Supplementary Information for Reciprocity and success in academic careers

Weihua Li,[1,2] Tomaso Aste,[1,2] Fabio Caccioli,[1,2] Giacomo Livan[1,2*]

[1] *Department of Computer Science, University College London, London WC1E 6EA, United Kingdom*

[2] *Systemic Risk Centre, London School of Economics and Political Sciences, London WC2A 2AE, United Kingdom*

[*] *Corresponding author. E-mail: g.livan@ucl.ac.uk.*

S. 1 Linear regression analysis

In S. 1, we detail the results of a few linear regression models for authors whose career started between 1970 and 1990. The dependent variable is the average number of citations per paper for models M1-2, while the dependent variable for models M3-4 is the average total number of citations per author. In both cases those quantities are computed at the 20th career year for each author. The average number of publications per author is used as an independent variable in all models. Moreover, the average number of citations at the 10th career year and the average number of citations at the 10th career year are used as independent variables for models M1-2 and M3-4, respectively. In all columns, variables such that $p < .05$ are highlighted with one asterisk, while variables such that $p < .01$ are highlighted with two asterisks. Excess reciprocity is used as an additional independent variable for models M2 and M4. As it can be seen from the R-squared values, the introduction of excess reciprocity as an additional independent variable does not provide any meaningful statistical improvement to the models. Yet, let us remark that in model M2, i.e., the only one where excess reciprocity is statistically significant, the relationship between

excess reciprocity and success (as measured by average citations) is aligned with the results we find in the main paper.

S. 1. Linear regressions models for the relationship between citations and excess reciprocity (1970-1990).

Variable	M1	M2	M3	M4
Intercept	1.97^{**}	2.31^{**}	-55.19^{**}	-51.55^{**}
Publications	0.02^{**}	0.02^{**}	3.73^{**}	3.74^{**}
Citations			2.37^{**}	2.36^{**}
Average Citations	2.05^{**}	2.03^{**}		
Excess reciprocity		-1.26^{**}		-15.11
R-squared	0.5715	0.5715	0.6969	0.6969

S. 2 Individual excess reciprocity for 1950-1970 and 1970-1990

We complement our analysis at the level of individual authors by considering authors who started their careers in two time periods: 1950-1970 and 1970-1990. In S. 1, plots in the left panel are for authors who started their career between 1950 and 1970; Plots in the right panel are for authors who started their career between 1970 and 1990. The plots in panels **d** and **e** are the same shown in Fig. 3 of the main paper, which we replot here for convenience. We refer the reader to the caption of such Figure for the details on such plots. The sample of authors starting their career in 1950-1990 contains $1,577$ authors, and the fraction of authors falling within each cluster are as follows: 3.0% in group 1, 8.2% in group 2, 28.3% in group 3, 62.1% in group 4, for authors who

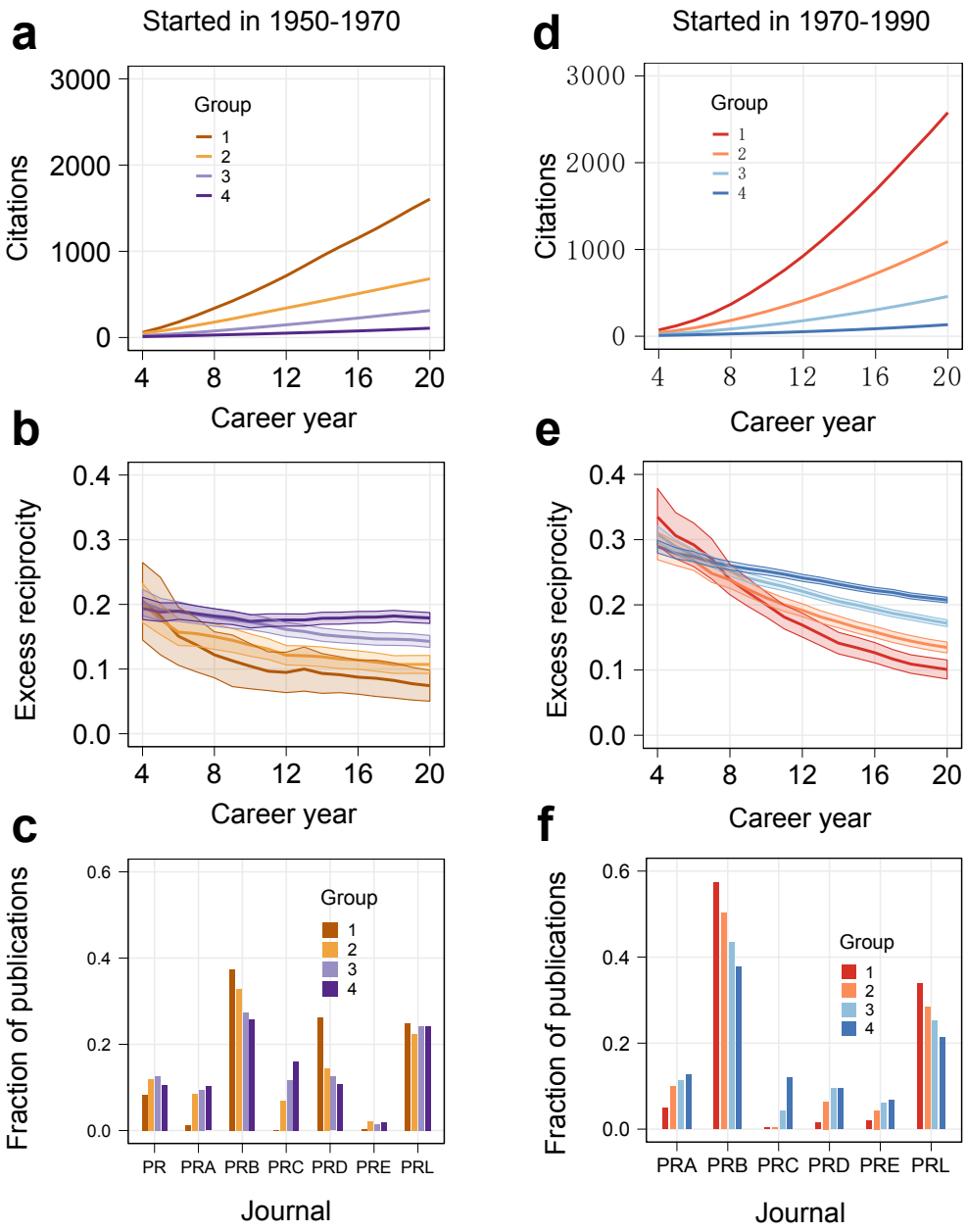
S. 2. Results of KS tests on excess reciprocity

group	group	p-value (yr. 4)	p-value (yr. 20)
1	2	0.181	$< 10^{-3}$
1	3	0.136	0.012
1	4	$< 10^{-3}$	$< 10^{-3}$
2	3	0.108	$< 10^{-3}$
2	4	$< 10^{-3}$	$< 10^{-3}$
3	4	0.001	$< 10^{-3}$

started their career between 1950 and 1970.

S. 3 Kolmogorov-Smirnov tests on excess reciprocity

In this Supplementary Note we report the results of two-sided Kolmogorov-Smirnov (KS) tests performed on the distributions of excess reciprocity for the groups of authors shown in Fig. 3 of the main paper. Namely, we consider the null hypothesis of excess reciprocities being drawn from the same distributions both, and run KS tests both at career year 4 and career year 20. The results are reported in Table S. 2. As it can be seen, at career year 4 the null hypothesis can be safely rejected for all pairs of groups involving group 4, i.e., the authors who eventually become the least successful. In all other cases, the null hypothesis cannot be rejected. On the other hand, at career year 20 the null hypothesis can be rejected in all cases at a 5% significance level.



S. 1. Illustration of the relationship between reciprocity and long-term career success (1950-1970 and 1970-1990). **a** Centroids of the clusters identified by k-means based on the cumulative number of citations received by authors. **b** Excess reciprocity, defined as per Eq. (1) in the main paper, within each cluster identified in **a**. Thick solid lines denote the average within the cluster, while ribbon bands denote 95% confidence level intervals. **c, f** Fraction of publications in each APS journal by authors from each cluster.

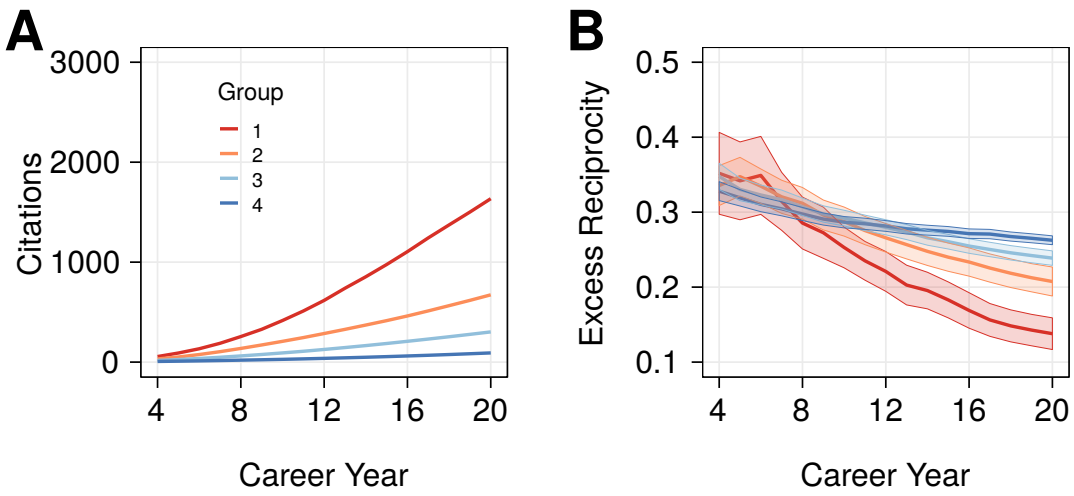
S. 4 Alternative null model specifications

We apply a number of additional robustness checks to corroborate the excess reciprocity analysis, focusing mainly on authors that started their careers in 1970-1990. In addition to the geodesic distance 1 null model, we also examine two additional null models based on community structure[1]. In S. 4 **c**, we show results based on the 47 communities in the citation network of authors identified by a *modularity-based* algorithm[2], and in S. 4 **d** we show the results obtained on the $2,458$ communities identified by the *InfoMap* algorithm[3]. As described in the Methods section of the main paper, these additional null models execute the previous rewiring procedures with constraints based on the time ordering of publications and on the community structures of author citation networks. After the system has reached equilibrium, we compute the annual excess reciprocity for authors who published at least 10 papers, published at least 1 paper every five years, and whose careers lasted at least 20 years. Then we use the k-means classification method to identify four clusters of authors based on their career citation dynamics in **a**, and see how it associates with their excess reciprocity. The results indicate that our excess reciprocity measure is robust in all null model settings, and that excess reciprocity is negatively correlated with success at the later career stages.

We also replicate our analysis based on a productivity threshold on the authors. In S. 5, we show results for null models where we select authors that have received at least 10 citations by their last career year. As most authors who have a consistently long career are able to achieve that, the sample contains $5,016$ authors, and the fractions of authors falling within each cluster are as follows: 1.3% in group 1, 7.0% in group 2, 24.9% in group 3, 66.8% in group 4. The negative relationship between excess reciprocity and success is still verified under this model setting.

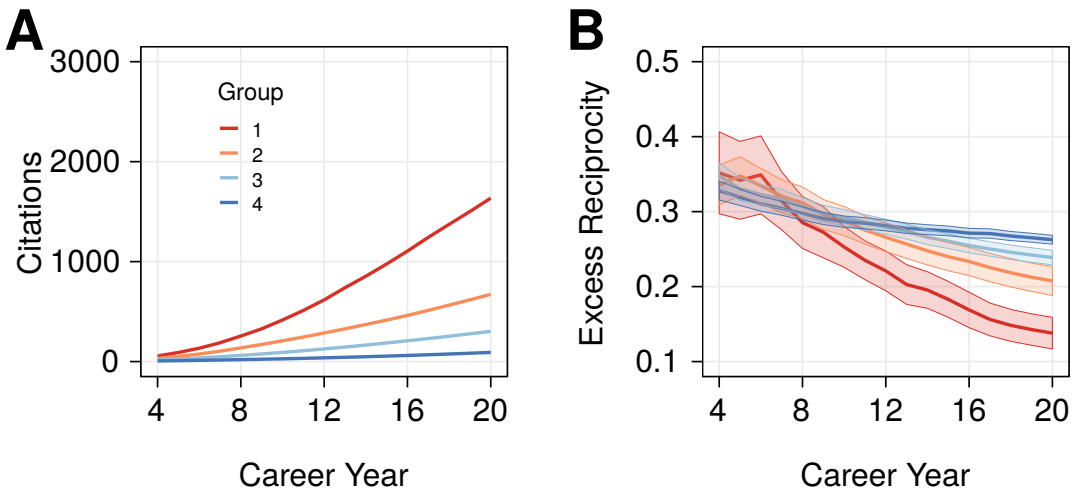
We then proceed to verify whether the participation in large research teams might skew our results by favoring authors with larger numbers of collaborators and coauthors. The analyses presented in the main paper already discount this potential effect into account, as they are based

only on papers with ten or less authors. S. 2 shows the author clusters and the relationship between excess reciprocity and long-term career success only based on papers published by three or less authors. As it can be seen, our main result still holds



S. 2. Excess reciprocity for papers with at most three authors whose career started in 1970-1990. (**A**) Centroids of k-means clusters on the total number of citations from the 4th to the 20th author career year. (**B**) Annual average excess reciprocity computed from the community–based null model, with 47 communities identified by the *modularity-based* algorithm (see A. Clauset, M. E. Newman, C. Moore, *Physical Review E* **70**, 066111 (2004)), for the four clusters of authors in **A**. The number of authors included here is $2,276$.

S. 3 shows instead the results we obtain when replicating our analyses on US-based authors only. The rationale behind this restriction was to discount potential geography-related biases. In particular, we wanted to assess whether the name disambiguation procedure we used to identify authors might have been biased, e.g., by the high frequency of very common Asian or Spanish / Latin American surnames. As shown in the Figure, all results are qualitatively unchanged when restricting the analysis to US-based authors only.



S. 3. Excess reciprocity for *US-based* authors whose career started in 1970-1990. (**A**) Centroids of k-means clusters on the total number of citations from the 4th to the 20th author career year. (**B**) Annual average excess reciprocity computed from the geodesic distance one null model for the four clusters of authors in **A**. (**C**) Annual average excess reciprocity computed from the community–based null model, with 47 communities identified by the *modularity-based* algorithm (see A. Clauset, M. E. Newman, C. Moore, *Physical Review E* **70**, 066111 (2004)), for the four clusters of authors in **A**. (**D**) Annual average excess reciprocity computed from the community–based null model, with $2,458$ communities identified by *infoMap* algorithm (see M. Rosvall, C. T. Bergstrom, *Proceedings of the National Academy of Sciences* **105**, 1118-1123 (2008)), for the four clusters of authors in **A**. In panels **B-D**, thick solid lines denote the average within the cluster, while ribbon bands denote 95% confidence level intervals.

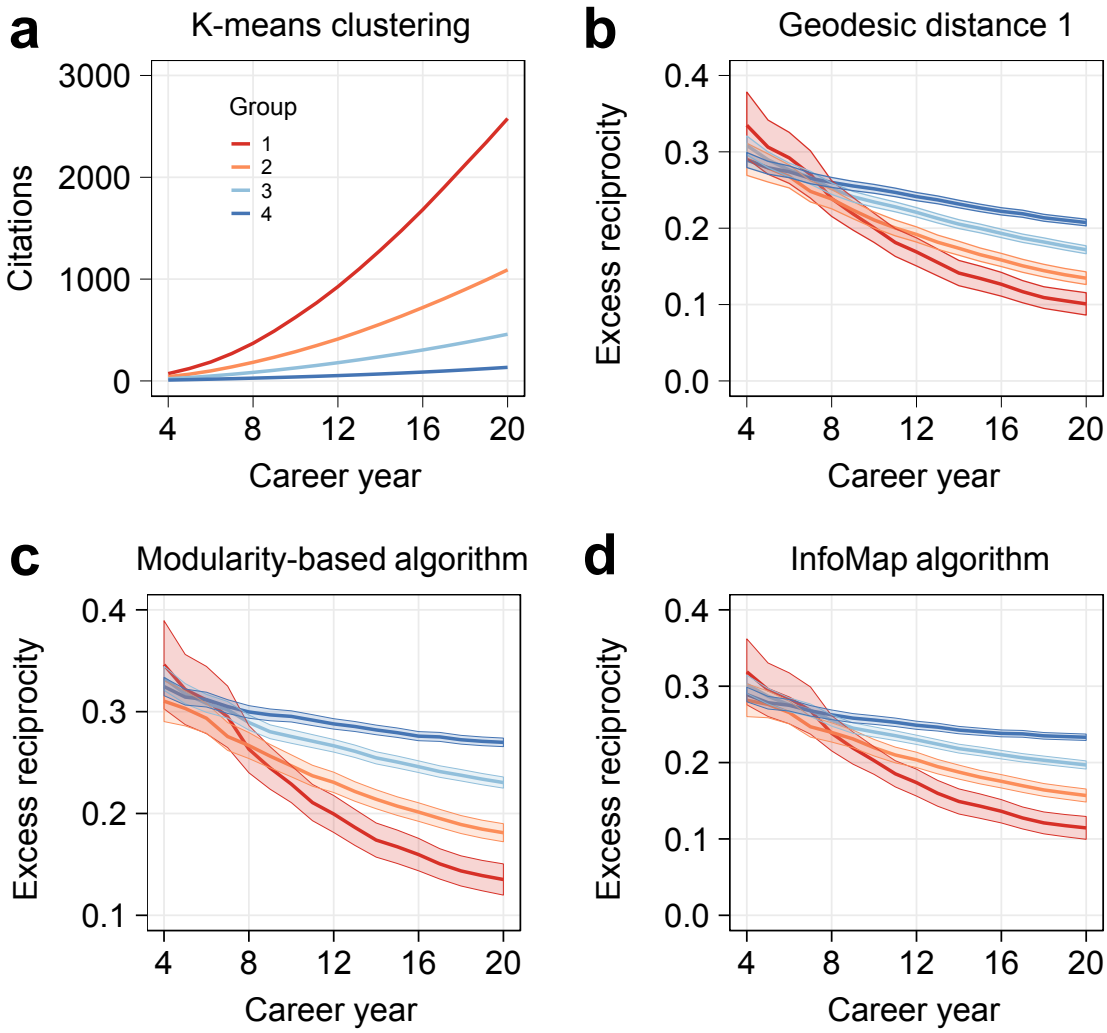
Finally, we study whether the clustering algorithms on author career trajectories affect on our main conclusions or not. In S. 6, we cluster authors based both on the quartiles of the distribution of citations accrued over the first 20 career years, and on the output of the Expectation-Maximization clustering algorithm [4] on annual citation counts from the 4th to the 20th career year For the the Expectation-Maximization clustering algorithm, the fractions of authors falling within each cluster are as follows: 6.0% in group 1, 23.5% in group 2, 41.8% in group 3, 28.6% in group 4. We find that these clustering methods do not change our conclusion that more successful careers are correlated with lower level of excess reciprocity.

S. 5 Productivity-based clustering

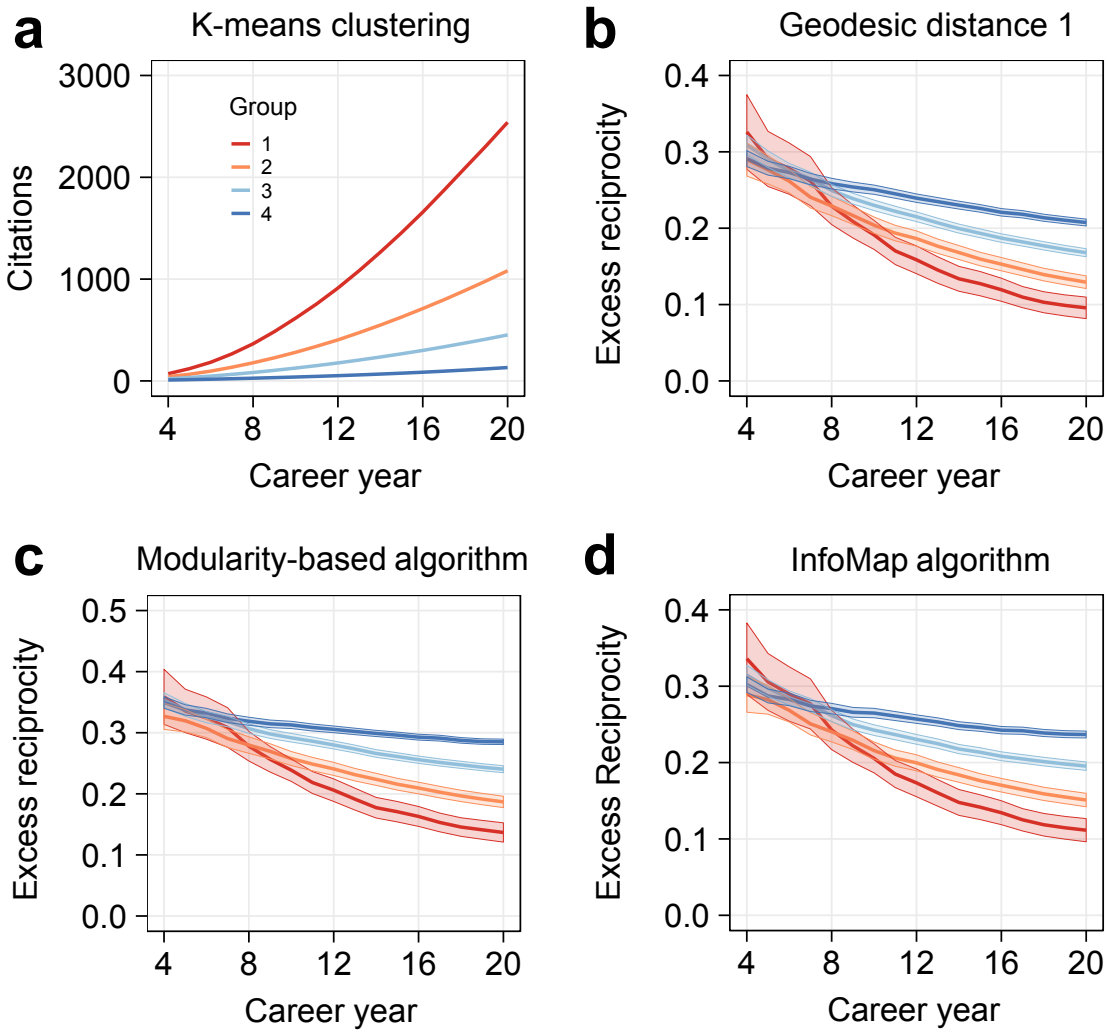
In Fig. S. 7 we provide evidence that the negative relationship between excess reciprocity and career impact is still observed when discounting productivity as a potential confounding factor. As a matter of fact, productivity (measured by the numbers of papers published) is strongly correlated with the number of citations accrued by an author. Therefore, clustering authors based on their productivity would yield results very close to those shown in the main paper. Thus, we proceeded to cluster authors based on their career trajectories in terms of citations accrued per paper. As it can be seen in Fig. S. 7, the aforementioned negative relationship is still present, and actually visible even from early career stages.

S. 6 Matched pair analysis

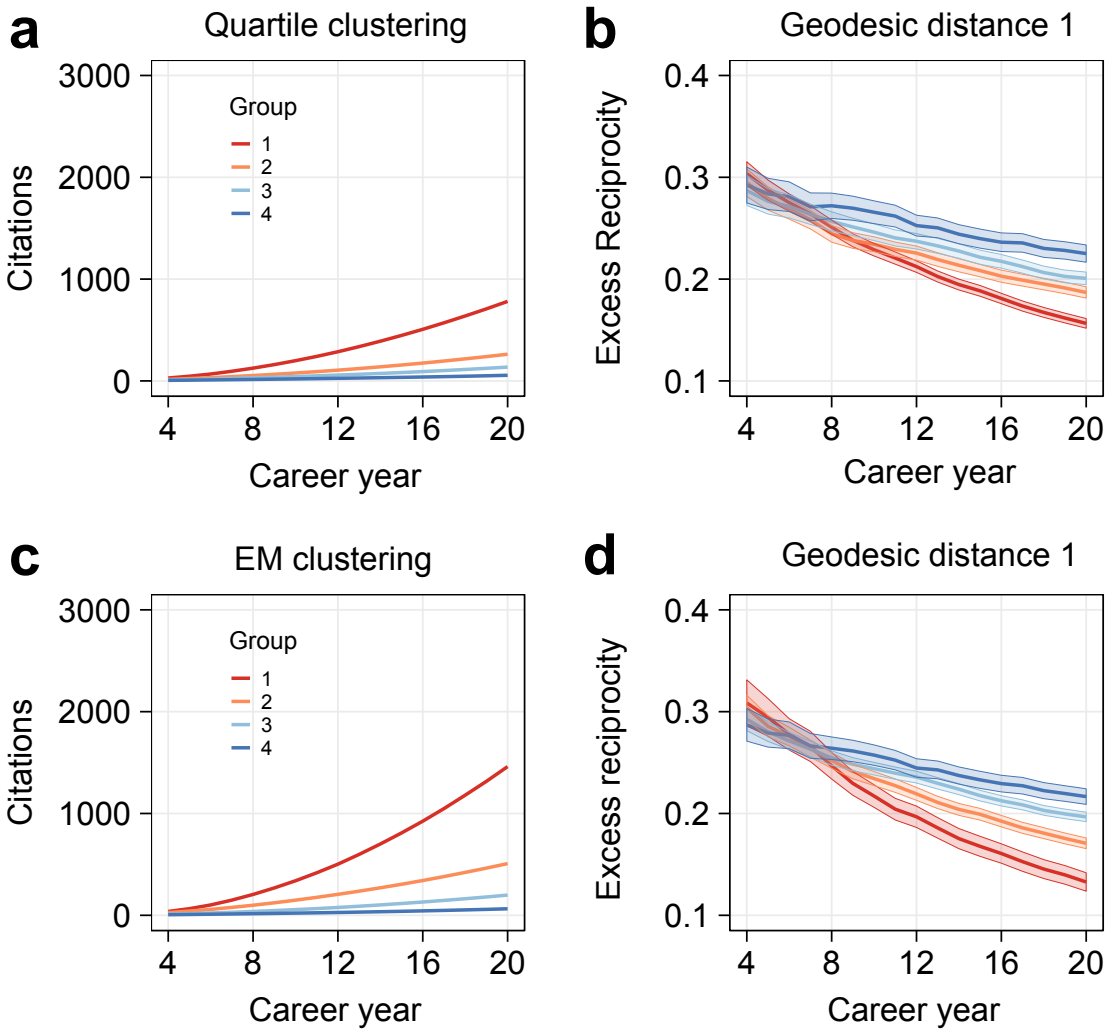
To conduct a matched pair analysis, we first select authors whose careers started in 1970-1990 [5]. We separate them into two groups, i.e., we consider authors who are in the top 25% in terms of excess reciprocity computed over the first 10 career years as the treatment group, and authors in the bottom 50% as the control group, which we then proceed to pair based on the number of publi-



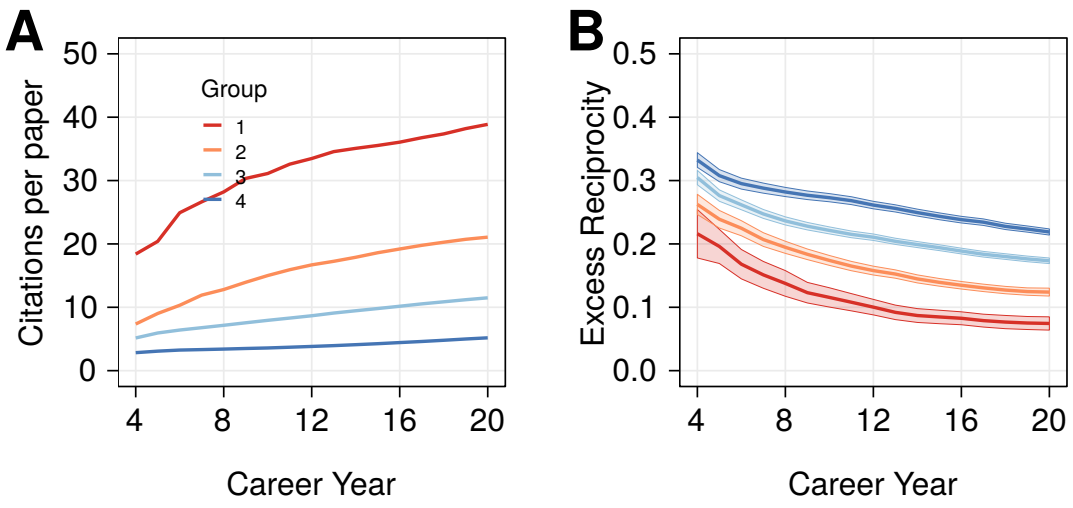
S. 4. Excess reciprocity for authors whose career started in 1970-1990, under three different null model settings. **a** Centroids of k-means clusters on the total number of citations from the 4th to the 20th author career year. **b** Annual average excess reciprocity computed from the geodesic distance 1 null model for the four clusters of authors in **a**. **c** Annual average excess reciprocity computed from the community–based null model with *modularity-based* algorithm. **d** Annual average excess reciprocity computed from the community–based null model *InfoMap* algorithm. In panels **b** and **d** thick solid lines denote the average within the cluster, while ribbon bands denote 95% confidence level intervals.



S. 5. Excess reciprocity for authors who started their careers in 1970-1990 and have at least 10 accrued citations by the 20th career year. **a** Centroids of k-means clusters on the total number of citations from the 4th to the 20th author career year after removing all self-citations. **a** Centroids of k-means clusters on the total number of citations from the 4th to the 20th author career year. **b** Annual average excess reciprocity computed from the geodesic distance 1 null model for the four clusters of authors in **a**. **c** Annual average excess reciprocity computed from the community–based null model with *modularity-based* algorithm. **d** Annual average excess reciprocity computed from the community–based null model *InfoMap* algorithm. In panels **b** and **d** thick solid lines denote the average within the cluster, while ribbon bands denote 95% confidence level intervals.

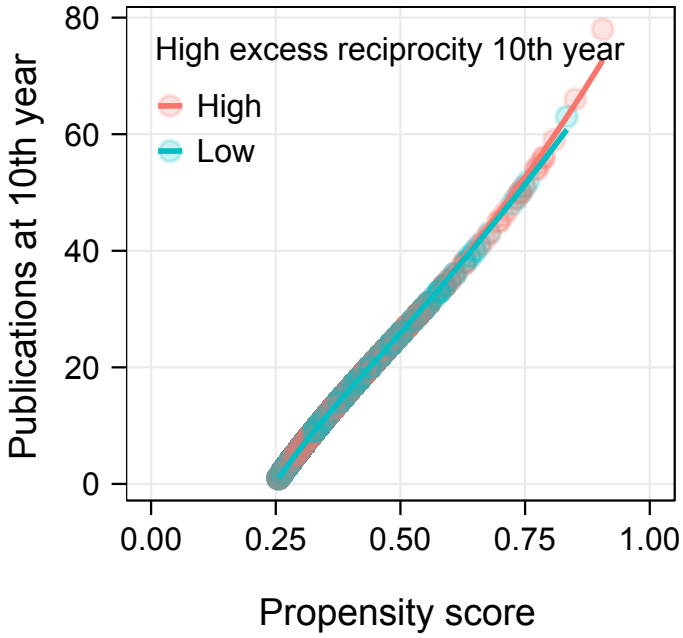


S. 6. Different clustering algorithms on citation patterns and excess reciprocity for authors whose career started in 1970-1990. **a** Average career trajectories for 4 groups of authors corresponding to the quartiles of the distribution of citations accrued over the first 20 career years. **b** Annual average excess reciprocity computed from the geodesic distance 1 null model for the four clusters of authors in **a**. **c** Average career trajectories for 4 groups of authors obtained with the Expectation-Maximization clustering algorithm. **d** Annual average excess reciprocity computed from the geodesic distance 1 null model for the four clusters of authors in **c**. In panels **b** and **d** thick solid lines denote the average within the cluster, while ribbon bands denote 95% confidence level intervals.



S. 7. Excess reciprocity for authors who started their careers in 1970-1990 and have at least 10 accrued citations by the 20th career year. **a** Centroids of k-means clusters on the average number of citations per published paper from the 4th to the 20th author career year. **b** Annual average excess reciprocity computed from the geodesic distance 1 null model for the four clusters of authors in **a**. In panel **b** thick solid lines denote the average within the cluster, while ribbon bands denote 95% confidence level intervals.

cations in APS journals in their first 10 career years. This leaves us with $2,536$ authors. S. 8 shows the mean number of publications at the 10th career year against the propensity score (estimated running a logit model), based on the different treatment status (red for high excess reciprocity, green for low excess reciprocity). The treatment and control groups have nearly identical means at each value of the propensity score. We then proceed to measure the effects of low / high excess reciprocity on success, quantified by the number of citations received between the 11th and 20th career years. We find that authors in the control (low excess reciprocity) group have received, on average, 331.6 citations, while authors in the treatment group (high excess reciprocity) have received 272.2 citations. The difference between such two values is significant ($p < 0.001$) under a t-test.

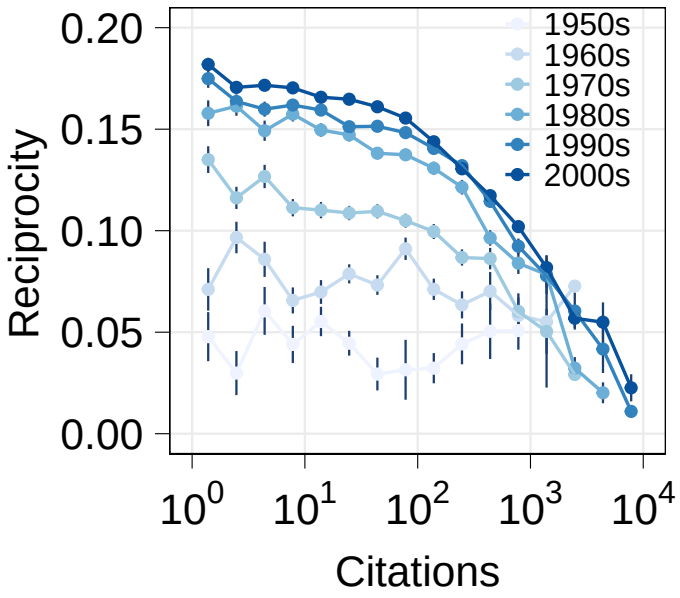


S. 8. Matched pair analysis of authors with high and low excess reciprocity.

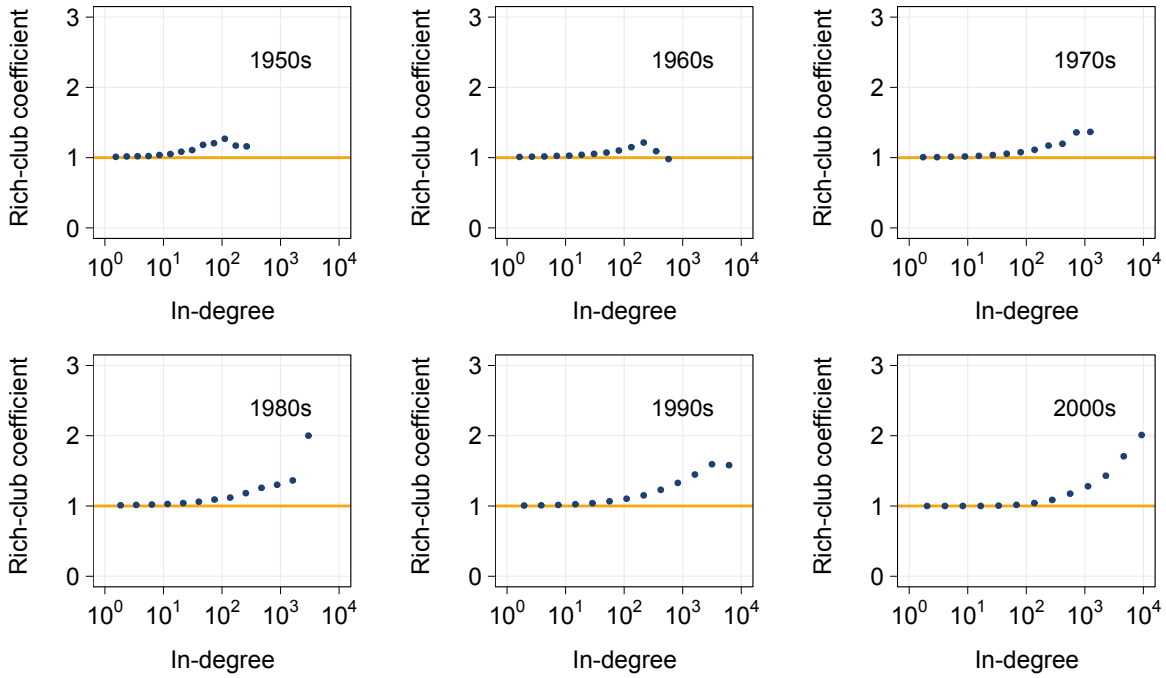
S. 7 Decade level analyses

In S. 9, we plot the excess reciprocity of groups of authors in six decades. It is evident that over time the overall level of excess reciprocity for low and mid-level impact authors has been steadily increasing, while the high impact group of authors did not change their behaviour substantially.

In S. 10, we plot the rich-club coefficient [6] for the citation networks of authors over six decades. The rich-club effect has become more pronounced over time, suggesting that high-impact authors tend to give more citations to already successful peers.



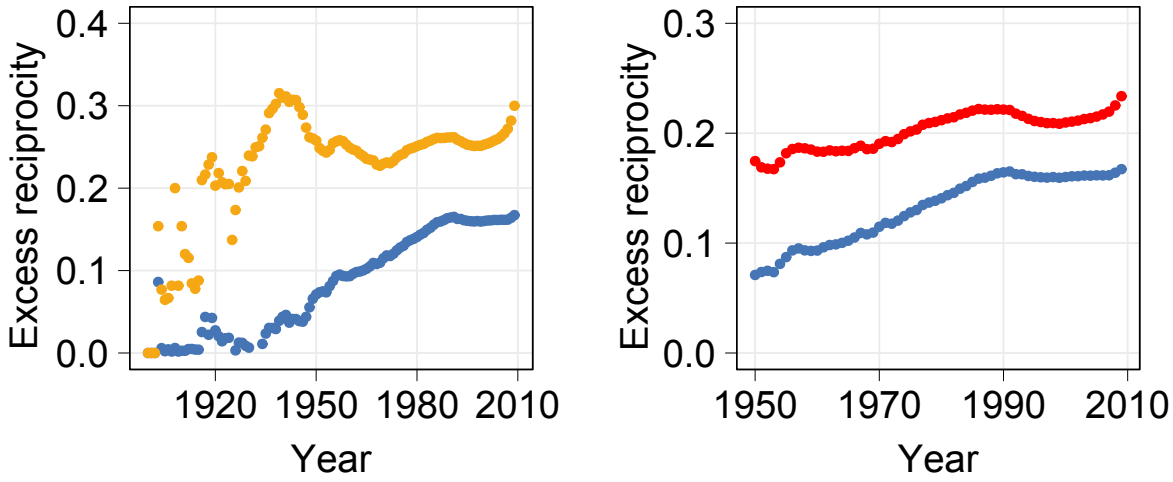
S. 9. Results at the level of groups of authors for all decades. Excess reciprocity as a function of the number of citations received for authors active in each decade from the 1950s to the 2000s, with standard error bars.



S. 10. The rich-club effect in the empirical citation networks for all decades. Rich-club coefficient of the APS author citation network with respect to the null model (see Materials and Methods Section of the main paper) for all decades from the 1950s to the 2000s.

S. 8 Null model constraints

S. 11 shows the effect of the constraints induced by our geodesic distance 1 null model. Blue dots in the right panel denote the the excess reciprocity ρ rescaled by such null model (according to Eq. (1) in the main text when generalized to the entire network), while red dots denote excess reciprocity when rescaled according to a null model with no constraints on the distance between authors (i.e., only with a time constraint based on the time ordering on publications, see the Methods section of the main paper). As it can be seen, the latter, less constrained, null model results in much lower levels of excess reciprocity.



S. 11. Additional results on network-wide excess reciprocity of the citation networks of authors. (Left panel): Orange dots are the reciprocity ρ_0 not rescaled by the null model (i.e., the fraction of reciprocated weight in the network), from 1900 to 2010. (Right panel):a Red dots denote the excess reciprocity rescaled by a null model accounting for the time dynamics of publications but not subject to the constraint on distance implemented in the null model used in the main paper.

Supplementary references

1. Guimera, R., Sales-Pardo, M. & Amaral, L. A. Classes of complex networks defined by role-to-role connectivity profiles. *Nat. Phys.* **3**, 63 (2007).

2. Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111 (2004).

3. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* **105**, 1118–1123 (2008).

4. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38 (1977).

5. Stuart, E. A., King, G., Imai, K. & Ho, D. Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* **42** (2011).

6. Opsahl, T., Colizza, V., Panzarasa, P. & Ramasco, J. J. Prominence and control: the weighted rich-club effect. *Phys. Rev. Lett.* **101**, 168702 (2008).