nature human behaviour

SUPPLEMENTARY INFORMATION

VOLUME: 1 | ARTICLE NUMBER: 0078

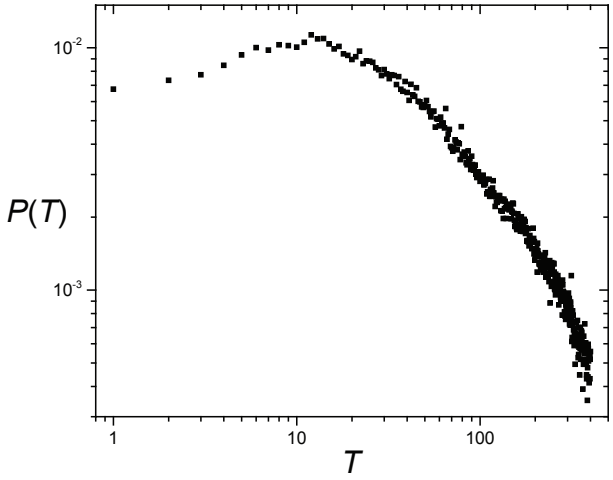
In the format provided by the authors and unedited.

Quantifying patterns of research-interest evolution

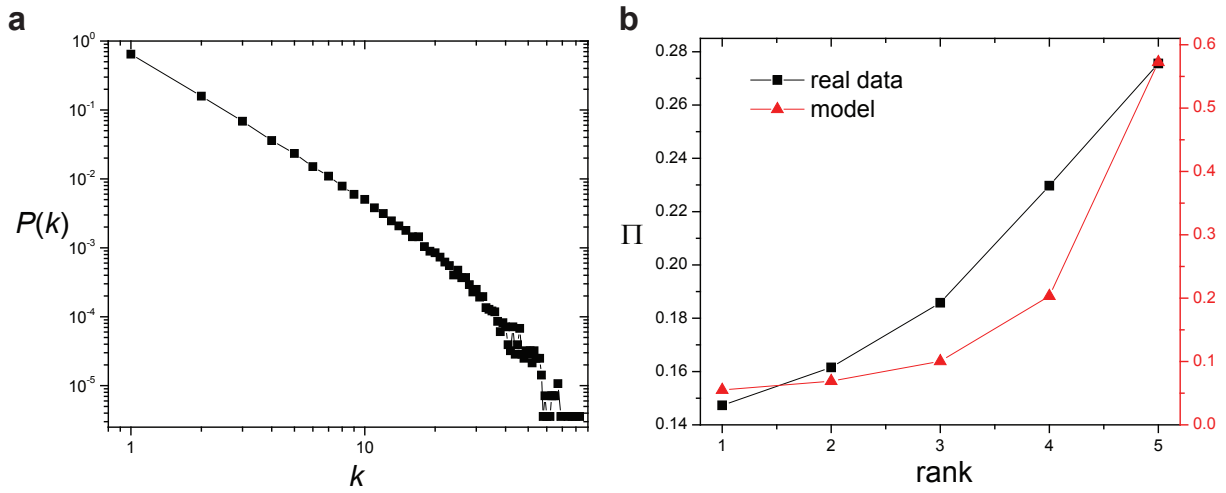
Tao Jia,* Dashun Wang,† Boleslaw K. Szymanski‡

* † ‡ To whom correspondence should be addressed: tjia@swu.edu.cn or dashun.wang@northwestern.edu or szymab@rpi.edu

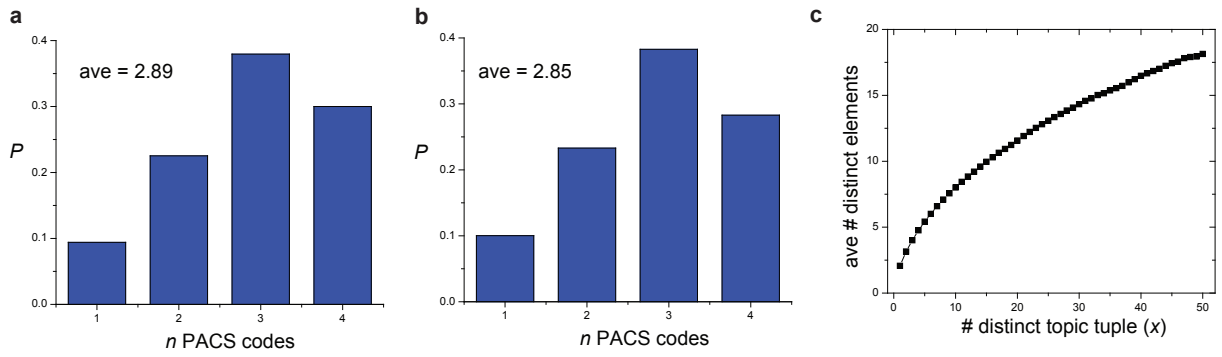
Supplementary Figure S1. An illustration of the creation of modified publication sequence in which one of the three feature (*Heterogeneity*, *Recency* and *Subject proximity*) is removed. T.T. is the topic tuple for short. (**a**) Removal of *Heterogeneity*. The modified sequence is generated by retaining only the first occurrence of each distinct topic tuple. The interest change is then measured in the modified sequence by taking the first and last m distinct topic tuples. Since the length of the modified sequence is shorter than that of the original sequence, any modified sequence with less than $2m$ distinct topic tuples will be dropped. A similar approach is also applied in which we fix the m papers but count each distinct topic tuple once in calculating elements of a topic vector. See more details in Note 7 "An alternative approach to measuring interest change without heterogeneity". (**b**) Removal of *Recency*. In the modified sequence, the usage of each distinct topic tuple is retained. However, each paper's (equivalently a topic tuple's) position is randomly shuffled. (**c**) Removal of *Subject proximity*. In the modified sequence, the use of each topic tuple and the order that each topic tuple is used are retained. However, each distinct topic tuple is replaced by one that is randomly drawn from all existing tuple tuples in our data set, i.e. T.T. i is replace by a randomly drawn topic tuple T.T. i'. This corresponds to the situation when a scientist adopts a new research subject, its topics are totally unrelated with the current research.



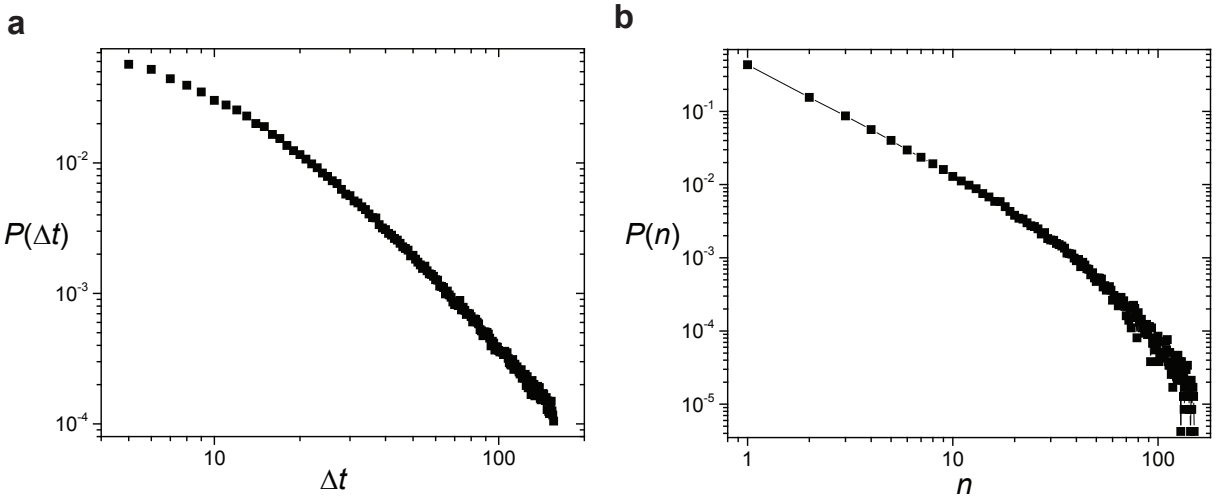
Supplementary Figure S2. The lifetime distribution of a scientist, which is based on statistics of all scientists in the data set. The lifetime T is measured as the time interval, in the unit of month, between an author's first and last paper in our data set.



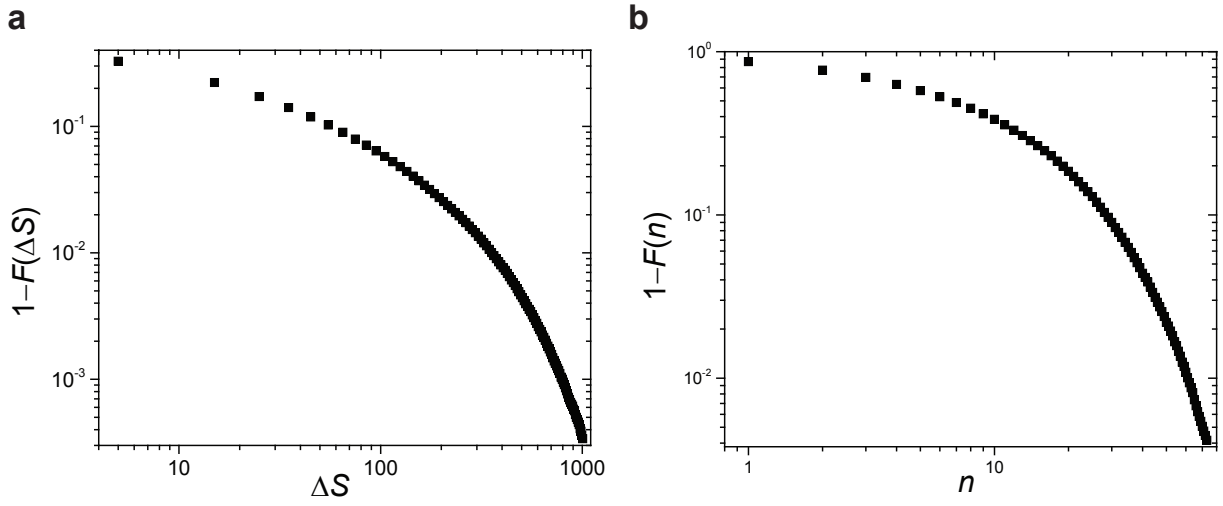
Supplementary Figure S3. The "seashore walk" reproduces heterogeneity and recency features empirically observed in real data (Figs. 2e and 2g of the main text). The variables applied here are the same as those in Fig. 3d of the main text: $p = 0.2$, $P(q) \sim q^{-2}$ and the log-normal distribution $P(S)$ with mean $\mu = 6$, standard deviation $\sigma = 3$ and cutoff $S_{\max} = 2,000$. (**a**) The probability $P(k)$ that a topic tuple is used k times in one's career follows a power-law distribution, which affirms the reproduction of heterogeneity feature. (**b**) The relationship between the probability to reuse a previously studied topic tuple Π and the rank of its first usage (rank 1 is assigned to the first distinct topic tuple used in an individual's career, *etc*). An individual in the "seashore walk" model is less likely to reuse an old topic tuple than a recent one, in line with the recency feature observed in real data.



Supplementary Figure S4. (**a-b**) The distributions of the number of PACS codes each topic tuple contains. They both peak at the value 3 and the average number of PACS codes in each topic tuple is ≈ 3. Cases when the number of PACS codes exceeds 4 are very rare and their probabilities are negligible. (**a**) The distribution based on the occurrence of author-tuple pairs, i.e. each topic tuple is counted once if it appears in an scientist's publication records. (**b**) The distribution based on the occurrence of topic tuple alone, i.e. each topic tuple from a paper is counted only once in calculating the distribution. (**c**) The average number of distinct topics increases monotonically with the number of distinct topic tuples used over a scientist's career, demonstrating the tendency that new topics are added to the current research. The x-axis is the number of distinct topic tuples x used in an individual's career and the y-axis is the average number of distinct topics included in the x distinct topic tuples. Note that distinct topic tuples can be generated via different combinations of current topics without introducing new ones. If that were the case, the number of distinct topics would not change as the number of distinct topic tuple increases. Hence, the observed monotonic increase indicates an important element of knowledge expansion that new topics are added to the current research agenda. This is implemented in our model by replacing one topic code in a topic pool when moving from one topic pool to the other.



Supplementary Figure S5. (**a**) The time interval between a scientist's successive publications follows a power-law distribution, documenting the burstiness in scientific publication. The Δt is measured in the unit of month. (**b**) The distribution of the number of papers authored by a scientist. (**a-b**) are based on statistics of all scientists in the data set.



Supplementary Figure S6. (**a**) The survival function related with $P(\Delta S)$ in Fig. 4c of the main text. The survival function is defined as $1 - F(\Delta S)$ is the cumulative distribution function of ΔS. (**b**) The survival function related with $P(n)$ in Fig. 4d of the main text. The survival function is defined as $1 - F(n)$ is the cumulative distribution function of n.

Supplementary Note 1

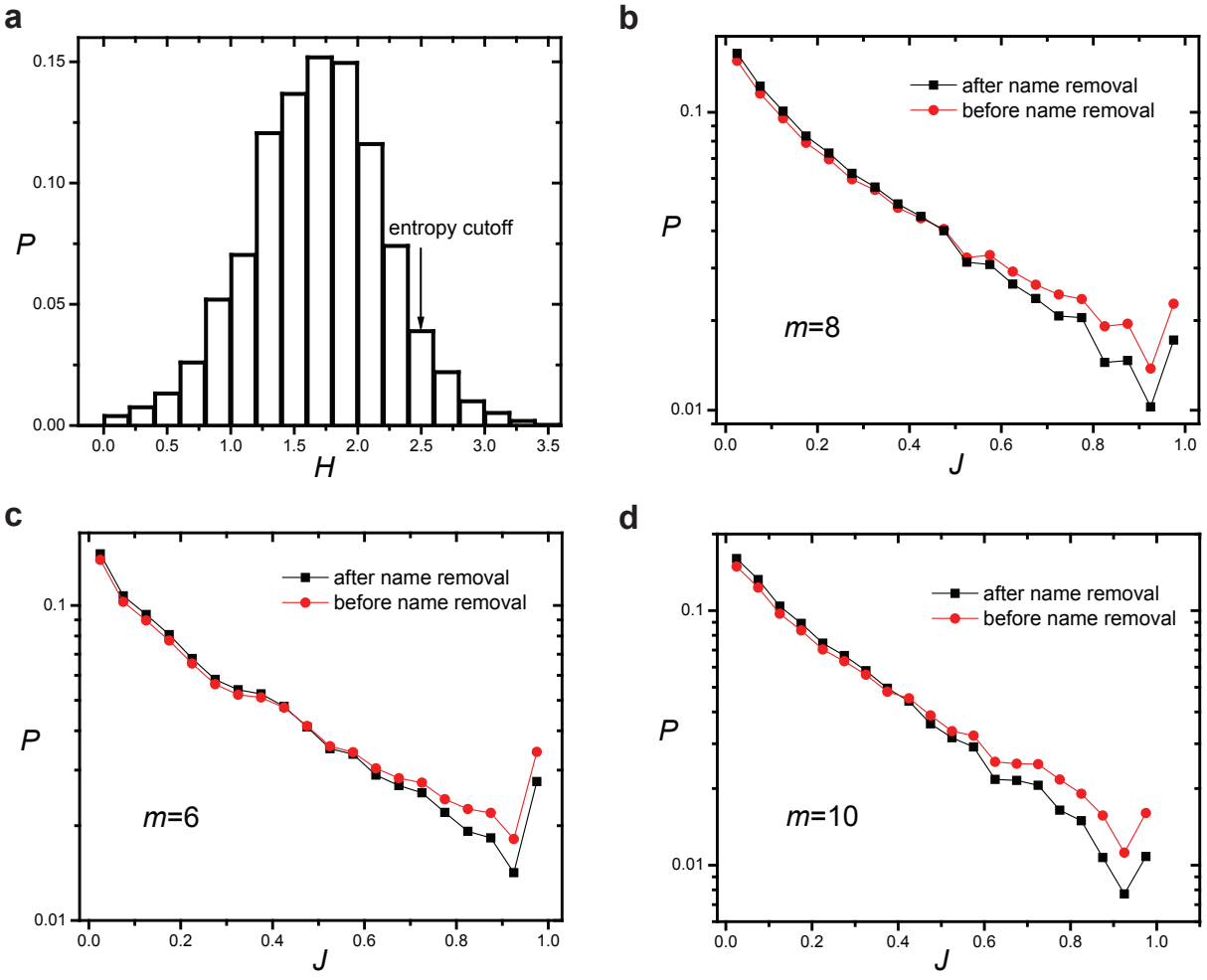
Data set. The APS publication data set is obtained directly at http://journals.aps.org/datasets. The original data set contains publication records from year 1893 to 2009. The PACS codes were originally developed by the American Institute of Physics (AIP) in 1975 and were soon applied by APS to classify topics of the articles published in the APS journals. Since then the fraction of APS publications containing PACS codes has steadily increased and exceeded 90% in 1985. As we concentrate only on papers with PACS codes in this work, we focus on publication records from year 1976 to 2009 and the majority of them are between year 1985 to 2009.

A PACS code is composed of six digits arranged in a hierarchical format in which the first digit identifies one of the ten top level terms and the second digit defines one of up to nine second level terms. We consider the 67 topics defined by the first two digits of the PACS code that can be found at https://www.aip.org/publishing/pacs/pacs-2010-regular-edition. PACS codes start with 35 and 99 are infrequently used, hence are not considered in our analysis.

We consider a sequence of papers with PACS codes published by a scientist as the proxy of this scientist's career record. The first paper of the sequence, sorted by the time of publication, is considered as the start of the scientist's career and the last paper as the end of the career. It is possible that a scientist's earliest publication is not characterized by PACS codes (before 1976) hence not included in the sequence analyzed. It is also likely that a significant number of scientists stay active and continue publishing after 2009 when the record ends. The term "career" only refers to available recored data, which may not be the full extent of a scientist's career.

The APS data set does not directly specify papers belonging to each author. Hence to analyze individual information, we have to group papers published by a single author. Yet, an author's name could be encoded in different forms and different authors may have the same first and last name. As a result, author names must be disambiguated. The detailed study about techniques of author name disambiguation is beyond the scope of this paper. Here we utilize the existing data set based on APS publications with author's name disambiguated. The one applied in our study is obtained from Ref[5] with a total of 237,038 distinct authors. It is reported to have 2% false positive rate (i.e. authors that are considered the same person while in reality they are not) and a 12% false negative rate (i.e. authors that

are wrongly categorized as different persons). This is the one of the few data sets available with sufficiently large number of author names encoded with reasonable accuracy. The same data set has also been used in other recent studies [6–8], suggesting its reliability.



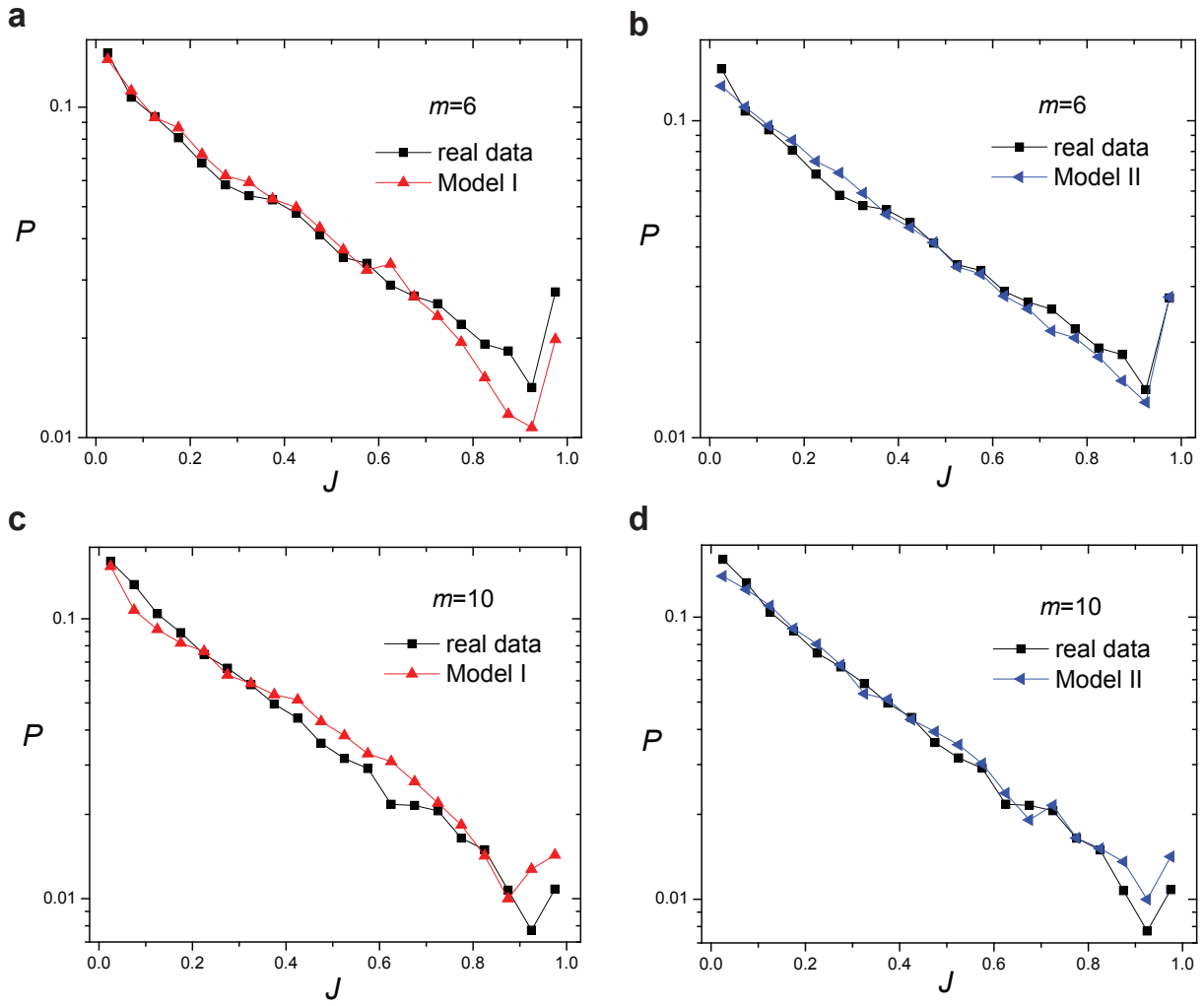
Supplementary Figure S7. (**a**) The entropy distribution of 15,582 authors in the original data set with $n \geq 16$ papers. We set an entropy cutoff value at $H = 2.5$ and remove 867 potential false positive author names. (**b-d**) The interest change distribution $P(J)$ before and after name removal retains the form of exponential distribution for all three different choices of m values ($m = 6, 8, 10$).

It is noteworthy that the macroscopic patterns emphasized in this study is not significantly affected by potential errors in author name disambiguation given the relatively large number of scientists analyzed. We find that after removing a few author names from the original data set whose topic vectors' Shannon entropies H are high, the interest change distributions $P(J)$ before and after name removal differ only slightly, but they are both char-

acterized by an exponential distribution. More specifically, we construct a topic vector g for each author based on all her publications. The Shannon entropy H, which can be considered as a quantification of the scientist's research diversity, is calculated as $H = \sum_{i=1}^{67} -g_i log(g_i)$, where g_i is the i^{th} element of topic vector g. From the original data set, we obtain 15,582 authors with $n \geq 16$ papers (the case of $m = 8$ studied in the main text), whose Shannon entropy H within population follows a Gaussian distribution (Fig. **S7**a). If an author name represents several distinct scientists (false positive), its likely that this author name will cover a wide range of research topics, hence this author will achieve a relatively high H. To diminish its impact on our observation of interest change distribution, we set a cutoff value at $H = 2.5$ and remove 867 authors with $H > 2.5$, which leaves the 14,715 authors for analysis as discussed in the main text. We further apply the same entropy cutoff value and extend this name removal procedure for other two cases discussed in this Supplementary Materials. For $m = 6$, we remove 944 authors with $H > 2.5$ from the 22,102 authors in the original data set who have $n \geq 12$ papers. For $m = 10$, we remove 790 authors from the 11,426 authors in the original data set. In all three cases, we find that the name removal only brings a small change to the interest change distribution of $P(J)$, which, however, does not change the exponential form of distribution (Figs. **S7**b-c).

Supplementary Note 2

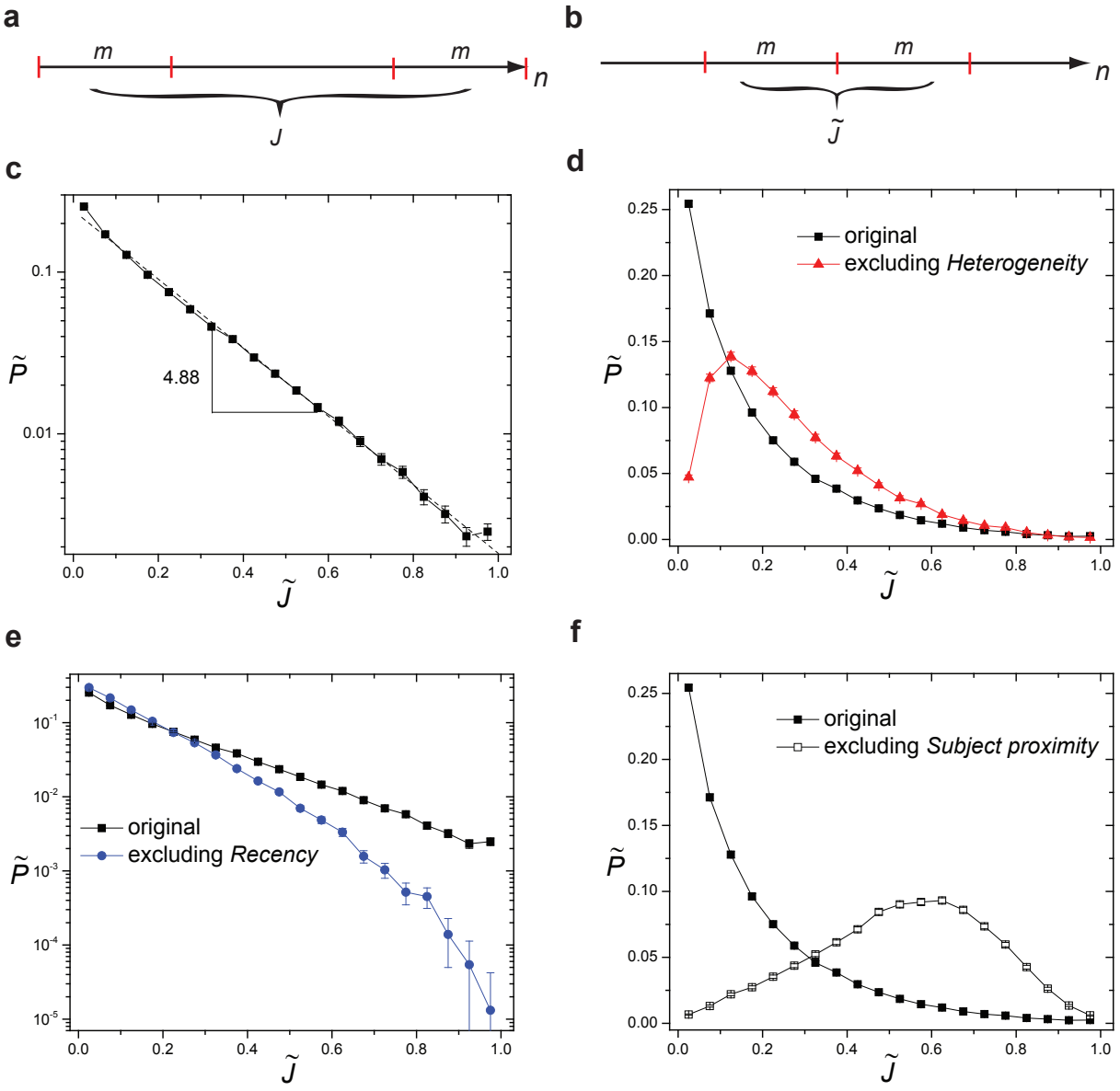
Results based on alternative choices of m. Choosing large m value shrinks the size of population since only scientists with $n \geq 2m$ papers are taken into account in our analysis. On the other hand, small m value makes the topic vector sensitive to noise. $m = 8$ chosen in the main text balances these two factors. For completeness, we investigate cases when $m = 6$ and $m = 10$, and find that our results are not affected by the choice of m (Fig. **S8**). There are 21,158 scientists with $n \geq 12$ and 10,636 scientists with $n \geq 20$, defining the size of populations in each analysis.



Supplementary Figure S8. For different values of m, the extent of interest change follows an exponential distribution. The distributions of J can be fairly captured by the Model I. The distributions of J generated by the Model II match closely with those in real data. The shell sequences are the same as those in Fig. 3 of the main text. They are based on parameters $p = 0.2$, $L = 35$, $P(q) \sim q^{-2}$ and the log-normal distribution $P(S)$ with mean $\mu = 6$, standard deviation $\sigma = 3$ and cutoff $S_{\max} = 2,000$.

Supplementary Note 3

Interest change based on two adjacent paper sets. In the main text, we measure the interest change using the first and last m papers of a scientist, capturing the research interest along the career (Fig. **S9**a). The corresponding interest change is denoted by J. To eliminate bias that may arise from the gap between the two sequences, we also consider a distinct but complementary metric, in which we use two adjacent m paper sequences starting at a randomly chosen paper (Fig. **S9**b). The interest change measured is denoted by measure $\tilde{J}$, capturing the change within a given number of papers. Since its distribution $\tilde{P}$ depends on the choice of the paper sequence for each scientist, we repeat the measurement of $\tilde{P}$ with 2000 randomly chosen paper sequences for each scientist and obtain the mean and standard deviation of the result. We find that the patterns observed are not affected by the metric applied. $\tilde{P}$ also follows an exponential distribution (Fig. **S9**c). The three features (*Heterogeneity*, *Recency* and *Subject proximity*) discussed in the main text are also important in shaping the distribution $\tilde{P}$ (Figs. **S9**d-f).

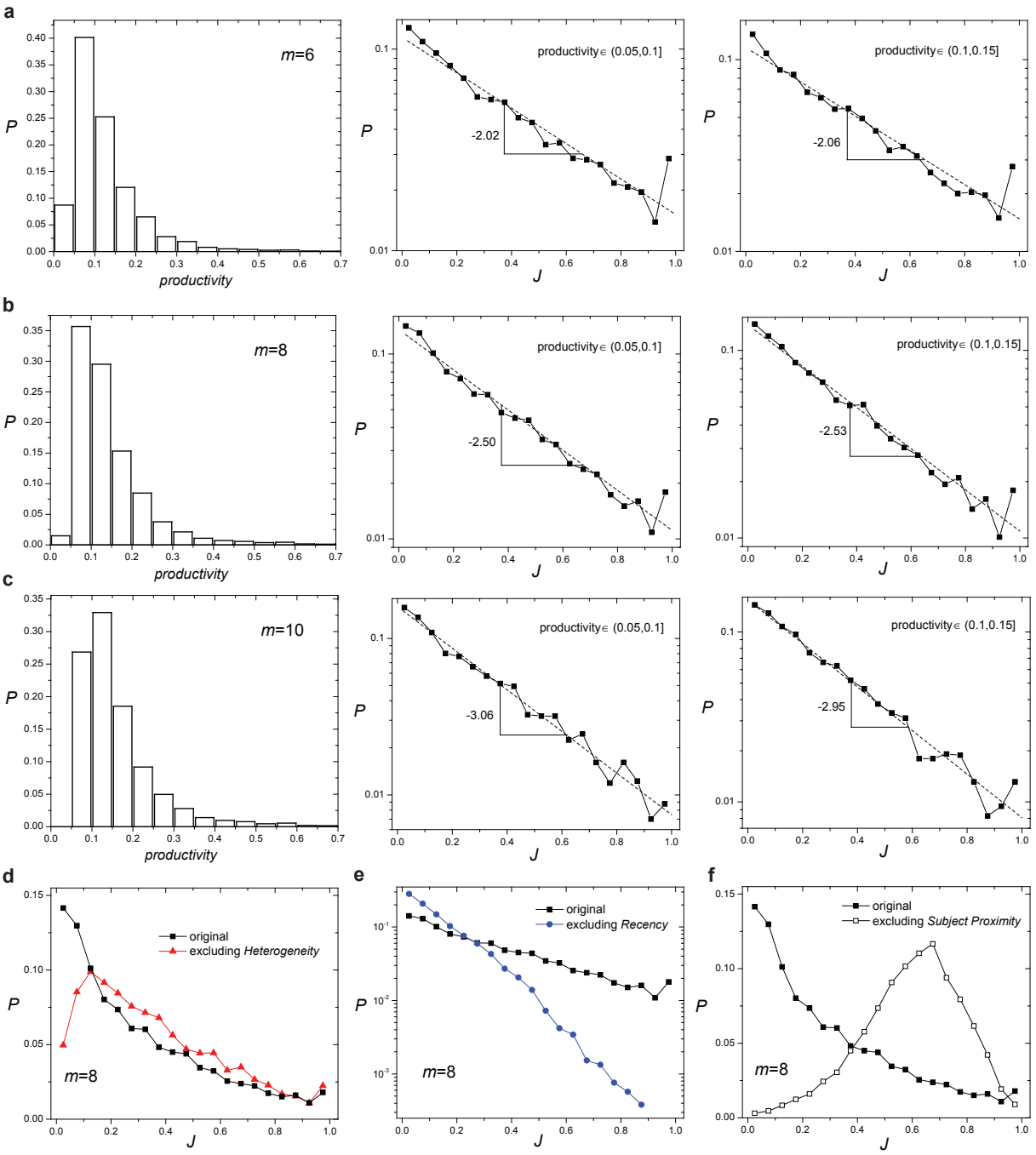


Supplementary Figure S9. (**a**) The measurement in the main text that takes the first and the last m papers in a scientist's publication sequence to obtain the interest change J and its distribution P. (**b**) An alternative measurement that uses two adjacent sequences of m papers randomly chosen from a scientist's publication sequence to obtain $\tilde{J}$. (**c**) The distribution of interest change $\tilde{P}$ can be well fitted with an exponential distribution. (**d-f**) The three features (*Heterogeneity*, *Recency* and *Subject proximity*) are essential to the distribution of $\tilde{J}$, without which $\tilde{P}(\tilde{J})$ will be different. In all cases, p-value equals 0 in the two-sample Kolmogorov-Smirnov test, indicating that the observed differences are statistically significant.

Supplementary Note 4

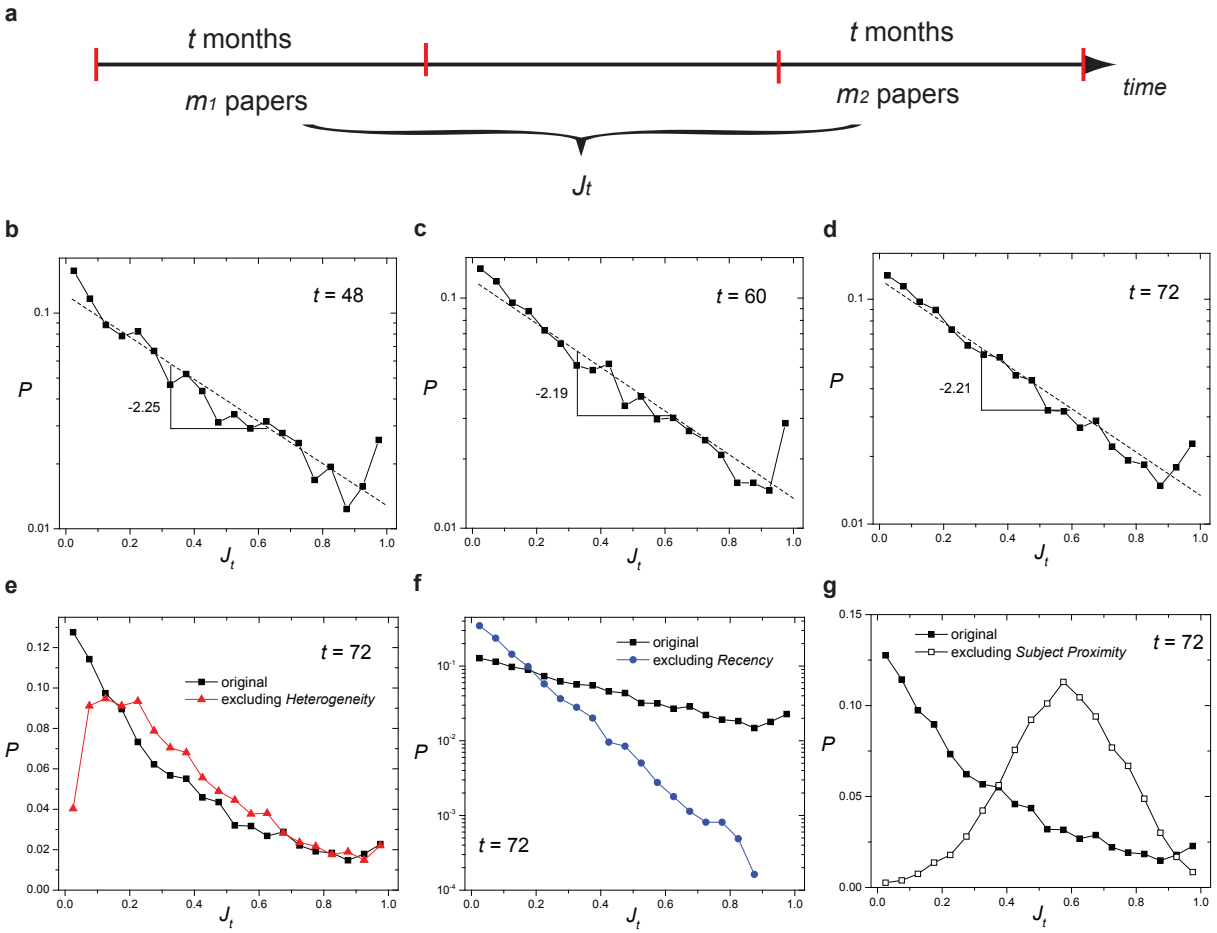
Interest change conditioning on productivity. Considering the fact that scientists publish at different speed, the m papers may cover different time scale for different scientists. Hence our conclusion may be affected by the different output between prolific and non-prolific scientists. To address this issue, we first find the distribution of scientists' productivities. The productivity is defined as $productivity = (n-1)/(t_n - t_1)$, where n is the total number of paper authored by a scientist, t_1 and t_n are the times when the first and last paper published, respectively. The unit of time is month. We focus on scientists who has published at least $2m$ papers ($n \geq 2m$) and chose three different m values ($m = 6, 8, 10$) as used in the main text and SI. Depending on the choice of m, the productivity distribution may peak at different region. But in general, the region $(0.5, 1]$ and $(1, 1.5]$ contain the first and second largest population, corresponding to groups of scientists who on average produce 0.9 and 1.5 papers per year. For scientists whose productivities fall into each region $(0.5, 1]$ and $(1, 1.5]$, we calculate their interest change J using the first and last m papers. In all cases analyzed, we find that the interest change distribution conditioned on productivity can be fitted with an exponential distribution, a similar finding to that using the whole population.

We further perform the three "experiment" and re-measure the interest change distribution in the modified sequences in which each of the three feature *Heterogeneity*, *Recency* and *Subject proximity* are eliminated. The interest change distribution in the modified sequences differ significantly from that of the original sequence, documenting the role of the three features in shaping the research interest evolution. Note that productivity is calculated based on number of papers, not on number of distinct topic tuples. Hence when calculating the interest change in the manipulated sequence where *Heterogeneity* is removed, we apply a different approach to measure the change. In that approach, we fix the m papers and count each distinct topic tuple only once in calculating the elements of a topic vector. See more details in Note 7 "An alternative approach to measuring interest change without heterogeneity".



Supplementary Figure S10. (**a-c**) The distribution of productivity for scientists who at least authored $2m$ papers ($n \geq 2m$). For each choice of m, we take the two largest groups of scientist and measure their interest changes using the first and last m papers in the career. The distribution of interest change can be well fitted with an exponential function. (**d-f**) The three features (*Heterogeneity*, *Recency* and *Subject proximity*) are essential to the distribution of J conditioned on productivity, without which the distribution will be different.

Supplementary Note 5

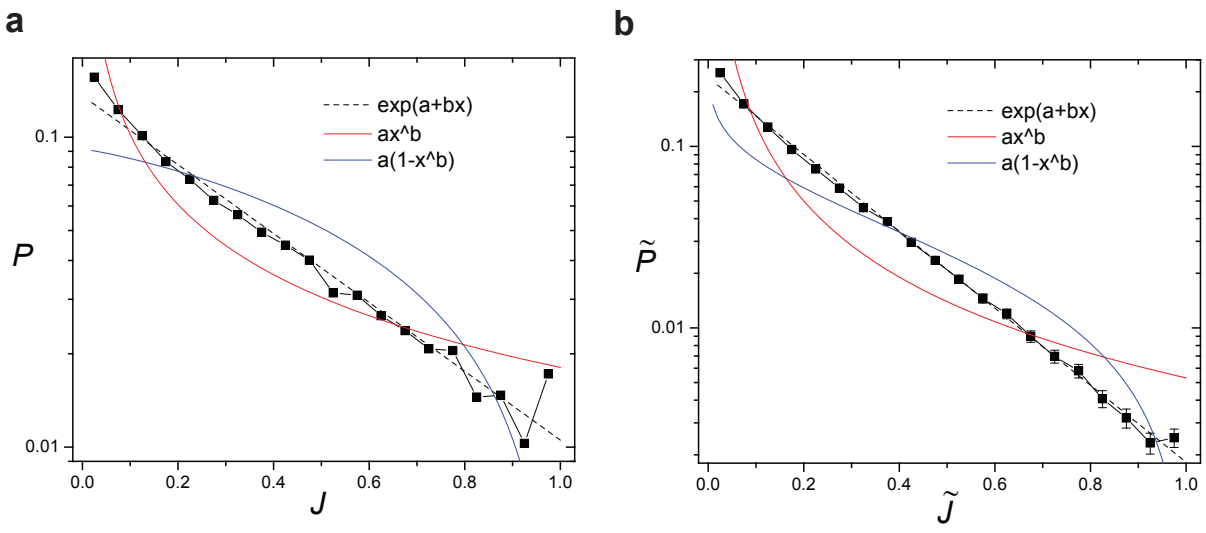


Supplementary Figure S11. (**a**) The illustration of the measure of J_t (**b-d**) The distribution of J_t based on different choices of t values: $t = 48$, $t = 60$ and $t = 72$. (**e-g**) The three features (*Heterogeneity*, *Recency* and *Subject proximity*) are essential to the distribution of J_t, without which the distribution will be different. The original distributions are based on $t = 72$.

Interest change based on two fixed periods of time. Research interest is associated with time. Hence it is important to quantify research interest for a given period of time and further quantify the corresponding change. Here we first identify two time periods: the first one covers t months after the publication of the first paper and the second one covers t months before the publication of the last paper. Correspondingly, we locate the $m1$ and $m2$ papers published in these two time periods, based on which we measure the interest change J_t (Fig. **S11**a). Authors whose career life span is less than $2t$ months are excluded from the analysis. To effectively calculate topic vectors, we also require that both $m1$ and $m2$ are not

less than 6. By choosing different t values ($t = 48$, 60 and 72), we obtain groups of authors of different sizes (3564, 5142 and 6152 scientists respectively), based on which we calculate the distribution of J_t. We find that $P(J_t)$ can be well fitted by functions with exponential form (Figs. **S11**b-d). We further perform similar "experiments" on the sequences with $t = 72$ and find that these three features (heterogeneity, recency and subject proximity) play similar roles in shaping the distribution of interest change as discovered from other measures (Figs. **S11**e-g).

Supplementary Note 6



Supplementary Figure S12. The distribution $P(J)$ and $\tilde{P}(\tilde{J})$ and different form of fitting functions. The exponential function gives the best fit compared with the other two.

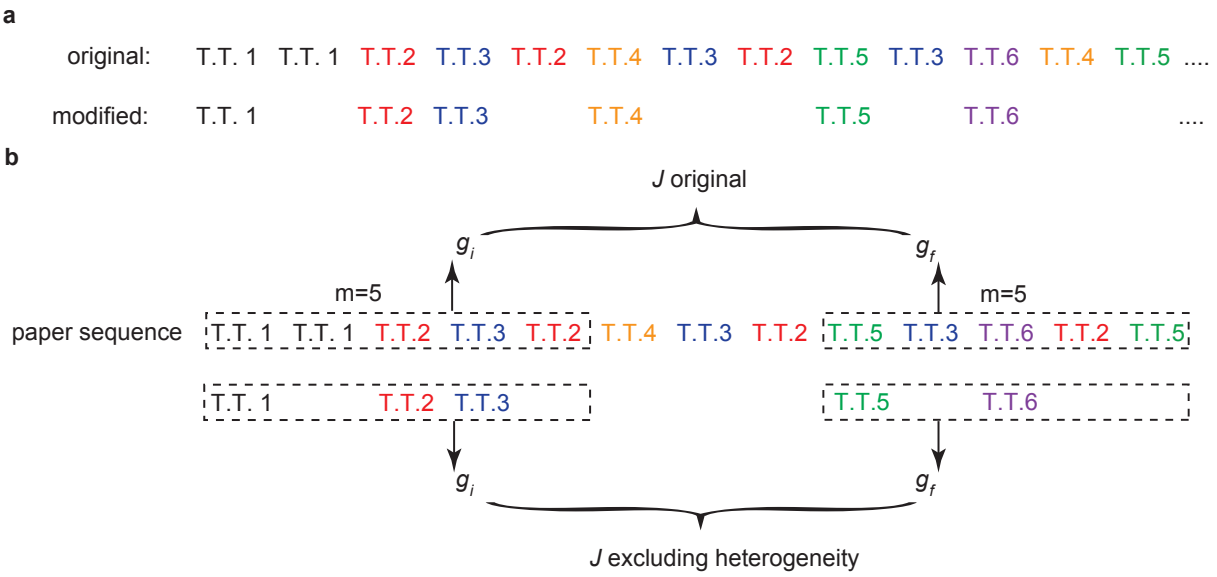
The function of the best statistical fit to interest change distribution. Because both J and $\tilde{J}$ are obtained from two sets of m papers and their distributions $P(J)$ and $\tilde{P}(\tilde{J})$ are both based on the whole population of 14,715 scientists, we would expect that they can be fitted by the same form of distribution. Here we use both $P(J)$ and $\tilde{P}(\tilde{J})$ to test our fitting functions. $P(J)$ and $\tilde{P}(\tilde{J})$ demonstrates an obvious linear form after taking the y-axis in log scale. Hence we choose a simple exponential function with two parameters $y = e^{a+bx}$ to fit the data. To check if other forms of fitting function can also fit the data well, we also perform the fit using two forms of power-law function with two parameters $y = ax^b$ and $y = a(1 - x^b)$. We constrain fitting parameters to be two because a numerical fit improves with the number of fitting parameters. To further check if the fit can be significantly improved by introducing additional fitting parameters, we test an exponential function with three parameters $y = e^{a+bx+cx^2}$. We use reduced chi-square (χ^2) as a quantification of how good the fit is. Since the value of $P(J)$ ($\tilde{P}(\tilde{J})$) for large J ($\tilde{J}$) is significantly smaller than that for small J ($\tilde{J}$), we calculate χ^2 using the log value of the fitting function and the data.

fitting function	χ^2 of fitting $P(J)$	χ^2 of fitting $\tilde{P}(\tilde{J})$
e^{a+bx}	**0.0187**	**0.00619**
$y = ax^b$	0.0857	0.357
$y = a(1 - x^b)$	0.278	0.216

Supplementary Table S1. The χ^2 of different fitting results

As shown in both Fig. **S12** and Table **S1**, the fitting functions with the power-law form systematically deviate from the data. The corresponding χ^2 values are several times or even several magnitudes larger than that of the exponential function. Hence, exponential function is a better choice than others. We further find that adding an additional parameter can only marginally reduce χ^2. Hence the exponential function e^{a+bx} adequately fits the distributions observed.

Supplementary Note 7



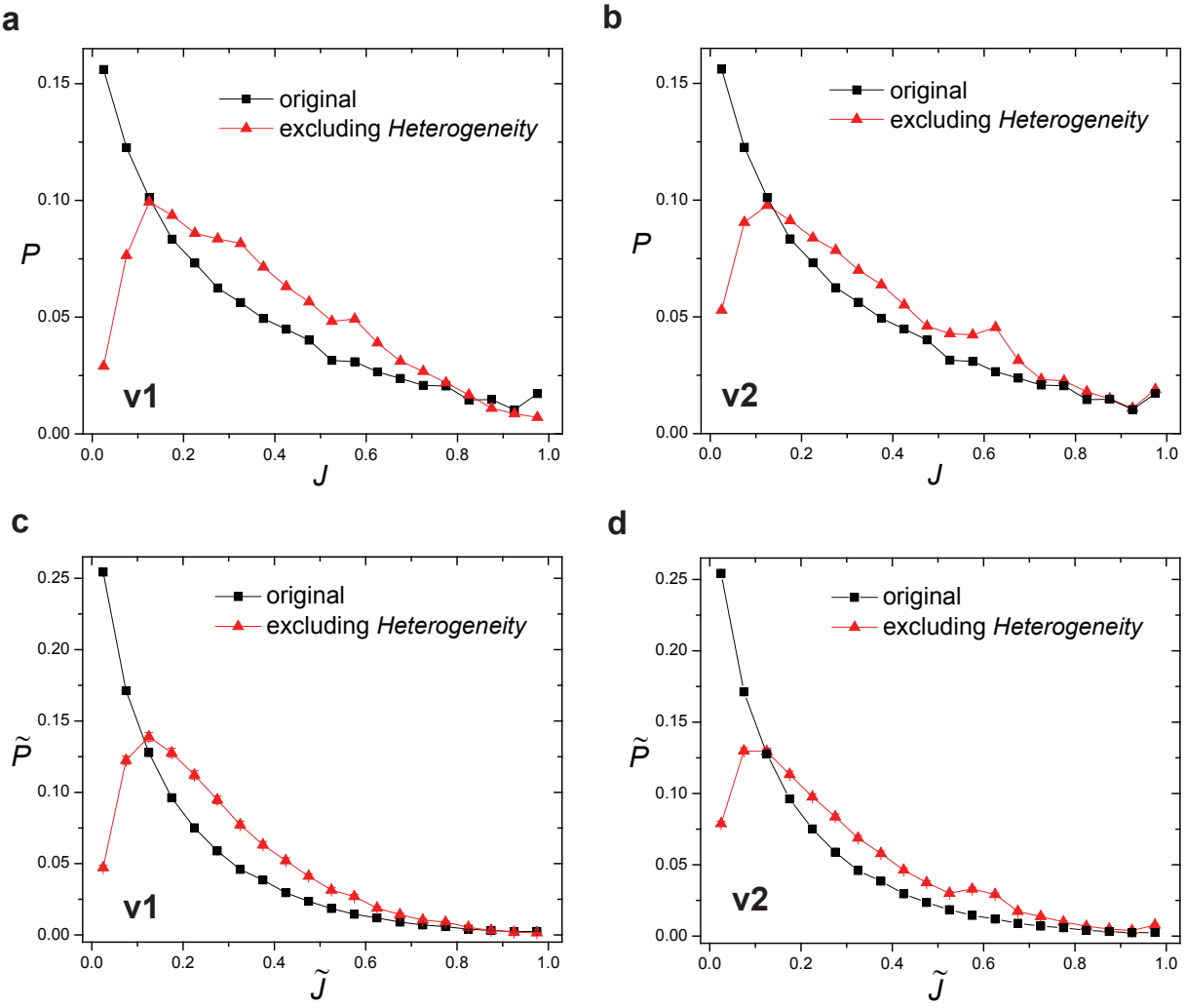
Supplementary Figure S13. (**a**) The method used in the main text to remove heterogeneity in the paper sequence. (**b**) An alternative approach. We keep track of the first and last m papers of each author and only consider a distinct topic tuple once when calculating the topic vector.

An alternative approach to measuring interest change without heterogeneity. In the main text, we remove heterogeneity in the publication sequence by keeping the first occurrence of each topic tuple (T.T. for short) and deleting its subsequent occurrences to generate a new topic tuple sequence for each author (Fig. **S13**a, which the same as Fig. S1a). We then use the first and last m topic tuples in the modified sequence to calculate topic vectors and the corresponding interest change. In this way, the metric is kept unchanged taking m topic tuples for each topic vector. The sample size, however, shrinks from around 15,000 to about 8,000 as only authors who used at least $2m$ distinct topic tuples (not $2m$ paper) are kept.

Alternatively, we have also done a different measure in which we keep track of the first and last m papers of each author. But we only consider a distinct topic tuple once when calculating the topic vector (Fig. **S13**b). In the example shown in Fig. **S13**b, two sets of papers are identified for $m = 5$. Because T.T. 1 (topic tuple 1) and T.T. 2 appear twice in the first set, g_i is calculated using 3 topic tuples. Likewise, because T.T. 1 and T.T. 2 are already counted in calculating g_i, they are not used again in calculating g_f. T.T. 5 is used

only once though it appears twice in this set. Hence, g_f is calculated using 2 topic tuples.

This alternative approach keeps the sample size unchanged. However, the metric changes as each topic vector may be based on fewer than m topic tuples. For the analyses performed in the main text and in S4, we can use either of them. Regardless of which method we choose, we arrived at the same conclusion about the impact of heterogeneity on the distribution of research interest change (Fig. **S14**). For the measure of J conditioning on productivity and J_t, we have to adopt the approach in Fig. **S13**b because the two sets of papers are fixed after conditioning on productivity or choosing the time-scale. The method used in the main text does not apply in these cases.



Supplementary Figure S14. (**a-b**) The distribution $P(J)$ obtained after eliminating heterogeneity using one of the two methods. (**c-d**) The distribution $\tilde{P}(\tilde{J})$ obtained after eliminating heterogeneity using one of the two methods. **v1** corresponds to the approach illustrated in Fig. **S13**a as well as Fig. S1a, that is taken in our main text and analyses in S4. **v2** corresponds to the approach illustrated in Fig. **S13**b. Our conclusion about the effect of heterogeneity on the research interest change distribution is independent of the method used to quantify it.

Supplementary Note 8

Implementation of the Model I. The simulation proceeds as follows.

1. Construct a 1-D lattice with 5000 sites. Divide the lattice into multiple blocks according to the length of topic pool L. Assign each block a topic pool. The pool codes varies from one to the other by replacing one existing code by a new one. For simplicity we use continuous numbers 1, 2, ... as artificial classification of topics similar to that given by the PACS code. We start at site coordinate 0. Sites [0,L-1] are given topic pool $\{1, 2, 3\}$, sites [L,2L-1] are given topic pool $\{2, 3, 4\}$, and so on.

2. Designate a site non-empty with probability p. If a site is non-empty, assign to it an integer number according to the distribution $P(q)$, representing the number of "shells" (the maximum number of papers) on that site. Randomly draw 3 times a topic code from the topic pool to which this site belongs. The combination of the 3 topic codes drawn is considered as the artificial topic tuple of the papers generated by this discovery.

3. For each independent walker, restore the site's "shell" number to the original state. Obtain the walker's number of steps S from the distribution $P(S)$. Randomly pick a starting point for the walker. Every time step the walker has 0.5 probability moving to the left and 0.5 probability moving to the right. When the walker reaches a site containing a "shell", the walker gains 1 paper and the "shell" number of the site is deducted by 1. The corresponding topic tuple is recorded. The walker stops at time step S. Periodic boundary is applied that connects the first and last site of the lattice.

We record 150 independent random walkers for every configuration of the 1-D lattice and we generate 400 random configurations of the 1-D lattice, with a total of 60,000 walkers recorded. We follow the same procedure in the analysis of the real data to calculate J and consider only the walkers with $n \geq 2m$ papers. For the case shown in the main text when $m = 8$, there are around 14,000 qualified walkers, a comparable size to the number of qualified authors in real data.

As topic tuples are represented as random combinations of topics, different locations on the lattice may be characterized by the same artificial topic tuple in the current model. We check and confirmed that this does not affect the heterogeneity or recency feature in the topic tuple sequence. Such duplication, caused by the simplicity of the model's setup, can also be eliminated easily by excluding repetitions in the topic tuple generation.

Given the complexity in the calculation of J, and the stochastic nature of the model

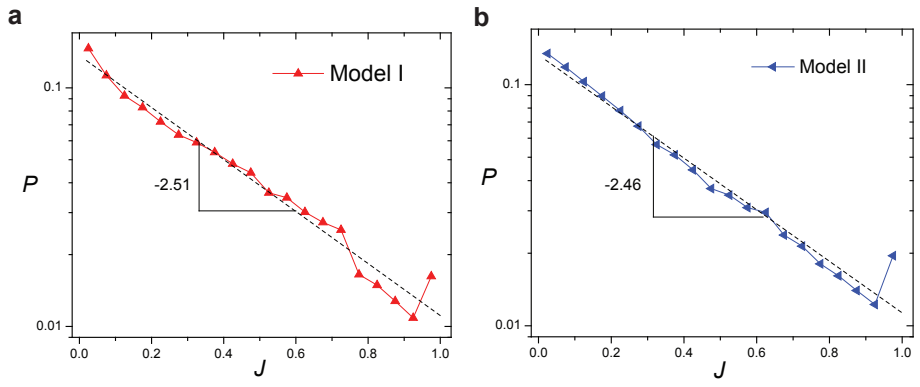
and the walker's transient state behavior we are interested in, the model is analytically intractable to the best of our effort. We choose and decide the variables manually. We first choose the variable $P(q)$ as it is most relevant to the distribution $P(k)$ observed empirically. Note that the distribution $P(q)$ and $P(k)$ are not the same. A random walker is likely to return and pick up all "shells" at a location, but this may not always be achieved during a limited number of moving steps. Hence we choose $P(q)$ only similar but not identical to $P(k)$. The rest of variables are not directly given by empirical observations. For simplicity, we first fix the $P(S)$. The log-normal distributed $P(S)$ has a fat-tail $\sim S^{-1}$ when $\sigma^2 > \mu$, hence we fix μ and σ and only change the cut-off value $S_{\max}$. In general, increasing $S_{\max}$ will grant more moving steps and walker has a longer moving range. Hence large J values are generated more frequently. Increasing L will make it more difficult for the walker to encounter new topics, which consequently makes large J value appear less frequently. We also need to balance the number of papers in each topic pool by tunning p value as well.

The variables in the main text are chosen such that both the Model I and the Model II can be validated by the same set of variables. As far as Model I is concerned, many choices of variables can qualitatively reproduce the distribution of J, with a few of the tested ones listed in Table **S2**.

	p	L	$P(q)$	$S_{\max}$
1	0.15	100	$P(q) \sim (1+q)^{-3}$, $q \in [1, 100]$	20000
2	0.15	80	$P(q) \sim (1+q)^{-2.5}$, $q \in [1, 100]$	10000
3	0.2	50	$P(q) \sim (1+q)^{-2.5}$, $q \in [1, 100]$	4000
4	0.3	60	$P(q) \sim (1+q)^{-3}$, $q \in [1, 100]$	6000
5	0.4	25	$P(q) \sim q^{-2}$, $q \in [1, 100]$	1000
6	0.25	30	$P(q) \sim q^{-2}$, $q \in [1, 100]$	2000
7	0.2	50	$P(q) \sim q^{-2.5}$, $q \in [1, 100]$	5000

Supplementary Table S2. Other variables that can qualitatively reproduce the distribution of J. In all cases, log normal distribution $P(S)$ with mean $\mu = 6$ and standard deviation $\sigma = 3$ is applied. Only the cutoff $S_{\max}$ varies.

Supplementary Note 9



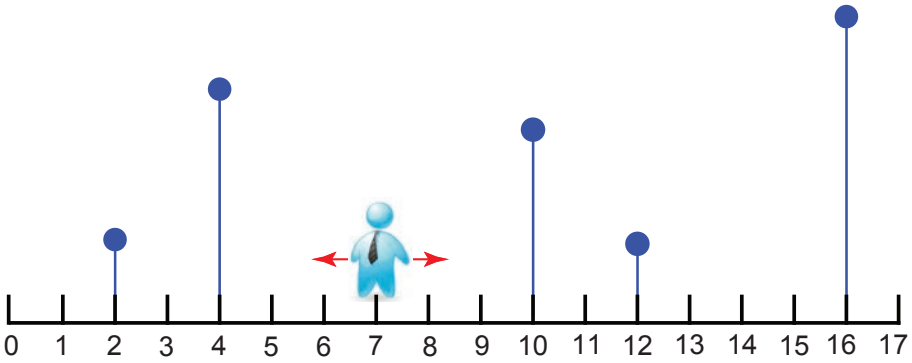
Supplementary Figure S15. The exponential fit of the two model results

Statistical analyses of model results. We use mean percentage error (MPE) and mean absolute percentage error (MAPE) to quantify how well our model reproduces empirically observed interest change distribution $P(J)$. MPE is defined as $\frac{100\%}{n}\sum_{i=1}^{n}\frac{a_i-f_i}{a_i}$ where n is the number of data points, a_i is the actual value of point i and f_i is the value of point i generated by the model (predicted value). MAPE is defined as $\frac{100\%}{n}\sum_{i=1}^{n}|\frac{a_i-f_i}{a_i}|$. We also report the MPE and MAPE based on the log value of a_i and f_i in Table **S3**. In general, both Model I and Model II well reproduce $P(J)$. Model II demonstrates an improved accuracy compared with that of Model I.

	MPE	MAPE	MPE based on log values	MAPE based on log values
Model I	2.79%	9.13%	-0.469%	2.72%
Model II	2.68%	6.04%	-0.461%	2.15%

Supplementary Table S3. Statistics of the model results

Supplementary Note 10



Supplementary Figure S16. An illustration of the "seashore walk"

The random mapping procedure for the Model II. The simulation proceeds as follows.

1. Construct a 1-D lattice with 5000 sites. If a site contains "shells" (the property granted with the probability p), assign the number of "shells" according to the distribution $P(q)$, representing the maximum number of papers that can be generated at this site.

2. For each independent walker, restore the site's "shell" number to the original state. Obtain the walker's number of steps S from the distribution $P(S)$. Randomly pick a starting point for the walker. Every time step the walker has 0.5 probability moving to the left and 0.5 probability moving to the right. When the walker reaches a site containing "shells", the walker gains 1 paper and the "shell" number of the site is deducted by 1. The walker stops after S steps. Periodic boundary is applied that connects the first and last site of the lattice.

The random walker generates a sequence of papers along the entire walk that can be labeled by the coordinate of the discoveries, i.e. the site on the lattice at which these papers are created (Fig. **S16**). For example, a walker generates a sequence (4, 4, 2, 2, 4, 4, 4, 10, 10, 10, 12, 12, 16, 16, 10, 10) with 16 elements within the lifetime steps, corresponding to 16 papers from 5 discoveries (4,2,10,12,16). From the data set, we randomly pick an author who used 5 distinct topic tuples in the career and randomly map the 5 topic tuples to locations (4,2,10,12,16). Therefore papers are characterized by real topic tuples, allowing us to calculate the interest change J with the same approach as for real data.

In the paper, we use the same sequence used in Model I. Particularly, we apply the variables $p = 0.2$, $P(q) \sim q^{-2}$ and a log normal lifetime distribution $P(S)$ with mean $\mu = 6$,

standard deviation $\sigma = 3$ and cutoff $S_{\max} = 2,000$. We record 150 independent random walkers for every configuration of the 1-D lattice and we generate 400 random configurations of the 1-D lattice, with a total of 60,000 walkers recorded. The J is calculated by considering only the walkers with $n \geq 2m$ papers. For the case shown in the main text when $m = 8$, there are around 14,000 qualified walkers, a number comparable to the number of qualified authors real data.

Supplementary Discussion 1

Difference between random mapping and duplicating data sequence. The random mapping procedure is non-trivial and quite different from duplicating the real sequences. First, the order of topic tuple usage is shuffled. A topic tuple that first appears in the middle of a real sequence can be the one first used in the model. Second, the number of papers published under each topic tuple is random and it can be different in the model and in the data. A frequently used topic tuple in the data can become one touched only occasionally in the model. Finally, the total number of papers by a given set of topic tuples can be different as well. We would like to particularly mention that in the random mapping procedure, an author is randomly picked from the complete ensemble of authors, not from the group where each has authored at least $2m$ papers ($n \geq 2m$). That is, a randomly picked author who used 5 distinct topic tuples in the career may have only published 10 papers and is not one of 14,715 scientists used to calculate the distribution of J.

Taken together, the random mapping procedure provides us with a simple way to have the topic tuples and their sequences in the model statistically similar to those of real data. It allows us to statistically infer subject proximity from individuals' moving trajectories among different research subjects. It also avoids the difficulties of rebuilding topic tuples. Indeed, topics within each topic tuple are correlated: some topics are likely to appear together while some others are rarely doing so. The sequences, however, are not duplications of real data.

Supplementary Discussion 2

Some assumptions imposed on the model. There are several assumptions in our model based on or motivated by empirical observations, such as a power-law distribution for shells at a site and a truncated log-normal distribution for a walker's time-steps. These assumptions are important. Yet, they alone are not sufficient to reproduce the characteristics of the research interest evolution. We would like to briefly show three examples. First, if the walker does not make unbiased random walk but only walks in one direction, it can only pick one shell at a site. Therefore, the number of papers under each topic tuple will not be a power-law even when the number of shells at each location follows power-law. Second, if the walker picks all shells at one location at a time, the number of papers under each topic tuple can be a power-law, but different topic tuples are segregated, differing from the actual situation. The recency feature will not be accounted for in this case, as there would be no re-use of different topic tuples. Finally, if the underlying space is multiple-dimensional instead of 1-D, the probability that the walker returns to a site visited is significantly smaller (under the current walking rules). This means that the random walker can be easily trapped in and wandering about some empty zone without publishing any papers (finding any shells). Hence the number of papers during a career will follow a different distribution even though the walker's time-steps still follow a truncated log-normal distribution.

We would like to further mention that some important characteristics, such as the recency feature and burstiness in publication, rely on the mechanism of the random walk, not on the assumptions about distribution of the shells or the distribution of time-steps. The random walker performing unbiased random walk is likely to return to a site recently visited, which helps it collect the shells at a site that contains them. However, when the shells at that site are exhausted, the walker returns with the same probability but will not gain any new shells in the subsequent visits. Such mechanism ensures the recency feature. The recency feature is generated by this mechanism even when a fixed and uniform number of shells (not power-law) are assigned to the site of lattice. The burstiness is also generated by the property of random walk. The 1-D lattice contains multiple piles of shells with a random number of empty sites in-between. Assume that a random walker is at a random position between the two piles of shells. The steps needed to produce the next paper depend on how soon the walker can reach either of the two piles of shells. This is equivalent to the random walker's first passage time steps to reach a boundary, which has been found to follow a power-law

distribution with exponential cutoff. This property in random walk eventually gives rise to the burstiness of publications generated by our model.

Finally, the truncated log-normal distribution for the walker's life-time steps helps to generate the power-law (with exponential cutoff) distributed number of papers. It, however, does not help reproduce other features such as the heterogeneity, the recency and the distribution of J. Indeed, the model can still reproduce these patterns with a fixed career span assigned to all walkers. A truncated log-normal distribution is chosen because it best reflects the actual life-time span of scientists.

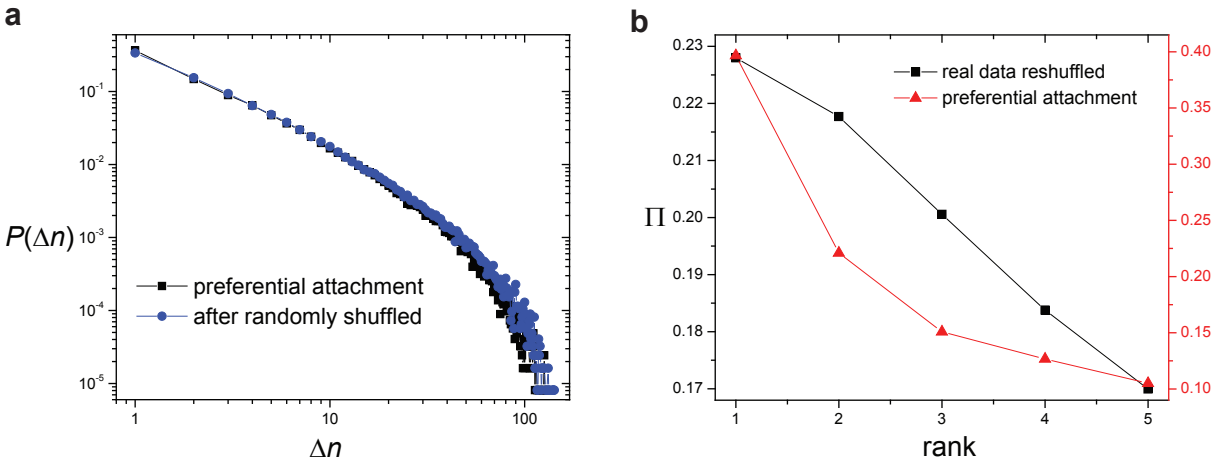
Taken together, some empirical based assumptions are important in our model. However, the systematic reproduction of empirical observations arises from the interplay of multiple factors. Missing either of them could invalidate the model.

Supplementary Discussion 3

The reproduction of the burstiness and heterogeneity in publication. The 1-D lattice contains multiple piles of shells with a random number of empty sites in-between. Assume that a random walker is at a random position between the two piles of shells. The steps needed to produce the next paper depend on how soon the walker can reach either of the two piles of shells. This is equivalent to the random walker's first passage time t to reach a boundary, which has been found to follow a power-law distribution with exponential cutoff. The head of the distribution scales as $\sim t^{-3/2}$. This property in random walk eventually gives rise to a power-law distributed (with exponential cutoff) number of steps for successive publications. Indeed, the numerical fit of $P(\Delta S)$ shows a scale factor -1.33 for the power-law head followed by an exponential cutoff, close to -1.5 that is theoretically predicted.

Our model generates a power-law distribution with exponential cutoff for number of shells picked by a walker. The same form of distribution is observed in real data for the number of papers authored by a scientist. Our model is able to reproduce this form of distribution due to a combination of several factors. First, the number of distinct sites visited by a random walker scales as $S^{1/2}$, where S is the number of steps the walker moves. Second, the probability that a site on the lattice contains a shell is uniform and does not depend on the distance from the starting point of the walker. Furthermore, the number of shells at each site follows an identical and independent distribution. Therefore, from a mean-field point of view, the number of shells picked by a walker should also scale as $S^{1/2}$. Finally, the time-step distribution $P(S)$ for each walker is set as a truncated log-normal distribution. As known, when the variance (σ^2) is much larger than the mean (μ), a log-normal distribution contains a fat tail scales approximately as S^{-1}. Combining all these factors, the distribution of number of shells (papers) $P(n)$ by each walker is expected to be power-law with exponent close to -0.5. $P(n)$ must have an exponential tail (exponential cutoff) because the distribution of time-step is truncated. The numerical fit of $P(n)$ shows a scale factor -0.41 for the power-law head followed by an exponential cutoff, which roughly agrees with the coarse-grained theoretical prediction.

Supplementary Discussion 4



Supplementary Figure S17. (**a**) $P(\Delta n)$ of sequences generated by preferential attachment model does not change significantly with or without random shuffling, implying that sequences by preferential attachment model do not contain temporal correlation. (**b**) Π is measured in sequences generated by preferential attachment model and reshuffled sequences of real data. The descending trend is observed in both cases, indicating that it is the temporal correlation that gives rise to the different behaviors between preferential attachment model and real data.

Preferential attachment and recency feature. The sequence generated by preferential attachment does not exhibit temporal correlation, as $P(\Delta n)$ is almost the same as that of the randomly reshuffled sequence (Fig. **S17**a), differing from the empirical observations (Fig. 2e of the main text). The recency feature, however, arises from the temporal correlation within the topic tuple sequence. If we randomly reshuffle the real sequence, the recency feature disappears and an individual is more likely to reuse the old topic tuple, the same trend as predicted by preferential attachment (Fig. **S17**b).

Supplementary Reference

[1] Deville, P. *et al.* Career on the move: Geography, stratification, and scientific impact. *Sci. Rep.* **4**, 4770 (2014).

[2] Shen, H.-W. & Barabási, A.-L. Collective credit allocation in science. *Proc. Natl. Acad. Sci.* **111**, 12325–12330 (2014).

[3] Sinatra, R., Deville, P., Szell, M., Wang, D. & Barabási, A.-L. A century of physics. *Nat. Phys.* **11**, 791–796 (2015).

[4] Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).

[5] Deville, P. *et al.* Career on the move: Geography, stratification, and scientific impact. *Sci. Rep.* **4**, 4770 (2014).

[6] Shen, H.-W. & Barabási, A.-L. Collective credit allocation in science. *Proc. Natl. Acad. Sci.* **111**, 12325–12330 (2014).

[7] Sinatra, R., Deville, P., Szell, M., Wang, D. & Barabási, A.-L. A century of physics. *Nat. Phys.* **11**, 791–796 (2015).

[8] Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).