is innovation getting harder? *Rev. Econ. Stud.* **76**, 283–317 (2009).
16. Page, S. E. *The Diversity Bonus: How Great Teams Pay off in the Knowledge Economy*, vol. 5 (Princeton University Press, 2019).
17. Dong, Y., Ma, H., Tang, J. & Wang, K. Collaboration diversity and scientific impact. Preprint at <https://arxiv.org/abs/1806.03694> (2018).
18. Yegros-Yegros, A., Rafols, I. & D’Este, P. Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity. *PLoS one* **10**, e0135095 (2015).
19. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph evolution: densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* **1**, 2–41 (2007).
20. Bhat, U., Krapivsky, P. L., Lambiotte, R. & Redner, S. Densification and structural transitions in networks that grow by node copying. *Phys. Rev. E* **94**, 062302 (2016).
21. Lambiotte, R., Krapivsky, P. L., Bhat, U. & Redner, S. Structural transitions in densifying networks. *Phys. Rev. Lett.* **117**, 218301 (2016).
22. Zipf, G. K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley Press, Inc., Cambridge, MA, 1949).
23. Lü, L., Zhang, Z.-K. & Zhou, T. Zipf’s law leads to Heaps’ law: Analyzing their relation in finite-size systems. *PLoS ONE* **5**, 1–11 (2010).
24. Simini, F. & James, C. Testing Heaps’ law for cities using administrative and gridded population data sets. *EPJ Data Sci.* **8**, 24 (2019).
25. Tria, F., Loreto, V., Seredvio, V. D. P. & Strogatz, S. H. The dynamics of correlated novelties. *Sci. Rep.* **4**, 5890 EP – (2014).
26. Sinha, A. et al. An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th international conference on world wide web*, 243–246 (ACM, 2015).
27. Batty, M. Rank clocks. *Nature* **444**, 592–596 (2006).
28. Iacopini, I., Di Bona, G., Ubaldi, E., Loreto, V. & Latora, V. Interacting discovery processes on complex networks. *Phys. Rev. Lett.* **125**, 248301 (2020).
29. Burghardt, K., Percus, A., He, Z. & Lerman, K. A model of densifying collaboration networks. Preprint at <https://arxiv.org/abs/2101.11056> (2021).
30. Gibrat, R. *Les inégalités économiques; applications: aux inégalités des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d’une loi nouvelle, la loi de l’effet proportionnel*. (Librairie du Recueil Sirey, Paris, 1931).
31. Eeckhout, J. Gibrat’s law for (All) cities. *Am. Econ. Rev.* **94**, 1429–1451 (2004).
32. Axtell, R. L. Zipf distribution of U.S. firm sizes. *Science* **293**, 1818–1820 (2001).
33. Herrmannova, D. & Knoth, P. An analysis of the Microsoft Academic Graph. *D-Lib Magazine* <http://www.dlib.org/dlib/september16/herrmannova/09herrmannova.html> (2016).
34. Depersin, J. & Barthelemy, M. From global scaling to the dynamics of individual cities. *Proc. Natl. Acad. Sci.* **115**, 2317–2322 (2018).
35. Keuschnigg, M. Scaling trajectories of cities. *Proc. Natl. Acad. Sci.* **116**, 13759–13761 (2019).
36. Ribeiro, F. L., Meirelles, J., Netto, V. M., Neto, C. R. & Baronchelli, A. On the relation between transversal and longitudinal scaling in cities. *PLOS ONE* **15**, 1–20 (2020).
37. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci.* **104**, 7301–7306 (2007).
38. Bettencourt, L. M. A. The origins of scaling in cities. *Science* **340**, 1438–1441 (2013).
39. Fix, B. Energy and institution size. *PLOS ONE* **12**, 1–22 (2017).
40. Bettencourt, L. M. A. et al. The interpretation of urban scaling analysis in time. *J. R. Soc. Interface* **17**, 20190846 (2020).

Acknowledgements

Research was funded by in part by DARPA under contract #W911NF1920271 and by the USC Annenberg Fellowship.

Author contributions

K.B., Z.H., A.G.P., and K.L. designed research; K.B. and Z.H. performed research; K.B. created and simulated the model; K.B. and Z.H. analyzed data; K.B., Z.H., A.G.P., and K.L. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42005-021-00693-2>.

Correspondence and requests for materials should be addressed to K.A.B.

Peer review information *Communications Physics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

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) 2021, corrected publication 2022